DSA4212 Assignment 1

March 24, 2022

Group Number: 34

Group Member 1: Shaunn Tan De Hui, A0087785H

Group Member 2: Zhang Shaoxuan, A0080411X

Group Member 3: Ong Jian Ying Gary, A0155664X

Group Member 4: Lu Zhengjie, A0067082E

1 Summary

In this Assignment, we attempt to perform a classification task on a subset of the CelebFaces Attributes Dataset (CelebA) dataset.

Comprising 20,000 images and the corresponding attributes of each picture, we attempt to perform classification of each image to the binary Gender class: {Male, Female}.

The training of models will be performed using the first 15,000 images, while the performance of each model will be assessed based on the remaining 5,000 images.

The organisation of the jupyter notebook & the accompanying report will be as follows: 1. provide summary information about the dataset, 1. discuss the preprocessing steps for the pictures before they are used for model training, 1. propose the models that we will experiment with, and select the best model to address questions relating to the task 1. address the following questions (part of Assignment 1): 1. how does the accuracy of the prediction change with the size of the training set? 2. how does the accuracy of the prediction change with resolution of the input images? 3. compare the performance of the model using colored vs grayscale images. 4. compare the performance of the models using various parts of the faces 5. compare ensemble models 6. identify a model with good performance if the data input is limited

Where relevant, the performance measured using accuracy of prediction on the test set will be presented.

2 Dataset

The CelebA dataset is a face attributes dataset, with the full dataset comprising 200 thousand images of celebrities and 40 attributes are accorded to image.

The dataset used for this classification task is a subset comprising 20 thousand images, of which the first 15,000 will be used for training, while the remaining 5,000 will be used to assess the performance of each model. Each image has dimensions 218px by 178px, with 3 RGB channels.

The classification task in this report is to determine is the individual in an image is "Male" or "Female".

Grouping by the set each observation belongs to, we noted that the proportion of Male vs Female images is balanced and are similar for both the train and test sets.

[4]:

Set	Male	Proportion
Test	-1	0.58
	1	0.42
Train	-1	0.58
	1	0.42

3 Preprocessing

Before any training can be performed, we will perform the following preprocessing steps in order:

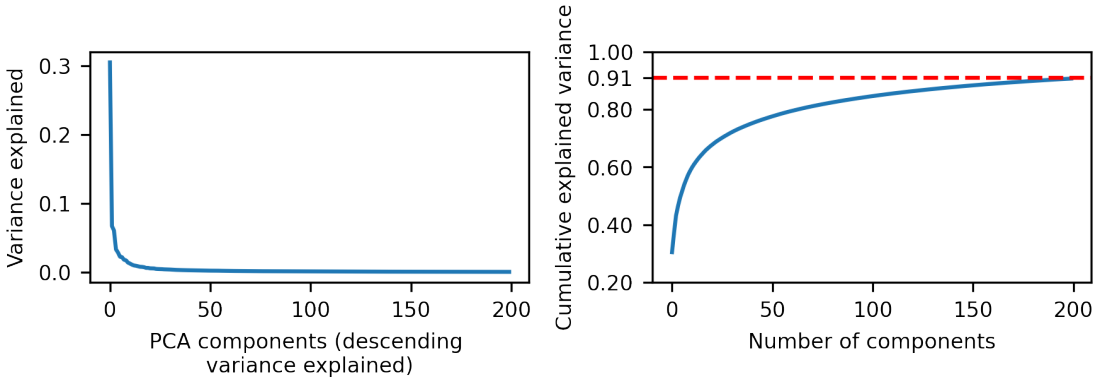
1. An image size parameter will be defined as a hyperparameter and all input images for training and testing will be reshaped into the specified dimension (e.g. 200px by 200px)
2. Images will be converted to grayscale
3. Enhance the contrast of each image using `skimage.exposure.equalize_adapthist`

We create utility functions to perform this preprocessing task and attempt to parallelize the reading in and processing of images using `joblib`.

