Sensitivity and Specificity

Although similar, *sensitivity* and *specificity* are not the same as *precision* and *recall*. Here are the definitions:

In the cancer example, sensitivity and specificity are the following:

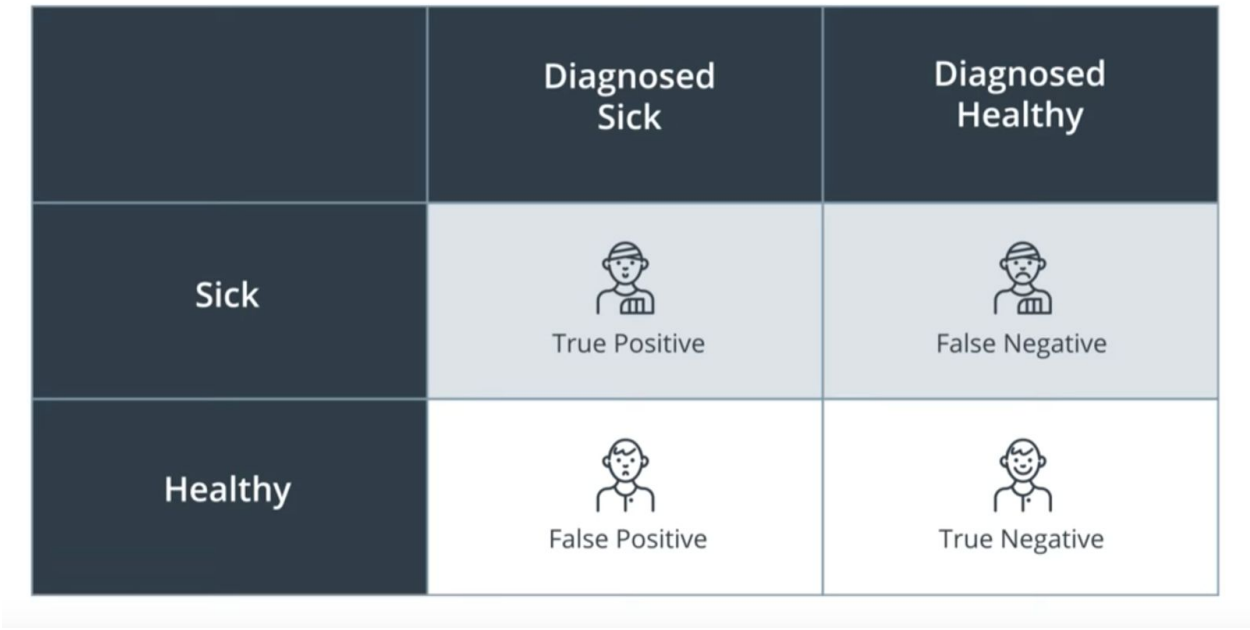
- Sensitivity: Of all the people **with** cancer, how many were correctly diagnosed?
- Specificity: Of all the people **without** cancer, how many were correctly diagnosed?

And precision and recall are the following:

- Recall: Of all the people who **have cancer**, how many did **we diagnose** as having cancer?
- Precision: Of all the people **we diagnosed** with cancer, how many actually **had cancer**?

From here we can see that Sensitivity is Recall, and the other two are not the same thing.

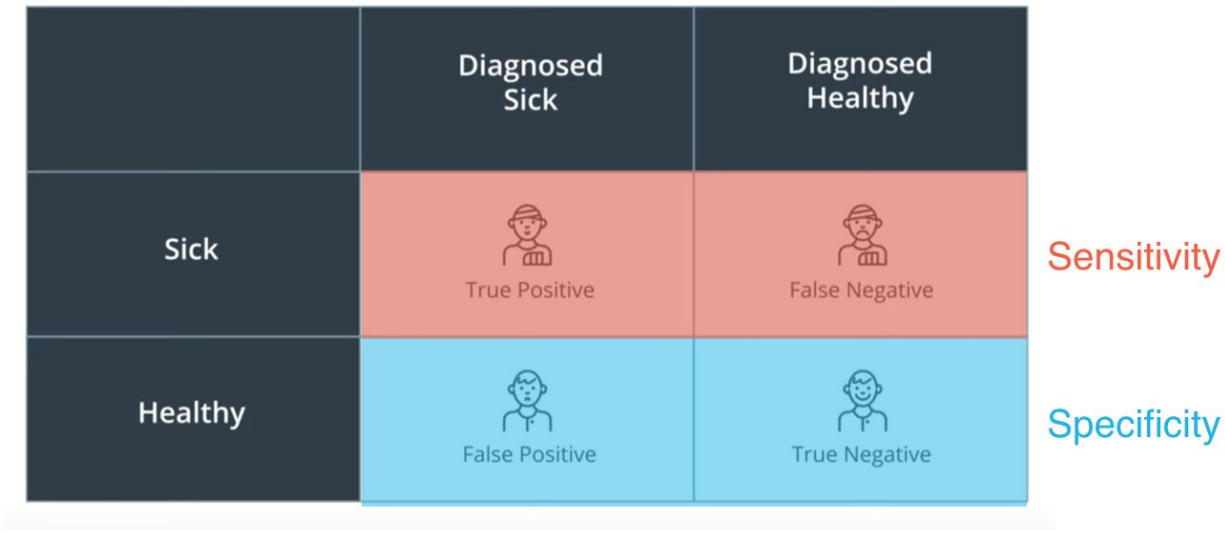
Trust me, we also have a hard time remembering which one is which, so here's a little trick. If you remember from Luis's Evaluation Metrics section, here is the confusion matrix:



Now, sensitivity and specificity are the rows of this matrix. More specifically, if we label

- TP: (True Positives) Sick people that we **correctly** diagnosed as sick.
- TN: (True Negatives) Healthy people that we **correctly** diagnosed as healthy.
- FP: (False Positives) Healthy people that we **incorrectly** diagnosed as sick.
- FN: (False Negatives) Sick people that we **incorrectly** diagnosed as healthy.

