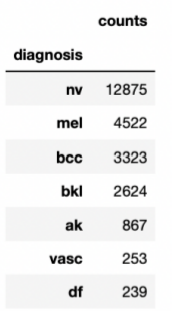
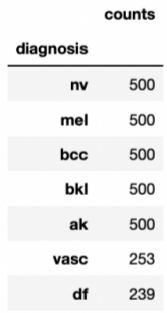
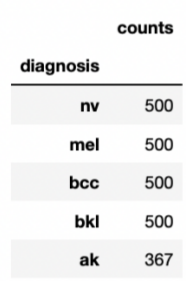
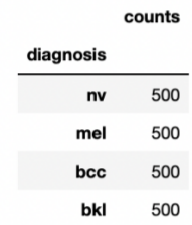
2018 Ensemble 3 Model Evaluation using Sampled 2019 Test Sets

Introduction

In this model evaluation, 5 unique test-sets generated from the 2019 data were created through sampling to facilitate the prediction of the 2018 Ensemble 3 model. The report includes descriptions of the sampling method, prediction results, and suggestions on improvements.

Sampling Method

The sampling method to create the 5 data sets that were used for predictions was performed on the data frame created from the 2019 ground truth CSV file. To match the prediction classes of the 2018 model, all images with the 'scc' diagnosis were removed from the data frame and the row index of the data frame was reset. For each of the classes, their corresponding row indexes were gathered separately as individual lists with each list representing a diagnosis class. The position of each row indexes in each of the lists was randomly shuffled by using the shuffle function in the 'random' library with the exception of images in the class 'vasc' and 'df' as all their corresponding images will be used in one of the 5 generated test data set. To split the randomly shuffled lists into individual partitioned lists with 500 images each, a function called 'chunks' was created to yield partitions of size 500 from each list. For the first 5 partitions ('ak' only has 2 partitions and 'vasc' and df' were not partitioned) created from each list that represents each diagnosis class, the partitions from each list were combined to form 5 unique lists of row index of the different diagnosis class. For example, the first partition from each list is combined to form a single list of row indexes of the different diagnosis classes. The same procedure applies to the remaining 4 partitions of each list. Using the 5 lists of row index created through the combination of the partitions, 5 data frames were created to represent the 5 unique test data set that will be used in the prediction. The number of images for each class in each data set is shown below. To verify that none of the 5 data set contains duplicates, all 5 data sets were compared against each other by putting the image names in each data set into 5 separate sets and looking if there exist any intersection between the different sets. The results showed no intersection for all combinations of the 5 data set pairs which indicates there are no images that exist in more than one data set.

Data set	Number of images in each diagnosis class
Original Data Set	
Test set 1	
Test set 2	
Test set 3, 4, 5	

diagnosis	counts
nv	12875
mel	4522
bcc	3323
bkl	2624
ak	867
vasc	253
df	239

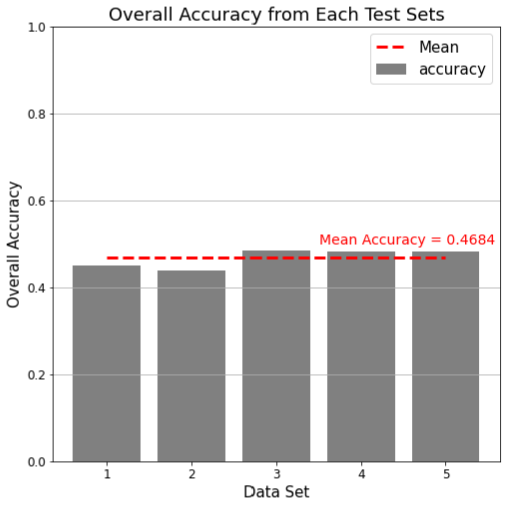
diagnosis	counts
nv	500
mel	500
bcc	500
bkl	500
ak	500
vasc	253
df	239

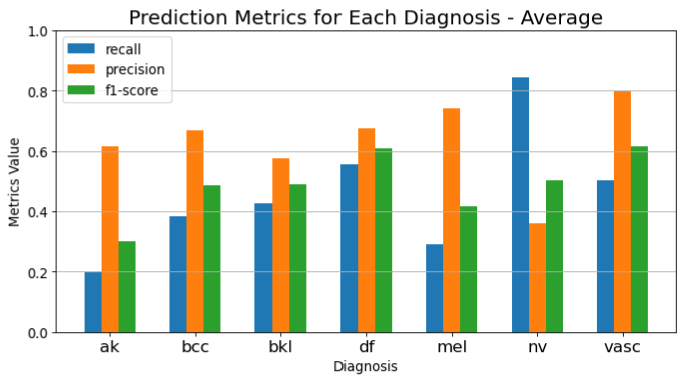
diagnosis	counts
nv	500
mel	500
bcc	500
bkl	500
ak	367