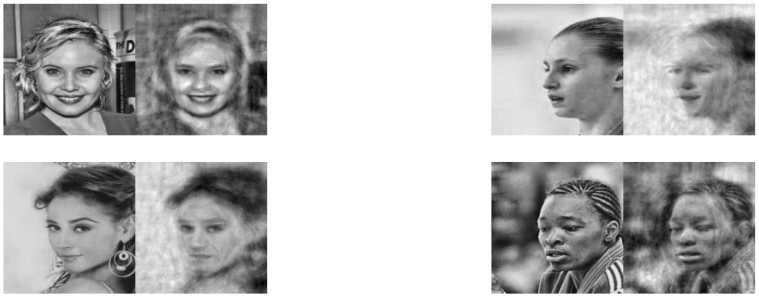
Finally, we perform dimensionality reduction using PCA. The PCA procedure will be performed from a randomly drawn sample of 1,000 images from the training set. The PCA process shall be performed using prebuilt functions in `scikit-learn`.

3.1 PCA

We define an arbitrarily selected 200 Principal Components to retain. A PCA transformer is fitted using the 1,000 randomly selected images. Each image is preprocessed according to the 3 steps mentioned previously.



With just 200 Principal Components, around 91% of the variance of 1,000 randomly selected samples is explained. Next, we visualise the impact of dimensionality reduction on a few randomly selected images.



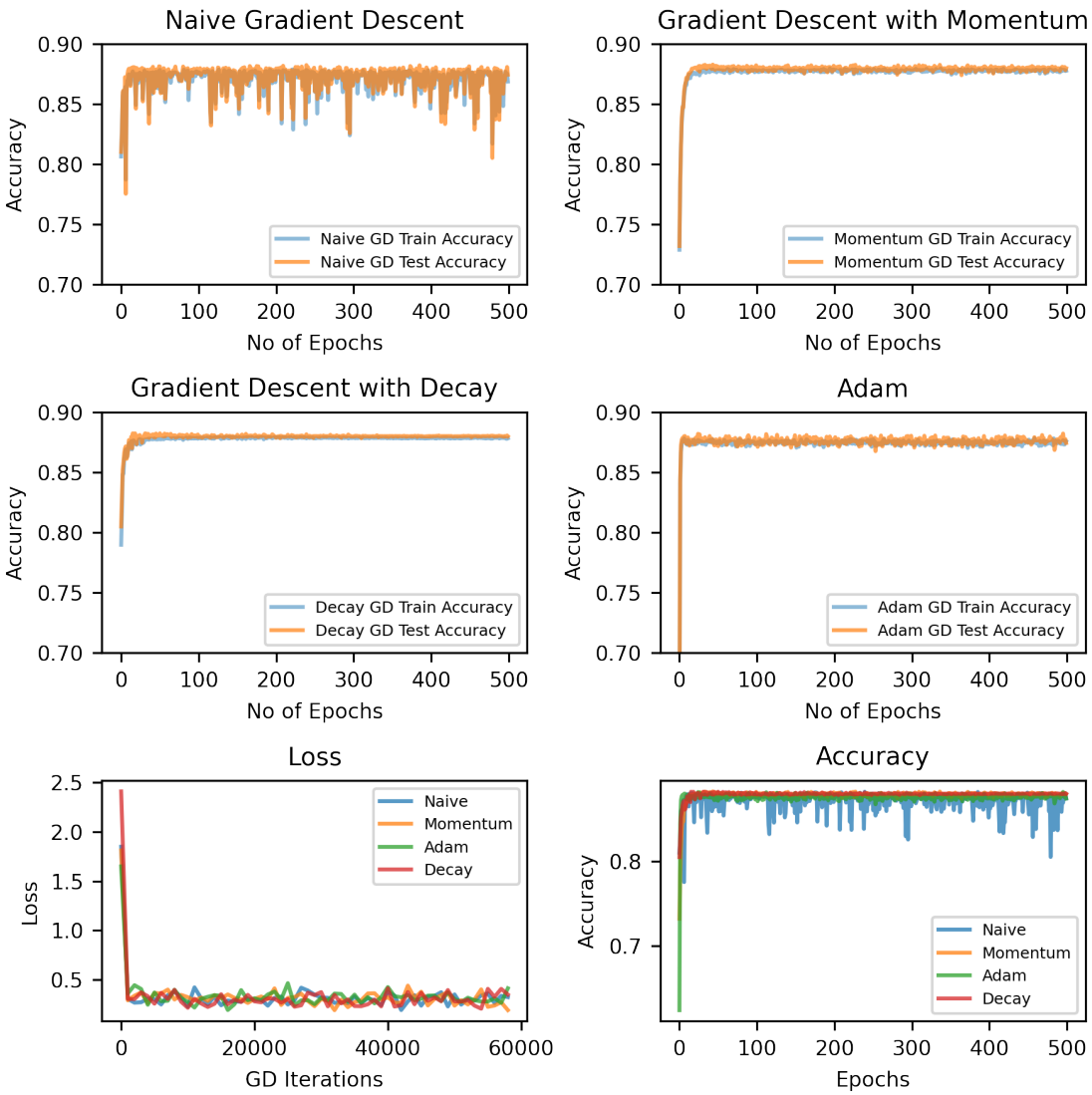
We noted that the dimensionality reduction procedure appears to retain mostly facial features. All images from the train and test sets are PCA transformed for use in training of the classification model.

```
Size of dimension reduced training dataset: 24.0 Megabytes
Size of dimension reduced test dataset: 8.0 Megabytes
```

