Disease Classification in Patient X-Rays using Deep Learning Approaches

Srinivas Natarajan
SCAI
Arizona State University
Tempe AZ US
snatar28@asu.edu

Saiteja Alaparthi
SCAI
Arizona State University
Tempe AZ US
salapart@asu.edu

Tanushi Ahuja
SCAI
Arizona State University
Tempe AZ US
tahuja4@asu.edu

Darshan Govindaraj
SCAI
Arizona State University
Tempe AZ US
dgovind5@asu.edu

ABSTRACT

The interpretation of X-rays is critical for identifying various medical conditions and their precursors. In this project, we explore the potential of using computer systems to aid in the diagnosis process using various deep learning approaches for efficient disease localization in patient X-rays. Specifically, we use standard boosting techniques, deep learning architectures, specialized architectures for the medical field, and transformer-based approaches. Our results demonstrate the effectiveness of deep learning approaches for disease localization in X-rays, with implications for improving the accuracy and efficiency of medical diagnoses.

1. Introduction

Medical imaging has revolutionized healthcare by enabling the early detection and diagnosis of a wide range of diseases. However, the increasing number of medical images generated has resulted in a growing demand for skilled radiologists and technicians, which has become a bottleneck in the healthcare system. This has highlighted the need for computer-aided diagnosis (CAD) systems to help clinicians make accurate diagnoses in a timely and cost-effective manner.

Deep learning approaches have shown great potential in the field of medical imaging, particularly in the localization of diseases in X-rays. By using various preprocessing techniques and classification algorithms, deep learning models can effectively identify and localize diseases in patient X-rays with a high degree of accuracy. This can have significant implications for improving the efficiency and accuracy of medical diagnoses, potentially leading to earlier detection and treatment of diseases.

In this project, we aim to explore the potential of deep learning approaches for disease localization in patient X-rays, with a focus on four different approaches: standard boosting techniques, generalized deep learning architectures, and specialized architectures for the medical field, and transformer-based approaches. Through this investigation, we hope to demonstrate the effectiveness of deep learning models in aiding clinicians in the diagnosis of medical conditions using X-rays.

2. Related Work

In the process of tackling this medical classification and localization problem, we encountered a few major types of approaches to create an accurate model. The main approach and a theme that is seen through most work done in this field involves the use of pre-trained architectures like Residual neural networks (ResNet38/50) [3]. These approaches are an industry standard due to the lack of large amounts of data to tackle medical problems, making transfer learning a more viable method. Most papers we came across either simply apply this model to approach this task or train very few layers at the end of the network based on the new data. On the other hand, we unfreeze a lot of layers of the models and add additional layers at the end before training, helping finetune the model.

Deviating a little further from standard CNNs, we see work done where conventional machine learning classifiers are combined with the architecture to provide an alternative approach [4]. This includes using SVMs and linear classifiers as substitutes for the final sigmoid layer of deep learning models. This approach does show some success but is lacking in the use of global context available in the image and in the speed of computation. This is where a variety of different models have been developed, making a medical system that is viable in real time. This also cuts down the computation costs associated with image classification. One such approach is the use of lightweight models like the MobileNetv2 [6] that are designed to run on smaller, less powerful systems. These types of models achieve a similar level of performance as the initial work [3],[5] mentioned but are far faster in run time and are less resource intensive. The other way to approach this is by

using modified architectures incorporating atrous convolutions as in [10]. This type of work on efficient systems was more important back when systems were less powerful but current systems have all but caught up in terms of prediction speed. Even larger models like ResNet 50 have pretty fast prediction times.

With the advancement of computing technology, it has become feasible to use a transformer based system consisting of two main components: encoder networks to extract information from the data and decoder systems that break down this encoding and classify the output. We see a growing usage of partial context systems such as residual attention networks [8] that use attention mechanisms to both category wise [8] and globally [9] to capture contextual information to be used for predictions. This also leads to the possibilities of full transformer networks for image classification tasks where originally they were limited to NLP problems. This trend was started with Microsoft's Swin transformer architecture which was created to tackle multiple modalities and not just NLP tasks. This can be seen in systems like RATCHET [5] where two separate architectures are combined (CNN+Transformer in this example) to offer better results than traditional methods. This is an approach that we aim to explore further in our project.