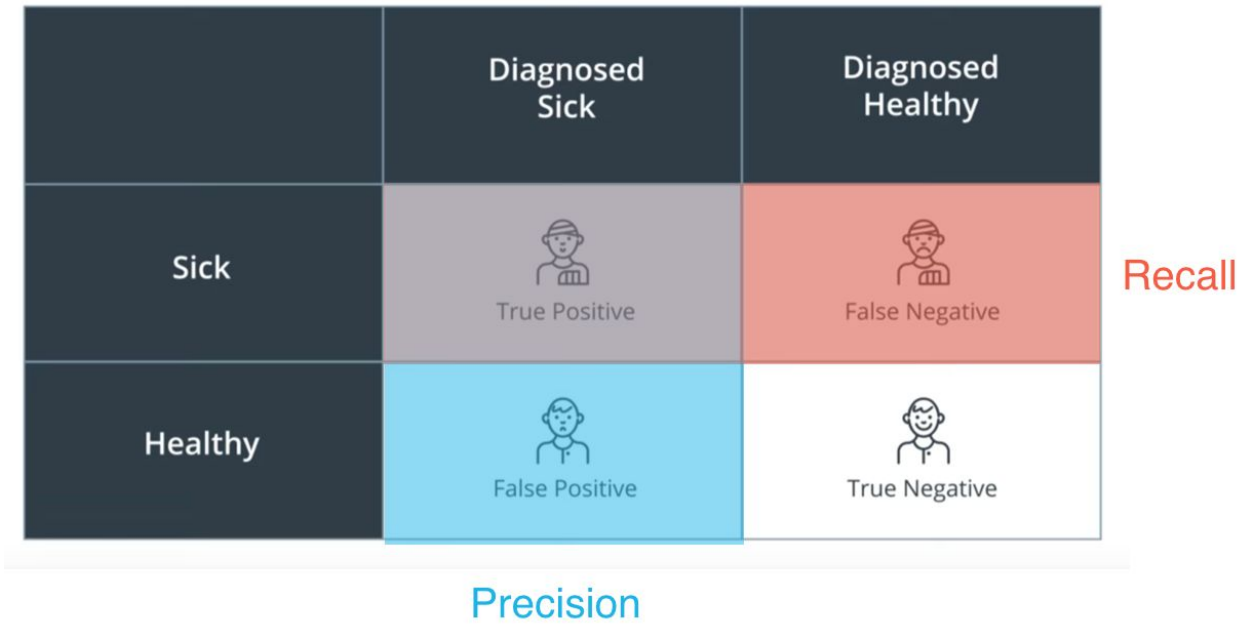
Then: Sensitivity = $\frac{TP}{TP+FN}$ And Specificity = $\frac{TN}{TN+FP}$



Sensitivity and Specificity

And precision and recall are the top row and the left column of the matrix:

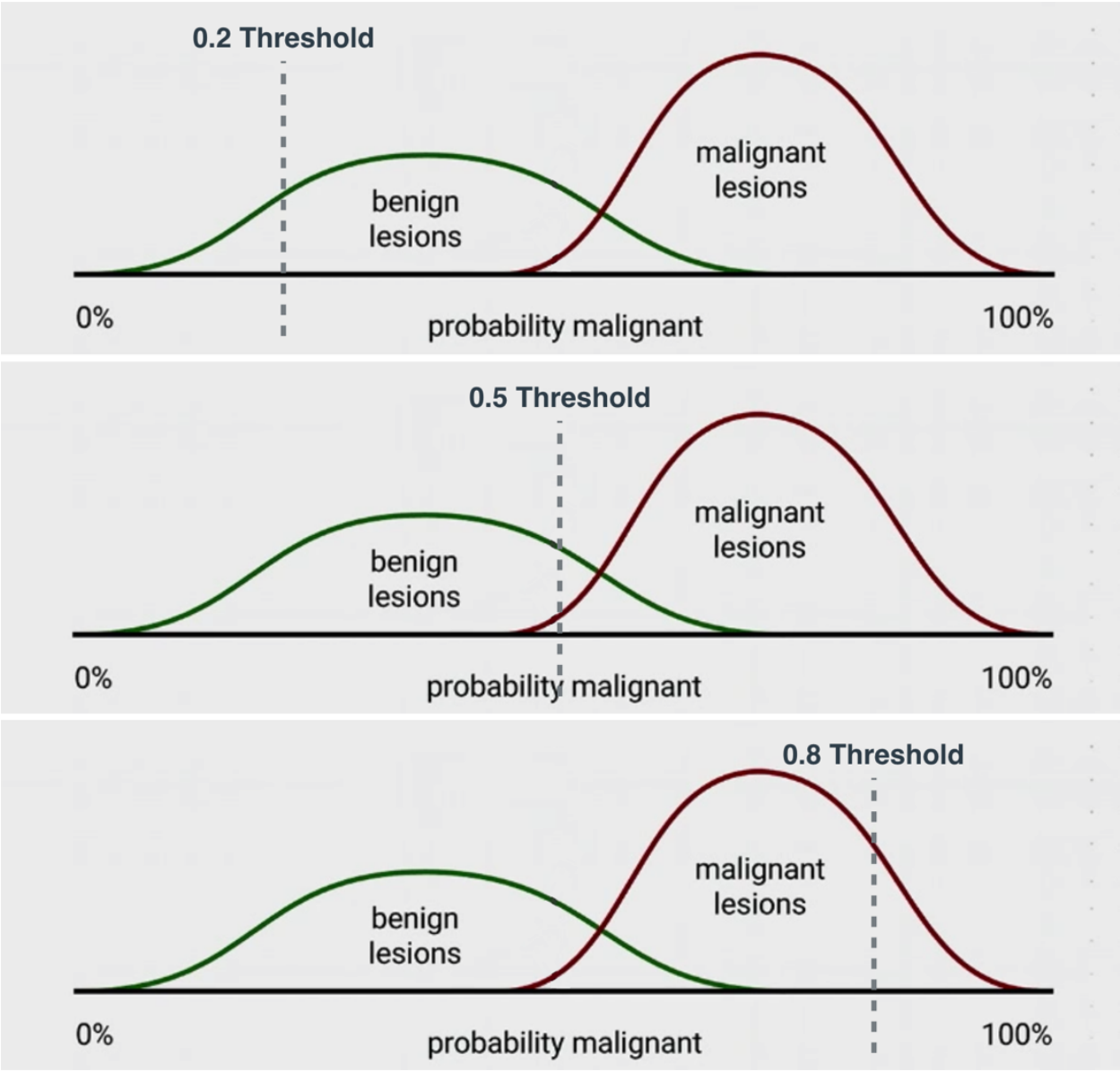
$$\text{Recall} = \frac{TP}{TP + FN} \text{ And Precision} = \frac{TP}{TP + FP}$$



Precision and Recall

The graph below is a histogram of the predictions our model gives in a set of images of lesions, as follows:

- Each point in the horizontal axis is a value
- p
- p from 0 to 1.
- Over each value
- p
- p, we locate all the lesions that our classifier predicted to have probability p of being malignant.