The final memory requirement for storing the transformed training and test datasets are 24MB and 8MB respectively, a substantial reduction from the actual input size.

4 Logistic Regression

We first performed a baseline logistic regression algorithm, utilising input images of size 200x200px, with 200 principal components and naive gradient descent. Then, we assess if alternative gradient descent algorithms are preferred over the naive approach. The model with the best performing gradient descent algorithm will be used for the subsequent experiments. At each step, we assess the performance of each model using accuracy of predictions on the test set.



We noted that from the above graphs that all four types of gradient descent algorithms proved to have similar generalization abilities for the unseen test dataset at about 88.0% accuracy.

Comparing between train and test accuracies for each algorithm, we noted that is little overfitting of the logistic regression models. One possible explanation for this great generalization capability is perhaps that the dimensional reduction process produced a robust representation of face images, and therefore produced similar outputs for both the training and the (unseen) test data.

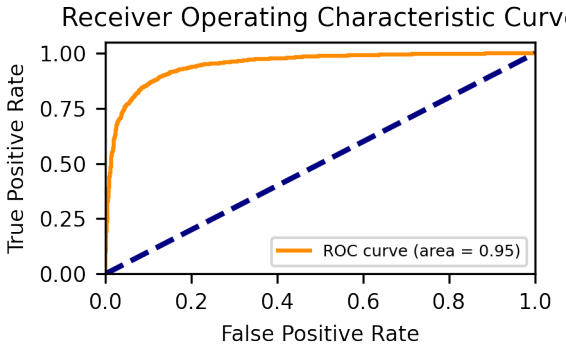
Another key observation to note is that the Naive GD algorithm has a larger variation in accuracy as compared to the rest.

Also, all 4 algorithms appear to reach convergece well within 100 epochs and the Adam algorithm converges the fastest.

Given the similar accuracy of all 4 GD algorithms for this task, we select and present the results

on the test set for the model fitted using **Gradient Descent with Momentum** as it performed similarly but remains parsimonious.

The ROC Curve and AUC of 0.95 is presented below.



5 Additional Tasks for DSA4212 Assignment 1

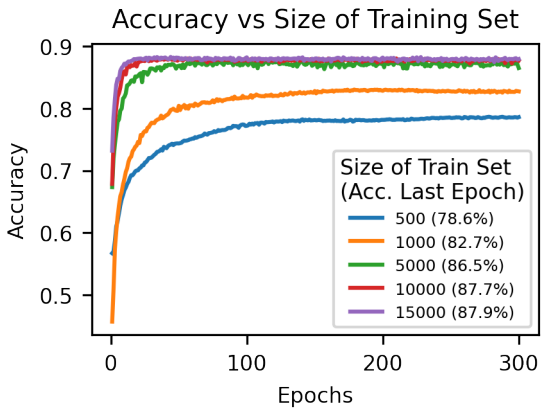
Next, we address the following questions that are part of Assignment 1:

1. how does the accuracy of the prediction change with the size of the training set?
2. how does the accuracy of the prediction change with resolution of the input images?
3. compare the performance of the model using colored vs grayscale images
4. compare the performance of the models using various parts of the faces
5. compare ensemble models
6. Report the error rate and AUC of your best model (when evaluated on the last 5,000 images)
7. identify a model with good performance if the data input is limited to first 200 images

5.1 Question 1

How does the accuracy (ie. tested on the last 5,000 images) depend on the size of the training set? Is it necessary to use all the training set, or does the accuracy stabilize before?