We also explore more novel approaches, like incorporation of patient metadata to augment image based prediction systems. This is because papers found a positive correlation between patient data like age, blood type and gender. But current work on this has yet to show better results than traditional image systems. A sample architecture has been mentioned below:

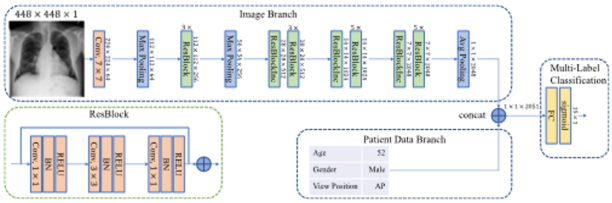

Fig. 1: Combined approach architecture

Finally, it is important to address any real world side-effect our project may have, such as intrinsic biases. This is why we also explored studies on existing datasets [7] to analyze how data can be skewed based on patient information and factors. This is essential to create a system that can help address any societal and demographic imbalances and make it overall, more objective.

3. Data

3.1. Exploratory Data Analysis

For this problem, we used the Chest X-ray 14 dataset created by the National Institute of Health. This is a dataset that consists of nearly 110,000 X-ray images with two main parts of metadata. Each x-ray can consist of multiple health conditions from the following classes: 'Atelectasis', 'Consolidation', 'Infiltration', 'Pneumothorax', 'Edema', 'Emphysema', 'Fibrosis', 'Effusion', 'Pneumonia', 'Pleural Thickening', 'Cardiomegaly', 'Nodule', 'Mass', 'Hernia' and 'No Findings'. It also consists of boundaries for each of these conditions for localization problems.

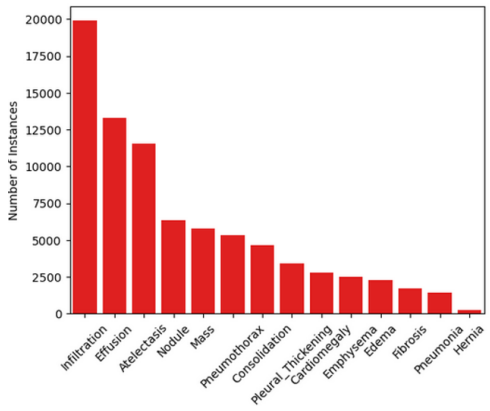

Fig. 2: Class distribution of the chest conditions

Visualizing the class distribution in the data, we can see that there is a significant imbalance between the majority and minority classes. There are three possible approaches forward: first is to ignore this problem and go ahead with model creation, although this may skew the results. Second, weigh each class inversely according to its difficulty to detect and select a certain number of instances from each class. The third is to reduce the number of overall instances by undersampling the data.

3.2. Data Processing

We reduce the overall number of instances by choosing only 40,000 X-rays, making sure to include all available instances of the minority classes. This proved to be the most effective method. The next step is to process our images and labels by defining custom data loaders. Each image is resized to a 128x128 dimension, grayscaled, and converted to tensors. More experimentation on the varying sizes of the image along with augmentations will be performed in the future and will be a part of our ablation study. We form labels for each image by one-hot encoding a vector of size 14, with "1"' representing the presence of the disease and "0"' representing its absence.

The data is then split in the ratio 80-10-10 for the training, testing, and validation sets, respectively. This is important because we use validation loss as a measure for our early stopping callback. This helps us prevent overfitting.

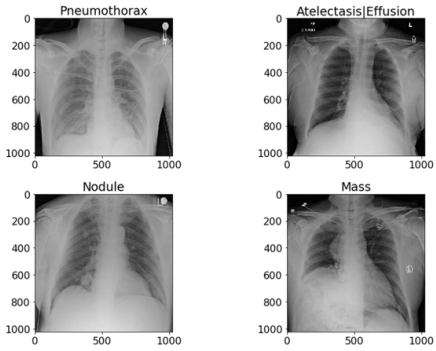

Fig. 3: Representative samples of some classes

4. Methods

The work done will be divided into two main sections: experimentation on the preprocessing performed on the image which involves various input sizes, channel sizes, data balancing techniques and augmentation techniques which will form our ablation study and different approaches to deep learning approaches that involve standard convolutional neural networks[3], combined approaches of neural networks and traditional classification approaches like Gaussian Mixture Models (GMMs) or K Means [4], modified approaches that focus on the speed of prediction by leveraging shallower networks and techniques like atrous layers [6,10] and the stretch goal for our project will be the use of the next generation transformer models that aim to bridge the gap between the text and image modalities.

4.1. Traditional Deep Learning

In this section, we will explore traditional deep learning models that use convolutions and their variations. This will include: ResNets and MobileNet. We modify the architecture of these models to include an average pooling layer. Due to overfitting concerns, we added dropout layers as these are relatively dense architectures.

