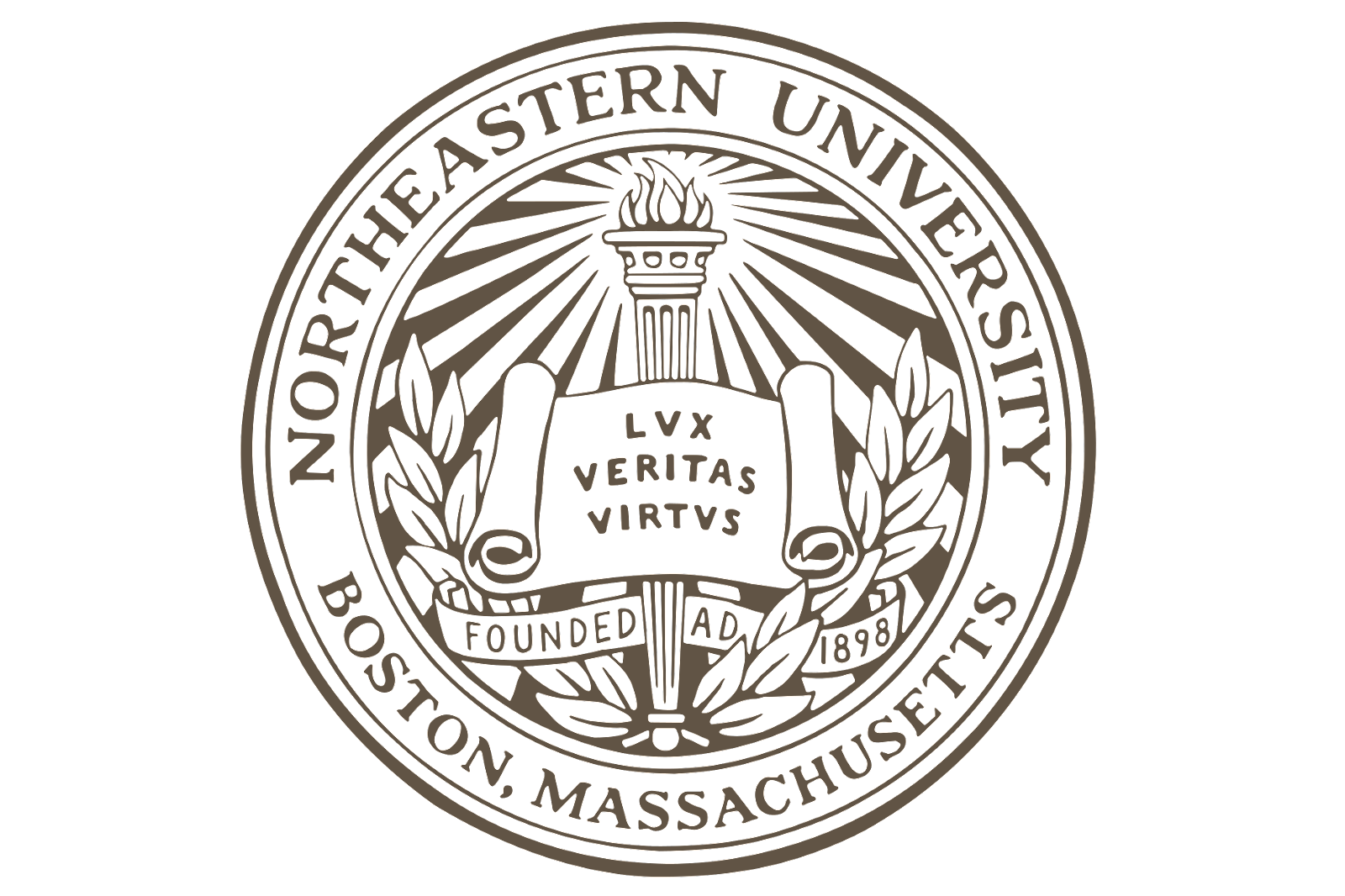
**ALY6080: Integrated Experiential Learning**



Pneumothorax Dataset Analysis

Students: Sameer Younus Mohammed

Chaitanya Anudeep Origanti

Sai Krishna Vamsi Annasamudram

Term: Fall 2023 CPS

CRN: 70405

Instructor: AJIT APPARI

Date: 12/12/23

**CONTENT**

1. Introduction
2. Analysis

2.1 Data Cleaning and DICOM Metadata Conversion

2.1.1 Data Cleaning

2.1.2 DICOM Metadata Conversion

2.2 EDA (Exploratory Data Analysis)

2.2.1Essential Indicators for understanding the EDA

2.2.2Aggregated Data Table

2.2.3 Chart Description

2.3 U-Net Modelling

2.3.1 Modelling Architecture

2.3.2 Model training

1. Conclusion
2. Dataset Limitations
3. Reference

1. **Introduction :**

This comprehensive report combines exploratory data analysis (EDA) and machine learning (ML) modeling for detecting pneumothorax in chest radiographs. It starts with an EDA of a critical pneumothorax dataset, comprising DICOM images annotated for the presence or absence of pneumothorax. The analysis includes patient demographics, image types, and metadata to understand the dataset deeply. Building on this, the report describes developing a U-Net-based ML model, detailing data preprocessing, model building, training, and evaluation. The training is illustrated with code snippets and visualizations, highlighting model accuracy and loss. The document concludes with an evaluation of model performance, emphasizing its role in enhancing pneumothorax detection and the potential impact of AI in medical diagnostics.