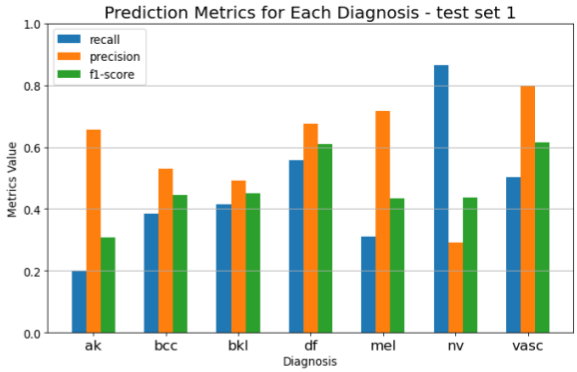
diagnosis	counts
nv	500
mel	500
bcc	500
bkl	500

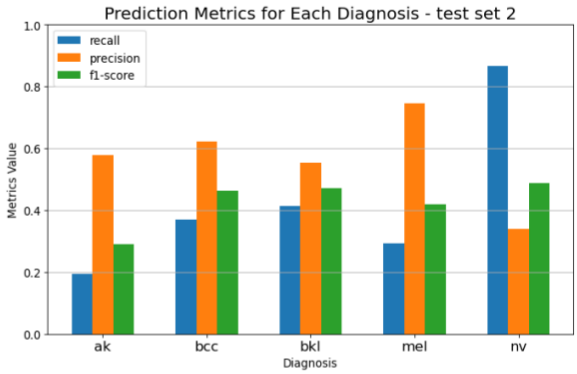
Prediction Results

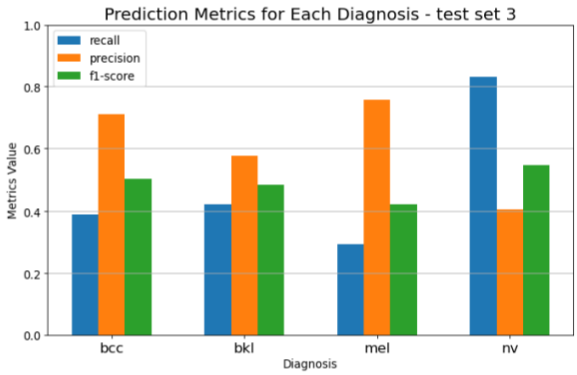
By using the five sampled test sets to perform predictions separately using the 2018 Ensemble 3, the results show an overall accuracy ranging from 43.98% to 48.35% with an average overall accuracy of 46.84%. Looking at other metrics, ‘nv’ class displays the highest recall and lowest precision out of all diagnosis classes for all 5 test sets. This can indicate that most of the misclassification of the other diagnosis classes is occurring with the ‘nv’ class and that a large proportion of the actual ‘nv’ images were correctly predicted. All other classes have precision higher than recall suggesting there is a higher proportion of the predicted diagnosis being actually correct than the proportion of actual diagnosis classes being actually correctly predicted. On average, ‘ak’ and ‘mel’ presented the lowest recall and f1-score, but the high precision for ‘mel’ suggests that if an image is predicted as ‘mel’, it is likely that it will be ‘mel’.

