Project Report

Definition

Project Overview

Skin cancer is the most prevalent type of cancer. Melanoma, specifically, is responsible for 75% of skin cancer deaths, despite being the least common skin cancer. The American Cancer Society estimates over 100,000 new melanoma cases will be diagnosed in 2020. It's also expected that almost 7,000 people will die from the disease. As with other cancers, early and accurate detection—potentially aided by data science—can make treatment more effective.

Currently, dermatologists evaluate every one of a patient's moles to identify outlier lesions or “ugly ducklings” that are most likely to be melanoma. Existing AI approaches have not adequately considered this clinical frame of reference. Dermatologists could enhance their diagnostic accuracy if detection algorithms take into account “contextual” images within the same patient to determine which images represent a melanoma. If successful, classifiers would be more accurate and could better support dermatological clinic work.

As the leading healthcare organization for informatics in medical imaging, the Society for Imaging Informatics in Medicine (SIIM)'s mission is to advance medical imaging informatics through education, research, and innovation in a multi-disciplinary community. SIIM is joined by the International Skin Imaging Collaboration (ISIC), an international effort to improve melanoma diagnosis. The ISIC Archive contains the largest publicly available collection of quality-controlled dermoscopic images of skin lesions. This dataset has been relesed on the competitive data science platfor Kaggle on this link

Source

Problem Statement

In this competition, participants will identify melanoma in images of skin lesions. In particular, they will use images within the same patient and determine which are likely to represent a melanoma. Specifically, participants need to predict a binary target for each image ie, the probability (floating point) between 0.0 and 1.0 that the lesion in the image is malignant (the target).

For this competition, we are going to build an image classifier using deep learning. We will need to begin with image pre-processing as we have images of varying sizes, for eg., 1024x1024x3 vs 512x512x3 etc. We can combine the results of the image classifier with a tabular data classifier on image metadata for ensembling.

For working with tfrecords, Tensorflow library will be a good choice to build our neural network. We can use stratified k-folds for model validation before making predictions on the test set. Since training a deep learning model on a large image dataset (~120 GB) is going to be a compute heavy task, Kaggle notebooks which offer free GPUs (and TPUs) can serve as the ideal solution for training this model. Additionally, pretrained models such as ImageNet might be explored to get a good score.

Metrics

Evaluation metric for this image classification Kaggle competition is Area under the ROC curve. An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

- True Positive Rate (TPR), or Recall, is defined as follows :

```
TPR = TP / (TP + FN)
```

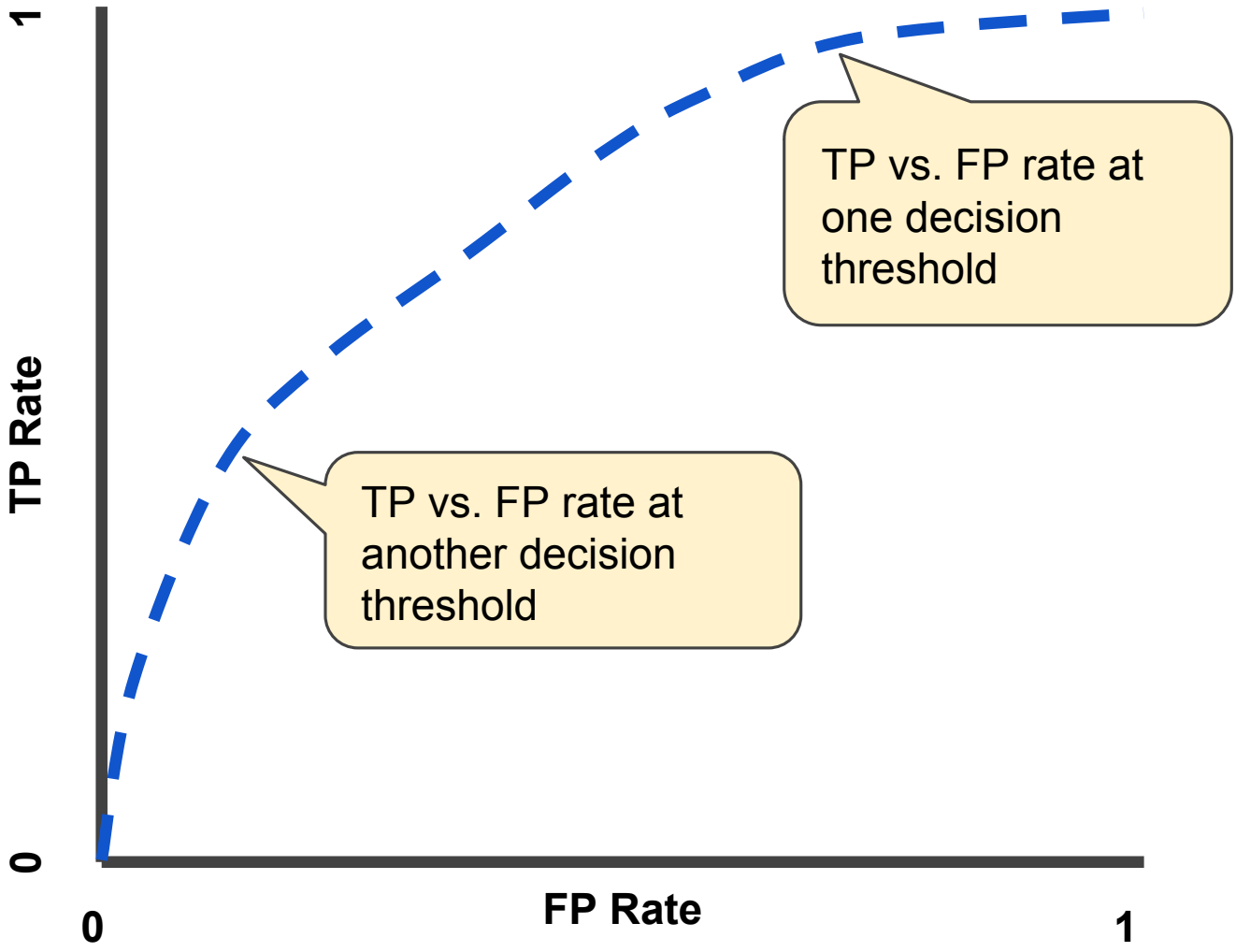
where TP = True Positives & FN = False Negtives

