Collection of all questions and answers asked in the google form as well as in the assignment.pdf:

Questions asked in Google form (Submitted the answers in the google form as well there as well): [I have included my earlier long answers and then due to Google form restrictions - it wanted 300 characters or less answer - I have included them below as well]

1. Why and what layers of the model did you fine-tune?

   Answer: For this assignment, and given the time, I only focussed on fine tuning the final classification layer of the pretrained ResNet model 50. The rationale being, that this is a transfer learning approach, where pre-trained convolutional layers extract the features and the last part (i.e. classifier head) is adapted for pneumonia detection (distinguishing between normal vs pneumonia inflicted) task.

   [Shortened]: For this assignment, and given the time, I only focussed on fine tuning the final classification layer of the pretrained ResNet model 50. Although other layers could be unfrozen gradually, but due to paucity of time, focused on the last one.

2. Describe how did you split your data for fine-tuning the model:

   Answer : The data was split into three classes for fine tuning the model - first into the training class, second the validation class, and final one was test class. This was done also per the details given in the kaggle link - where for each of the class, indicated data was present. The split were as follows: 3882 images for training, 624 for validation and 524 for testing each of size 28x28.

   [Shortened]: The data was split into three classes for fine tuning the model - first into the training class, second the validation class, and final one was test class. The split were as follows: 3882 images for training, 624 for validation and 524 for testing each of size 28x28.

3. How is your selected evaluation metric relevant to the clinical use case?

   Answer: The evaluation metrics I used were : Precision, Recall, F1 score, AUC-ROC (curve). Each of these metrics were used for the following primary reason:

   1. Precision and Recall: Crucial for medical diagnoses, high recall ensures pneumonia cases are not ignored, white precision minimizes false alarms e.g Normal cases being treated as Pneumonia inflicted.)
   2. F1-Score: Ensures there is a balance between precision and recall values, since both the false positive and false negatives have clinical consequences and economic repercussions.

3. AUC-ROC: It provides a visual depiction as well as provides a quantitative assessment for the model's ability to distinguish between normal and pneumonia cases across different threshold values.

[Shortened]: Precision, Recall : Crucial for medical diagnoses, high recall ensures pneumonia cases are not ignored. F1-Score: Ensures there is a balance between precision and recall values.AUC-ROC: It provides a visual depiction as well as provides a quantitative assessment

4. Take one example of a mis-classified case in your test set. Explain why the model failed and what you would try next?

   Answer: As per the final model run, and post generation of the confusion matrix, there were 117 false positive cases (normal cases identified as pneumonia). During intermediate analysis, a normal chest X-ray was incorrectly identified (or classified) as pneumonia. Primary reasons for it being:

   1. Class imbalance: In the provided dataset, there were more pneumonia cases (images) [3,494] as compared to the normal (or healthy) cases [only 388].
   2. Due to high number of pneumonia images, resulting class imbalance, model memorized pneumonia ‘better’ and was overfit - hence even detecting those in the normal cases as well.

Things to try next (In order to reduce this):

1. Augment more normal cases, hence have a balanced training dataset.
2. Add annotations with consultation with radiologists in the provided images, to guide and improve model’s attention to relevant areas [this way, hopefully reduce the False Positives, False Negatives]
3. Use ensemble methods - the methodology of combining multiple models - in order to reduce individual model biases and improve classification strength.

[Shortened]: 117 false positives (normal X-rays misclassified as pneumonia - in initial runs) due to class imbalance: 3,494 pneumonia vs 388 normal images. Model overfit pneumonia. To fix: augment normal data, add expert annotations, use ensembles. Goal: reduce errors, save time/costs, aid COVID detection.

5. Model will assist in reducing time, save costs, increase pneumonia detection rates (particularly during covid outbreaks)

[Shortened]: Model will help in reducing time, save costs, increase pneumonia detection

Questions asked in Assignment:

Based on retrospective analysis of my notebook, here are the answers:

a. Three Appropriate Metrics and Justifications:

The evaluation metrics I used were : Precision, Recall, F1 score, AUC-ROC (curve). Each of these metrics were used for the following primary reason:

1. Precision and Recall: Crucial for medical diagnoses, high recall ensures pneumonia cases are not ignored, white precision minimizes false alarms e.g Normal cases being treated as Pneumonia inflicted).
2. F1-Score: Ensures there is a balance between precision and recall values, since both the false positive and false negatives have clinical consequences and economic repercussions.
3. AUC-ROC: It provides a visual depiction as well as provides a quantitative assessment for the model's ability to distinguish between normal and pneumonia cases across different threshold values.

b. Class Imbalance Detection and Mitigation:

Class imbalance detected post running the model on the test dataset and then seeing the model performance evaluation metrics. [388 normal vs 3,494 pneumonia cases]. Furthermore 0 false negatives showing up repeatedly, followed by 1.00 precision for Normal cases.

Mitigation strategies applied: Utilized weighted loss function, weighted random sampler, enhanced data augmentation technique (used geometric transformations).

c. Overfitting Prevention Measures:

1. I applied data augmentation techniques, and utilized transforms such as RandomResizedCrop, RandomHorizontalFlip, RandomRotation to training data.
2. Furthermore, I also applied regularization techniques such as L2 Weight Decay, Early stopping (with patience = 3, triggered at the epoch run at 4, to prevent overtraining of the ResNet50 model).
3. Fine tuned final classification layer to leverage learned features. However, more improvement with this respect could be done, such as more layers could be unfrozen and utilized. However due to paucity of time, restricted myself to this.

Question : A short note on the hyperparameters used:

I implemented and optimized different hyperparameters, especially post realizing the class imbalance issue that manifested during the initial training, implementation of the ResNet 50 model. First started with a conservative learning rate of 1e^(-4); batch size of 32 and Adam optimizer with L2 regularization (using a e^(-5) weight decay).

Post identification of the class imbalance, and to remedy the overfitting of the model, decided to further adjust the training related hyper-parameters by setting a maximum of 20 epochs, but with early stopping with patience = 3, which halted training at epoch 4 itself, to prevent overfitting (meaning, to prevent the model learning/memorizing the training data itself too well). Finally, addressed the class imbalance again by adjusting CrossEntropyLoss with class weights as [1.7,1.0] and WeightedRandomSampler. Furthermore, did a couple of more adjustments with the hyperparameters as I continued to refine the model to improve performance - especially with class imbalance plaguing it.