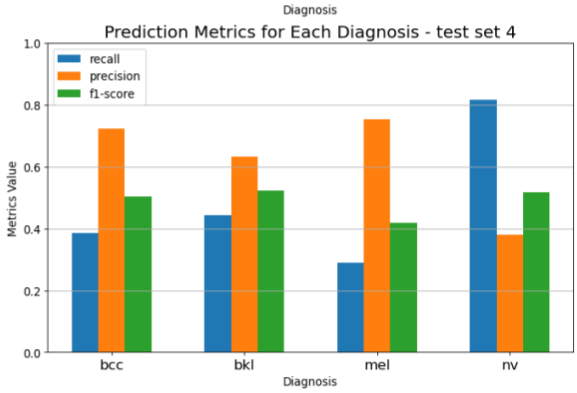


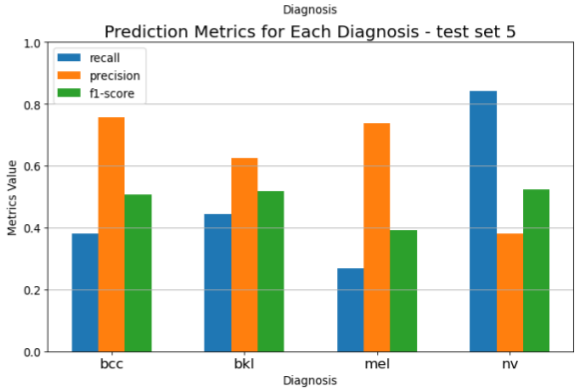



Prediction Metrics for Each Diagnosis - test set 1
recall
precision
f1-score
Metrics Value
1.0
0.8
0.6
0.4
0.2
0.0
ak
bcc
bkl
df
mel
nv
vasc
Diagnosis


Prediction Metrics for Each Diagnosis - test set 2
recall
precision
f1-score
Metrics Value
1.0
0.8
0.6
0.4
0.2
0.0
ak
bcc
bkl
mel
nv
Diagnosis


Prediction Metrics for Each Diagnosis - test set 3
recall
precision
f1-score
Metrics Value
1.0
0.8
0.6
0.4
0.2
0.0
bcc
bkl
mel
nv
Diagnosis


Prediction Metrics for Each Diagnosis - test set 4
recall
precision
f1-score
Metrics Value
1.0
0.8
0.6
0.4
0.2
0.0
bcc
bkl
mel
nv
Diagnosis

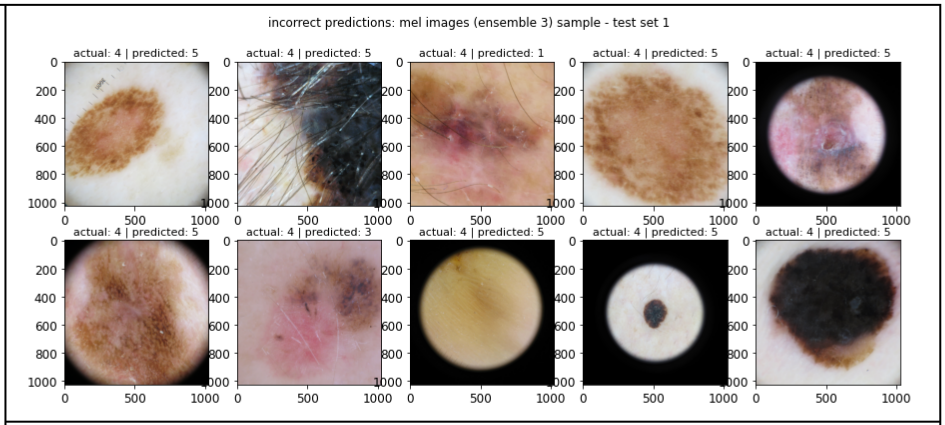
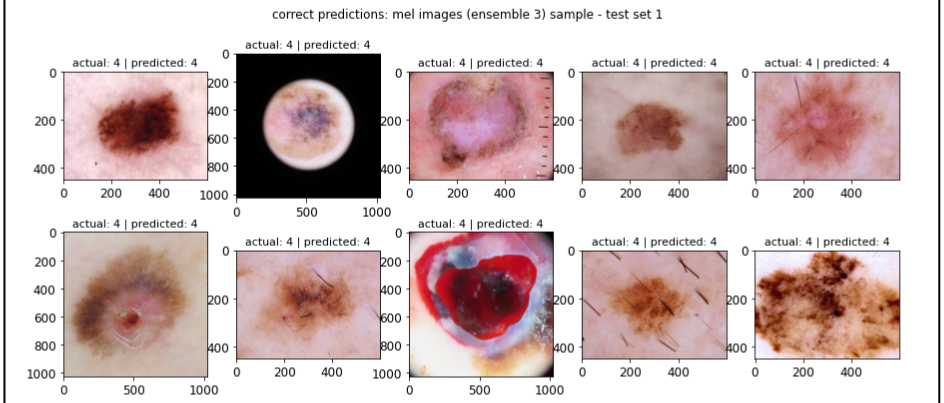
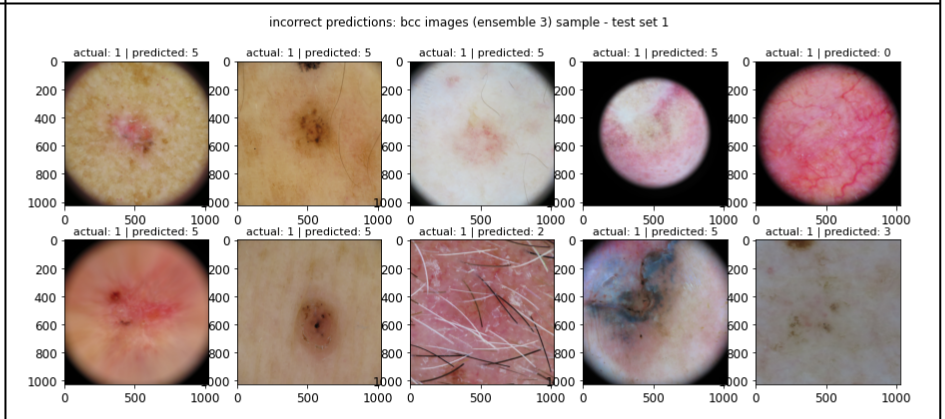
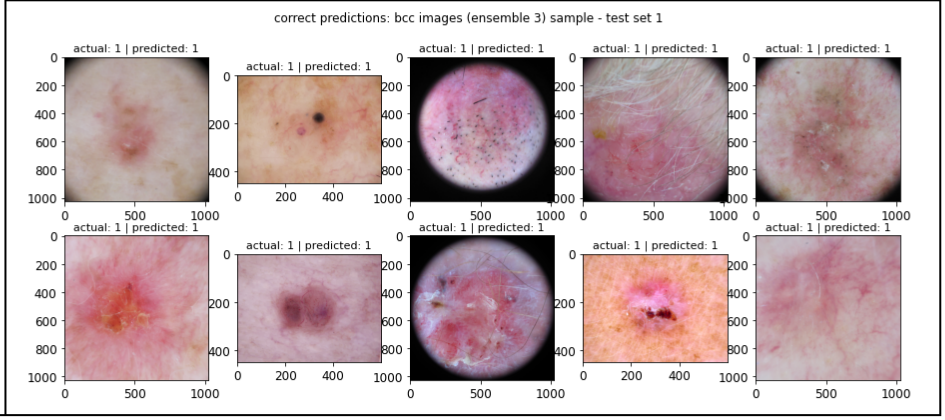

