Activity08: Understanding Classification Error

Goal

In this activity you will practice calculating the ROC curve and computing the confusion matrix.

Instructions

You NEED to hand in your solution for this Activity**,** please see the due date on Canvas.

Solution key to the questions will be released after the deadline.

Submission format: Please upload a single PDF file of this document after inserting your answers in the space after each question, which shows all the steps you took, intermediate calculations, and the code you wrote and used.

You can upload extra .m files as supplementary document in a .zip file beside the pdf, but it's not going to add any extra grade.

I. Evaluate an AI-based COVID-19 Diagnosis System

Note: The information provided here is based on real research, however, given the seriousness of COVID19, please assume the information here is hypothetical and may contain errors and should not be used for any purpose beyond this assignment.

The current gold-standard for diagnosis of COVID-19 is real-time polymerase chain reaction (RT-PCR) lab test [Bai 2020]. However, lab resources are expensive, limited and time consuming. A quick, cheaper and non-invasive alternative may be to perform CT imaging and use the features such as peripheral distribution, ground-glass opacity and vascular thickening of the CT images for diagnosis [Bai 2020].

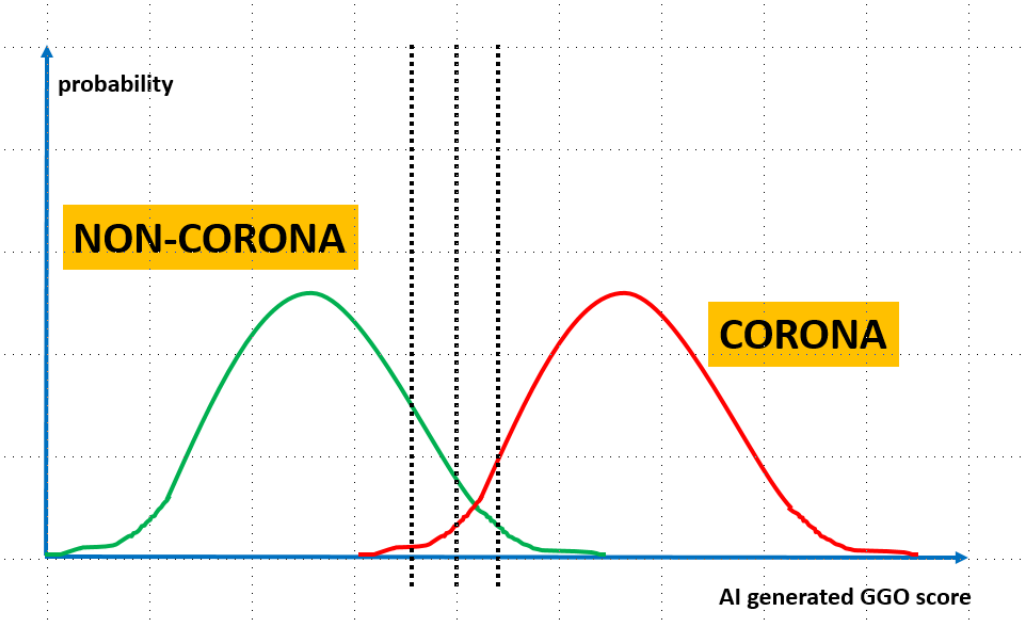
Assume the scientists designed an alternative AI system, which takes in a CT image, recognizes the ground-glass opacity (GGO) feature, and performs the diagnosis in a few seconds. However, there is a trade-off between efficiency and accuracy, so we have to evaluate how much we can trust the system.

(Simulated) Dataset: 100 patients were both tested by RT-PCR and the CT-based AI system: 51 patients were diagnosed by RT-PCR (the gold-standard) as positive (True) while 49 tested negative (False). The raw GGO values were collected from the AI system before making any thresholding.

The data is saved in data\GGO_value.mat and data\diagnosis.mat respectively.

