Efficacy of Densely Connected Convolutional Neural Networks & ResNet in Classifying Chest X-rays

June Yun, Jonathan Lin

December 2023, Emory University, Dr. Yuanzhe Xi

Abstract: This study investigates the use of deep learning algorithms, specifically ResNet and DenseNet, for the automated interpretation of chest X-rays (CXR), a crucial diagnostic tool in detecting conditions like pneumonia, pneumothorax, and tumors. Addressing the gap between the high volume of CXR images and the limited number of radiologists, the project explores the potential of artificial intelligence to streamline and enhance image analysis in radiology. The results demonstrate that both ResNet and DenseNet models achieve high AUC scores, with DenseNet-121 slightly outperforming ResNet50V2. Moreover, a new model is proposed that outperforms both by at least 5.1% and rivals the performance of the previously developed CheXNet and radiological experts.

Introduction

A chest X-ray is one of physicians' most widely used diagnosis tools. Some common conditions diagnosed by CXRs include heart and lung diseases. Pneumonia, by itself, has been reported to be a large contributor to deaths around the globe. For reference, in 2017, it was responsible for 15% of all child deaths.[20] Further, the presence of a condition in a patient's heart and lung are often significant comorbidities for other conditions such as cancer and other chronic illnesses. Thus, it is obviously of utmost importance to not only be able to catch conditions that are already present in a patient, but indicate them as quickly as possible for adequate treatment to be accounted for.

Besides the obvious need for accurate interpretations of these images, computer vision can also potentially lighten the workload on radiologists. The discrepancy between supply and demand has a rippling effect on the rest of the healthcare pipeline and can delay much-needed care (outpatient exams take anywhere from 1-6 weeks to be read [18]). Certain experts cite the need to reduce the medical images taken at visits. While noble, this would take away from much of what physicians are trained to do. Think about it like this: when you visit a physician, you present with a certain complaint and associated symptoms. The provider is meant to take the present illness's history and use their experience to develop a work-up for a medical decision. Simply taking this information at face value and making a diagnosis is similar to asking a chef to make a meal by smell alone. Sounds impossible, right? However, by ruling certain conditions out, providers can strengthen their certainty in a diagnosis just like with a cookbook or recipe, chefs can perfect their dishes. Additional ideas have come in the form of hiring more foreign experts and raising radiology salaries to incentivize professional activity.[5] So, what is the solution that most are leaning on? Many experts have looked to, you guessed it, AI.

In general, CXRs as an imaging option are considered non-invasive and cost-effective. However, the interpretation requires quite a deep wealth of knowledge in pathology and anatomical features to distinguish physiological abnormalities and further integrate this knowledge into a patient's past medical history / presenting chief complaint. The recent pandemic only exacerbated the fact that the interpretation of CXRs might not have been as expedient as necessary in times of medical emergency. Specifically, lung conditions that COVID-19 compounded upon were of pertinent focus.[20] Computational methods have already been shown to have the potential to play a significant role in decision-making. One study revealed that a user-crafted deep learning model reached an accuracy of 96.4% with a recall (true positive rate) of 99.62% on certain disease labels.[4] Further studies also found that CheXNeXt's achieved radiologist-level performance or even exceeded them on most pathologies including atelectasis (alveoli deflation in lungs) and pleural effusion.[13]

The recent surge in deep learning and accumulation of mass amounts of healthcare data ranging from MRI images to EHR records extracted by efficient NLP algorithms will continue to shape the landscape of biomedical imaging identification. However, finding a model that will consistently perform accurately across the board has proven a challenge. Another prior study deployed GoogLeNet in a similar classification task and did not find it very accurate (overall approx. 70% accuracy).[16] However, this was done in a study that was extracting labels and reading images which tends to make the task much harder as opposed to classifying off of explicitly-labeled images like the ones in this study. Nevertheless, the current state of computer vision in healthcare, in general, is incredibly encouraging and deserves further note. A rather shallow dive into the current literature, even, would prove the previous statement.

Methods

In this study, two overarching types of deep neural networks were investigated: DenseNets and ResNet. Both offer relatively robust algorithms for classifying images. However, they differ in their specific architecture and numerical algorithm. Prior research involving both types of convolutional neural networks has shown promise in the medical imaging field. However, further research comparing the two and implementing them on this array of images has yet to be completed. A basic overview of the process can be seen in Figure 2.