Prediction Metrics for Each Diagnosis - test set 5
recall
precision
f1-score
Metrics Value
1.0
0.8
0.6
0.4
0.2
0.0
bcc
bkl
mel
nv
Diagnosis

Conclusion

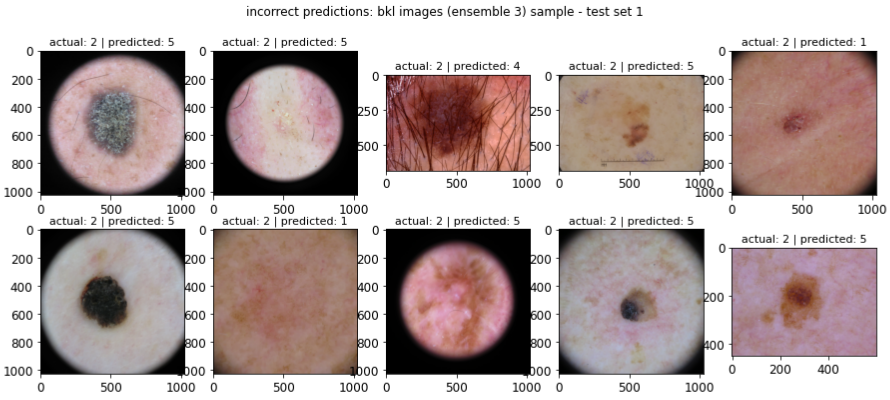
From the prediction results we can see that the predictions with the 5 test set yielded an average overall accuracy of 46.84%. On average of the 5 test data sets, the high recall and low precision shows that the actual ‘nv’ images were mostly correctly predicted but images that were predicted as ‘nv’ contains a large proportion of incorrect predictions. The low recall and high precision of all the other classes shows that the actual images of those classes were mostly incorrectly classified and are likely to be classified as ‘nv’, but if an image is predicted as one of those classes it is likely to be a correct prediction. Even though methods to alleviate the imbalance classes issue were implemented in the data that was used to train the model, the model still tends to have predictions that are more lenient on the ‘nv’ class.