To address this question we compare model performance when the training set size is reduced from 15,000 to 500, 1,000, 5,000 and 10,000 by randomly selecting the required number of images for training, then assessing performance on the 5,000 images test set.



Comparing the accuracy of each model at convergence, we noted that the accuracy of the trained model predicting on the test set increases with the size of the training set provided. A model fitted using just 5,000 training samples has nearly the same performance as the model fitted using 15,000 samples.

5.2 Question 2

How does the accuracy depend on the resolution of the input image?

To understand the impact of reducing/increasing input image size, we modified the size of input images and assessed the performance of the fitted models at the following input sizes:

1. 50 x 50
2. 100 x 100
3. 150 x 150

The performance of these 3 models are then compared to the baseline model (Logistic Regression fitted with GD with momentum and 200x200px input).

We observed significantly lower test set accuracy with smaller input images. This is expected, as in the process of scaling down the images, we are discarding image details that could be used by the model for classification.

The test set accuracy increases as in the input image size increases. Further, we note that a model fitted with 100x100px image as input has similar performance as a model fitted with 200x200px inputs.

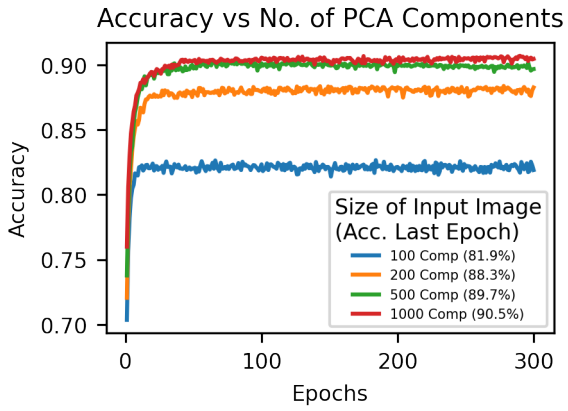
5.3 Question 3

Is it necessary to use colored images (or black & white images are enough)? Is it helpful to increase the contrast of the images? Other preprocessing ideas?

5.3.1 Adjusting Number of PCA Components

The performance of our models may be limited by the arbitrarily chosen 200 PCA components. Setting the baseline as the 150x150px model (since it performs similarly to the 200x200px model),

we compare the performance of models fitted with different number of PCA components.

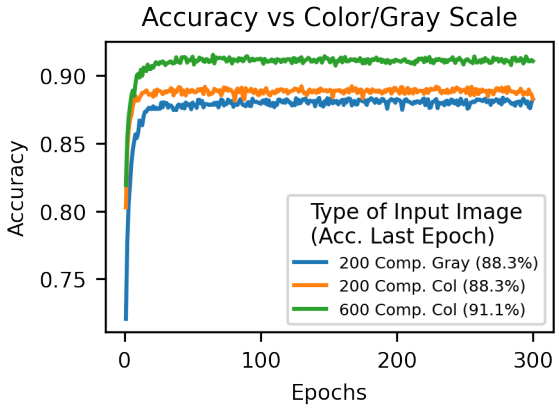


As expected, the accuracy of predictions on the test set increases as the number of PCA components in the model increases since more variations in the input data is retained for training. The model with 1,000 PCA components performed the best with almost 91% accuracy on the test set.

```
Size of train set with 100 PCA components: 12.0 Megabytes
Size of train set with 200 PCA components: 24.0 Megabytes
Size of train set with 500 PCA components: 60.0 Megabytes
Size of train set with 1,000 PCA components: 120.0 Megabytes
```

5.3.2 Utilising Colour Images

Next, we attempt to fit models utilising colour images as input.

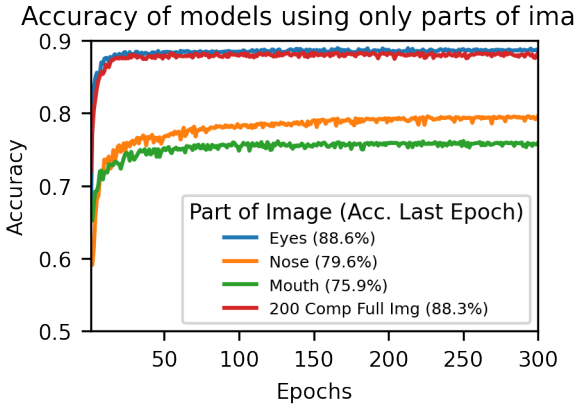


We noted a slight improvement in test accuracy when color images are utilised as inputs. The model fitted with 150x150px colour images, with 600 PCA components performed the best among all models tested thus far.