- False Positive Rate (FPR) is defined as follows :

```
FPR = FP / (FP + TN)
```

where, FP = False Positives & TN = True Negatives

An ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives.



Source: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

AUC is desirable for the following two reasons:

- AUC is scale-invariant. It measures how well predictions are ranked, rather than their absolute values.
- AUC is classification-threshold-invariant. It measures the quality of the model's predictions irrespective of what classification threshold is chosen.

Analysis

Data Exploration

The dataset consists of images in :

```
DICOM format
JPEG format in JPEG directory
TFRecord format in tfrecords directory
```

Additionally, there is a metadata comprising of train, test and submission file in CSV format.

File Sizes

Total size of the dataset (Images + Files) - 108.19 GB

The sizes of the CSV files are shown below :

- train.csv - 1.96 MB - (33126 records, 8 columns)
- test.csv - 479 KB - (10982 records, 5 columns)
- sample_submission.csv - 161 KB (10982 records, 2 columns)

Column description

The description for columns in these 3 files are :

- image_name - unique identifier, points to filename of related DICOM image
- patient_id - unique patient identifier
- sex - the sex of the patient (when unknown, will be blank)
- age_approx - approximate patient age at time of imaging
- anatom_site_general_challenge - location of imaged site
- diagnosis - detailed diagnosis information (train only)
- benign_malignant - indicator of malignancy of imaged lesion
- target - binarized version of the target variable, the value 0 denotes `benign`, and 1 indicates `malignant`

Sample values from train & test CSVs

Missing Values

Imputing missing values

Summary statistics of metadata

Exploratory Visualization

A visualization has been provided that summarizes or extracts a relevant characteristic or feature about the dataset or input data with thorough discussion. Visual cues are clearly defined.

Algorithms and Techniques

Algorithms and techniques used in the project are thoroughly discussed and properly justified based on the characteristics of the problem.

Benchmark

A good baseline model can be created by using 3 features, namely, age, sex and the location of the images site. We can calculate the grouped mean value for each combination of these features in the train set and use that to make predictions on the test set.

Predictions using this simple mean value of the target variable gives an Area under ROC value of `0.699` on the public leaderboard !!

Our final classifier should be able to beat atleast this benchmark to be deemed useful.

The code for the baseline model is added in the Github repo (melanoma-simple-baseline.ipynb).

Methodology

Data Preprocessing

All preprocessing steps have been clearly documented. Abnormalities or characteristics of the data or input that needed to be addressed have been corrected. If no data preprocessing is necessary, it has been clearly justified.

Implementation

The process for which metrics, algorithms, and techniques were implemented with the given datasets or input data has been thoroughly documented. Complications that occurred during the coding process are discussed.

Refinement

The process of improving upon the algorithms and techniques used is clearly documented. Both the initial and final solutions are reported, along with intermediate solutions, if necessary.

Results

Model Evaluation and Validation

The final model's qualities—such as parameters—are evaluated in detail. Some type of analysis is used to validate the robustness of the model's solution.

Justification

The final results are compared to the benchmark result or threshold with some type of statistical analysis. Justification is made as to whether the final model and solution is significant enough to have adequately solved the problem.