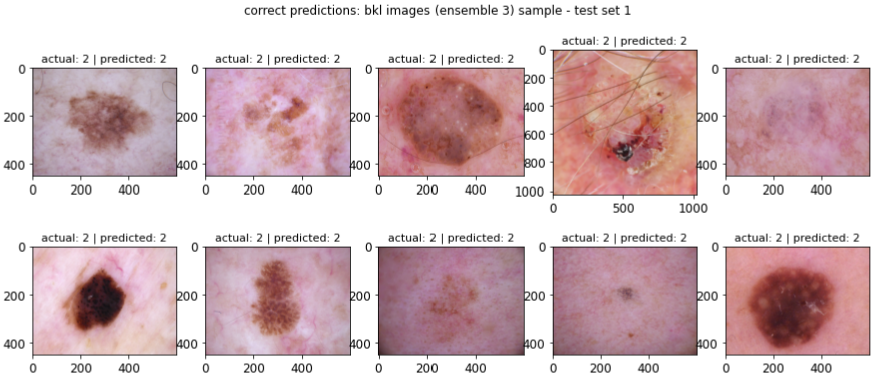
One suggestion that may improve the accuracy of the model is to implement a method to remove vignette borders for testing images with vignette borders before performing predictions. Another suggestion is to implement a more generalized image pre-processing method when training the data in which the same pre-processing method can then be applied to the test images before making predictions.

Image Samples of Prediction Results

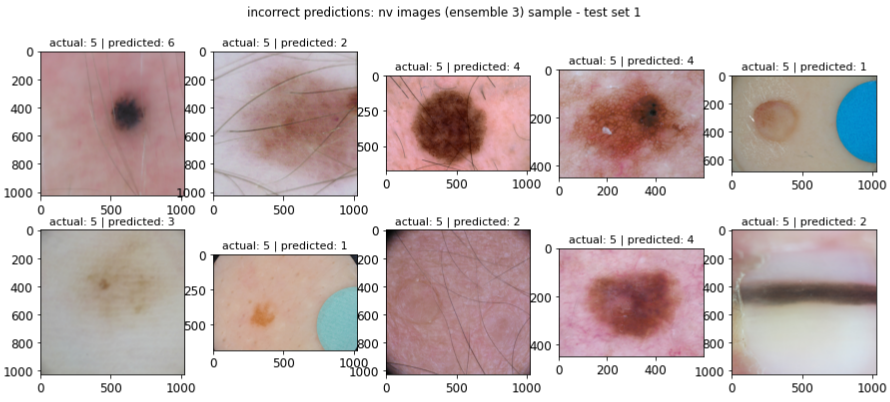
mel	
	
bcc	
	

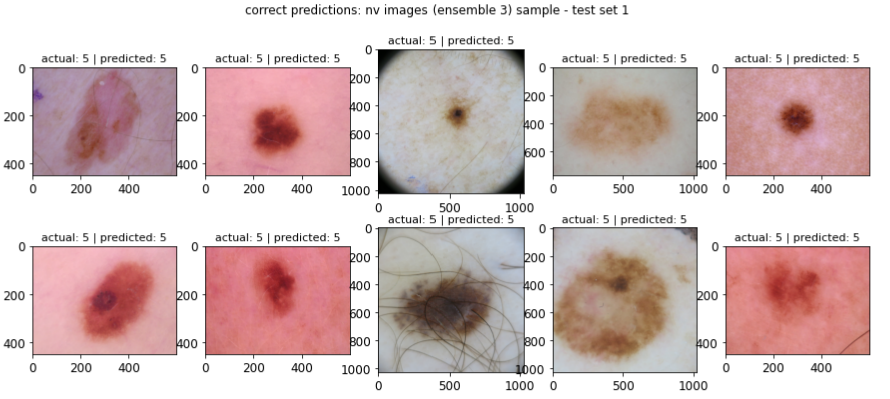
bkl


incorrect predictions: bkl images (ensemble 3) sample - test set 1
actual: 2 | predicted: 5
actual: 2 | predicted: 5
actual: 2 | predicted: 4
actual: 2 | predicted: 5
actual: 2 | predicted: 1
actual: 2 | predicted: 5
actual: 2 | predicted: 1
actual: 2 | predicted: 5
actual: 2 | predicted: 5
actual: 2 | predicted: 5