Dataset

The dataset used in this study is one of the largest accumulations of labeled CXRs (Figure 1). The journey to reach the capability of automated medical image reading remains cloudy and arduous. Thus, this

project is an effort to thoroughly examine this dataset through the lens of CNNs and produce classifications based on already well-trained models. Use of ResNet and a densely connected CNN (Densenet) will be used to prove current, state-of-the-art models' validity in classifying these important images.

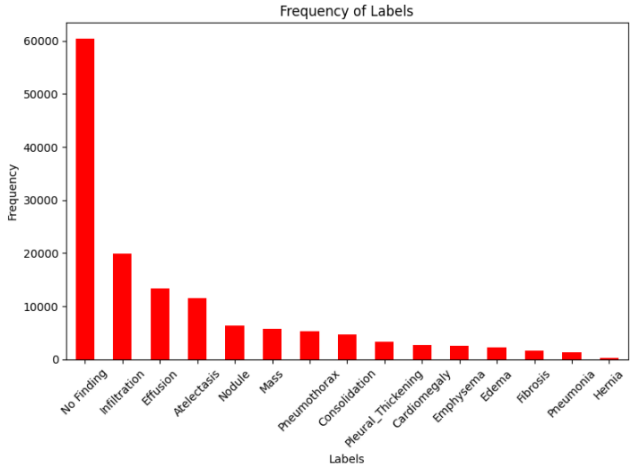


Figure 1: Observed frequency of each condition (some observations elect multiple conditions!).

Much of the preprocessing done for other models was unnecessary in this dataset. Each image was already made into the same dimensions (224 x 224). The dataset is also already standardized in such a way that each image fits into the frame and further quality considerations were not necessary. Further, each image is also a front-facing image. Patients ranged just above 30,000 unique patients and images were collected from 15 years between 1992 to 2015.[17] For each patient, there are multiple images for their various number of checkups within this period. Further, there are 15 possible labels for each image with some images being host to multiple labels. For the research project, we removed all images that were labeled with the "No Finding" class during the preprocessing steps. Example images and labels are provided in Figure 3.

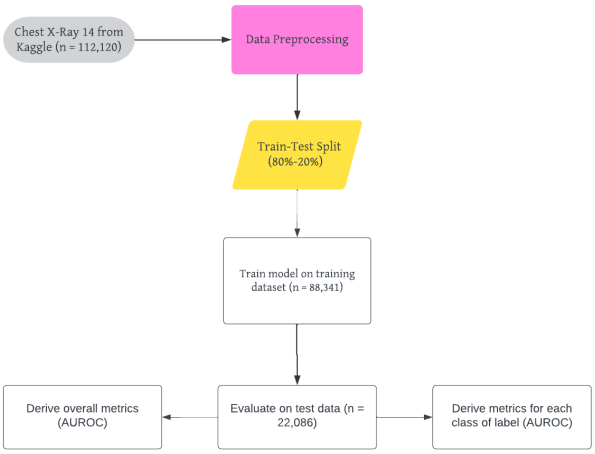


Figure 2: Proposed Flowchart of Method Pipeline (Details will be explored below).

Labels

Labels were reported to be extracted by natural language processing and the study-proposed pipeline yielded 0.9 for precision, recall, and F1 scores.[17] The accuracy in this process accounts for much of the uncertainty that corpora of medical images often have. A recent study focused on radiology report-extraction literature. It was reported that approximately 17% of studies found F1 scores greater than 0.85.[3] Thus, the evaluation metric in this dataset suggests a fairly high level of accuracy relative to the current capabilities.