Question 1 [3 points]

Assume the probability of positive and negative patients follow Gaussian distributions (see the two schematic plots below). Notice there is overlap between the two distributions (which means if we take different thresholds, we'll obtain different prediction results).



a) Using MATLAB, load the data and find the mean and standard deviation (std) of the Gaussian that models the positive distribution for 51 subjects.
b) Find the mean and std of the Gaussian that models the negative distribution for 49 subjects by MATLAB.
c) Show the plot of the two distributions in MATLAB (using the mean and std values found in parts a and b). Label your axes to obtain a figure similar to schematic plot above. Hint: use MATLAB's **normpdf** function.

Question 1. Your Answers:

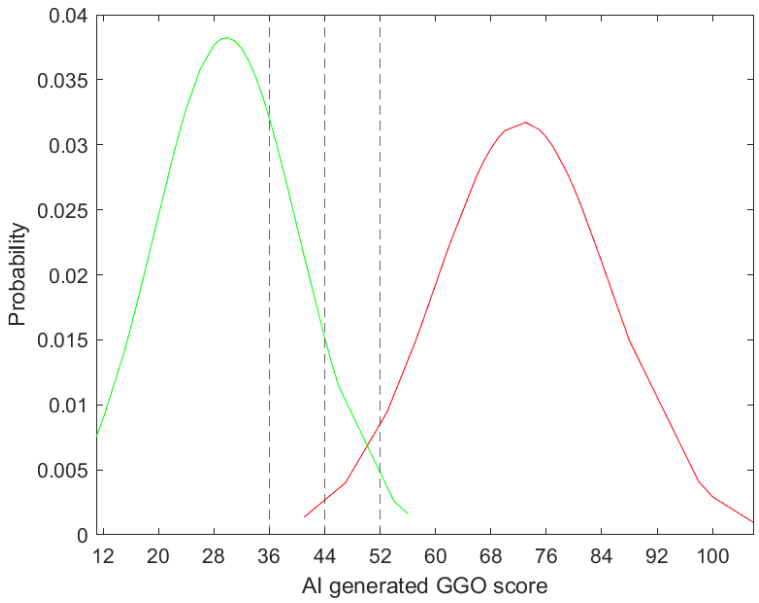
a) Mean of positive subjects: 72.5686

   Std of positive subjects: 12.5686

b) Mean of negative subjects: 29.7959

   Std of negative subjects: 10.4323

c) Paste plot here:



Paste Code Here:

Loading data:

```
GGOVals = load('GGO_value.mat').GGO_values;
diagnosis = load('diagnosis.mat').diagnosis;
```

Calculating mean and std:

```
positiveGGOValues = [];
negativeGGOValues = [];

for i = 1:100
    if diagnosis(i) == 1
        positiveGGOValues = [positiveGGOValues GGOVals(i)];
    else
        negativeGGOValues = [negativeGGOValues GGOVals(i)];
    end
end

meanPositiveGGOVals = mean(positiveGGOValues)
stdPositiveGGOVals = std(positiveGGOValues)

meanNegativeGGOVals = mean(negativeGGOValues)
stdNegativeGGOVals = std(negativeGGOValues)
```

Choose the threshold range and step
Build the distribution function

```
positiveGGOValues = sort(positiveGGOValues);
negativeGGOValues = sort(negativeGGOValues);

figure(1);
% Positive
positiveY =
normpdf(positiveGGOValues,meanPositiveGGOVals,stdPositiveGGOVals
);

% Negative
negativeY =
normpdf(negativeGGOValues,meanNegativeGGOVals,stdNegativeGGOVals
);

thresholdRange = [11 106];
thresholdValues = [];
step = 8;
```

Plot your figures

```
plot(positiveGGOValues,positiveY,'r');

hold on

plot(negativeGGOValues,negativeY,'g');

xlim(thresholdRange);
xline(36,'--');
xline(44,'--');
xline(52,'--');

for g = 0 : 11
   thresholdValues = [thresholdValues g*step+12];
end

xticks(thresholdValues);
xlabel('AI generated GGO score') ;
ylabel('Probability') ;
```

Question 2 [3 points]

Given the 2 Gaussian distributions in Question 1, the goal is to construct the corresponding ROC curve.

a) Choose your threshold values to construct the ROC curve. Make sure your choice contains at least 10 different values.
b) Plot the ROC curve with TP rate (TPR) in percentage along the vertical axis, vs. FPR along the horizontal, using both the erf table and the **normcdf** function.

