Deep Learning for Chest X-Ray Medical Diagnosis of Lung Cancer

1 Problem Definition and Understanding

Lung cancer diagnosis is crucial for timely treatment, yet accurately interpreting chest X-rays remains a challenging task. By harnessing the power of deep learning, specifically through the use of Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs), this innovative system automates the diagnostic process. Utilizing the NIH Clinical Center dataset, which comprises 100,000 annotated chest X-ray images, the technology aims to achieve high accuracy and generalization in disease prediction. Early detection, enabled by this system, improves patient outcomes and enhances healthcare efficiency. Its scalability and accessibility make it applicable across diverse clinical settings, ensuring broader access to accurate diagnosis. This represents a significant step towards leveraging technology to revolutionize lung cancer diagnosis, bridging the gap between innovation and clinical practice.

The following 14 conditions can be present in a chest X-ray: `Cardiomegaly, Emphysema, Effusion, Hernia, Infiltration, Mass, Nodule, Atelectasis, Pneumothorax, Pleural Thickening, Pneumonia, Fibrosis, Edema, and Consolidation`. Each chest X-ray can exhibit one or more of these conditions simultaneously, and the presence of one condition does not necessarily exclude the presence of another. The labels are medically relevant and crucial for accurate diagnosis, but they require the model to learn from multiple labels per image rather than a single disease classification.

Predicting the absence of labels requires sophisticated balancing of label frequencies. However, in our case, since the ”No Findings” label skews the data significantly, we forgo it entirely to minimize incorrect predictions.

2 Data Preparation and Preprocessing

	Image Index	Patient ID	Cardiomegaly	Emphysema	Effusion	Hernia	Infiltration	Mass	Nodule	Atelectasis	Pneumothorax	Pleural_Thickening	Pneumonia	Fibrosis	Edema	Consolidation	FilePath
57468	00014285_001.png	14285	0	0	0	0	0	0	0	0	0	0	0	0	0	0	../input/data/images_007/images/00014285_001.png
896	00000213_005.png	213	0	0	0	0	0	0	0	0	0	0	0	0	0	0	../input/data/images_001/images/00000213_005.png
84198	00020767_000.png	20767	0	0	0	0	0	0	1	0	0	0	0	0	0	0	../input/data/images_009/images/00020767_000.png
69659	00017215_000.png	17215	0	0	0	0	0	0	0	0	0	0	0	0	0	0	../input/data/images_008/images/00017215_000.png
95650	00025234_013.png	25234	0	0	1	0	1	0	0	0	0	0	0	0	0	0	../input/data/images_011/images/00025234_013.png
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
78787	00019390_003.png	19390	1	0	0	0	0	0	0	1	0	0	0	0	0	0	../input/data/images_009/images/00019390_003.png
108716	00029594_000.png	29594	0	0	0	0	0	0	0	0	0	0	0	0	0	0	../input/data/images_012/images/00029594_000.png
68818	00017028_000.png	17028	0	0	0	0	0	0	1	0	0	1	0	0	0	0	../input/data/images_008/images/00017028_000.png
82099	00020243_009.png	20243	0	0	0	0	1	0	0	0	0	0	0	0	0	0	../input/data/images_009/images/00020243_009.png
88762	00022072_008.png	22072	0	0	0	0	0	0	0	0	0	0	0	0	0	0	../input/data/images_010/images/00022072_008.png

5000 rows × 17 columns

The dataset comprises various elements including images, labels, follow-up visits, age, gender, and view position. For the purpose of this project, only the images and labels are utilized for disease prediction. Patient age, gender, and view position are excluded from the analysis as they do not influence the disease identification process on the X-ray images. The focus is strictly on the image content and the multi-label classification applied to these images.

Data Preprocessing: The following steps were taken to process our input data frame:

- The ChestXRay dataset has a corresponding dataset that explicitly labels bad labels, which are filtered out through an image_label_map.
- Unwanted columns can be dropped; in our case, we made the decision to drop ’No Finding’, allowing us to focus on the analysis of pathologies. Since this is a multi-label classification, we can determine the lack of a disease solely by low AUC or prediction scores for each class.