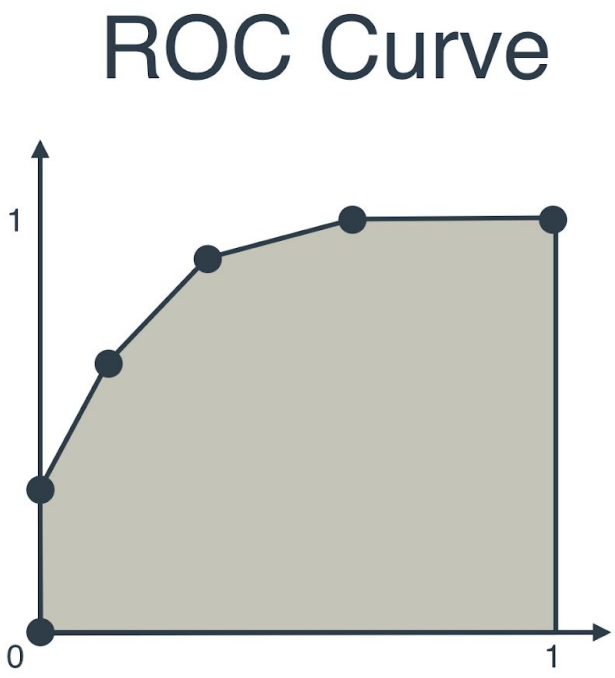
Here we have graphed the thresholds at 0.2, 0.5, and 0.8. Notice how:

- At 0.2, we classify every malignant lesion correctly, yet we also send a lot of benign lesions for more testing.
- At 0.5, we miss some malignant lesions (bad), and we send a few benign lesions for more testing.

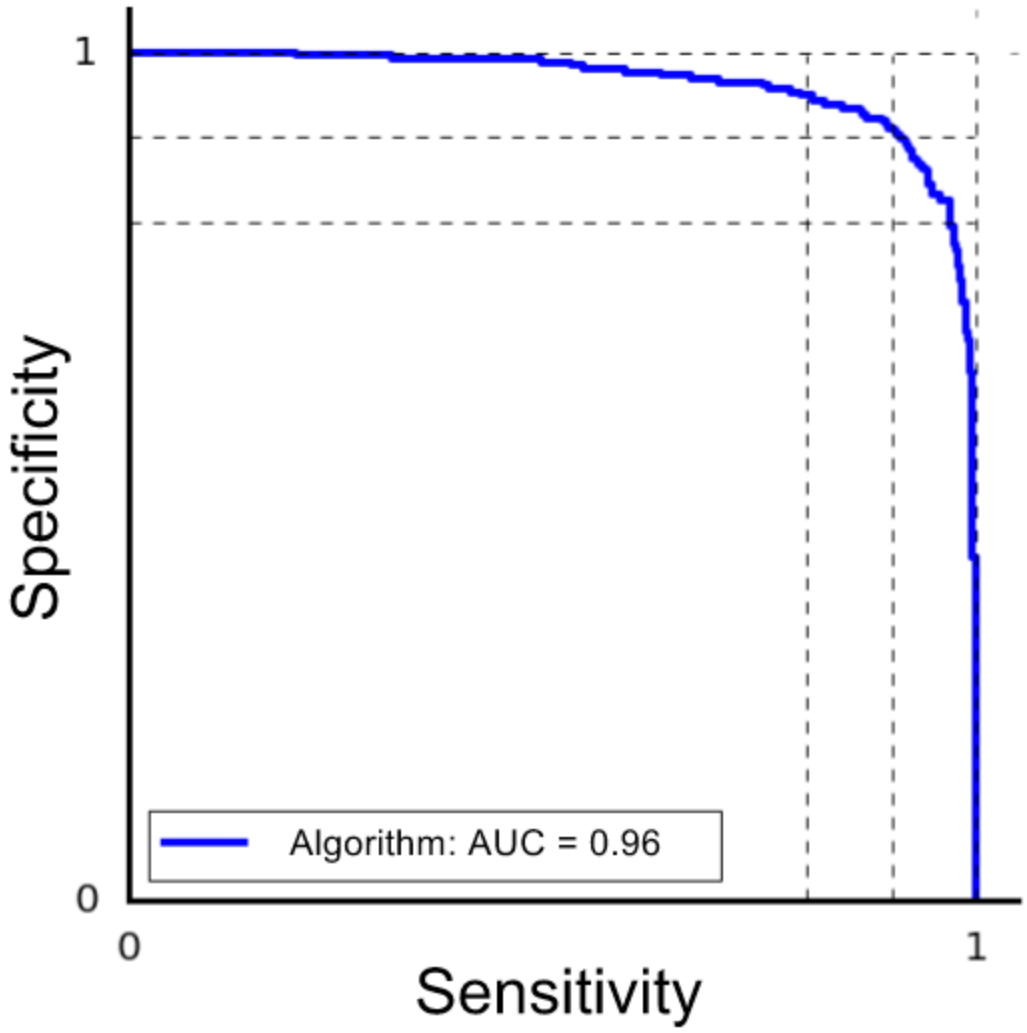
- At 0.8, we correctly classify most of the benign lesions, but we miss many malignant lesions (very bad).

So in this case, it's arguable that 0.2 is better.

The curves have been introduced as follows, where in the horizontal axis we plot the True Positive Rate, and in the vertical axis we plot the False Positive Rate.

