2018 Model Ensembles Evaluation Using 2019 Training Data

Overview

In this model evaluation, the test data set used is extracted from the 2019 training data set. The idea is to have equal number of images in each class for the testing. This process is done by randomly sampling 100 images from the 2019 data for each of the classes. The shuffling process allows for a different test set to be used for testing each time. Through testing with multiple randomly generated test sets, we are getting in range of 70% to 85% for overall accuracy for ensemble 3. The best result that we got from a test set is 88% overall accuracy for ensemble 3. The high accuracy we get from this test set is most likely due to that this test set contains lesser images with hair and other artifacts that can affect the prediction process.

Summary of Performance of the 3 ensembles

The table shows the best result we get from the multiple randomly generated test data set. Ensemble 3 performs the best out of the 3 ensembles with 88% in overall accuracy.

ensemble	Ensemble Description	2019 Data Set Best Overall Accuracy
1	Ensemble of Base Model	65 %
2	Ensemble of Modified Base Model (Dropout 25%, 1 extra CNN Layer)	13%
3	Ensemble of Modified Base Model (Dropout 40%, 1 extra CNN Layer)	88%

2019 Sampled Data Prediction Metrics

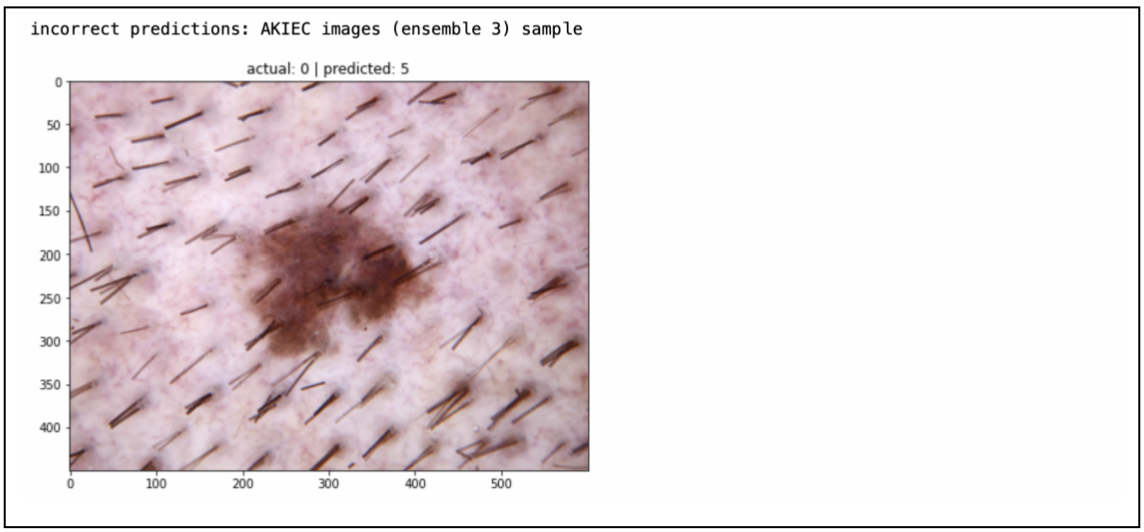
The following table shows the confusion matrix and classification report for the three ensembles using the dataset sample that gave the best result. From the confusion matrix, we can see that for ensemble 3, some images still tend to be misclassified as nevus, but the predictions for the minority classes (classes other than nevus) are showing a more optimal results both in their respective precision and recall.

Ensemble	Result
1	ensemble 1 Confusion Matrix [[69 1 2 0 0 28 0] [0 84 2 0 0 14 0] [0 2 40 0 0 58 0] [1 5 0 54 0 40 0] [0 0 1 0 36 63 0] [0 0 0 0 0 100 0] [0 9 0 0 1 15 75]] Classification Report precision recall f1-score support akiec 0.99 0.69 0.81 100 bcc 0.83 0.84 0.84 100 bkl 0.89 0.40 0.55 100 df 1.00 0.54 0.70 100 mel 0.97 0.36 0.53 100 nv 0.31 1.00 0.48 100 vasc 1.00 0.75 0.86 100 accuracy 0.65 700 macro avg 0.86 0.65 0.68 700 weighted avg 0.86 0.65 0.68 700
2	ensemble 2 Confusion Matrix [[0 0 31 69 0 0 0] [0 0 27 73 0 0 0] [0 0 20 80 0 0 0] [0 0 30 70 0 0 0] [1 0 11 88 0 0 0] [0 0 24 76 0 0 0] [0 0 12 88 0 0 0]] Classification Report precision recall f1-score support akiec 0.00 0.00 0.00 100 bcc 0.00 0.00 0.00 100 bkl 0.13 0.20 0.16 100 df 0.13 0.70 0.22 100 mel 0.00 0.00 0.00 100 nv 0.00 0.00 0.00 100 vasc 0.00 0.00 0.00 100 accuracy 0.13 700 macro avg 0.04 0.13 0.05 700 weighted avg 0.04 0.13 0.05 700