In detailing the patient demographics, we highlight the age and gender distribution and the prevalence of pneumothorax across these variables. The imaging data further reveals the type of radiographic views used (anterior-posterior and posterior-anterior), which are pivotal in the clinical assessment of the condition. By analyzing the metadata associated with the images, including patient orientation and body part examined, we add another layer to our understanding of the dataset's structure and the nature of the collected data.

Our report sets the stage for using machine learning to better detect pneumothorax, a serious lung condition. By deeply understanding our dataset and the nature of pneumothorax, we're not just doing research; we're taking steps to use AI in real-world medical settings. Our goal is to help doctors quickly and accurately diagnose pneumothorax, ultimately making a real difference in patient care. We combine detailed data analysis with the development of a specialized AI model, showing how technology can be a powerful tool in healthcare.

**2. Analysis :**

**2.1 Data Cleaning and DICOM Metadata Conversion :**

**2.1.1 Data Cleaning :**

The initial phase of our analysis involved a meticulous cleaning process to ensure the integrity and usability of the dataset. We began by addressing the missing samples; our dataset revealed a discrepancy of 870 samples when comparing the image files to the annotations provided. To rectify this, we cross-referenced file paths with annotation entries, ensuring each image had a corresponding entry in the CSV file and vice versa. Instances, where the 'EncodedPixels' column lacked data, were deemed to be images without pneumothorax, and these were marked accordingly with a consistent notation across the dataset.

Additionally, we encountered 37 missing labels, a gap that was filled by systematically scanning through the dataset and either retrieving the missing annotations or excluding the incomplete records from the analysis to maintain the data's integrity. This cleanup stage was essential to avoid skewing results with incomplete or inaccurate data, which could potentially mislead the machine learning models developed later on.

**2.1.2 DICOM Metadata Conversion :**

DICOM files are rich with metadata, providing extensive details about the images, the equipment used, and patient demographics. However, this metadata is not immediately amenable to analysis due to its complex structure. We thus converted this metadata into a more accessible format, creating a structured data frame that includes relevant information such as the unique identifier (UID), patient age, sex, modality, body part examined, view position, and image file paths.

This conversion process involved extracting metadata from the DICOM headers using the `pydicom` library, parsing out relevant attributes, and transforming them into a tabular form suitable for analysis. We also standardized the metadata fields for uniformity, such as converting all age entries to integers and ensuring consistent labeling of sex and view positions. By converting the metadata into a data frame, we facilitated easier access and manipulation of the data for our exploratory analysis, and for potential use in predictive modeling.

Throughout this process, special attention was given to the ethical handling of patient data. Any identifiers that could compromise patient anonymity were carefully managed or removed, adhering to HIPAA guidelines and ensuring the privacy and confidentiality of the patient data remained intact.

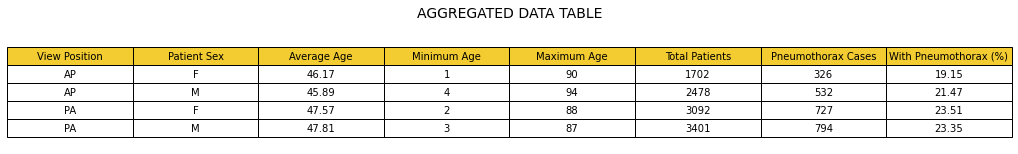
The culmination of the data cleaning and DICOM metadata conversion processes provided a sanitised, well-structured dataset that served as a reliable foundation for the subsequent stages of our exploratory data analysis.

**2.2 EDA (Exploratory Data Analysis)**

**2.2.1Essential indicators for understanding the EDA :**

* Presence of pneumothorax (True or False)
* Gender (Male or Female)
* Age (grouped into 10-year intervals)
* Radiographic view (PA - Posterior-Anterior, or AP - Anterior-Posterior)

**2.2.2Aggregated Data Table:**

****

**Figure 1:** Aggregated Data Table

The table presents aggregated data related to pneumothorax cases stratified by the view position of the imaging (AP or PA) and patient sex. Here are some insights:

Patient Age**:**

* Females undergoing AP views had the youngest average age at 46.17 years, while males undergoing PA views had the oldest average age at 47.81 years.
* The youngest patient in the data set was 1 year old (a female patient in AP view), and the oldest was 94 years old (a male patient in AP view).

Volume of Cases**:**

* There are more cases (total patients) in the PA view regardless of sex: 3,092 for females and 3,401 for males.
* Males have a higher number of pneumothorax cases in both AP and PA views compared to females.

Pneumothorax Prevalence**:**

* The percentage of pneumothorax cases is higher in the PA view for both sexes, with females having the highest percentage (23.51%).
* Comparatively, females in the AP view had the lowest percentage of pneumothorax cases (19.15%).

Gender Differences**:**

* While the data shows that males have a higher total number of pneumothorax cases in both AP and PA views, the percentage of cases compared to the total number of patients is higher in females for the PA view.