Note:

- The ROC should show the operating points for equally-separated thresholds.
- Do not use the raw data to calculate the operating points, instead use the Gaussian distributions.
- To calculate needed integrals (i.e. CDF), refer to the lecture slides and make use of the values given in the provided file: **erf_tables.pdf**. Note: for $x < 0$, erf is negative and equal to -erf(-x) as read from the table.
- Use MATLAB to plot and make sure that the operating points are clearly visible. Double check your answers using MATLAB's built-in function **normcdf** to calculate the integrals over a Gaussian distribution. You can use a finer threshold grid so the ROC curve will look smoother.

Question 2. Your Answers:

a) Choose threshold Values of the ROC Operating points:

The threshold values of the ROC Operating points used with the erf table will come from the range starting from GGO value 12 to GGO value 100 in increments of 8. More specifically, the values will be:

12, 20, 28, 36, 44, 52, 60, 68, 76, 84, 92, 100

To obtain a smoother ROC curve, a finer threshold grid was chosen for use with MATLAB's normcdf function. The values used again come from the range starting from GGO value 12 to GGO value 100 in increments of 1. More specifically, the values will be:

12, 13, 14, 15, …, 96, 97, 98, 99, 100

b) Explain how you used erf table for three example thresholds values:

For example, using 12 as the threshold.

To calculate TN. First step would be to plug in 12 for x, the mean of the negative subjects 29.7959 for μ, and the standard deviation of the negative subjects 10.4323 for σ into erf $\left(\frac{x-\mu}{\sigma * \sqrt{2}}\right)$.

Then, we would calculate the equation inside the brackets, leaving us with erf(-1.206215…), however, since the erf table provided only goes up to hundredth digits of x, we only need erf(-1.20).

Now, we need to look up -1.20 in the erf table. Since erf(-x) = -erf(x), we actually need to find 1.20 in the erf table, which turns out to be 0.91031. Adding back the negative, we obtain -0.91031.

Now that we have the erf value, we just need to plug it back into $normcdf(x|\mu,\sigma) = \frac{1}{2}[1 + \text{erf}\left(\frac{x-\mu}{\sigma*\sqrt{2}}\right)]$. So, after plugging in -0.91031 for $\text{erf}\left(\frac{x-\mu}{\sigma*\sqrt{2}}\right)$, we need to add 1 to -0.91031 to obtain 0.08969, then multiply by ½ to obtain **0.044845**, which is our normcdf and which is also our TN.

To obtain FP, we just need to do 1 – 0.044845 = **0.955155**.

Now, to obtain the TP, we do a similar process. Again, plug in 12 for x, but mean of the positive subjects 72.5686 for μ, and standard deviation of the positive subjects 12.5686 for σ into $\text{erf}\left(\frac{x-\mu}{\sigma*\sqrt{2}}\right)$ to obtain -3.41.

Then, we need to find 3.4 in the erf table, which is greater than what the table has therefore we choose the largest x value in the table which is 3.2 and use that the error function value of 3.29, which is 1.00. Adding back the negative, we obtain -1.00.

Now, plugging back into $normcdf(x|\mu,\sigma) = \frac{1}{2}[1 + \text{erf}\left(\frac{x-\mu}{\sigma*\sqrt{2}}\right)]$. So, after plugging in -1.00 for $\text{erf}\left(\frac{x-\mu}{\sigma*\sqrt{2}}\right)$, we need to add 1 to -1.00 to obtain 0.00, then multiply by ½ to obtain **0.00**, which is our normcdf and which is also our FN.