- Filepath listing: Images are given their full path for easy access across multiple folders. Since the images are extracted and stored in separate bins.
- Furthermore, the labels are encoded across multiple features, thereby accomplishing encoding without having to do encoding within the class feature 'Finding Labels'.

	Image Index	Patient ID	Cardiomegaly	Emphysema	Effusion	Hernia	Infiltration	Mass	Nodule	Atelectasis	Pneumothorax	Pleural_Thickening	Pneumonia	Fibrosis	Edema	Consolidation
57468	00014285_001.png	14285	0	0	0	0	0	0	0	0	0	0	0	0	0	0
896	00000213_005.png	213	0	0	0	0	0	0	0	0	0	0	0	0	0	0
84198	00020767_000.png	20767	0	0	0	0	0	0	1	0	0	0	0	0	0	0
69659	00017215_000.png	17215	0	0	0	0	0	0	0	0	0	0	0	0	0	0
95650	00025234_013.png	25234	0	0	1	0	1	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
81585	00020125_005.png	20125	0	0	0	0	0	0	0	1	0	0	0	0	0	0
41480	00010752_013.png	10752	0	0	0	0	0	0	0	1	0	0	0	0	0	0
20160	00005403_001.png	5403	0	0	0	0	0	0	0	0	1	0	0	0	0	0
76978	00018964_008.png	18964	0	0	0	0	0	1	1	0	0	0	0	0	1	0
37042	00009805_004.png	9805	0	0	0	0	0	0	0	0	0	0	0	0	0	0

3 Model Selection and Development

The images we used outline the architecture of DenseNet-121 for multi-label chest X-ray image classification, prompting a comparison with VGG16, ResNet50, and Google Vision Transformer (ViT) to highlight each model's strengths:

- **DenseNet-121**: is efficient in feature reuse and propagation, reducing the number of parameters and preventing overfitting, making it ideal for limited data and detailed feature tasks.
- **VGG16**: stands out for its simple, uniform structure of 3x3 convolutional layers, making it easy to implement and scale, although it is heavier on parameters and computational demands.
- **ResNet50**: utilizes residual blocks to train deep networks by solving the vanishing gradient problem, offering a balance between performance and computational efficiency.
- **ViT base patch 16-224-in21k**: differs by using self-attention mechanisms instead of convolutional layers, capturing global dependencies in images and scaling well with more data.

4 Generalization

4.1 Random Sampling vs Sequence Selection:

- This line of code is used to randomly select 5000 rows from the `all_ray_df` DataFrame. In terms of simplicity, sequence selection would select 5000 entries, regardless of distribution, which does not properly address the problems that the dataset might inherently have. However, we could also try to explore the dataset and impute values, regardless of the distribution, to get a more balanced dataset.
- Instead, we opted for random sampling, especially when we require a representative subset of the data that reflects the overall dataset's distribution without any bias towards order or position within the DataFrame. By doing random sampling, we minimize the risk of introducing sample bias based on the order of the data, which might be influenced by the way it was collected or arranged.
- Medical Images often run into a problem where data is often sparsely collected. Here we see outliers in the number of images collected for each condition, "No Finding", "Infiltration", "Effusion" , "Atelectasis", and several other categories. This distribution highlights significant class imbalance.

4.2 Implications of Imbalance

Based on the imbalance observed, we can infer that the following problems may be encountered:

- **Model Bias Towards Majority Class**
  - `"No Finding"` is the majority class that relays no information towards our classification task. So we drop this class.
  - `"Pneumonia"` or `"Hernia"` has a very low count and class representation which can result in higher misclassification.
- **Poor Generalization**
  - *Overfitting to Majority Class:* If most of the training data represents one class, the model might perform well on training data but poorly on unseen data.
- **Evaluation Metrics Misleading**
  - Accuracy might not be a reliable metric.

