8. **Measuring Performance**

When creating different machine learning models, it is important to keep in mind which kind of way the model’s performance will be measured. In some cases, it might be smart to measure how well the model predicts positives also called recall or sensitivity. In some cases, Specificity is a great measure of how well the model correctly predicts negatives.

However, in the case of this project, we found f1-score to be the best way to measure our performance. The f1-score is a combination of both recall and specificity and will see how many correct predictions the model made from the entire dataset.

F1-Score worked great in our project, since the goal was to find how many images the model managed to correctly predict. The way we gathered data for generating performance measurements was that we created a model, trained it and then we used the models predict function to predict a certain number of images. For the graph data that we generated, we used approximately 2000 images (since each prediction was way faster) and for the entire images we predicted around 200 images (due to the slow prediction times). Then we used scikit-learn’s classification\_report function to compare all the predictions to the actual labels of the same images. The output of this function looks something like the table shown in table-1 below:

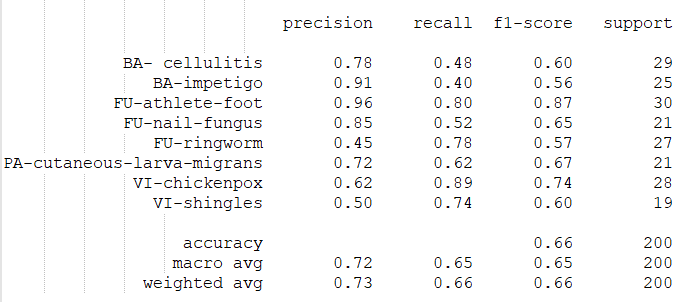


table-1

As shown in table-1 there is also a column for support, which is how many predictions of the certain class that has been done, in the example we got an accuracy of 0.66 % which was not quite the performance we wanted to get considering the importance of a high accuracy when dealing with patients and disease treatments.