Finally, TP = 1 – FN, therefore 1.00 – 0.00 = 1.00. As a result, TP = **1.00**.

Now, we just need to plot the point (0.955155, 1.00).

For another example, using 52 as the threshold.

To calculate TN. First step would be to plug in 52 for x, the mean of the negative subjects 29.7959 for μ, and the standard deviation of the negative subjects 10.4323 for σ into $\text{erf}\left(\frac{x-\mu}{\sigma*\sqrt{2}}\right)$ to obtain 1.51.

Now, we need to look up 1.51 in the erf table, which turns out to be 0.96728.

Now that we have the erf value, we just need to plug it back into $normcdf(x|\mu,\sigma) = \frac{1}{2}[1 + \text{erf}\left(\frac{x-\mu}{\sigma*\sqrt{2}}\right)]$. So, after plugging in 0.96728 for $\text{erf}\left(\frac{x-\mu}{\sigma*\sqrt{2}}\right)$, we need to add 1 to 0.96728 to obtain 1.96728, then multiply by ½ to obtain **0.98364**, which is our normcdf and which is also our TN.

To obtain FP, we just need to do 1 – 0.98364 = **0.01636**.

Now, to obtain the TP, we do a similar process. Again, plug in 52 for x, but mean of the positive subjects 72.5686 for μ, and standard deviation of the positive subjects 12.5686 for σ into $\text{erf}\left(\frac{x-\mu}{\sigma*\sqrt{2}}\right)$ to obtain -1.16.

Then, we need to find 1.16 in the erf table, which is 0.89910, adding back the negative -0.89910.

Now, plugging back into $normcdf(x|\mu,\sigma) = \frac{1}{2}[1 + \text{erf}\left(\frac{x-\mu}{\sigma*\sqrt{2}}\right)]$. So, after plugging in -0.89910 for $\text{erf}\left(\frac{x-\mu}{\sigma*\sqrt{2}}\right)$, we need to add 1 to obtain 0.1009, then multiply by ½ to obtain **0.05045**, which is our normcdf and which is also our FN.

Finally, TP = 1 – FN, therefore 1.00 – 0.05045 = 0.94955. As a result, TP = **0.94955**.

Now, we just need to plot the point (0.01636, 0.94955).

For the last example, using 92 as our threshold.

To calculate TN. First step would be to plug in 92 for x, the mean of the negative subjects 29.7959 for μ, and the standard deviation of the negative subjects 10.4323 for σ into $\text{erf}\left(\frac{x-\mu}{\sigma*\sqrt{2}}\right)$ to obtain 4.216.

Now, we need to look up 4.22 in the erf table, which is greater than what the table has therefore we choose the largest x value in the table which is 3.2 and use that the error function value of 3.29, which is 1.00.

Now that we have the erf value, we just need to plug it back into $normcdf(x|\mu,\sigma) = \frac{1}{2}[1 + \text{erf}\left(\frac{x-\mu}{\sigma*\sqrt{2}}\right)]$. So, after plugging in 1.00 for $\text{erf}\left(\frac{x-\mu}{\sigma*\sqrt{2}}\right)$, we need to add to obtain 2.00, then multiply by ½ to obtain **1.00**, which is our normcdf and which is also our TN.

To obtain FP, we just need to do 1.00 – 1.00 = **0.00**.

Now, to obtain the TP, we do a similar process. Again, plug in 92 for x, but mean of the positive subjects 72.5686 for μ, and standard deviation of the positive subjects 12.5686 for σ into $\text{erf}\left(\frac{x-\mu}{\sigma*\sqrt{2}}\right)$ to obtain 1.093.

Then, we need to find 1.09 in the erf table, which is 0.87680.

Now, plugging back into $normcdf(x|\mu,\sigma) = \frac{1}{2}[1 + \text{erf}\left(\frac{x-\mu}{\sigma*\sqrt{2}}\right)]$. So, after plugging in 0.87680 for $\text{erf}\left(\frac{x-\mu}{\sigma*\sqrt{2}}\right)$, we need to add 1 to obtain 1.87680, then multiply by ½ to obtain **0.9384**, which is our normcdf and which is also our FN.