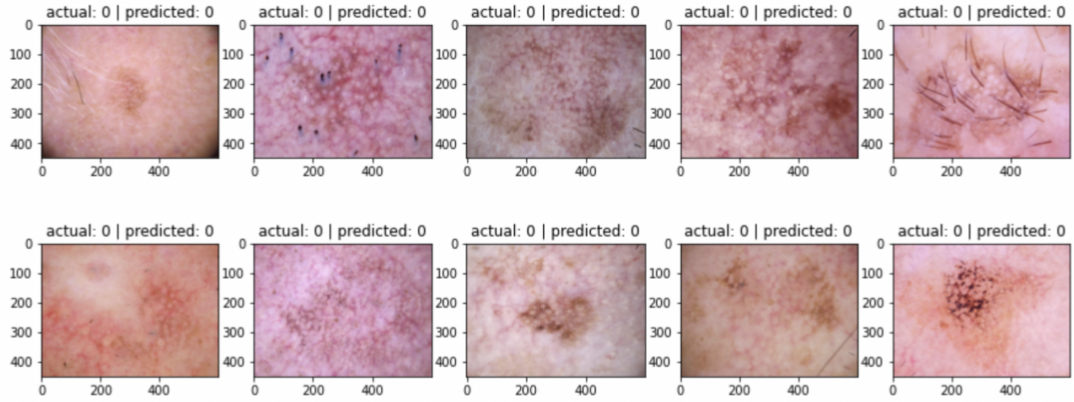
3	ensemble 3 Confusion Matrix [[99 0 0 0 0 1 0] [2 94 2 0 0 2 0] [4 1 76 1 5 13 0] [1 1 0 93 0 5 0] [2 1 1 0 77 19 0] [0 1 0 0 1 98 0] [1 9 0 0 1 10 79]] Classification Report precision recall f1-score support akiec 0.91 0.99 0.95 100 bcc 0.88 0.94 0.91 100 bkl 0.96 0.76 0.85 100 df 0.99 0.93 0.96 100 mel 0.92 0.77 0.84 100 nv 0.66 0.98 0.79 100 vasc 1.00 0.79 0.88 100 accuracy 0.88 700 macro avg 0.90 0.88 0.88 700 weighted avg 0.90 0.88 0.88 700

2019 Data Prediction result samples

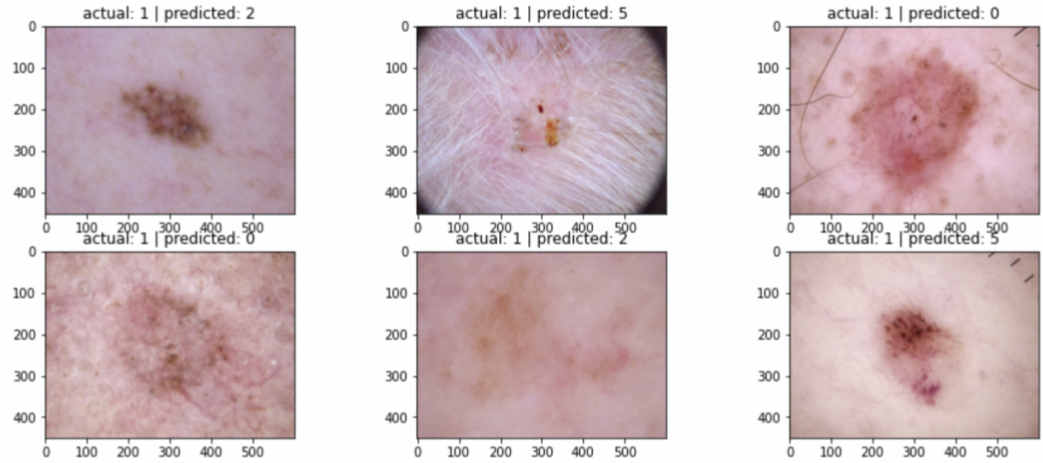
From the prediction result, we can see that the incorrect predictions for the minority classes (classes other than nevus) tends to contain a large amount of hair or other artifacts. For example, the only incorrect prediction in the AKIEC class is of an image with a significant amount of hair.