Moving forward, we will try to address these problems through our analysis and methods used on the dataset before training:

4.3 Trying to Address Imbalance

- Creating a custom `ChestXRayDataset` class in Python for handling the dataset with PyTorch environment and implementing oversampling to address class imbalance.
  - **Data Handling `'__getitem__'`:** Ensuring that each data point handles loading and transforming images efficiently.
  - Default transform method specified unless mentioned exclusively.
- **Oversampling minority classes:**
  - Dynamically adjusts the dataset to balance class distribution.
  - Calculate the deficit and oversample the minority classes by sampling with replacement (No loss of valuable data).
- **Image Transformations:**
  - Resize inputs to `256x256` - Standardization is crucial to ensure that all images are the same size.
  - `RandomCrop 224`: Randomly crops a `224x224` pixel area from the resized image.
  - `RandomHorizontalFlip`: Randomly flips images horizontally (50% probability).
  - `transforms.ToTensor()`: Converts PIL Image or NumPy ndarray.
  - `transforms.Normalize()`: Normalize an image by the mean and standard deviation.
- **Class Weights Calculation:**
  - `class_counts` presumably contain the count of positive sample for each class in the dataset. The count is then multiplied by the total number of classes.
  - Since our problem deals with multi-label classification, therefore using `BCEWithLogitsLoss` combined with a sigmoid activation.

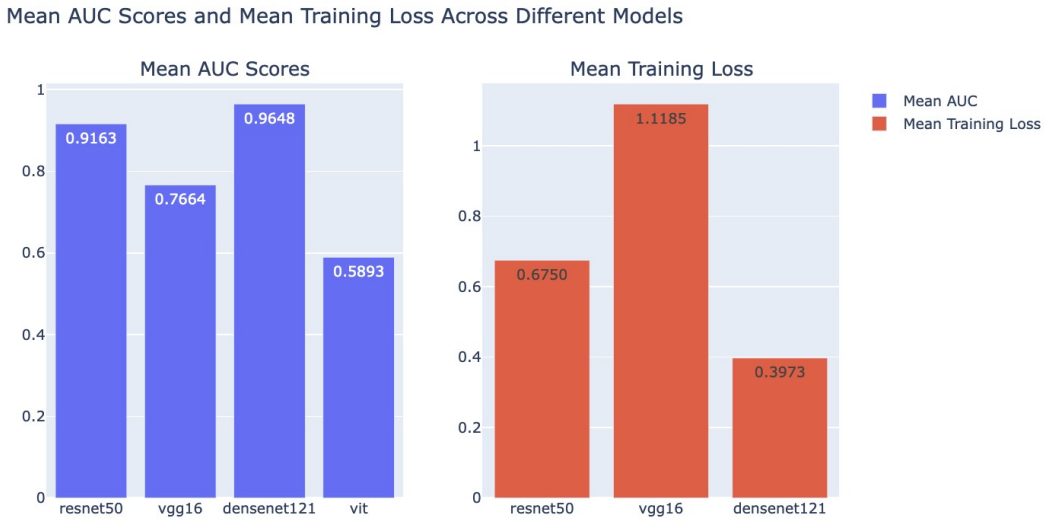
5 Evaluation and Performance Metrics

Analysis of Model Performance with Respect to These Metrics:

Table 1: Model Evaluation Metrics

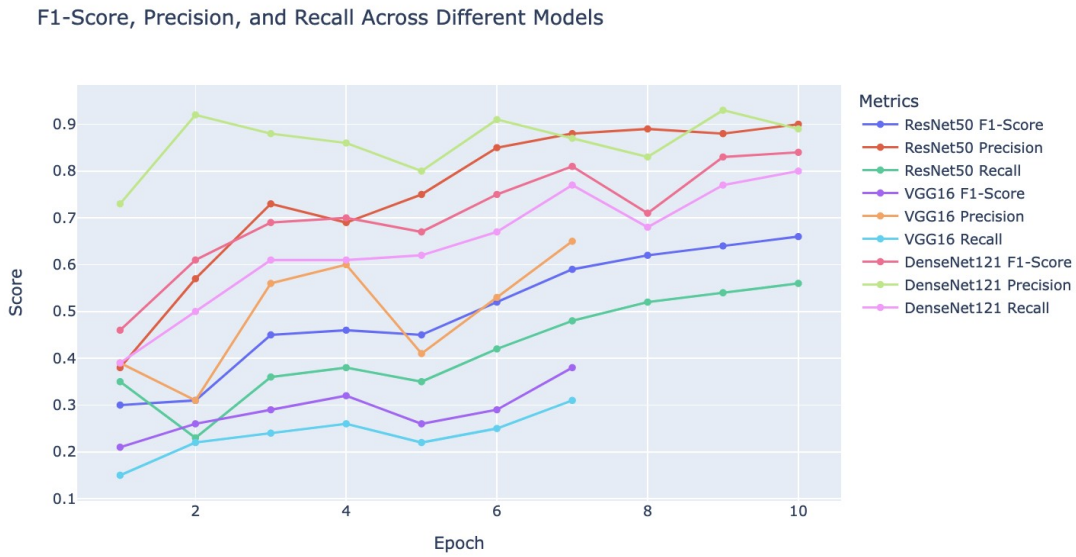
Epoch	Train Loss	Train Acc	Val AUC	Val Acc
		Resnet50		
1/10	1.2566	0.87	0.7503	0.83
2/10	0.9028	0.89	0.7706	0.86
3/10	0.7660	0.89	0.8293	0.90
10/10	0.4362	0.92	0.9154	0.93
		VGG16		
1/10	1.8729	0.86	0.6591	0.87
7/10	0.7471	0.89	0.7711	0.89
		Densenet121		
1/10	0.8607	0.89	0.8381	0.90
10/10	0.1635	0.96	0.9648	0.96
		VIT		
1/10	2.2006	0.84	0.6156	0.84
4/10	2.1765	0.84	0.5823	0.84

- **DenseNet121** and **ResNet50** exhibit better generalization capabilities as indicated by their high validation accuracy and AUC scores. DenseNet121, in particular, shows the best balance between recall and precision, leading to a higher F1-score.
- **VGG16**, while showing good training accuracy, has variable validation accuracy and moderate AUC, suggesting potential overfitting or suboptimal tuning for the task.
- **ViT** demonstrates significant challenges in adapting to the dataset with the lowest scores in AUC and validation accuracy, indicating possible overfitting and a need for further adjustments or more diverse training data.



- ResNet50 and DenseNet121 are both strong performers in terms of AUC and maintain low and stable training losses, indicating they are likely well-tuned for the dataset and problem.
- VGG16, while better than ViT in terms of AUC, still shows less capability in class discrimination than ResNet50 and DenseNet121. This might be due to inherent architectural limitations or less optimal tuning for the specific task.

- ViT shows the least promising results in terms of AUC, despite a sharp decrease in training loss, which might suggest overfitting or that the model is not suitable for the task without significant adjustments or more training data.



- DenseNet121 seems to achieve the best balance between recall and precision, leading to a higher F1-score, which indicates good model performance especially in scenarios where both precision and recall are important.
- VGG16 underperforms in these metrics compared to the other two, highlighting potential issues in model configuration, training process, or dataset compatibility.
- ResNet50, while showing good and balanced improvement, seems a bit conservative in predicting the positive class but improves steadily.

6 Conclusion

Our analysis of the NIH Clinical Center chest X-ray dataset using deep learning models like DenseNet-121, VGG16, ResNet50, and Vision Transformer demonstrates significant potential in automating lung cancer diagnosis. DenseNet-121 emerged as particularly effective, achieving the best balance between precision and recall, essential for accurate and reliable medical predictions. This study illustrates the power of advanced machine learning in enhancing diagnostic processes, offering a promising avenue for improving patient outcomes in clinical settings.