5.4 Question 4

What if one only uses the area around the eyes? Around the mouth? The hair? The ears? Etc..

We attempt to predict the gender of the person in each image using only the image data around the eyes, nose, or mouth. Since the eyes, nose and mouth only datasets are much smaller than the original image, we did not perform PCA. The performance of each model is then compared to the baseline 150x150px 200 PCA components model.



We noted that among models trained on images of eyes, mouth or nose, the highest test accuracy that was be achieved with an **eyes-only** model. The accuracy of the eyes-only model is higher than the 88% accuracy on our next-best model, Logistic Regression with GD with momentum using the full image.

We can infer from this result that eyes of a person in each image has significant informative value in the classification of gender, as compared to other features like eyes and nose.

5.5 Question 5

Is it useful to use an ensemble of models (eg. for example, you can use a different model for each part of the face, and then try to find a way to ensemble these models)?

We utilise the eyes, nose and mouth models to create a Voting Classifier and predict the labels of the test set, with the ensemble model outputting the class (Male/Female) that is chosen by the majority of the classifiers. Then, we compare the accuracy of the voting classifier against the PCA model which utilized full images and each component model of the voting classifier.

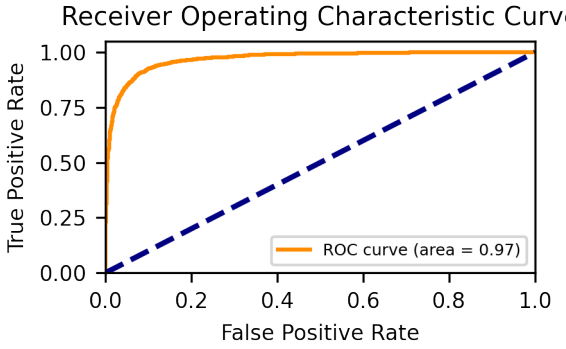
```
The voting model has a test accuracy of: 86.28%
```

We noted from the results,that a voting classifier using three models for eyes, nose and mouth has lower test accuracy (~86%) than an eyes-only model (~89%).

5.6 Question 6

Report the error rate and AUC of your best model (when evaluated on the last 5,000 images)

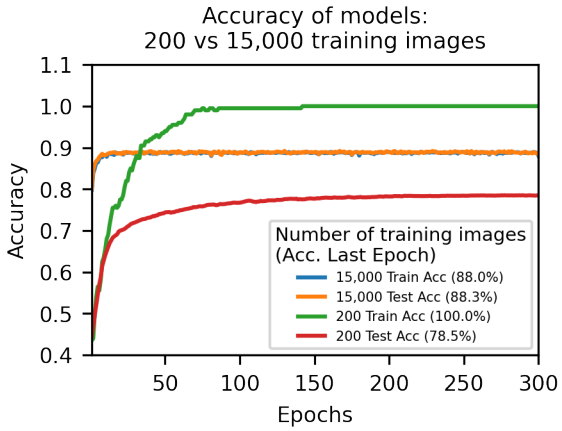
Our PCA model with 600 PCA components generated from colour images performed the best, with an error rate of ~8.9%. The ROC curve and AUC is presented below.



5.7 Question 7

Suppose now that you can only use 1% of the data, i.e. only the first 200 images, to train your model. What is the best model you can come up with? Is it helpful to use data-augmentation strategies? Is it helpful to use regularization strategies? Ensembling? Report the error rate and AUC of your best model (when evaluated on the last 5,000 images).

Since we are only training on 200 training samples, we will be limited to 200 PCA components. Thus, when training the limited training samples model, we will compare its performance to the next most similar model that had 200 PCA components and 150x150px color images (trained for question 3).