Deep-learning Architectures

ResNet and DenseNet are both thoroughly-trained models that saw their conception recently. The original ResNet50 is a 50-layer convolutional neural network that was based on the ImageNet dataset.[10] ResNet50V2 is slightly mutated and displayed the best performance on the ImageNet dataset.[11] Prior research with this model has shown quite extensive accuracy in discriminating pneumonia CXRs from normal CXRs (binary classification found an AUC of approx. 0.96). The V2 mutates this architecture slightly to include 48 convolutional layers, 1 MaxPool layer, and 1 average layer. So, what do these layers mean? What do these ResNets look like? Residual networks were originally developed to address the issue of vanishing gradients that might occur during the training of a deep neural

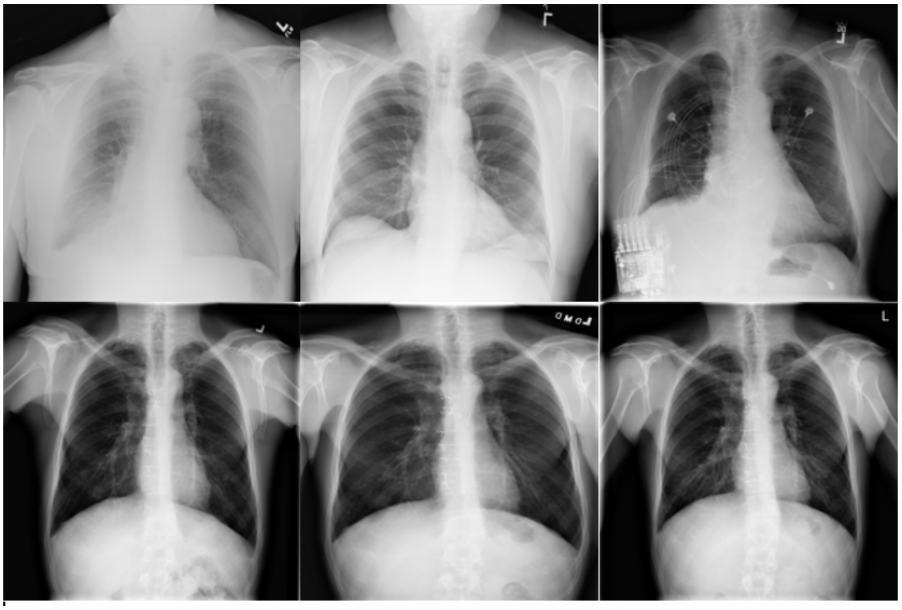


Figure 3: Examples of randomly extracted CXRs with labels as follows (from left to right, top to bottom): No finding, no finding, atelectasis + effusion, pleural thickening, pleural thickening, fibrosis + infiltration, fibrosis + infiltration, and pleural thickening.

network. The error backpropagating back to the earlier layers becomes significantly harder when the networks are deeper (more layers). Instead of the traditional, direct flow that is characteristic of regular networks, residual networks utilize skip connections that act as shortcuts by residual blocks. Thus, the depth of these networks is not as limited, and more "learning" can be done. These residual blocks can be further explored.

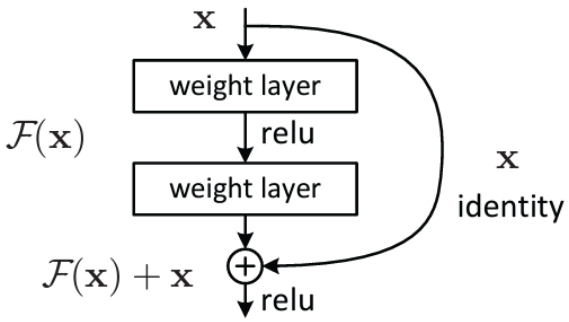


Figure 4: Depiction of a basic shortcut connection (residual block).[6]

Figure 4 depicts an identity mapping from F(x) to F(x) + x, and, in general, the gist of these residual blocks is that pushing this residual, x, to 0 is much easier than fitting a stack of non-linear layers. These layers then simply add the outputs to the already stacked layers and allow the network to skip certain blocks. The premise is incredibly interesting and finding a pertinent deep-learning article that does not reference some sort of residual learning is incredibly rare nowadays.

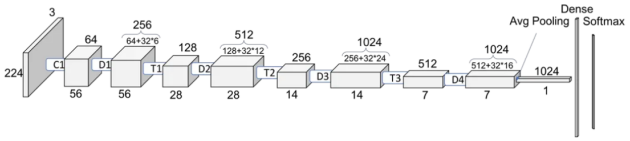


Figure 5: DenseNet-121 structure that has 4 dense blocks and 3 transition blocks.[6]