Finally, TP = 1 – FN, therefore 1.00 – 0.9384 = 0.0616. As a result, TP = **0.0616**.

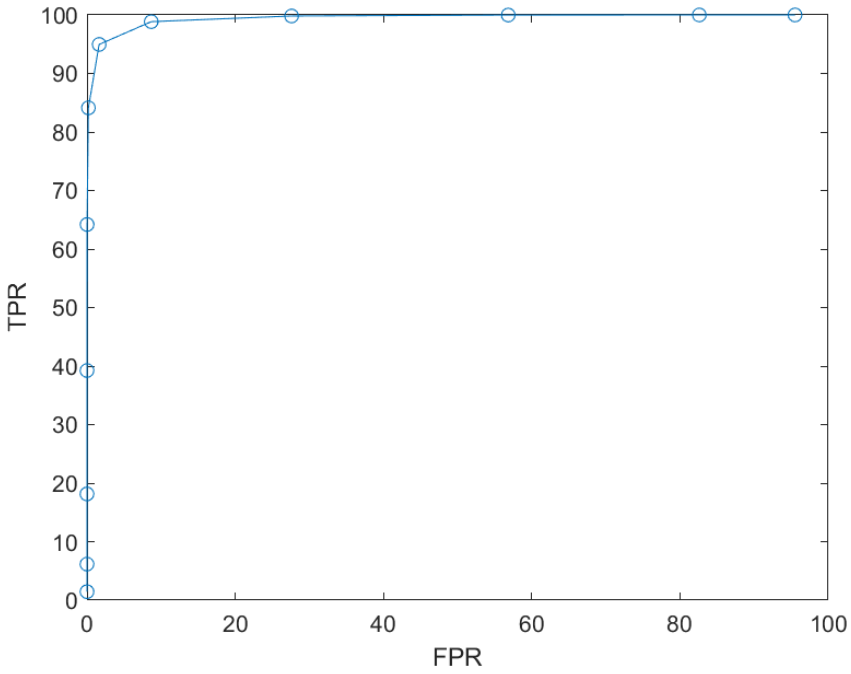
Now, we just need to plot the point (0.00, 0.0616).

Paste MATLAB Code for plotting the ROC curve (use scattered points instead of line segments) Here:

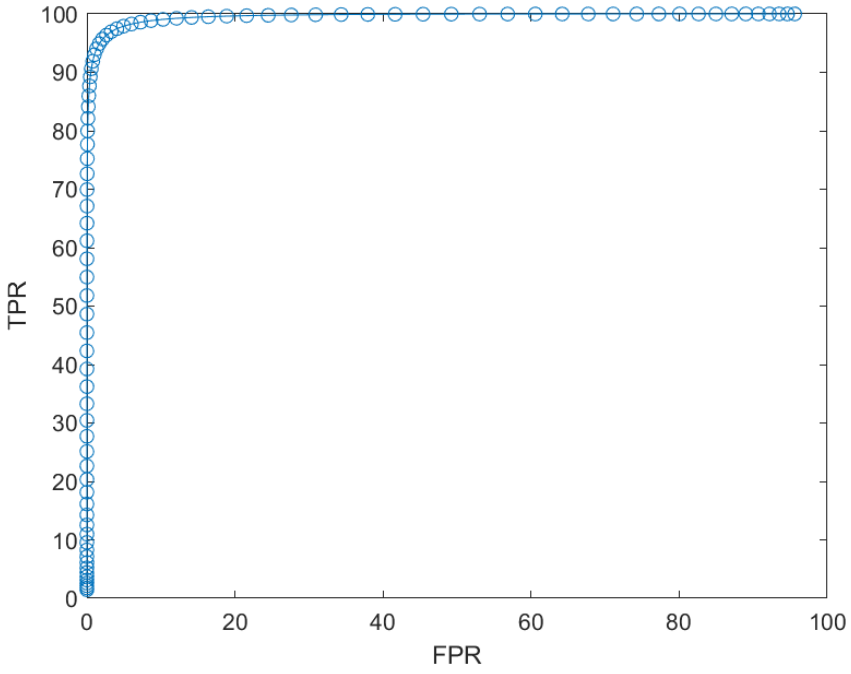
```
% Using the erf table

figure(2);

TPR = [1.0000, 1.0000, 0.9998, 0.9982, 0.9885, 0.9496, 0.8413,
0.6419, 0.3924, 0.1815, 0.0616, 0.0145];
FPR = [0.9552, 0.8261, 0.5683, 0.2760, 0.0867, 0.0164, 0.0019,
0.0001, 0.0000, 0.0000, 0.0000, 0.0000];

for i = 1 : 12
    TPR(i) = TPR(i)*100;
    FPR(i) = FPR(i)*100;
end


plot(FPR, TPR, '-o');
xlabel('FPR') ;
ylabel('TPR') ;


% Using the normcdf function

TNR = [];
FPR = [];
FNR = [];
TPR = [];

for threshold = 12:100
    TNR = [TNR normcdf(threshold, meanNegativeGGOVals,
stdNegativeGGOVals)*100];
    FPR = [FPR (100 - TNR(threshold-11))];
    FNR = [FNR normcdf(threshold, meanPositiveGGOVals,
stdPositiveGGOVals)*100];
    TPR = [TPR (100 - FNR(threshold-11))];
end

plot(FPR, TPR, '-o');
xlabel('FPR') ;
ylabel('TPR') ;
```