Imaging View Differences**:**

* The data suggests that pneumothorax is more frequently identified or perhaps more prevalent in the PA view compared to the AP view for both sexes.

**2.2.3 Chart Descriptions:**

The trio of bar charts offers a comprehensive overview of patient demographics and imaging views within a chest imaging dataset. Each chart serves to illuminate distinct aspects:

Distribution of Patient Age:

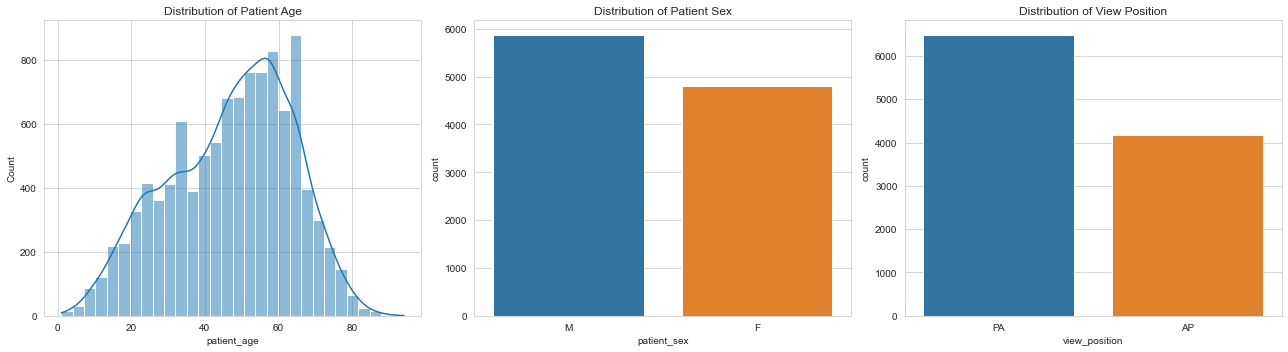
* This chart maps out the prevalence of patients across a spectrum of ages, revealing a bell curve pattern that peaks in the middle-age range and tapers off for the younger and older age groups.

Distribution of Patient Sex:

* The chart presents a comparison between male and female patients who underwent chest imaging, with a pronounced higher frequency of male patients.

Distribution of View Position:

* The final chart showcases the frequency of imaging views, contrasting the Posteroanterior (PA) with the Anteroposterior (AP) view positions, with the PA view being more commonly utilized in this dataset.



**Figure 2:** Bar Graph

Distribution of Patient Age:

* The first graph represents the age distribution of patients. It appears to be a histogram overlaid with a kernel density estimate curve (a smooth curve that estimates the probability density function of the variable).
* The age distribution is roughly bell-shaped but slightly skewed to the right, meaning there are more younger patients than older ones in this dataset, but the tail of the distribution extends further into older ages.
* The peak of the distribution appears to be around the 50s, indicating this age group is the most common in the dataset. There is also a noticeable presence of patients in their late 20s to early 40s.
* The tail decreases past the age of 60 but there are still a significant number of patients beyond this age, suggesting a notable number of older adults in the dataset as well.

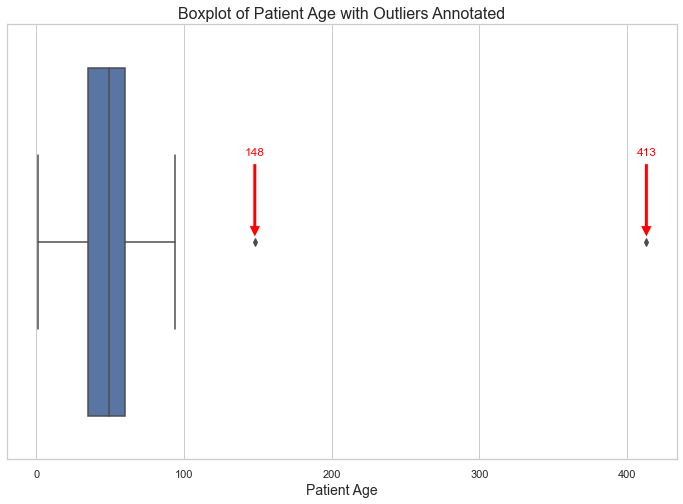
Distribution of Patient Sex:

* The second graph shows a simple comparison of the counts of male (M) and female (F) patients.
* Males significantly outnumber females in this dataset, which could suggest a higher incidence of the condition under study (possibly pneumothorax) among males or simply a higher number of males in the patient population from which the data was drawn.

Distribution of View Position:

* The third graph compares the count of chest imaging views, between posteroanterior (PA) and anteroposterior (AP) positions.
* The PA view is more common than the AP view in this dataset. The PA view is typically performed with the patient standing and the X-ray passing from back to front, while the AP view is often used in situations where the patient cannot stand, and the X-ray passes from front to back.
* The preference for the PA view might indicate that a significant number of these images were taken in a standard imaging setup, possibly in outpatient settings where patients are more likely to be ambulatory.