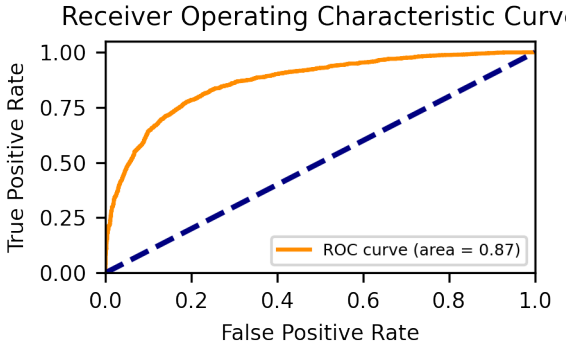


We see that the model trained with just 200 training images gets overfitted to the training set, with 100% accuracy on the training set, but test accuracy that is only 78.2%.

As there were only 200 images to work with, it becomes a challenge to extract principal components

from the training set that are also releveant to other samples, thus explaining the poor accuracy on the test set.

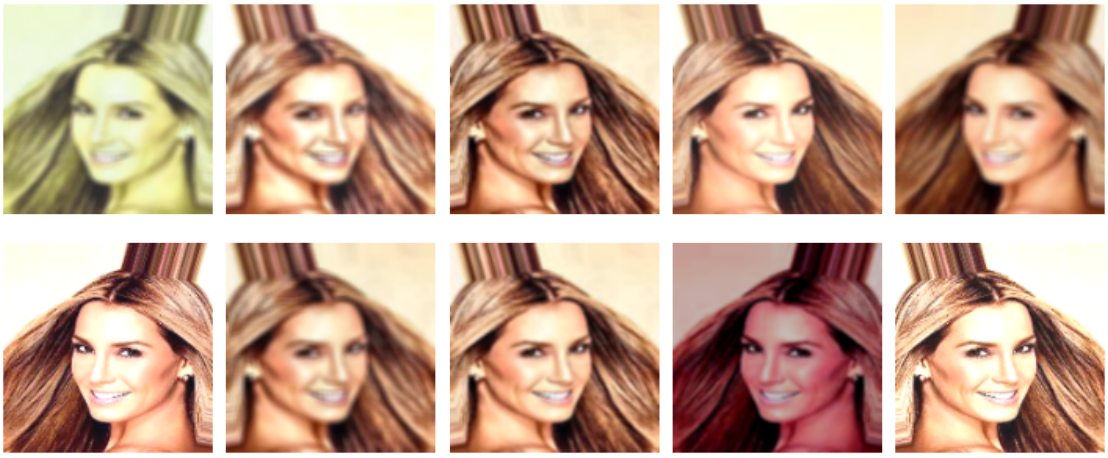
The ROC curve and AUC (0.86) of the limited training samples model is presented in the chart below.



Let us try to improve the performance of this model using data augmentation via the Albumentations package.

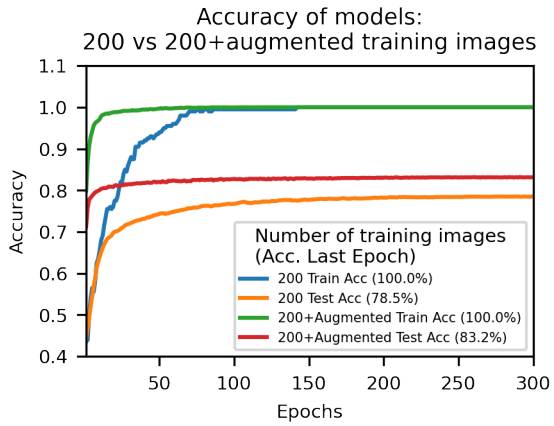
Because of the nature of logistic regression, it would not be advisable to use image augmentations that rearrange the spatial position of the pixels in the photos too drastically (e.g. rotating 90 degrees). Instead, we choose to use only transformations that change the colour (FancyPCA, HueSaturationValue), contrast (CLAHE, RandomBrightnessContrast) and sharpness (Blur) of the image, as well as the horizontal flip, which maintains the general position of the salient parts of the faces (eyes, ears, nose, mouth, etc.).

We can visualise the first 10 augmentations for the first training image.



As with the non-data-limited case, we construct a PCA decomposition using a random choice of 1000 of the (augmented) training images. From these 1000 images, we generate 1000 principal

components.



We see that the test accuracy is around 83%, which is an improvement from the figure without augmentation (78.5%).

An explanation for this would be that the data augmentation allowed for a more robust set of principal components to be generated, which manages to capture the structure of human faces while not being overly biased due to variations in lighting or colour balance.

The ROC curve and AUC (0.9) of the model using data augmentation is presented below.