Paste ROC Figure here:

Using erf table:

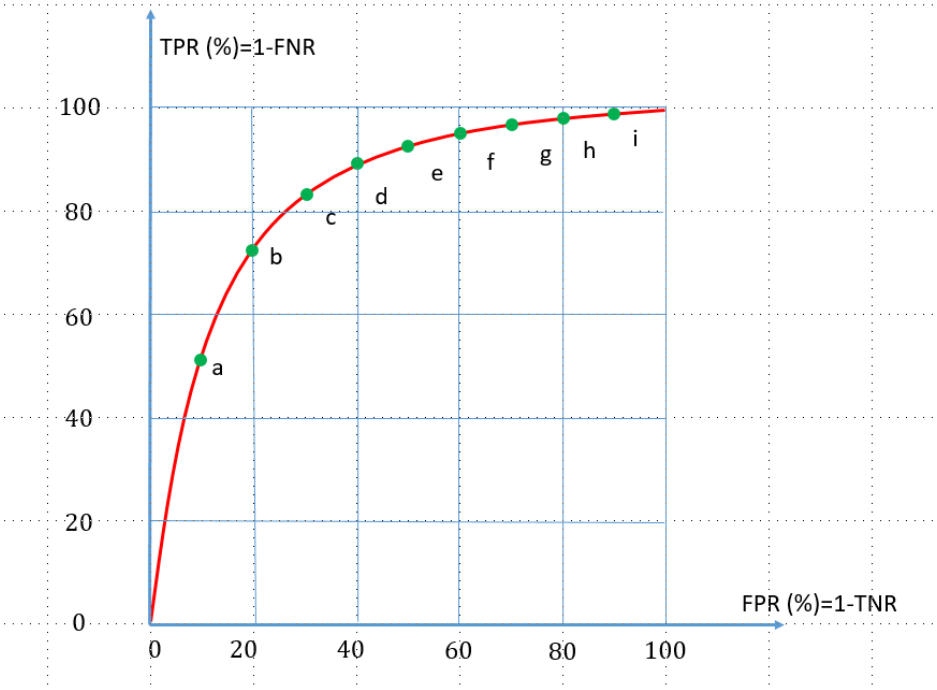


Using Normcdf:



Question 3 [4 points]

Now we look at a dataset collected by a deployed beta-version of the AI system, which resulted in the following ROC curve:



Your task is to control the mis-diagnosis ratio, by reaching a trade-off between TPR and FPR.