The other algorithm of choice for this study was DenseNet. This algorithm is significantly more intuitive, however, performs worse per prior research, specifically on medical image classification. DenseNets and ResNets are frequently uttered in the same sentence as both seek to solve similar problems involving the disappearing gradient due to deeper neural networks. So, what's the solution in this case? Simplify connectivity patterns and connect every single layer directly to each other. While a normal convolutional

neural network would be expected to have a normal amount of connections, DenseNets connect each layer to all of its preceding layers and every subsequent layer. The model can also be described as "thin" as each layer quite literally becomes slimmer within the dense blocks due to the small number of filters. The model of choice in this study is DenseNet-121 which is perhaps the most simple of the DenseNet variants that were trained on the ImageNet dataset (Figure 5). An easy way to mathematically visualize this algorithm is concatenation. While ResNets combines features through summation, DenseNets do so on concatenation. Thus, an alternate composite equation as opposed to the one offered by ResNets might look something like[15]:

$$x_l = H_l([x_0, x_1, ..., x_{l-1}])$$

Transition layers are then used to address the differing sizes of feature maps that are passed from block to block. Computationally, DenseNets might be more costly than ResNets as each layer receives and passes feature maps to all other layers. However, the DenseNet architecture is pretty exciting.

Metrics and Evaluations Used

The overall metric of choice was the area under the receiving operating characteristic curve. This error metric can quantify a classifier's performance across various thresholds of sensitivity and fall-out and maps the overall discriminative ability of the model at different points in this threshold pairing. Thus, one can also glean the sensitivity and false positive values from this AUC value.[1] The metric is most frequented in biomedical research in binary classification tasks, however, can be applied to multiclass labels. Further, this metric emphasizes the magnitude of the difference between false positives and false negatives. Accuracy simply does not discriminate and views both situations as similar. However, this is often not the case in a medical setting. False negatives are usually much more momentous mistakes than false positives. Additionally, the imbalance in this dataset suggests that a metric that is less affected by imbalance should be selected. Specifically, in this project, each class of the 15 possible diagnoses was given its own AUC, and further evaluation was performed by averaging these AUC values. For further intuition, a receiver operating characteristic curve with a slope of 1 indicates a 50-50 chance of a correct classification while a graph with a vertex at (0,1) depicts flawless classification (Figure 6).

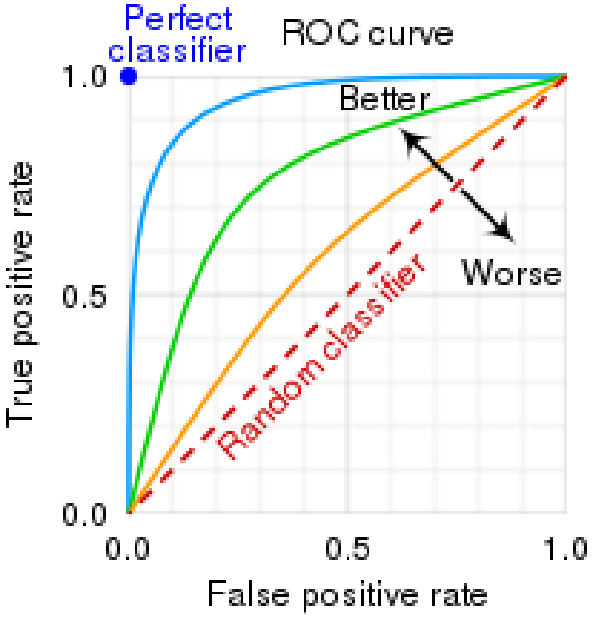


Figure 6: Example of ROC curves that are equivalents of random guesses and a perfect classifier.[2]

Loss and Optimizer

The ADAM optimizer and binary cross-entropy loss function were used for this study. ADAM was presented as a stochastic optimizer and well-suited for problems with large datasets and many parameters.[8] The most notable features of this algorithm include the adaptive learning rates that allow different parameters to have different learning rates and the utilization of momentum that accumulates from past gradients. Prior research has shown ADAM to possess incredibly efficient optimization and fast convergence with minimal tuning required.

Binary cross-entropy loss was chosen as the loss function due to its application in multi-class classification tasks. The metric treats each class independently as a binary classification problem and is especially helpful when multiple classes can be assigned to a single observation as is the case with this study. It is often used in conjunction with the sigmoid activation function for the same reason. Likewise, the AUC calculated additionally treats each class independently and evaluates labels in a more "binary" manner.