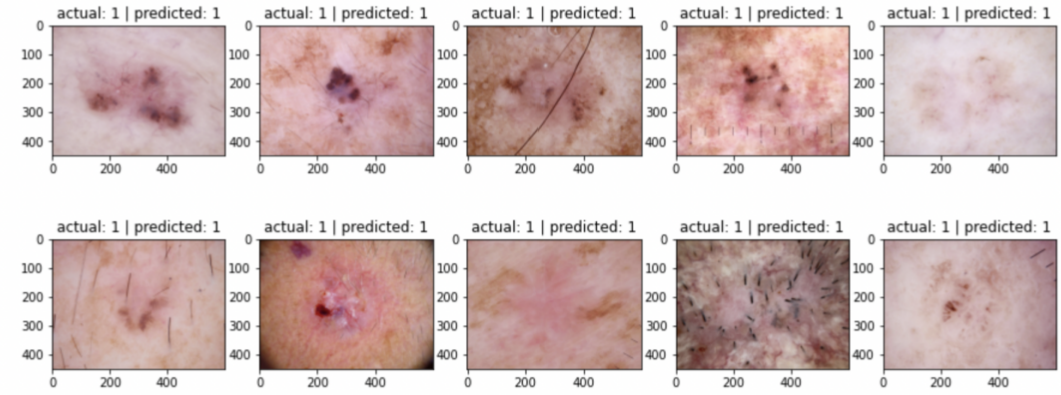
correct predictions: AKIEC images (ensemble 3) sample



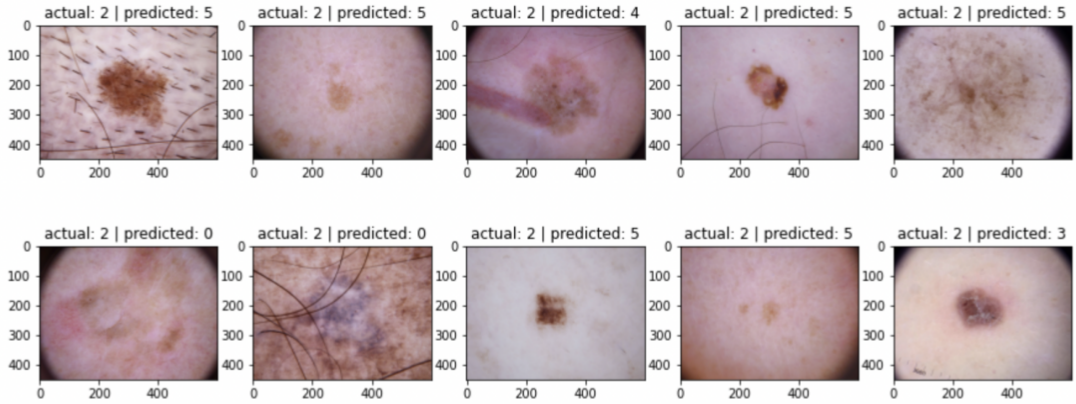
incorrect predictions: BCC images (ensemble 3) sample



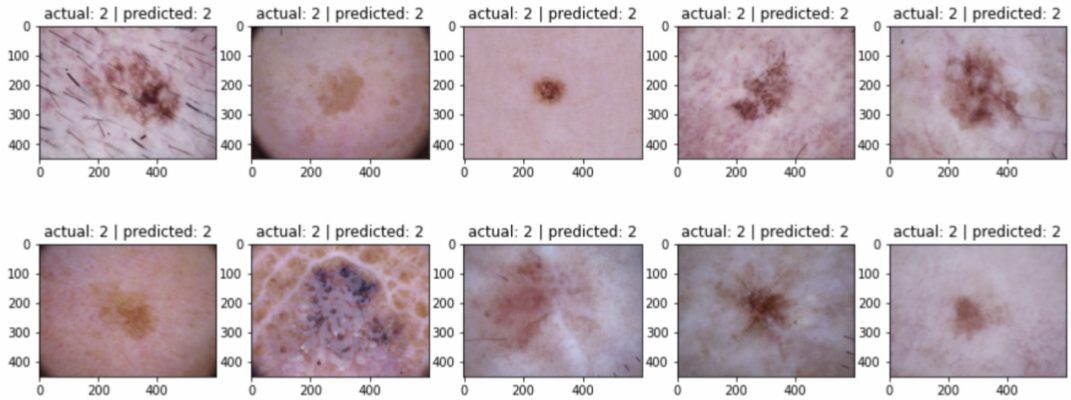
correct predictions: BCC images (ensemble 3) sample