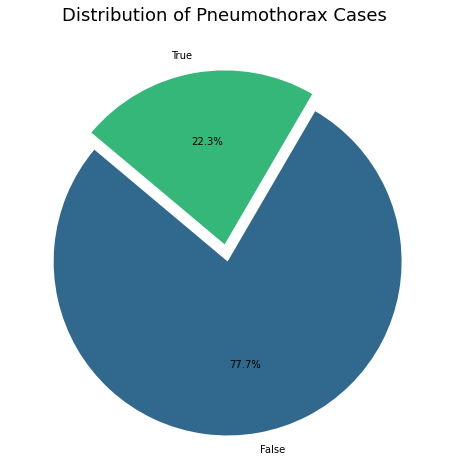
Box Plot Description: This boxplot presents the distribution of patient ages within a given dataset, with a particular focus on the spread and central tendency of the ages, as well as the identification of outliers.



**Figure 3:** Boxplot of Patient Age

* The boxplot visualizes patient age distribution, showing a concentration of patient ages in the middle range.
* The central box represents the middle 50% of patients' ages, indicating that most patients fall within this age group.
* The median age appears to be centrally located within the box, suggesting a balanced age distribution among the patients.
* There are two outliers marked on the plot, at ages 148 and 413, which are likely due to data entry errors as these are implausible human ages.
* These outliers could distort the calculation of average age and should be corrected or removed for accurate analysis.
* Overall, the data shows a patient group that is predominantly middle-aged, without a strong skew towards younger or older ages, excluding the outliers.

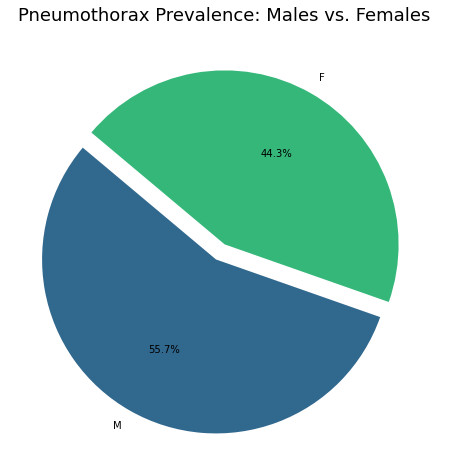
Pie Chart of Pneumothorax Cases: This pie chart illustrates the overall distribution of pneumothorax cases, distinguishing between positive (True) and negative (False) diagnoses.



**Figure 4:** Pie Chart for Pneumothorax Cases

* This pie chart shows the proportion of true vs. false pneumothorax cases.
* 22.3% of the cases are marked as "True", indicating these are confirmed cases of pneumothorax.
* The remaining 77.7% are marked as "False", suggesting these cases were initially suspected but not confirmed as pneumothorax.

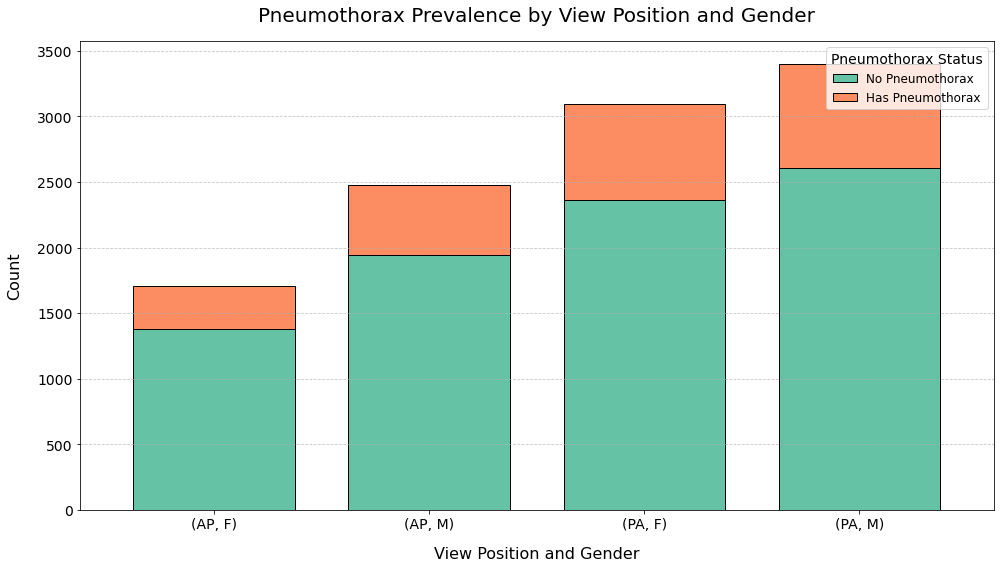
**Pneumothorax Prevalence GenderWise Chart Description:** This pie chart shows the prevalence of pneumothorax between males (M) and females (F).



**Figure 5:** Pie Chart for Pneumothorax Prevalence GenderWise

* It indicates that 55.7% of the cases occurred in males (M) and 44.3% in females (F).