Table 1: Comparison of AUC scores

Pathology	DenseNet	ResNet	Modified ResNet
Atelectasis	0.737	0.760	0.802
Cardiomegaly	0.901	0.793	0.898
Effusion	0.824	0.828	0.875
Infiltration	0.679	0.652	0.710
Mass	0.782	0.766	0.839
Nodule	0.721	0.668	0.727
Pneumonia	0.659	0.665	0.717
Pneumothorax	0.821	0.781	0.871
Consolidation	0.696	0.751	0.793
Edema	0.840	0.806	0.887
Emphysema	0.857	0.775	0.905
Fibrosis	0.756	0.723	0.801
Pleural Thickening	0.713	0.717	0.772
Hernia	0.797	0.704	0.746
Mean	**0.770**	**0.742**	**0.810**

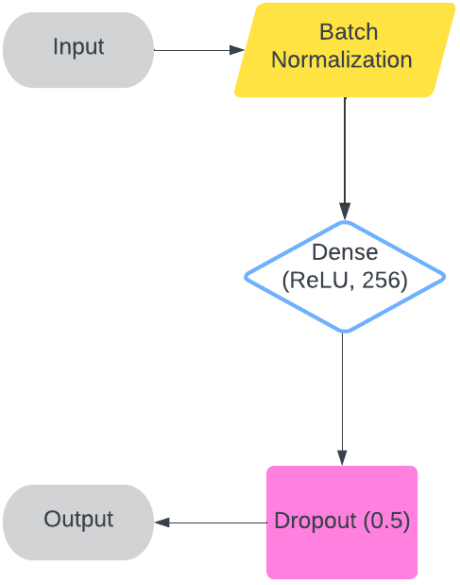
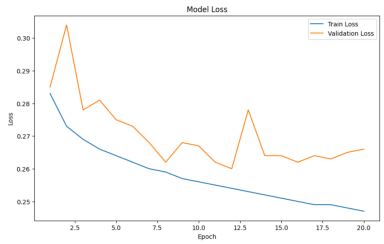
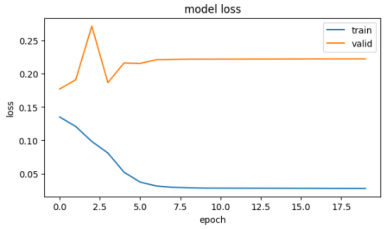


Figure 7: Basic block to combat the extreme overfitting of the naive ResNet.

Results

DenseNet-121 and ResNet50V2 were trained and tested as separate models. For both, an initial learning rate was set at 1e-4 because these are both trained models and require fine-tuning to the dataset of choice. Further, the learning rates were reduced by a factor of 0.1 when the loss plateaued with a minimum learning rate of 1e-8. Other configurations included the number of epochs that were run on each model to train it (20 for both) and standard β-coefficients for the ADAM optimizer ($\beta 1 = 0.9$, $\beta 2 = 0.999$) The results in terms of AUC are shown in Table 1. The training loss compared to the validation loss per epoch is also shown in Figure 8.





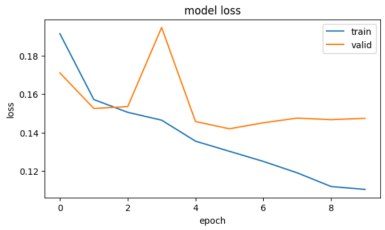


Figure 8: From top to bottom: DenseNet validation vs. train loss, original ResNet, and our modified ResNet.

The graph for the ResNet depicts some sort of overfitting. Thus, a modified algorithm was chosen by adding a basic block depicted in Figure 8 and freeing up the last 10 layers of the ResNet architecture to fit the dataset's more complex features. This elevated the performance of the ResNet to higher than that of the DenseNet by approximately 5.2 % and higher than the regular ResNet by approximately 9.2%. Thus, the relatively new architectures can perform quite similarly on the ChestXRay-14 image classification data (95% CI for regular ResNet: [0.710, 0.774], DenseNet121: [0.728, 0.812], modified ResNet: [0.770, 0.850]). A separate study found that unassisted radiologists had an average AUC of $\tilde{0}.713$ across various CXR-diagnosed conditions.[21] This provides a further benchmark into just how powerful of an assist these new algorithms can be. Figure 9 depicts a couple of CXRs with associated prediction probabilities as well as their clinically correct diagnoses.