1. In particular, you need to tune certain hyper-parameters (e.g. threshold T) for the decision system so that such that: $FNR \leq 20\%$ with the lowest possible FPR. Among the 9 possible operating points (green dots) in the figure above, which one would you choose to satisfy the requirement? Please explain.

2. Assume 20 AI-diagnosed patients had OGG values:
   $V = [70, 60, 30, 80, 40, 20, 50, 90, 85, 45, 75, 65, 55, 15, 35, 45, 45, 65, 65, 75]$.
   Then these 20 patients underwent the more reliable RT-PCR test, which returned, 72 hours later, the following diagnoses, which we regard as "truth":
   C = [P, P, N, P, N, N, N, P, P, N, P, P, P, N, N, N, N, P, P, N].
   where (P: positive, i.e. COVID19; N: negative, i.e. non-COVID19)

   Choose the threshold so that the AI-classification results would have $FNR \leq 10\%$ and $FPR \leq 40\%$? Justify your choice. Note: Calculate FNR using the data points and not a fitted, Gaussian or other, distribution.

3. Using the threshold, you chose in question 2. Answer the following questions:
   a. How many were misdiagnosed?

   b. How many sick patients were diagnosed as healthy?

c. How many healthy patients were diagnosed as sick?

d. What's the false negative ratio?

4. Calculate entries of 2x2 confusion matrix for the 10 patients, use the number of patient in the entries, e.g. number of patients that are N but were misdiagnosed as P, etc.

5. Now draw another confusion matrix and enter the percentages instead, i.e. out of 100% negative cases, what percent were correctly classified as N, etc.

Question 3. <u>Your Answers:</u>

1. Choose operating point

I would choose operating point c because as given in the diagram, TPR is equal to 1 – FNR, in other words, FNR is equal to 1 – TPR; so, to satisfy the $FNR \leq 20\%$ condition, we would need to pick a point where TPR is greater than or equal to 80%. But to also satisfy the "lowest possible FPR" condition, we would try to find a point where the FPR percentage is lowest, in other words, as left to the diagram as possible. The point that satisfies both conditions would be operating point c.

2.

a. List and sort the positive and negative OOG values (you can use MATLAB command sort here)

Positive OOG Values = 55 60 65 65 65 70 75 80 85 90

Negative OOG Values = 15 20 30 35 40 45 45 45 50 75

b. How to choose a threshold so that FNR<=10%?

Since we have the OOG values sorted in the above questions, we just need to look at the "Negative OOG Values" list and find a number such that one or less of the values is greater than that number.

c. How to choose a threshold so that FPR<=40%?

Again, we have the OOG values sorted, so all we need to do is look at the "Positive OOG Values" list and find a number such that four or less of the values in that list are less than the chosen number.

d. What's your choice of the final threshold?

There are many choices for a threshold, but my final choice of the threshold is 55; therefore, any value having a value greater than or equal to 55 will be considered to be a positive.

3.

a. Misdiagnosed number

One patient was misdiagnosed to be negative, while none were misdiagnosed to be positive.

b. Sick diagnosed as healthy

Only one sick patient was diagnosed as healthy.

c. Healthy diagnosed as sick

There were no healthy patients diagnosed as sick.

d. FNR

FNR = Σ False negative / Σ Condition positive

= **1 / 11** = 0.091 = 9.1%

4.

a. Confusion matrix in number

	True Condition		
		Condition positive	Condition negative
Predicted Condition	Predicted condition positive	10	0
	Predicted condition negative	1	9

b. Confusion matrix in percentage

	True Condition		
		Condition positive	Condition negative
Predicted Condition	Predicted condition positive	100%	0%
	Predicted condition negative	10%	90%

References

[Bai 2020] H. X. Bai et al., “Performance of radiologists in differentiating COVID-19 from viral pneumonia on chest CT”, Radiology, 2020. DOI: https://doi.org/10.1148/radiol.2020200823