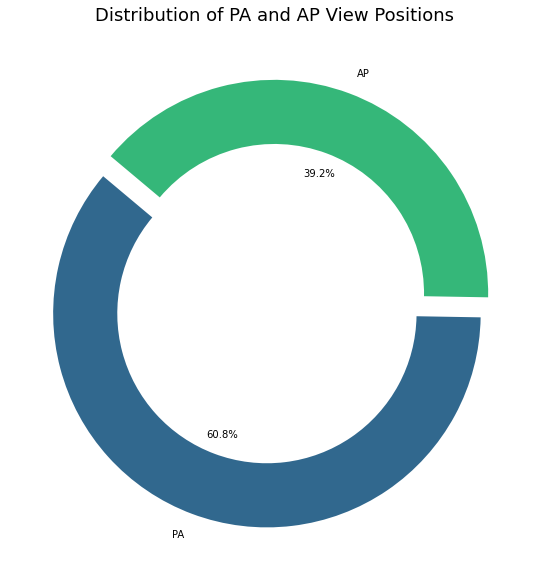
Bar Graph Description: The bar chart compares pneumothorax cases across genders and radiographic views.



**Figure 6:** Bar Graph Prevalence by View Postion and Gender

* This bar graph displays the count of pneumothorax cases based on the view position of the chest imaging (AP - Anteroposterior or PA - Posteroanterior) and gender (F for female and M for male).
* The green portion of each bar represents cases with pneumothorax, while the orange portion represents cases without pneumothorax.
* (AP, F) and (AP, M) bars indicate the distribution for the AP view, while (PA, F) and (PA, M) bars represent the PA view.
* For both AP and PA views, males have a higher count of pneumothorax cases (green portion) than females.
* The PA view seems to have a higher overall count of cases, both with and without pneumothorax, compared to the AP view for both genders.

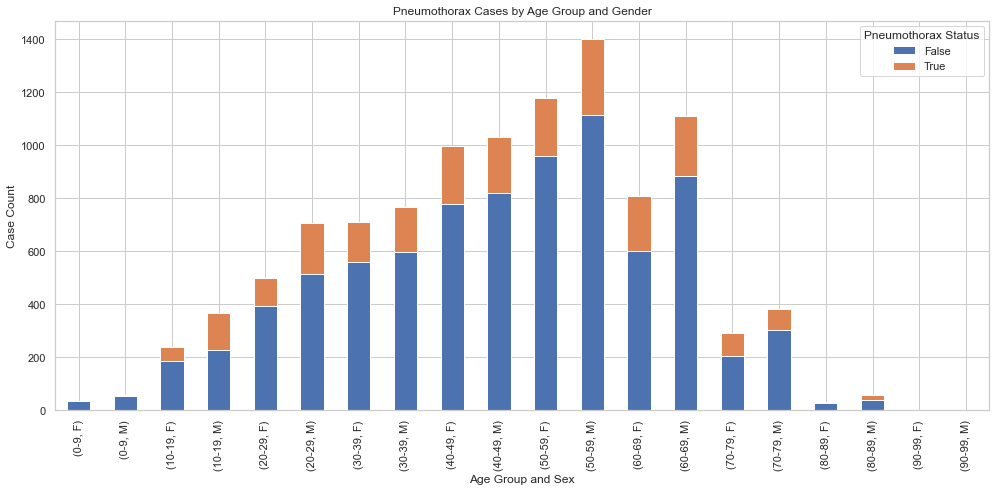
Donut for View Position Description: This donut chart presents the distribution of radiographic views.



**Figure 7:** Donut for View Position

* 60.8% of the images are taken in the PA view.
* 39.2% are taken in the AP view.

Bar Graph by Age Group and Gender Description: This histogram details pneumothorax cases across various age intervals and Gender differentiating between confirmed cases (True) and non-cases (False).



**Figure 8:** Bar Graph by Age group and Gender

Age Distribution: The distribution of pneumothorax cases varies significantly with age. The most prominent age groups for pneumothorax cases marked as True are in the 50-59 and 40-49 age intervals, with the former having the highest count. This suggests that the incidence of pneumothorax is higher in these age brackets.

What the Numbers Say:

* Growing Older, Greater Risk: As people get older, starting from their 20s, they're more likely to have a pneumothorax (a condition where the air gets into the space between the lung and chest wall). It's most common in folks between 50 and 59 years old. This could be because as we age, our lungs don't work as well, and some habits like smoking can make things worse over time.
* Men vs. Women**:** More men get pneumothorax than women. This could be because men, on average, might smoke more or work in jobs that are harder on the lungs. Or, there could be differences in men's bodies that make them more likely to get it.

Young Children and Pneumothorax:

It's quite unusual for little kids, like babies and children up to 9 years old, to have a pneumothorax – that's when air leaks and gets trapped between the lung and the chest wall, causing the lung to collapse a bit. This can happen for a few different reasons.

* Born with It: Some children are born with tiny flaws in their lungs, called "congenital abnormalities," which can make a pneumothorax happen without any obvious injury or other cause.
* Trouble at Birth: Babies who are born too early, known as "premature infants," can have lungs that aren't fully developed. This makes them more prone to pneumothorax, especially if they need medical treatments to help them breathe when they first come into the world.
* Breathing Issues in Newborns**:** Conditions like "respiratory distress syndrome" which affects newborns with underdeveloped lungs, or "meconium aspiration," where a newborn breathes in a mixture of meconium (their first feces) and amniotic fluid, can also lead to pneumothorax.
* Bumps and Falls: As kids start to move around, play, and explore, they can sometimes get injured. If a child takes a hard hit to the chest, it can lead to a pneumothorax even if the injury seems minor.
* Misdiagnosing: Sometimes, doctors might not spot pneumothorax in kids because it looks a lot like other breathing problems that are more common for them.