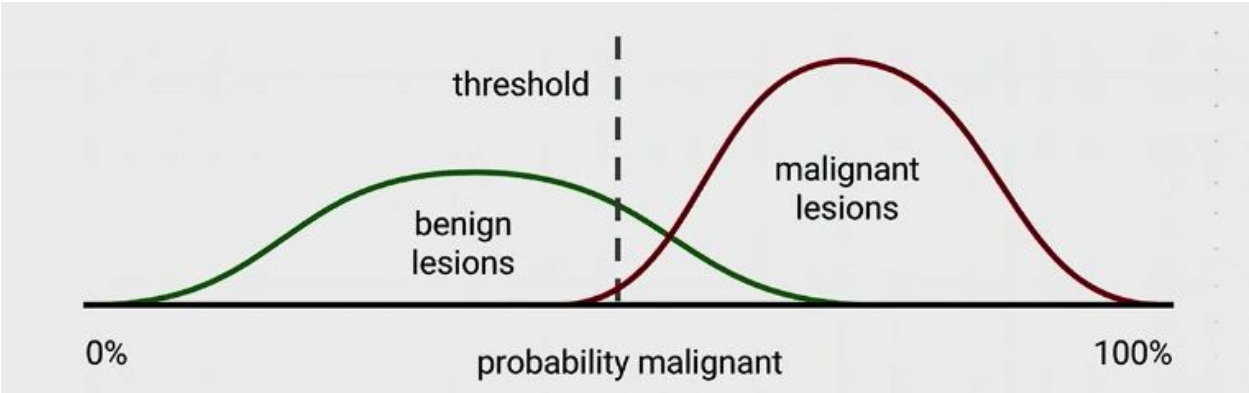


However, you'll see that in this section, I will use a different ROC Curve. The one I use looks like I flipped it sideways, like this:



And there's a really cool reason why I use this one. And it's because it's the curve we get when we plot the sensitivity in the horizontal axis, and the specificity in the vertical axis!

Let me be more specific (yes pun intended). Let's use the same histogram as in the last section.



Recall that the values in the horizontal axis are all the possible thresholds. For any threshold

p

p between 0 and 1, the verdict of the model will be the following: *"Any lesion to the left of this threshold will be considered benign, and any lesion to the right of this threshold will be considered malignant, and sent for more tests."*

Now, for this particular model, we calculate the sensitivity and specificity as follows:

- Sensitivity: Out of all the malignant lesions, what percentage are to the right of the threshold (correctly classified)?
- Specificity: Out of all the benign lesions, what percentage are to the left of the threshold (correctly classified)?

And we plot that point, where the coordinates are (Sensitivity, Specificity). If we plot all the points corresponding to each of the possible thresholds between 0% and 100%,

we'll get the ROC curve that I drew above. Therefore, we can also refer to the ROC curve as the *Sensitivity-Specificity Curve*.

And finally, here's a little animation of the ROC curve getting drawn, as the threshold moves from 0 to 1.

Type 1 and Type 2 Errors

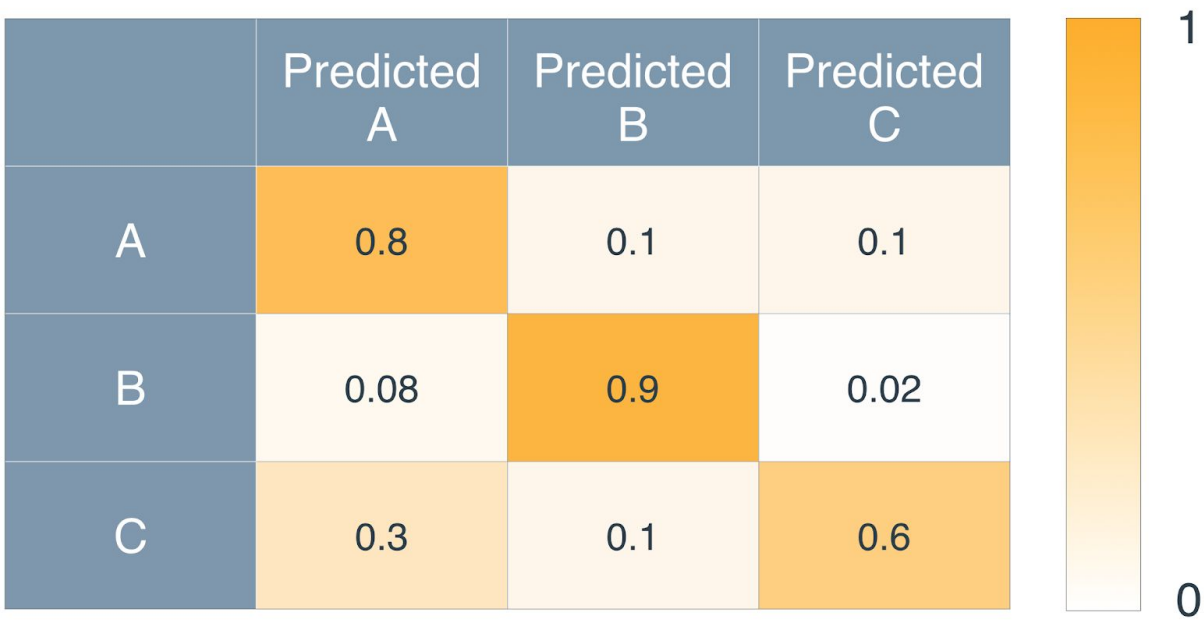
Sometimes in the literature, you'll see False Positives and True Negatives as Type 1 and Type 2 errors. Here is the correspondence:

- **Type 1 Error (Error of the first kind, or False Positive):** In the medical example, this is when we misdiagnose a healthy patient as sick.
- **Type 2 Error (Error of the second kind, or False Negative):** In the medical example, this is when we misdiagnose a sick patient as healthy.

But confusion matrices can be much larger than

2×2

2×2. Here's an example of a larger one. Let's say we have three illnesses called A, B, C. And here is a confusion matrix:



A confusion matrix for three types of illnesses: A, B, and C

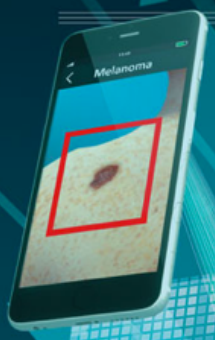
As you can see, each entry in the i-th row and the j-th column will tell you the probability of the patient having illness i and getting diagnosed with illness j.

For example, from the entry on the second row and the first column, we can determine that if a patient has illness B, the probability of getting diagnosed with illness A is exactly 0.08.

Useful Resources

nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE



LESIONS LEARNT

Artificial intelligence powers detection of skin cancer from images PAGES 36 & 115

NATURE.COM/NATURE
2 February 2017 £10
Vol. 542, No. 7639

Here's our publication in Nature.