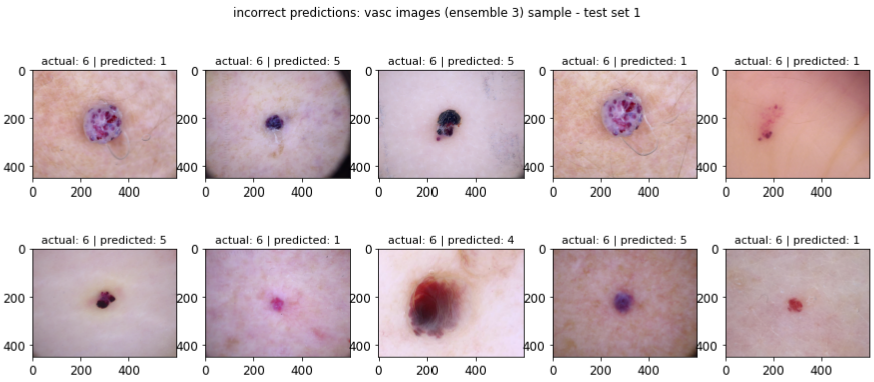

correct predictions: bkl images (ensemble 3) sample - test set 1
actual: 2 | predicted: 2
actual: 2 | predicted: 2
actual: 2 | predicted: 2
actual: 2 | predicted: 2
actual: 2 | predicted: 2
actual: 2 | predicted: 2
actual: 2 | predicted: 2
actual: 2 | predicted: 2
actual: 2 | predicted: 2
actual: 2 | predicted: 2

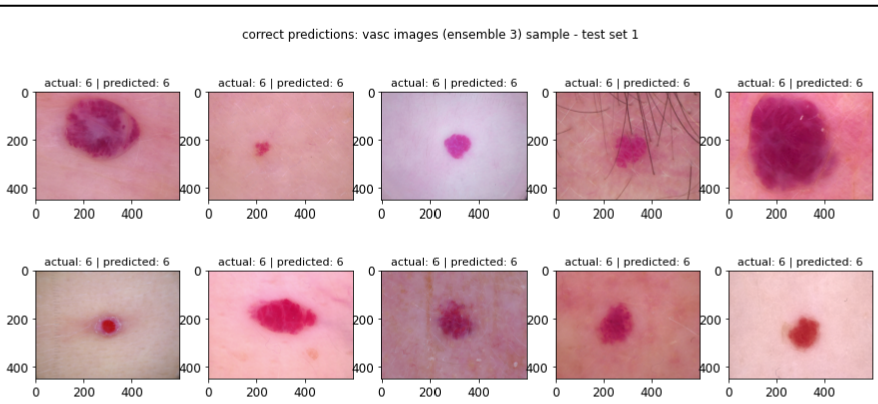
nv


incorrect predictions: nv images (ensemble 3) sample - test set 1
actual: 5 | predicted: 6
actual: 5 | predicted: 2
actual: 5 | predicted: 4
actual: 5 | predicted: 4
actual: 5 | predicted: 1
actual: 5 | predicted: 3
actual: 5 | predicted: 1
actual: 5 | predicted: 2
actual: 5 | predicted: 4
actual: 5 | predicted: 2

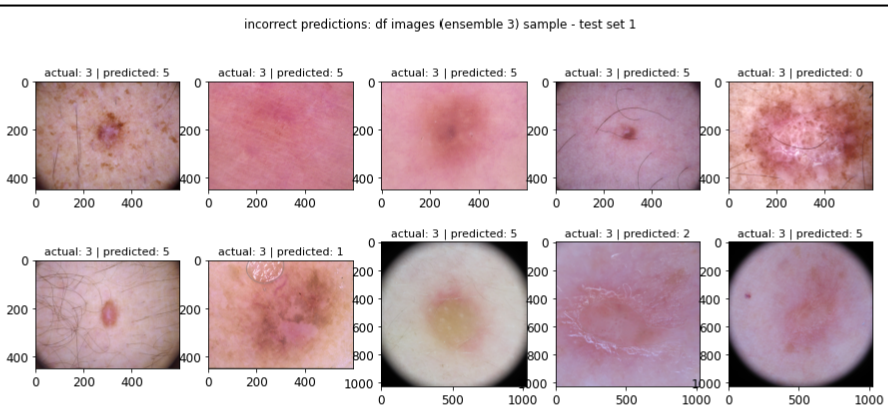

correct predictions: nv images (ensemble 3) sample - test set 1
actual: 5 | predicted: 5
actual: 5 | predicted: 5
actual: 5 | predicted: 5
actual: 5 | predicted: 5
actual: 5 | predicted: 5
actual: 5 | predicted: 5
actual: 5 | predicted: 5
actual: 5 | predicted: 5
actual: 5 | predicted: 5
actual: 5 | predicted: 5