Layer (type)	Output Shape	Param #
resnet50 (Model)	(None, 4, 4, 2048)	23581440
global_average_pooling2d_2 (	(None, 2048)	0
dropout_3 (Dropout)	(None, 2048)	0
dense_3 (Dense)	(None, 512)	1049088
dropout_4 (Dropout)	(None, 512)	0
dense_4 (Dense)	(None, 13)	6669

Total params: 24,637,197
Trainable params: 24,584,077
Non-trainable params: 53,120

Fig. 4: ResNet50 Model Summary

Layer (type)	Output Shape	Param #
mobilenet_1.00_128 (Model)	(None, 4, 4, 1024)	3228288
global_average_pooling2d_1 (	(None, 1024)	0
dropout_1 (Dropout)	(None, 1024)	0
dense_1 (Dense)	(None, 512)	524800
dropout_2 (Dropout)	(None, 512)	0
dense_2 (Dense)	(None, 13)	6669

Total params: 3,759,757
Trainable params: 3,737,869
Non-trainable params: 21,888

Fig. 5: Mobile Net Model Summary

4.2. Mixed Learning

This section will look into the use of deep learning networks to generate a set of features that will be used by more traditional classifiers. For each image, information is extracted by a convolutional neural network, and the dimensionality is reduced using PCA before being passed as features to models like Random Forest, GMMs, and K Means for classification. This approach is something suggested in previous research [4], but we will expand upon the SVM approach taken.

4.3. Transformer Models

This section will be part of our stretch goals, where we will try to expand upon recent advances in transformer technology, converting them from an architecture initially suited for only natural language tasks and bridge the modality to image tasks as well. For this purpose, we will be basing our work on the RATCHET transformer system [5] and Microsoft's new Swin architecture, which uses sliding windows to retain attention in images.

For practical purposes, we will use pre-trained network encoder weights as the default 'ImageNet' weights. The model consists of three overall sizes, Swin-T (0.25x size), Swin-S (0.5x size), and Swin-L (2x size). Increasing the size of the model comes at the risk of overfitting due to the amount of data present for the minority classes, so this will have to be monitored carefully.

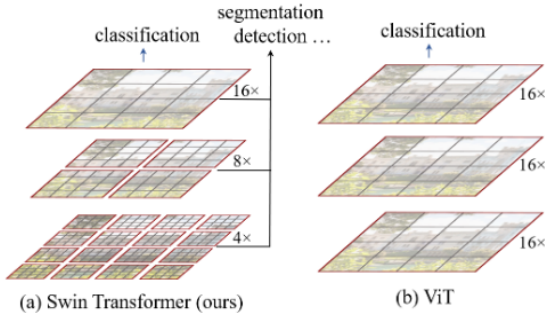

Fig. 6: Swin Transformer Architecture vs. Traditional Vision Transformers

5. Results

Our results will consist of benchmarking standard models and contrasting them with the improved model we have constructed. We will measure the AUC and F-score performances of the models and aim to create a novel system that can achieve a near state of the art (SoTA) performance.

We modified the existing ResNet and MobileNet models with additional Global average pooling layers, Dropouts and a final output layer of 14 classes. They were trained with the 'Adam' optimizer with a learning rate of 0.01 and a Softmax activation layer as the output.