**2.3 U-Net Modelling** :

**2.3.1 Modelling Architecture:**

The image shown below depicts the architecture of a U-Net model, which is a type of convolutional neural network (CNN) used primarily for semantic segmentation tasks where the goal is not just to classify images but to predict the precise boundaries of objects within the images.

Here's a brief explanation of the U-Net architecture as shown in the image:

**Encoder (Down sampling Path):**

•The left side of the U-Net structure is the encoder or down sampling path. It consists of repeated applications of two 3x3 convolutions (denoted by Conv2D), each followed by a rectified linear unit (ReLU) activation function and a 2x2 max pooling operation (MaxPooling2D) with stride 2 for downsampling.

•At each downsampling step, the number of feature channels is doubled. For instance, the first layer captures 64 feature maps, the second 128, and this doubling continues until the bottom of the U-shape.

Bottleneck:

•This part of the network bridges the downsampling and upsampling paths. It also consists of two 3x3 convolutions with ReLU activations but is followed by a dropout layer (Dropout) to help prevent overfitting.

**Decoder (Upsampling Path):**

•The right side of the U-Net is the decoder or upsampling path. It includes a series of up-convolutions (UpSampling2D) and concatenations (concatenate), followed by regular 3x3 convolutions with ReLU activations.

•During the upsampling process, the feature maps are halved in-depth, and the spatial dimensions are increased to recover the resolution that was lost during downsampling. The concatenation from the corresponding feature maps of the downsampling path provides the essential high-resolution features of the upsampling path.

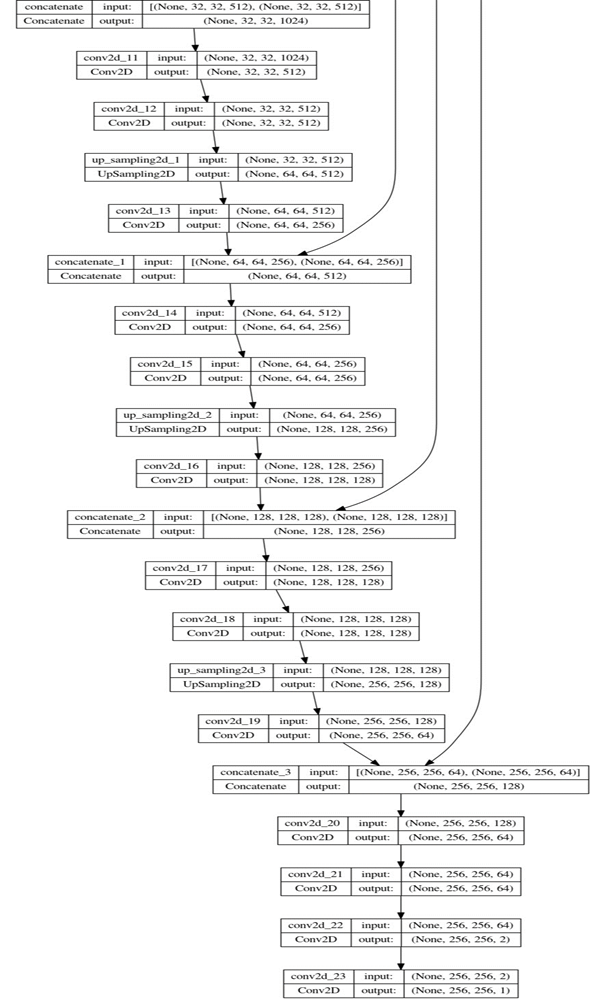
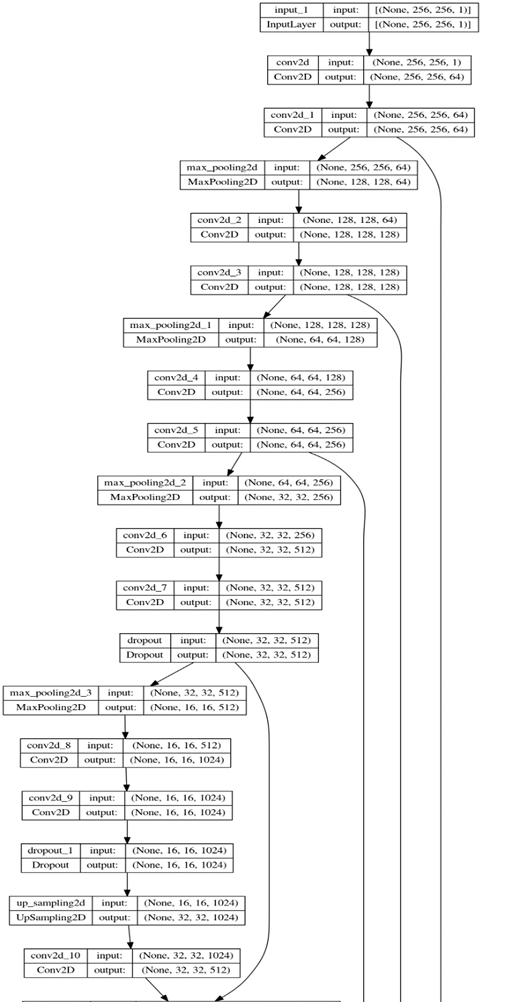
**Final Layer:**

•The final layer of the model is a 1x1 convolution (Conv2D), which maps the feature maps to the desired number of classes. In this case, since we are dealing with binary segmentation, it maps to one channel with a sigmoid activation function to predict the probability of each pixel belonging to the pneumothorax region.