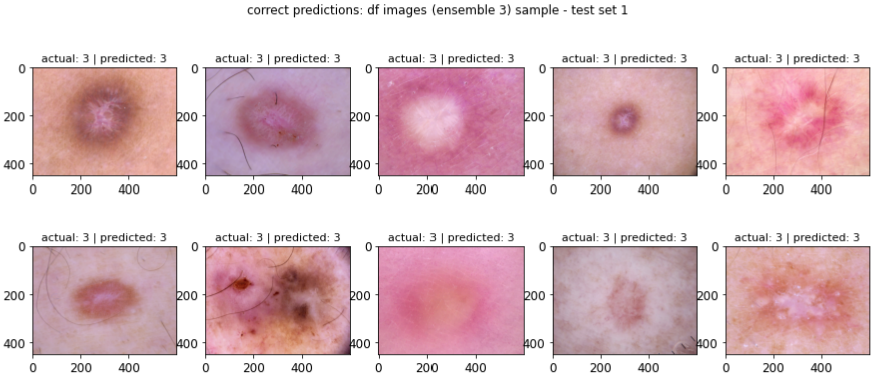
vasc

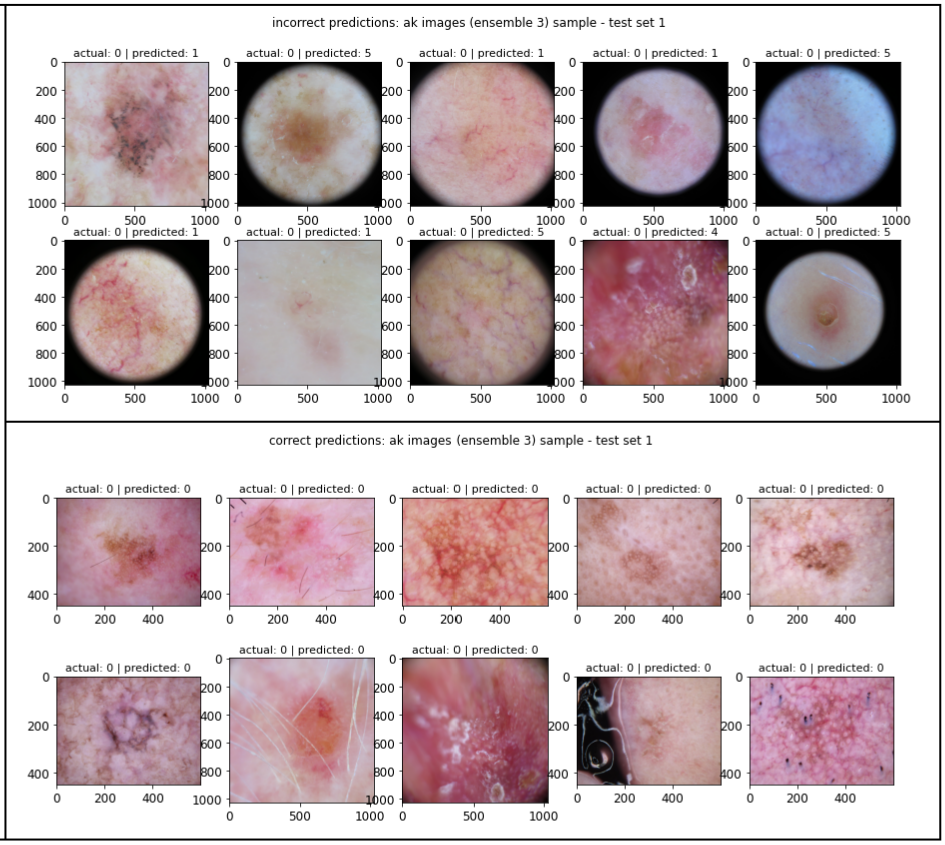

incorrect predictions: vasc images (ensemble 3) sample - test set 1
actual: 6 | predicted: 1
actual: 6 | predicted: 5
actual: 6 | predicted: 5
actual: 6 | predicted: 1
actual: 6 | predicted: 1
actual: 6 | predicted: 5
actual: 6 | predicted: 1
actual: 6 | predicted: 4
actual: 6 | predicted: 5
actual: 6 | predicted: 1


correct predictions: vasc images (ensemble 3) sample - test set 1
actual: 6 | predicted: 6
actual: 6 | predicted: 6
actual: 6 | predicted: 6
actual: 6 | predicted: 6
actual: 6 | predicted: 6
actual: 6 | predicted: 6
actual: 6 | predicted: 6
actual: 6 | predicted: 6
actual: 6 | predicted: 6
actual: 6 | predicted: 6

df


incorrect predictions: df images (ensemble 3) sample - test set 1
actual: 3 | predicted: 5
actual: 3 | predicted: 5
actual: 3 | predicted: 5
actual: 3 | predicted: 5
actual: 3 | predicted: 0
actual: 3 | predicted: 5
actual: 3 | predicted: 1
actual: 3 | predicted: 5
actual: 3 | predicted: 2
actual: 3 | predicted: 5