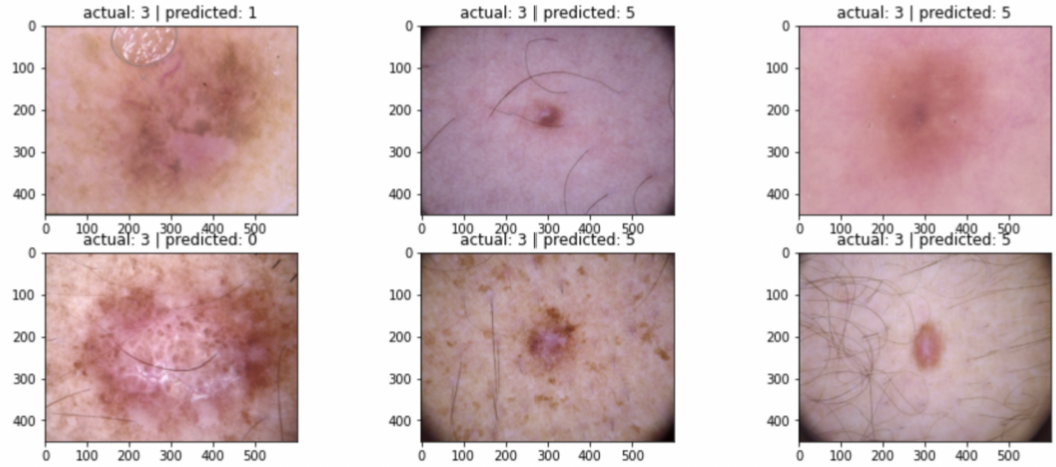
incorrect predictions: BKL images (ensemble 3) sample



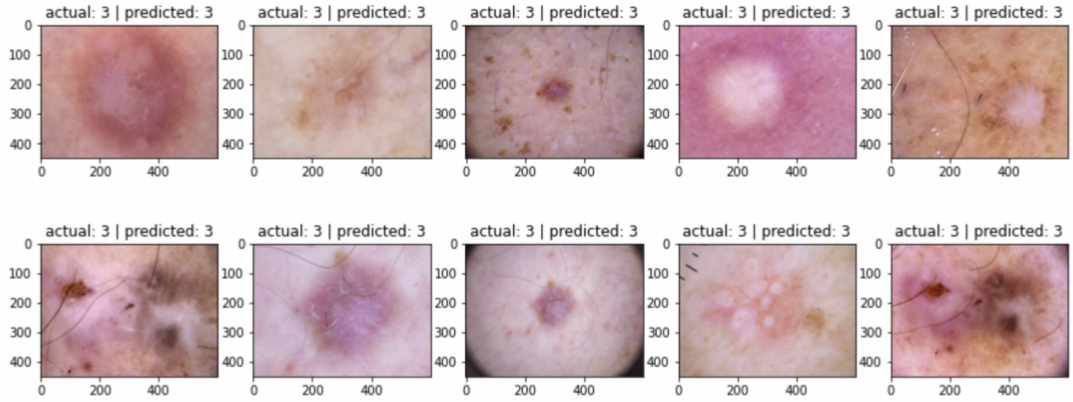
correct predictions: BKL images (ensemble 3) sample