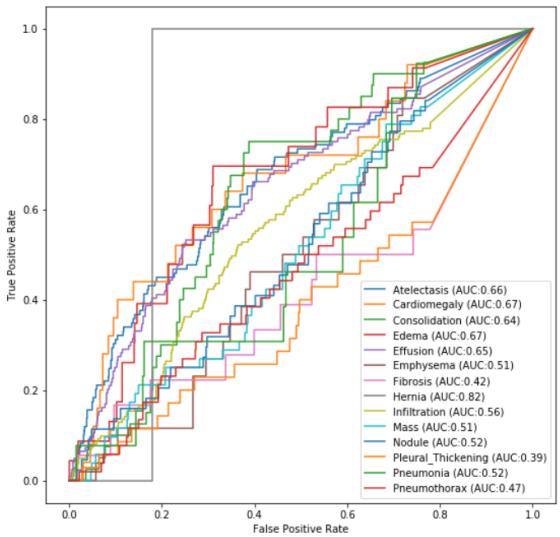

Fig. 7: AUC scores for the modified ResNet50 model

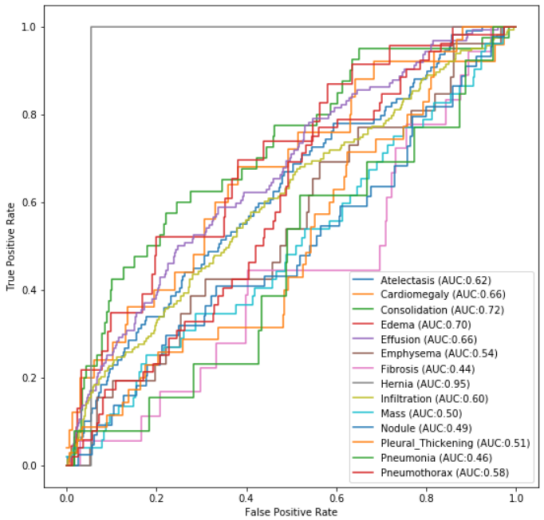

Fig. 8: AUC scores for the modified MobileNet model

Disease / Model	Mobile Net	ResNet 50	Dense Net	Swin - L
Atelectasis	0.75	0.76	0.834	0.84
Cardiomegaly	0.85	0.83	0.913	0.93
Consolidation	0.70	0.68	0.818	0.87
Edema	0.84	0.87	0.914	0.92
Effusion	0.82	0.83	0.905	0.82
Emphysema	0.87	0.90	0.936	0.93
Fibrosis	0.79	0.79	0.786	0.94
Hernia	0.67	0.70	0.665	0.85
Infiltration	0.76	0.76	0.886	0.71
Mass	0.68	0.76	0.790	0.89
Nodule	0.72	0.71	0.813	0.87
Pleural Thickening	0.66	0.66	0.807	0.90
Pneumonia	0.83	0.88	0.906	0.93
Pneumothorax	0.75	0.76	0.834	0.85

Table 1: Comparison of AUC scores of model

With the complexity and quantity of data available in the NIH Chest X-ray dataset, we can see from the results that having a deeper network can improve results. We compare three types of neural networks and the best performing transformer model. The MobileNet model represents a lightweight, shallow model that focuses on fast processing times over performance. The Resnet 50 model is a standard for image analysis tasks but has been fine tuned for this particular task. Coupled with dropout layers, it prevents overfitting. The Densenet 121 is the deepest model we use, and modifications to its structure help prevent overfitting, which is an inherent risk with its depth. This is the best performing model on our test set.

We also experimented with various batch sizes, color channels, image sizes, and learning rates. We tested their effectiveness based on their convergence rates and the loss over epochs.

5.1. Image Size

Image size is an important factor, as larger images allow us to extract more information, but this comes at the cost of computational and memory overhead. This is why we conducted a study of the various sizes of images to feed to the model and measured how the model performed based on its convergence and loss. We use standard image sizes ranging from 1024x1024 to 64x64. We also include a more irregular 224x224 as this is the standard window size for vision transformer models.

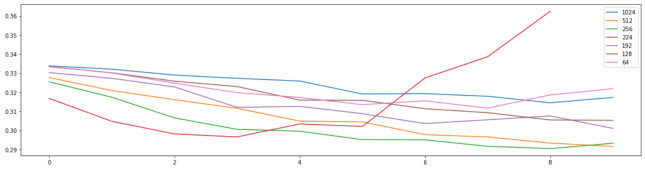

Fig. 9: Loss over epochs for different image sizes

We see that the best performing sizes are 512x512 and 256x256, with both of them having pretty similar results. We went forward with 256x256 images as it saves on the memory and time we need to compile and train our models. For the Swin Transformers, we are limited to a size of 224x224 due to pretrained weight constraints.

5.2. Color Channels

The two main color modes we have available are the standard RGB and a grayscaled version of the image dataset. While normal image processing tasks preferably use grayscaled images to scale down on computational cost, medical scans are more sensitive to changes in intensity. This can be seen in CT scan images, where the intensity measured in Hounsfield units is of great significance.

Image Size/ Color Channel	224x224	256x256
RGB	0.340	0.3385
Grayscaled	0.33	0.33823

Table 2: Comparison of model losses

Although the loss difference is miniscule, we opt to use RGB images for our models, as this also enables us to use pretrained networks like ResNet50 on the ImageNet data.

6. Discussion

We can see from the results why convolutional neural networks have been the de facto standard in image analysis for a good reason. Convolutional neural networks pre-trained on existing data prove to be the most accurate and reliable method. But we also find that Vision based transformers have improved leaps and bounds from their inception, with the larger versions of the Swin transformer. (Swin-L) proving to have SoTA results. It is also clear that these transformer models are far more sensitive to tuning of hyperparameters like window size, patch size, and model depth. This can be seen in the difference between the various sizes of Swin transformers, from the tiny Swin-L (0.25x Swin-B) to the enormous Swin-L (2x Win-B). But they also offer near SoTA performance, with an average AUC of 0.81, approaching the best results to date.