Other articles about our work:

- Fortune Magazine
- Bloomberg
- BBC
- Wall Street Journal
- Forbes
- Scientific American

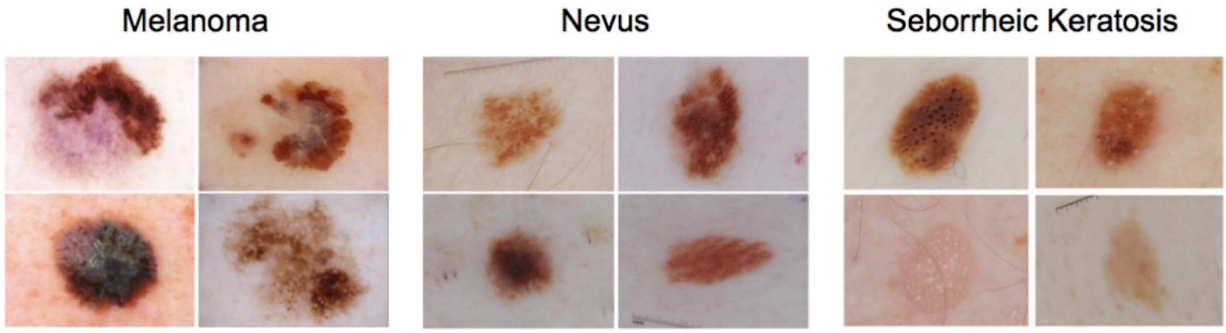
Mini Project: Dermatologist AI

Introduction

In this mini project, you will design an algorithm that can visually diagnose melanoma, the deadliest form of skin cancer. In particular, your algorithm will distinguish this malignant skin tumor from two types of benign lesions (nevi and seborrheic keratoses).

The data and objective are pulled from the 2017 ISIC Challenge on Skin Lesion Analysis Towards Melanoma Detection. As part of the challenge, participants were tasked to design an algorithm to diagnose skin lesion images as one of three different skin

diseases (melanoma, nevus, or seborrheic keratosis). In this project, you will create a model to generate your own predictions.



Getting Started

Clone the repository and create a `data/` folder to hold the dataset of skin images.
```
git clone https://github.com/udacity/dermatologist-ai.git
mkdir data; cd data
```

1.

Create folders to hold the training, validation, and test images.
```
mkdir train; mkdir valid; mkdir test
```

2.
3. Download and unzip the training data (5.3 GB).
4. Download and unzip the validation data (824.5 MB).
5. Download and unzip the test data (5.1 GB).
6. Place the training, validation, and test images in the `data/` folder, at `data/train/`, `data/valid/`, and `data/test/`, respectively. Each folder should

contain three sub-folders (`melanoma/`, `nevus/`, `seborrheic_keratosis/`), each containing representative images from one of the three image classes.

You are free to use any coding environment of your choice to solve this mini project! In order to rank your results, you need only use a pipeline that culminates in a CSV file containing your test predictions.

Create a Model

Use the training and validation data to train a model that can distinguish between the three different image classes. (*After training, you will use the test images to gauge the performance of your model.*)

If you would like to read more about some of the algorithms that were successful in this competition, please read this article that discusses some of the best approaches. A few of the corresponding research papers appear below.

- Matsunaga K, Hamada A, Minagawa A, Koga H. "Image Classification of Melanoma, Nevus and Seborrheic Keratosis by Deep Neural Network Ensemble". International Skin Imaging Collaboration (ISIC) 2017 Challenge at the International Symposium on Biomedical Imaging (ISBI).
- Daz IG. "Incorporating the Knowledge of Dermatologists to Convolutional Neural Networks for the Diagnosis of Skin Lesions". International Skin Imaging Collaboration (ISIC) 2017 Challenge at the International Symposium on Biomedical Imaging (ISBI). (github)

- Menegola A, Tavares J, Fornaciali M, Li LT, Avila S, Valle E. "RECOD Titans at ISIC Challenge 2017". International Skin Imaging Collaboration (ISIC) 2017 Challenge at the International Symposium on Biomedical Imaging (ISBI). (github)

While the original challenge provided additional data (such as the gender and age of the patients), we only provide the image data to you. If you would like to download this additional patient data, you may do so at the competition website.

All three of the above teams increased the number of images in the training set with additional data sources. If you'd like to expand your training set, you are encouraged to begin with the ISIC Archive.

Evaluation

Inspired by the ISIC challenge, your algorithm will be ranked according to three separate categories.

Category 1: ROC AUC for Melanoma Classification

In the first category, we will gauge the ability of your CNN to distinguish between malignant melanoma and the benign skin lesions (nevus, seborrheic keratosis) by calculating the area under the receiver operating characteristic curve (ROC AUC) corresponding to this binary classification task.

If you are unfamiliar with ROC (Receiver Operating Characteristic) curves and would like to learn more, you can check out the documentation in scikit-learn or read this Wikipedia article.

The top scores (from the ISIC competition) in this category can be found in the image below.