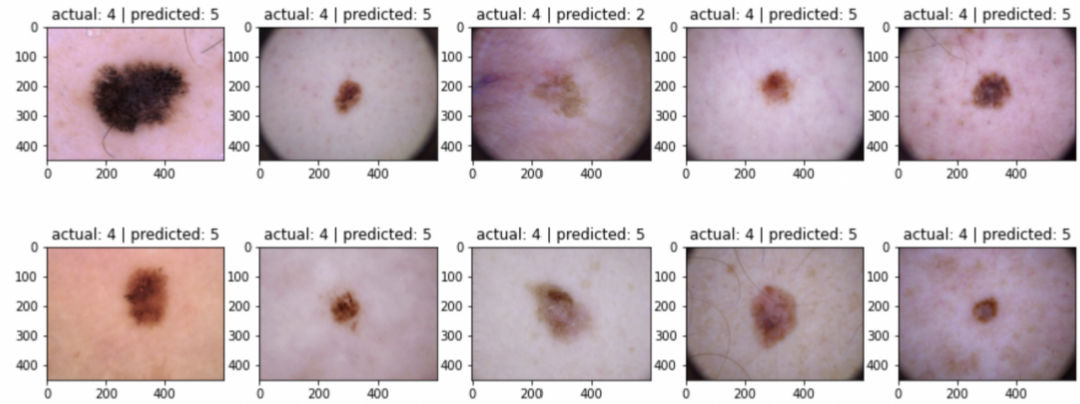
incorrect predictions: DF images (ensemble 3) sample



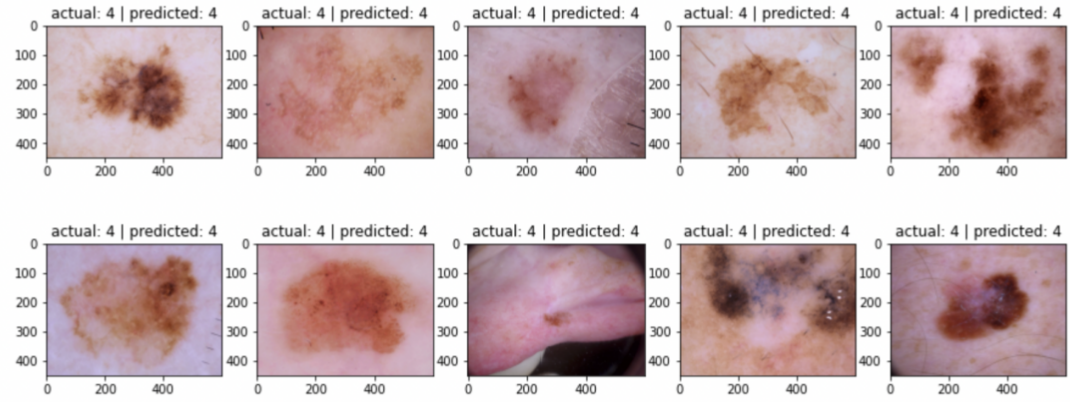
correct predictions: DF images (ensemble 3) sample



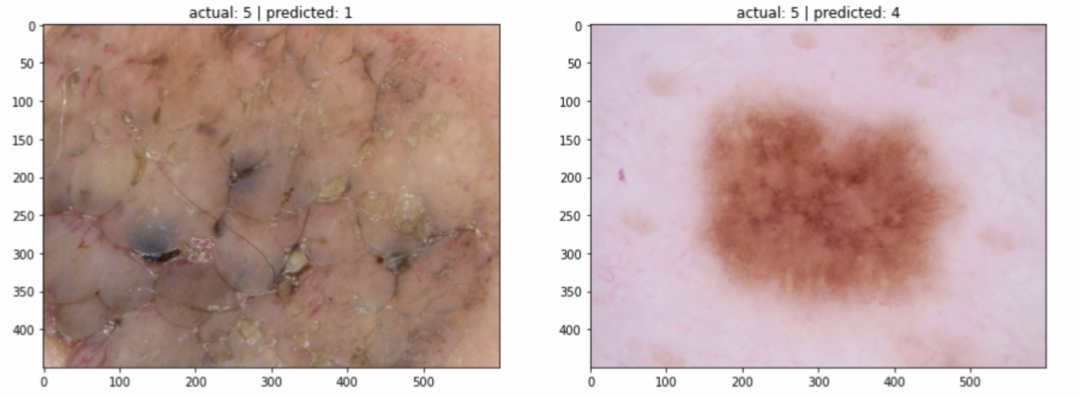
incorrect predictions: Melanoma images (ensemble 3) sample