While the inclusion of metadata involving a patient's records can seem intuitively productive, current systems have not been able to leverage them. This presents a future avenue for exploration to improve systems to combat serious medical issues.

7. Conclusion

This project emphasizes the potential of machine learning systems and their potential to augment medical professionals in the diagnosis of serious throasic issues. While current datasets are highly skewed, resembling anomaly detection tasks, improvements in the organization of data and the availability of medical information have greatly increased the feasibility of computer aided diagnosis. The use of transformer based models that can leverage context from various parts of the image has revolutionized the ability to accurately predict medical conditions.

In conclusion, we have created and fine tuned both models that have been the standard for a decade and contrasted them to more recent developments in deep learning to address a pressing issue in current society. We created a model capable of identifying and localizing serious thoracic issues with an accuracy rivaling state of the art current systems.

In the future, we can expand this system to include patient data and previous medical conditions to better use medical history in the diagnosis of diseases, as doctors do. This system can also be augmented to use better medical techniques used by professionals to improve its performance and learn from its mistakes.

REFERENCES

[1] Giełczyk A, Marciniak A, Tarczewska M, Lutowski Z (2022) Pre-processing methods in chest X-ray image classification. PLoS ONE 17(4): e0265949. https://doi.org/10.1371/journal.pone.0265949

[2] Ivo M. Baltruschat, Hannes Nickisch, Michael Grass, Tobias Knopp, and Axel Saalbach. 2019. Comparison of deep learning approaches for multi-label chest X-ray classification. Scientific Reports 9, 1 (2019). DOI:http://dx.doi.org/10.1038/s41598-019-42294-8

[3] I. M. Baltruschat, H. Nickisch, M. Grass, T. Knopp, and A. Saalbach, "Comparison of deep learning approaches for multi-label chest X-ray classification," *Nature News*, 23-Apr-2019. https://www.nature.com/articles/s41598-019-42294-8.

[4] I. Allaouzi and M. Ben Ahmed, "A Novel Approach for Multi-Label Chest X-Ray Classification of Common Thorax Diseases," in *IEEE Access*, vol. 7, pp. 64279-64288, 2019, doi: 10.1109/ACCESS.2019.2916849. https://ieeexplore.ieee.org/abstract/document/8719904

[5] B. Hou, G. Kaissis, R. Summers, and B. Kainz, "Ratchet: Medical Transformer for chest X-ray diagnosis and reporting," *arXiv.org*, 15-Sep-2021 https://doi.org/10.48550/arXiv.2107.02104.

[6] A. Souid, N. Sakli, and H. Sakli, "Classification and predictions of lung diseases from chest X-rays using MobileNet V2," *MDPI*, 18-Mar-2021. https://www.mdpi.com/2076-3417/11/6/2751.

[7] L. Seyyed-Kalantari, G. Liu, M. McDermott, I.Y. Chen, and M. Ghassemi, "Chexclusion: Fairness gaps in deep chest X-ray classifiers," *arXiv.org*, 16-Oct-2020. https://doi.org/10.48550/arXiv.2003.00827. https://arxiv.org/abs/2003.00827v2

[8] Q. Guan and Y. Huang, "Multi-label chest X-ray image classification via category-wise residual attention learning," *Pattern Recognition Letters*, vol. 130, pp. 259–266, 2020. https://www.sciencedirect.com/science/article/abs/pii/S0167865518308559

[9] Yu, K., Ghosh, S., Liu, Z., Deible, C., and Batmanghelich, K. "Anatomy-Guided Weakly-Supervised Abnormality Localization in Chest X-rays", *arXiv.org*, 25-Sep-2022, https://doi.org/10.48550/arXiv.2206.12704, https://arxiv.org/abs/2206.12704

[10] Ö. Özdemir and E. B. Sönmez, "Weighted Cross-Entropy for Unbalanced Data with Application on COVID X-ray images," *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, Istanbul, Turkey, 2020, pp. 1-6, doi: 10.1109/ASYU50717.2020.9259848.

[11] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021, August 17). *Swin Transformer: Hierarchical vision transformer using shifted windows*. arXiv.org. Retrieved April 13, 2023, from https://arxiv.org/abs/2103.14030