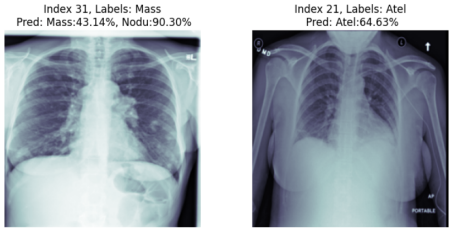


Figure 9: Prediction probabilities for predicted class and true classification from the original dataset

Discussion and Future Work

The mean AUC of 0.810 for our modified ResNet almost rivals that of cutting-edge networks such as CheXNet (mean AUC 0.84). Thus, this network also likely rivals the performance of radiology experts. The newer model significantly outperformed in every pathology besides nodules and hernias (> 0.025 AUC difference). Additionally, the fact that the addition of the basic block performed at the greatest capacity out of the three indicates that mixing the architectures of residual learning and dense layers might be a good idea moving forward. The highest classification accuracies in the modified model include emphysema, cardiomegaly, effusion, pneumothorax, and edema (AUC ≥ 0.85).

One of the more applicable future directions to the scope of this project is the maximization of AUC which can be referred to as deep AUC maximization (DAM). A recent publication cited an AUC margin loss that builds off of the AUC squared loss formulation.[19] This metric performed slightly better than past loss metrics and proves that some sort of implementation in training deep learning algorithms on medical imaging datasets might warrant consideration.

Further, the models that were trained demonstrated significant incapability when it came specifically to pneumonia, infiltration, and nodules (concerning AUC). Interestingly enough, prior sources in the introduction and recent research by Rajpurkar Et al. show an AUC > 0.75 in classifying the three conditions.[12] Since the two architectures of choice were trained on the ImageNet dataset, it can also be reasoned that a model strictly trained on CXRs might perform slightly better (e.g. CheXNet). However, this would require much diligence in terms of gathering the correct resources and time in preparing the training data. This model fine-tuning and image gathering could also be accomplished given more time.

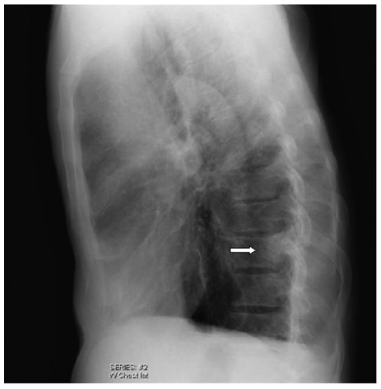


Figure 10: Depiction of a nodule displayed by a side-chest view that might otherwise be missed if shown by a frontal view.

An additional consideration is the history of the patient and other perspectives of the chest cavity. Prior research has already shown that radiologists perform worse without prior knowledge of the patient's past pathologies or additional views of the chest (side views).[14] This notion can be generalized to deep learning models as well. Thus, integrating this previous information somehow into future predictive algorithms would likely afford much more robust classifications. Figure 10 shows a view that might disrupt our current model from Raoof et al.

Contributions

This research project was a collaborative effort between June Yun and Jonathan Lin, who contributed significantly to different aspects of the study. June Yun was mainly responsible for the implementation and fine-tuning of the DenseNet architecture and literature review of material related to the model. Jonathan took charge of implementing and optimizing the ResNet architecture, specifically the ResNet50V2 variant as well as doing relevant literature reviews. Both collaborated on the report and GitHub repository.

Code

Code used in analysis and visualization can be found in the GitHub repository at this link.

References

[1] Anonymous. *Classification: ROC Curve and AUC — Machine Learning — Google for Developers.* Google, 2023. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

[2] Anonymous. *Receiver Operating Characteristic.* Wikipedia, December 3, 2023. https://en.wikipedia.org/wiki/Receiver_operating_characteristic

[3] Casey, Arlene, et al. *A Systematic Review of Natural Language Processing Applied to Radiology Reports - BMC Medical Informatics and Decision Making.* BioMed Central, June 3, 2021. https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-021-01533-7

[4] Chouhan, Vikash, et al. *A Novel Transfer Learning Based Approach for Pneumonia Detection in Chest X-Ray Images.* MDPI, January 12, 2020. https://www.mdpi.com/2076-3417/10/2/559