Rank	User	Title	Organization	Documentation	Date	Score
1	RECOD Titans	release (rc36xtrm) "alea jacta est"	RECOD Titans / UNICAMP		Wed, 1 Mar 2017, 11:42:07 pm	0.874
2	Lei Bi	EResNet (single scale w/o attributes)	USYD-BMIT		Wed, 1 Mar 2017, 8:04:42 pm	0.870
3	Kazuhisa Matsunaga	ResNet ensemble with normalized image	Casio and Shinshu University joint team		Wed, 1 Mar 2017, 11:18:03 pm	0.868
4	monty python	gpm-LSSSD	Multimedia Processing Group - Universidad Carlos III de Madrid		Wed, 1 Mar 2017, 12:57:35 pm	0.856
5	T D	Last Minute Submission!!!!	University of Guelph - MLRG		Wed, 1 Mar 2017, 11:55:50 pm	0.836
6	Xulei Yang	multi-task deep learning model for skin lesion segmentation and classification-3	Institute of High Performance Computing + National Skin Center, Singapore		Tue, 28 Feb 2017, 6:34:10 pm	0.830
7	Rafael Sousa	Araguaia Medical Vision Lab - GooglAlexNet	Universidade Federal de Mato Grosso		Wed, 1 Mar 2017, 3:26:22 pm	0.805
8	x j	finalv_L2C1_trir	CVI		Wed, 1 Mar 2017, 11:17:56 am	0.804
9	Cristina Vasconcelos	comb	icuff		Tue, 28 Feb 2017, 1:11:21 am	0.791
10	C V	all	icuff		Tue, 28 Feb 2017, 1:06:44 am	0.789
11	Euijoon Ahn	DeepAhn	USYD-BMIT		Wed, 1 Mar 2017, 10:30:13 am	0.786
12	Balázs Harangi	Ensemble of deep convolutional neural networks	University of Debrecen		Wed, 1 Mar 2017, 8:25:16 pm	0.783
13	Matt Berseth	Final Classification Submission	NLPLOGIX / WISEEYE.AI		Tue, 28 Feb 2017, 6:32:47 am	0.782
14	INESC TECNALIA	Final	INESC TEC Porto / TECNALIA		Wed, 1 Mar 2017, 7:05:40 pm	0.765
15	Dylan Shen	task3_final_RQ	Computer Vision Institute, Shenzhen University		Wed, 1 Mar 2017, 9:20:22 pm	0.759
16	Vic Lee	task3_final_Alice	Computer Vision Institute, Shenzhen University		Wed, 1 Mar 2017, 9:11:31 pm	0.757
17	Masih Mahbod	Skin Lesion Classification Using Hybrid Deep Neural Networks	IPA		Wed, 1 Mar 2017, 12:51:43 pm	0.715
18	Dennis Murphree	Transfer Learning from Inception	Dennis Murphree		Wed, 1 Mar 2017, 11:06:33 pm	0.684
19	Hao Chang	MYBrainAI	Yale		Wed, 1 Mar 2017, 11:53:55 pm	0.636
20	Jaisakthi S.M.	Lesion Classification	SSNMLRG		Wed, 1 Mar 2017, 9:25:02 pm	0.623
21	Wenhao Zhang	testPhase	CSMedical		Wed, 1 Mar 2017, 7:08:07 pm	0.500
22	Wiselin Jiji	Dr JIJI P2 Test	Dr Sivanthi Aditanar College of Engineering		Thu, 2 Mar 2017, 12:46:52 am	0.495
23	Yanzhi Song	submit of yanzhi	song		Wed, 1 Mar 2017, 8:05:13 am	0.475

Category 2: ROC AUC for Melanocytic Classification

All of the skin lesions that we will examine are caused by abnormal growth of either melanocytes or keratinocytes, which are two different types of epidermal skin cells. Melanomas and nevi are derived from melanocytes, whereas seborrheic keratoses are derived from keratinocytes.

In the second category, we will test the ability of your CNN to distinuish between melanocytic and keratinocytic skin lesions by calculating the area under the receiver

operating characteristic curve (ROC AUC) corresponding to this binary classification task.

The top scores in this category (from the ISIC competition) can be found in the image below.

Rank	User	Title	Organization	Documentation	Date	Score
1	monty python	gpm-LSSSD	Multimedia Processing Group - Universidad Carlos III de Madrid		Wed, 1 Mar 2017, 12:57:35 pm	0.965
2	Kazuhisa Matsunaga	ResNet ensemble with normalized image	Casio and Shinshu University joint team		Wed, 1 Mar 2017, 11:18:03 pm	0.953
3	RECOD Titans	release (rc36xtrm) "alea jacta est"	RECOD Titans / UNICAMP		Wed, 1 Mar 2017, 11:42:07 pm	0.943
4	Xulei Yang	multi-task deep learning model for skin lesion segmentation and classification-3	Institute of High Performance Computing + National Skin Center, Singapore		Tue, 28 Feb 2017, 6:34:10 pm	0.942
5	T D	Last Minute Submission!!!!	University of Guelph - MLRG		Wed, 1 Mar 2017, 11:55:50 pm	0.935
6	Lei Bi	EResNet (single scale w/o attributes)	USYD-BMIT		Wed, 1 Mar 2017, 8:04:42 pm	0.921
7	C V	all	icuff		Tue, 28 Feb 2017, 1:06:44 am	0.911
8	Cristina Vasconcelos	comb	icuff		Tue, 28 Feb 2017, 1:11:21 am	0.911
9	Masih Mahbod	Skin Lesion Classification Using Hybrid Deep Neural Networks	IPA		Wed, 1 Mar 2017, 12:51:43 pm	0.908
10	Dylan Shen	task3_final_RQ	Computer Vision Institute, Shenzhen University		Wed, 1 Mar 2017, 9:20:22 pm	0.886
11	Euijoon Ahn	DeepAhn	USYD-BMIT		Wed, 1 Mar 2017, 10:30:13 am	0.885
12	INESC TECNALIA	Final	INESC TEC Porto / TECNALIA		Wed, 1 Mar 2017, 7:05:40 pm	0.881
13	Vic Lee	task3_final_Alice	Computer Vision Institute, Shenzhen University		Wed, 1 Mar 2017, 9:11:31 pm	0.875
14	Balázs Harangi	Ensemble of deep convolutional neural networks	University of Debrecen		Wed, 1 Mar 2017, 8:25:16 pm	0.867
15	x j	finalv_L2C1_trir	CVI		Wed, 1 Mar 2017, 11:17:56 am	0.855
16	Rafael Sousa	Araguaia Medical Vision Lab - GooglAlexNet	Universidade Federal de Mato Grosso		Wed, 1 Mar 2017, 3:26:22 pm	0.840
17	Matt Berseth	Final Classification Submission	NLPLOGIX / WISEEYE.AI		Tue, 28 Feb 2017, 6:32:47 am	0.827
18	Dennis Murphree	Transfer Learning from Inception	Dennis Murphree		Wed, 1 Mar 2017, 11:06:33 pm	0.817
19	Wenhao Zhang	testPhase	CSMedical		Wed, 1 Mar 2017, 7:08:07 pm	0.817
20	Hao Chang	MYBrainAI	Yale		Wed, 1 Mar 2017, 11:53:55 pm	0.774
21	Jaisakthi S.M.	Lesion Classification	SSNMLRG		Wed, 1 Mar 2017, 9:25:02 pm	0.687
22	Wiselin Jiji	Dr Jiji P2 Test	Dr Sivanthi Aditanar College of Engineering		Thu, 2 Mar 2017, 12:46:52 am	0.498
23	Yanzhi Song	submit of yanzhi	song		Wed, 1 Mar 2017, 8:05:13 am	0.456

Category 3: Mean ROC AUC

In the third category, we will take the average of the ROC AUC values from the first two categories.