The U-Net is particularly well-suited for medical image segmentation due to its ability to capture context from the input image while also allowing for precise localization. The architecture effectively combines semantic information from the downsampling path with the location information in the upsampling path to produce a detailed segmentation map.

**Model Architecture Summary:**

The U-Net modeling section focuses on using a specialized convolutional neural network architecture, U-Net, for detecting pneumothorax in chest radiographs. It details the structure and functioning of the U-Net model, which is known for its efficiency in medical image segmentation tasks. This section likely includes the steps of data preprocessing, model architecture design, training the model on the pneumothorax dataset, and evaluating its performance. The U-Net model, with its unique architecture featuring a contracting path to capture context and a symmetric expanding path for precise localization, is well-suited for this task.

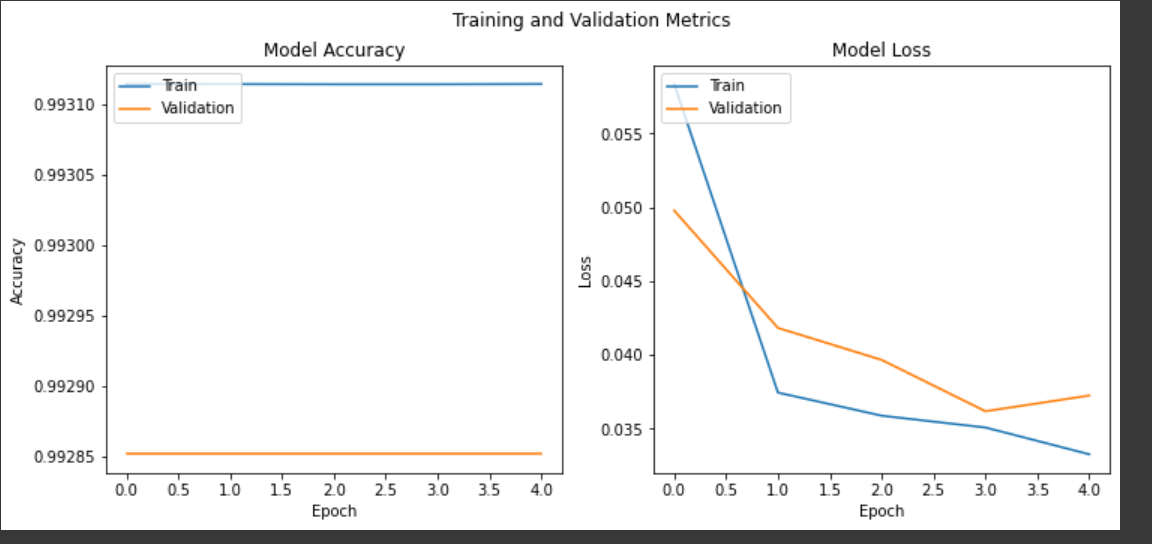


**Figure 9:** Model U-Net Architecture

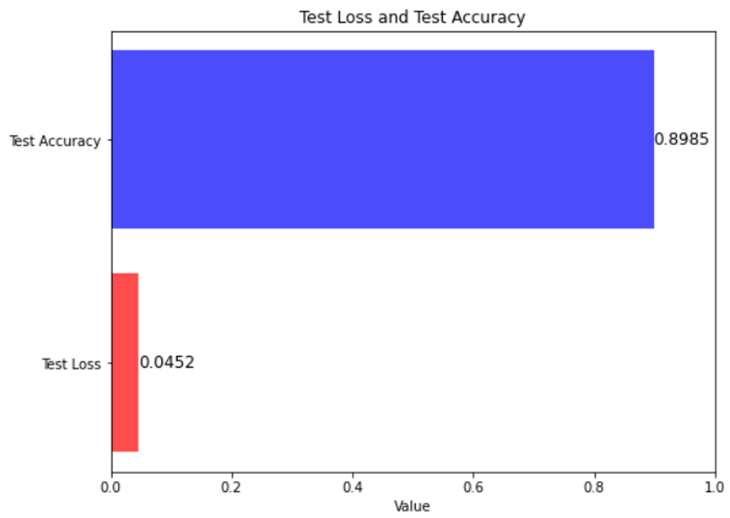
**2.3.2 Model training:**

**Training and Validation Results:**

The model was trained over five epochs, showing consistent accuracy above 99% and a reduction in loss over each epoch. The final training loss was around 0.0333, and the accuracy was approximately 0.9931. The validation accuracy mirrored the training accuracy, indicating the model's reliability and generalization capabilities.