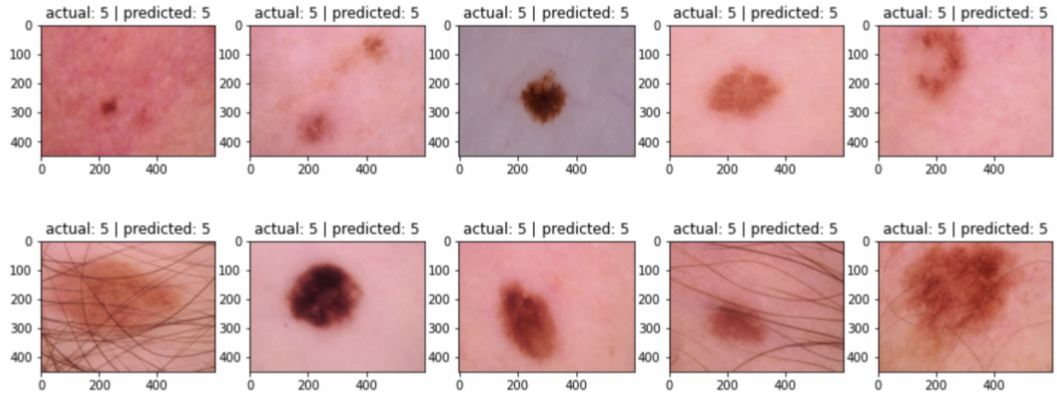
correct predictions: Melanoma images (ensemble 3) sample



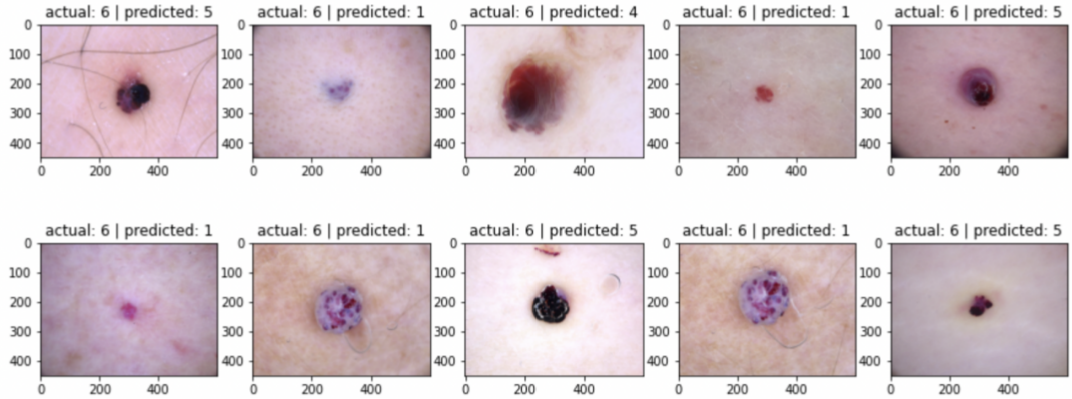
incorrect predictions: Nevus images (ensemble 3) sample



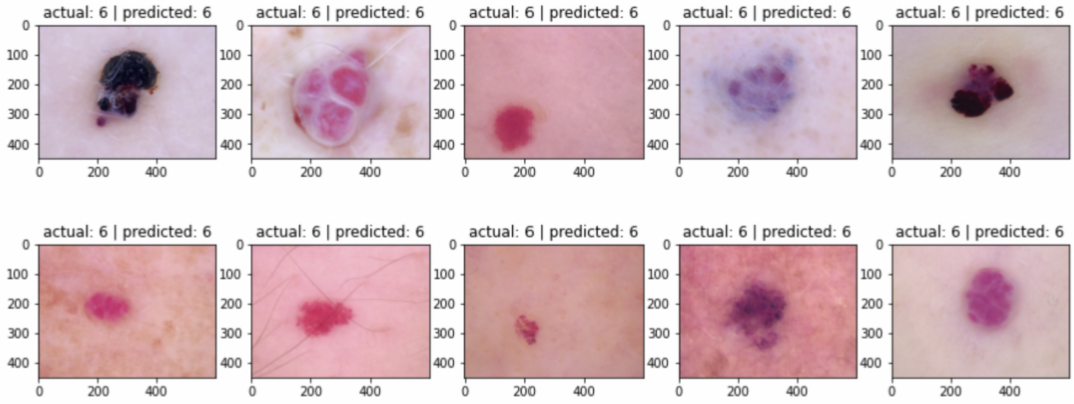
correct predictions: Nevus images (ensemble 3) sample



incorrect predictions: VASC images (ensemble 3) sample



correct predictions: VASC images (ensemble 3) sample



Conclusion

Among the various randomly generated sample dataset that was used, Ensemble 3 performs the best out of the 3. Ensemble 3 appears to perform the best when the testing data set contains images of less noise (hairs and artifacts).