The top scores in this category (from the ISIC competition) can be found in the image below.

Rank	User	Title	Organization	Documentation	Date	Score
1	Kazuhisa Matsunaga	ResNet ensemble with normalized image	Casio and Shinshu University joint team		Wed, 1 Mar 2017, 10:18:03 pm	0.911
2	monty python	gpm-LSSSD	Multimedia Processing Group - Universidad Carlos III de Madrid		Wed, 1 Mar 2017, 11:57:35 am	0.910
3	RECOD Titans	release (rc36xtrm) "alea jacta est"	RECOD Titans / UNICAMP		Wed, 1 Mar 2017, 10:42:07 pm	0.908
4	popleyi .	EResNet (single scale w/o attributes)	USYD-BMIT		Wed, 1 Mar 2017, 7:04:42 pm	0.896
5	Xulei Yang	multi-task deep learning model for skin lesion segmentation and classification-3	Institute of High Performance Computing + National Skin Center, Singapore		Tue, 28 Feb 2017, 5:34:10 pm	0.886
6	T D	Last Minute Submission!!!!	University of Guelph - MLRG		Wed, 1 Mar 2017, 10:55:50 pm	0.886
7	Cristina Vasconcelos	comb	icuff		Tue, 28 Feb 2017, 12:11:21 am	0.851
8	Cristina Vasconcelos	all	icuff		Tue, 28 Feb 2017, 12:06:44 am	0.850
9	Euijoon Ahn	DeepAhn	USYD-BMIT		Wed, 1 Mar 2017, 9:30:13 am	0.836
10	x j	finalv_L2C1_trir	CVI		Wed, 1 Mar 2017, 10:17:56 am	0.829
11	Balázs Harangi	Ensemble of deep convolutional neural networks	University of Debrecen		Wed, 1 Mar 2017, 7:25:16 pm	0.825
12	INESC TECNALIA	Final	INESC TEC Porto / TECNALIA		Wed, 1 Mar 2017, 6:05:40 pm	0.823
13	Rafael Sousa	Araguaia Medical Vision Lab - GooglAlexNet	Universidade Federal de Mato Grosso		Wed, 1 Mar 2017, 2:26:22 pm	0.823
14	Dylan Shen	task3_final_RQ	Computer Vision Institute, Shenzhen University		Wed, 1 Mar 2017, 8:20:22 pm	0.823
15	Vic Lee	task3_final_Alice	Computer Vision Institute, Shenzhen University		Wed, 1 Mar 2017, 8:11:31 pm	0.816
16	Masih Mahbod	Skin Lesion Classification Using Hybrid Deep Neural Networks	IPA		Wed, 1 Mar 2017, 11:51:43 am	0.811
17	Matt Berseth	Final Classification Submission	NLPLOGIX / WISEEYE.AI		Tue, 28 Feb 2017, 5:32:47 am	0.804
18	Dennis Murphree	Transfer Learning from Inception	Dennis Murphree		Wed, 1 Mar 2017, 10:06:33 pm	0.750
19	Hao Chang	MYBrainAI	Yale		Wed, 1 Mar 2017, 10:53:55 pm	0.705
20	Wenhao Zhang	testPhase	CSMedical		Wed, 1 Mar 2017, 6:08:07 pm	0.658
21	Jaisakthi S.M.	Lesion Classification	SSNMLRG		Wed, 1 Mar 2017, 8:25:02 pm	0.655
22	Wiselin Jiji	Dr Jiji P2 Test	Dr Sivanthi Aditanar College of Engineering		Wed, 1 Mar 2017, 11:46:52 pm	0.497
23	Yanzhi Song	submit of yanzhi	song		Wed, 1 Mar 2017, 7:05:13 am	0.465

Getting your Results

Once you have trained your model, create a CSV file to store your test predictions. Your file should have exactly 600 rows, each corresponding to a different test image, **plus** a header row. You can find an example submission file (`sample_submission.csv`) in the repository.

Your file should have exactly 3 columns:

- `Id` - the file names of the test images (in the **same** order as the sample submission file)
- `task_1` - the model's predicted probability that the image (at the path in `Id`) depicts melanoma

- `task_2` - the model's predicted probability that the image (at the path in `Id`) depicts seborrheic keratosis

Once the CSV file is obtained, you will use the `get_results.py` file to score your submission. To set up the environment to run this file, you need to create (and activate) an environment with Python 3.5 and a few pip-installable packages:

```
conda create --name derm-ai python=3.5
source activate derm-ai
pip install -r requirements.txt
```

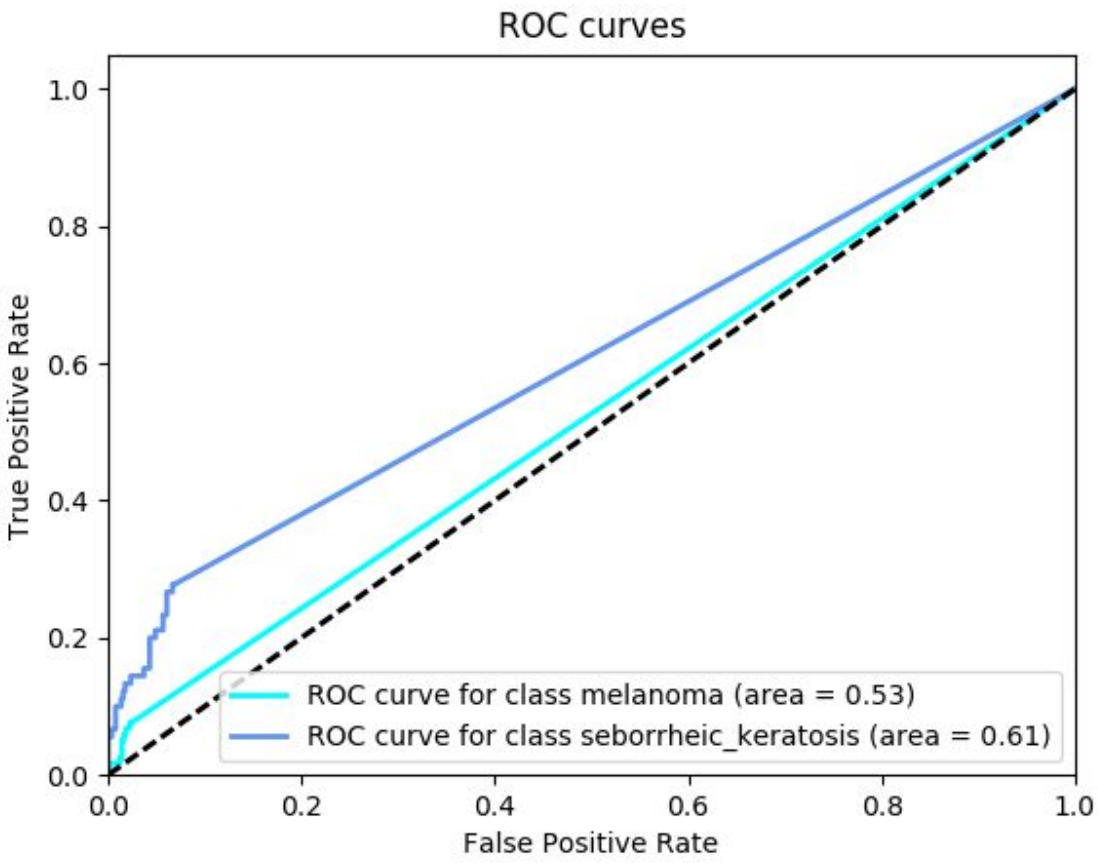
Once you have set up the environment, run the following command to see how the sample submission performed:

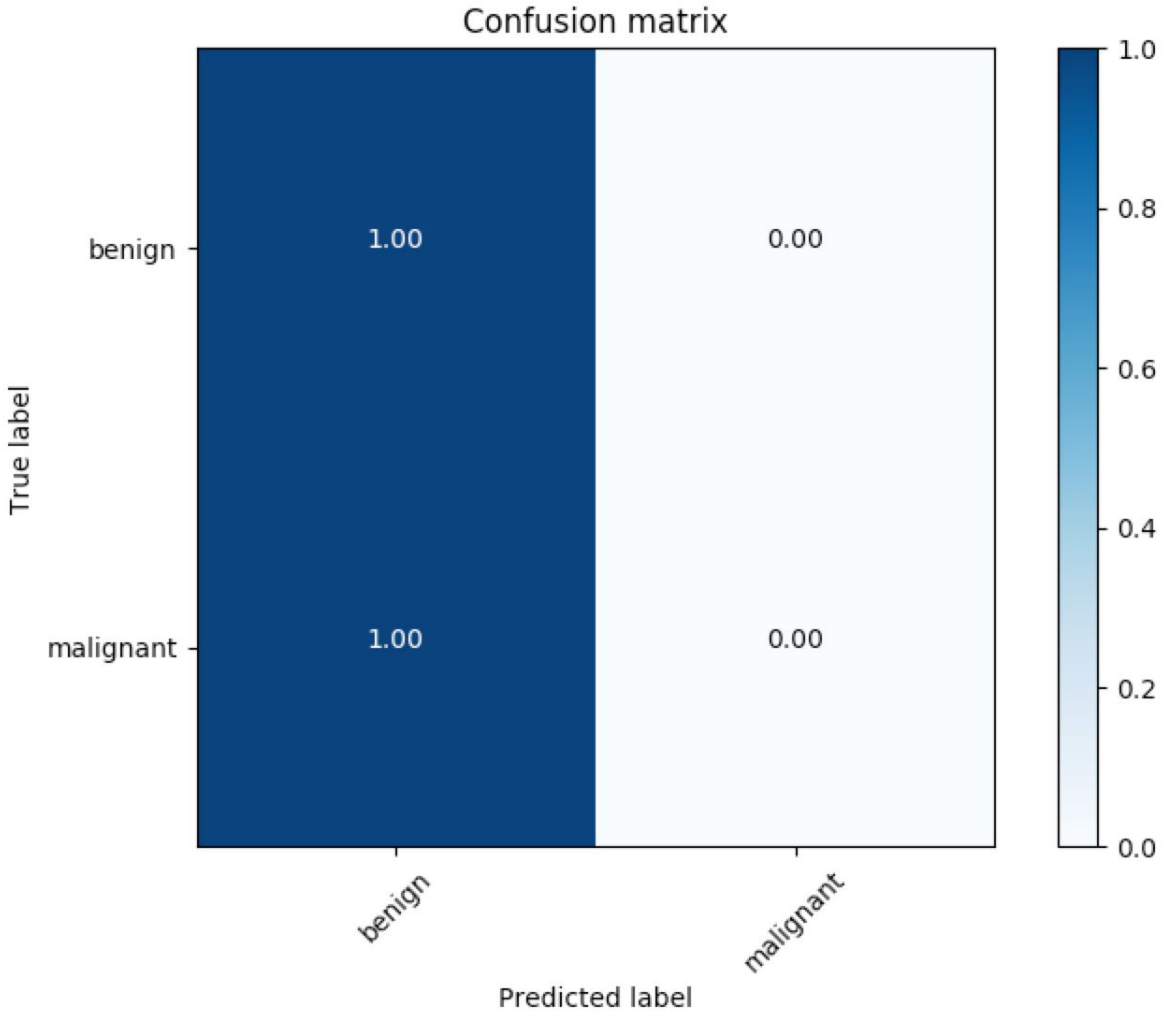
```
python get_results.py sample_predictions.csv
```

Check the terminal output for the scores obtained in the three categories:

```
Category 1 Score: 0.526
Category 2 Score: 0.606
Category 3 Score: 0.566
```

The corresponding **ROC curves** appear in a pop-up window, along with the **confusion matrix** corresponding to melanoma classification.





As you can see from the confusion matrix, the sample submission currently predicts that most of the images in the test dataset correspond to benign lesions. Let's see if your model can improve these results, towards better detecting cancer!

The code for generating the confusion matrix assumes that the threshold for classifying melanoma is set to 0.5. To change this threshold, you need only supply an additional

command-line argument when calling the `get_results.py` file. For instance, to set the threshold at 0.4, you need only run:

```
python get_results.py sample_predictions.csv 0.4
```

To test **your own** submission, change the code to instead include the path to **your** CSV file.