correct predictions: df images (ensemble 3) sample - test set 1
actual: 3 | predicted: 3
actual: 3 | predicted: 3
actual: 3 | predicted: 3
actual: 3 | predicted: 3
actual: 3 | predicted: 3
actual: 3 | predicted: 3
actual: 3 | predicted: 3
actual: 3 | predicted: 3
actual: 3 | predicted: 3
actual: 3 | predicted: 3

ak	

Confusion Matrix and Metrics

Data set 1

```
Confusion Matrix
[[101  67  70  11  14 229   8]
 [ 23 192  43  19  16 195  12]
 [ 12  32 207  13  13 223   0]
 [  4  13  11 133   1  76   1]
 [ 10  16  50   6 156 253   9]
 [  2   9  35   6  13 433   2]
 [  2  33   5   9   5  72 127]]
Classification Report
              precision    recall  f1-score   support

          ak       0.66      0.20      0.31       500
         bcc       0.53      0.38      0.45       500
         bkl       0.49      0.41      0.45       500
          df       0.68      0.56      0.61       239
         mel       0.72      0.31      0.43       500
          nv       0.29      0.87      0.44       500
        vasc       0.80      0.50      0.62       253

    accuracy                           0.45      2992
   macro avg       0.59      0.46      0.47      2992
weighted avg       0.57      0.45      0.45      2992
```

Data set 2	Confusion Matrix [[71 53 48 15 7 170 3] [24 185 43 24 16 196 12] [20 32 206 10 12 215 5] [0 0 0 0 0 0 0] [7 17 51 7 146 262 10] [1 11 25 13 15 433 2] [0 0 0 0 0 0 0]] Classification Report precision recall f1-score support ak 0.58 0.19 0.29 367 bcc 0.62 0.37 0.46 500 bkl 0.55 0.41 0.47 500 df 0.00 0.00 0.00 0 mel 0.74 0.29 0.42 500 nv 0.34 0.87 0.49 500 vasc 0.00 0.00 0.00 0 accuracy 0.44 2367 macro avg 0.40 0.30 0.30 2367 weighted avg 0.57 0.44 0.43 2367
Data set 3	Confusion Matrix [[0 0 0 0 0 0 0] [36 195 56 26 14 166 7] [22 47 211 14 17 187 2] [0 0 0 0 0 0 0] [6 17 53 5 147 258 14] [2 15 47 4 16 416 0] [0 0 0 0 0 0 0]] Classification Report precision recall f1-score support ak 0.00 0.00 0.00 0 bcc 0.71 0.39 0.50 500 bkl 0.57 0.42 0.49 500 df 0.00 0.00 0.00 0 mel 0.76 0.29 0.42 500 nv 0.41 0.83 0.54 500 vasc 0.00 0.00 0.00 0 accuracy 0.48 2000 macro avg 0.35 0.28 0.28 2000 weighted avg 0.61 0.48 0.49 2000

Data set 4	Confusion Matrix [[0 0 0 0 0 0 0] [22 193 42 17 11 206 9] [14 40 222 10 10 203 1] [0 0 0 0 0 0 0] [13 20 50 5 144 261 7] [5 14 37 8 26 408 2] [0 0 0 0 0 0 0]] Classification Report precision recall f1-score support ak 0.00 0.00 0.00 0 bcc 0.72 0.39 0.50 500 bkl 0.63 0.44 0.52 500 df 0.00 0.00 0.00 0 mel 0.75 0.29 0.42 500 nv 0.38 0.82 0.52 500 vasc 0.00 0.00 0.00 0 accuracy 0.48 2000 macro avg 0.36 0.28 0.28 2000 weighted avg 0.62 0.48 0.49 2000
Data set 5	Confusion Matrix [[0 0 0 0 0 0 0] [21 190 45 20 13 202 9] [22 25 222 10 17 204 0] [0 0 0 0 0 0 0] [10 15 55 7 134 278 1] [1 21 33 6 18 421 0] [0 0 0 0 0 0 0]] Classification Report precision recall f1-score support ak 0.00 0.00 0.00 0 bcc 0.76 0.38 0.51 500 bkl 0.63 0.44 0.52 500 df 0.00 0.00 0.00 0 mel 0.74 0.27 0.39 500 nv 0.38 0.84 0.52 500 vasc 0.00 0.00 0.00 0 accuracy 0.48 2000 macro avg 0.36 0.28 0.28 2000 weighted avg 0.62 0.48 0.49 2000