[5] Fornell, Dave. *Mismatch between Radiologist Shortages, Rising Exam Volumes a Growing Concern in Medical Imaging.* Radiology Business, December 1, 2023. https://radiologybusiness.com/topics/healthcare-management/healthcare-staffing/mismatch-between-radiologist-shortages-rising-exam-volumes-growing-concern-medical-imaging

[6] He, Kaiming, et al. *Deep Residual Learning for Image Recognition - Arxiv.Org.* arXiv, December 10, 2015. https://arxiv.org/pdf/1512.03385.pdf

[7] Huang, Gao, et al. *Densely Connected Convolutional Networks — IEEE ... - IEEE Xplore.* IEEE Xplore, 2017. https://ieeexplore.ieee.org/document/8099726

[8] Kingma, Diederik P., and Jimmy Ba. *Adam: A Method for Stochastic Optimization.* arXiv, January 30, 2017. https://arxiv.org/abs/1412.6980

[9] Meedeniya, Dulani, et al. *Chest X-Ray Analysis Empowered with Deep Learning: A Systematic Review.* Applied Soft Computing, September 2022. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9393235/

[10] Prusty, Sashikanta, et al. *ResNet50V2: A Transfer Learning Model to Predict ... - IEEE Xplore.* IEEE Xplore, 2022. https://ieeexplore.ieee.org/document/10076678

[11] Rahimzadeh, Mohammad, and Abolfazl Attar. *A Modified Deep Convolutional Neural Network for Detecting COVID-19 and Pneumonia from Chest X-Ray Images Based on the Concatenation of Xception and Resnet50v2.* Informatics in Medicine Unlocked, May 26, 2020. https://www.sciencedirect.com/science/article/pii/S2352914820302537

[12] Rajpurkar, Pranav, et al. *CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning*. arXiv, December 25, 2017. https://arxiv.org/abs/1711.05225

[13] Rajpurkar, Pranav, et al. *Deep Learning for Chest Radiograph Diagnosis: A Retrospective Comparison of the CheXNeXt Algorithm to Practicing Radiologists*. PLOS Medicine, November 20, 2018. https://journals.plos.org/plosmedicine/article?id=10.1371%2Fjournal.pmed.1002686

[14] Raoof, Suhail, et al. *Interpretation of Plain Chest Roentgenogram*. Chest, February 2012. https://pubmed.ncbi.nlm.nih.gov/22315122/

[15] Ruiz, Pablo. *Understanding and Visualizing DenseNets*. Medium, October 18, 2018. https://towardsdatascience.com/understanding-and-visualizing-densenets\protect\penalty\z@-7f688092391a

[16] Shin, Hoo-Chang, et al. *Learning to Read Chest X-Rays: Recurrent Neural Cascade Model for Automated Image Annotation*. arXiv, March 28, 2016. https://arxiv.org/abs/1603.08486

[17] Wang, Xiasong, et al. *Chestx-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks On ....* CVPR, 2017. https://openaccess.thecvf.com/content_cvpr_2017/papers/Wang_ChestX-ray8_Hospital-Scale_Chest_CVPR_2017_paper.pdf

[18] White, Ben. *The Radiologist Shortage Is Here: Ben White*. Ben White — Medicine & Miscellany, October 17, 2023. https://www.benwhite.com/radiology/the-coming-radiologist-shortage-is-here/

[19] Yuan, Zhuoning, et al. *Large-Scale Robust Deep AUC Maximization: A New Surrogate Loss and Empirical Studies on Medical Image Classification*. CVF Open Access, January 1, 1970. https://openaccess.thecvf.com/content/ICCV2021/html/Yuan_Large-Scale_Robust_Deep_AUC_Maximization_A_New_Surrogate_Loss_and_ICCV_2021_paper.html

[20] Zhou, Tao, et al. *Dense Convolutional Network and Its Application in Medical Image Analysis*. BioMed Research International, April 25, 2022. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9060995/

[21] Seah, Jarrel, et al. *Effect of a Comprehensive Deep-Learning Model on the Accuracy of Chest X-Ray Interpretation by Radiologists: A Retrospective, Multireader Multicase Study*. The Lancet Digital Health, July 1, 2021. https://www.sciencedirect.com/science/article/pii/S2589750021001060.