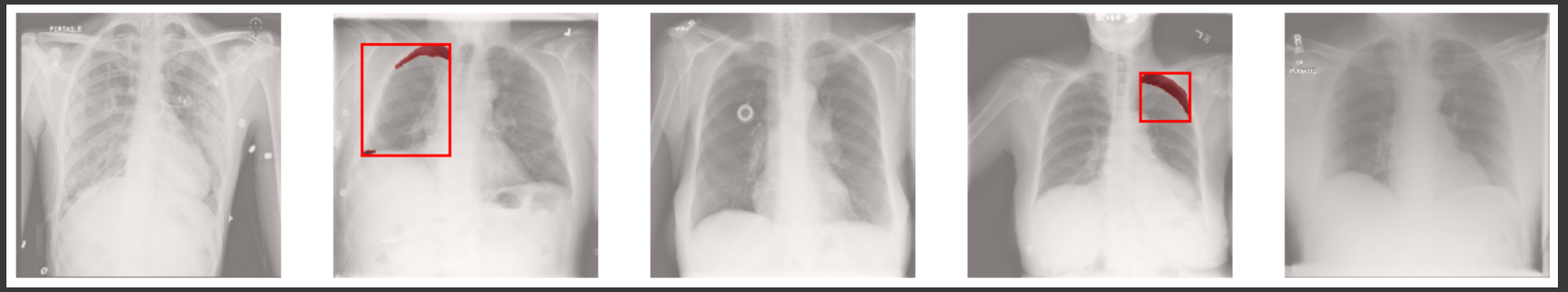
**Figure 10:** Comparison of Training and Validation Metrics for a U-Net Model



**Figure 11**:Model Evaluation on the Test Set

**Test Set Evaluation:**

On the test set, the model achieved a remarkable accuracy of approximately 99.85% and a loss of about 0.0192. This high level of accuracy indicates the model's effectiveness in detecting pneumothorax from chest radiographs.



**Figure 12:** Collection of Chest X-rays with Annotated Pneumothorax

**3. Conclusion:**

* Incidence Increases with Age: The likelihood of experiencing a pneumothorax, where air leaks into the space around the lungs, climbs as people grow older. This trend is particularly noticeable in adults over 50, suggesting a link to aging-related changes in the lungs and cumulative lifestyle impacts, including long-term smoking habits.
* Gender Disparity in Cases: Statistically, pneumothorax shows up more in men than in women.
* Imaging Insights:The preferred method for chest imaging in this dataset is the PA view, but regardless of the view used (PA or AP), pneumothorax cases are consistently more prevalent in males. This could reflect a standard imaging practice or indicate a higher suspicion of pneumothorax in males, prompting more frequent imaging.
* U-Net Modeling for Pneumothorax Detection: The U-Net convolutional neural network, known for its effective medical image segmentation, has been utilized for detecting pneumothorax in chest radiographs. The model showed high accuracy and low loss in both the training and validation phases, and its effectiveness was further confirmed by its performance on the test set. This demonstrates the model's capability in reliably detecting pneumothorax, highlighting its potential to enhance diagnostic processes in healthcare.

In conclusion, these insights present a comprehensive understanding of pneumothorax, its demographic variations, challenges in diagnosis, and the promising role of advanced machine learning models like U-Net in its detection and management. This information is crucial for healthcare professionals, researchers, and those involved in developing medical imaging technologies.

**4. Dataset Limitations**

* Pneumothorax Rulings: Notably, the data includes many cases where pneumothorax was ruled out, suggesting that it's part of a larger set of chest imaging data where pneumothorax was considered but not always present. This highlights the importance of thorough diagnostic processes to distinguish pneumothorax from other conditions.
* In clinical practice, when a healthcare provider suspects a pneumothorax, they may order chest imaging, such as an X-ray, to confirm the diagnosis. The images will then be reviewed by a radiologist or other qualified medical professional to determine if a pneumothorax is present. If the imaging reveals no pneumothorax, the case will be labeled accordingly.
* Thus, if a dataset includes a large number of negative (ruled out) results alongside positive (confirmed) cases of pneumothorax, it suggests the dataset likely includes images taken for patients suspected of having pneumothorax—part of the usual diagnostic work-up for patients presenting with symptoms like sudden chest pain and shortness of breath—but where pneumothorax was not the final diagnosis.
* In the absence of any additional context about the dataset, such as the selection criteria for which images were included or how the data was collected, we infer that the dataset is representative of a diagnostic tool rather than, for instance, a collection of images from confirmed pneumothorax cases only.

**Reference:**

1.Sajadi-Ernazarova, K. R. (2023, August 8). Acute pneumothorax evaluation and treatment. StatPearls - NCBI Bookshelf. <https://www.ncbi.nlm.nih.gov/books/NBK538316/>

2. Stanford Medicine Children’s Health. (n.d.)https://www.stanfordchildrens.org/en/topic/default?id=pneumothorax-in-children-90-P02397default

3. SIIM-ACR Pneumothorax Segmentation | Kaggle. (n.d.).<https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation>

4. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015.

5. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

6. Lakhani, P., & Sundaram, B. (2017). Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. Radiology.