Dependable AI - Course Project

Jaimin Sanjay Gajjar (B20AI014)

Mukul Shingwani (B20AI023)

A Comprehensive Approach to Image and Video DeepFake Detection

Deepfakes refer to artificially generated images, videos, and audio that are made to appear as if they are genuine. They are created using deep learning algorithms and are becoming increasingly sophisticated, making it difficult for humans to distinguish between real and fake content. This has raised concerns about the potential misuse of deepfakes for propaganda, misinformation, and other malicious purposes.

To address this challenge, researchers have been working on developing deepfake detection methods that can accurately identify fake content. However, one of the major challenges in deepfake detection is ensuring that these methods can generalize to new and unseen deepfakes. This is because deepfakes can be generated using a variety of techniques, and new methods are constantly being developed to improve their quality.

To enhance the generalization of deepfake detection methods, researchers have proposed various techniques, such as using larger and more diverse datasets, training on a combination of real and fake data, and incorporating domain-specific knowledge into the detection process. These approaches aim to improve the robustness and accuracy of deepfake detection models, making them more effective in detecting and mitigating the risks associated with deepfakes.

Dataset Links

1. FakeCelebAV - https://drive.google.com/file/d/1x0h3mhmfqWErN9xAq7mUfn6EcbUPIDMa/view (Got after request in google form)
2. FFHQ - https://github.com/NVlabs/ffhq-dataset
3. CelebA - https://www.kaggle.com/datasets/jessicali9530/celeba-dataset

4. **DeepFakeDetection (FaceForensics++) -** https://github.com/ondyari/FaceForensics

DeepFake Detections

Method 1: Unmasking DeepFakes with Simple Features

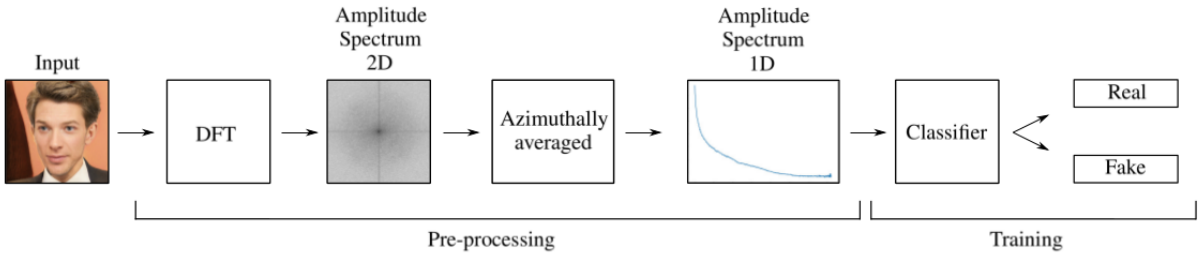
Reference -

Unmasking DeepFakes with simple Features

Deep generative models have recently achieved impressive results for many real-world applications, successfully generating high-resolution and diverse samples from complex datasets. Due to this...

https://arxiv.org/abs/1911.00686

- Deep generative models have led to a proliferation of fake digital contents, particularly fake face images (DeepFakes), raising concerns about spreading misinformation and deception.
- Researchers developed a simple method for detecting DeepFakes based on frequency domain analysis and basic classification.
- The method requires only a few annotated training samples and can achieve high accuracies even in unsupervised scenarios.
- Researchers created a new benchmark dataset, Faces-HQ, for evaluating their approach on high-resolution face images.
- The method achieved a perfect classification accuracy of 100% with just 20 annotated samples on Faces-HQ dataset.
- The method achieved high accuracies in detecting manipulated videos in the FaceForensics++ dataset.
- This research provides a promising solution to the growing concern of fake face images and can potentially be applied in real-world applications to combat misinformation and deception.



Datasets

1. **Detection Faces-HQ**

   Faces-HQ is a benchmark dataset consisting of 1,000 high-resolution face images that researchers created to evaluate the performance of their approach for detecting DeepFakes. The dataset includes both real and fake images, with resolutions ranging from 1024x1024 to 2048x2048 pixels.

2. **CelebA**

   CelebA (Celebrities Attributes) is a popular benchmark dataset in computer vision research, consisting of over 200,000 facial images of celebrities. The dataset is used for a wide range of tasks, including face detection, face recognition, and facial attribute analysis. Each image in the dataset is annotated with 40 attributes, such as gender, age, and facial hair, making it a useful resource for training and evaluating machine learning models. The images in CelebA have varying resolutions and backgrounds, and the dataset also includes images with occlusions and poses to provide a diverse set of images for research purposes.

3. **DeepFakeDetection (FaceForensics++)**

   DeepFakeDetection (FaceForensics++) is a benchmark dataset consisting of over 1,000 videos of real and fake faces that have been manipulated using deep learning techniques. The dataset is annotated with ground truth labels and is widely used for evaluating the performance of deepfake detection methods.

Method

Frequency Domain Analysis

1. Discrete Fourier Transform: DFT decomposes a discrete signal into sinusoidal components of various frequencies. It is the digital version of the continuous Fourier Transform and can be computed for 2-dimensional data of size M × N.

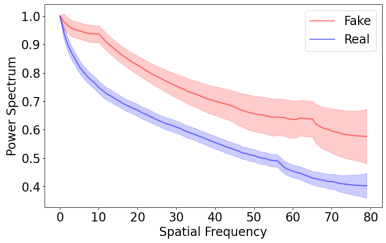
2. Azimuthal Average: After performing a Fourier Transform on an image, the resulting information is still in 2D. To extract a 1D representation of the FFT power spectrum, azimuthal averaging is applied. This is a compression technique that averages similar frequency components to reduce the number of features without losing important information. Additionally, this compression leads to a more robust representation of the input.

Classifier Algorithms

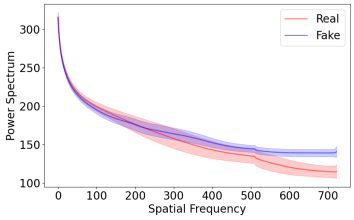
1. Logistic Regression
2. SVM
3. K-Means Clustering

Results

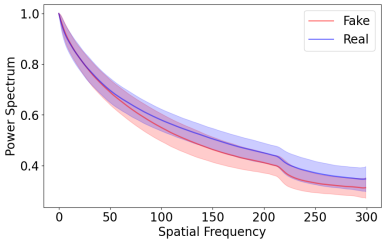
- Here in CelebA dataset, the real v/s fake images are very distinguishable on the basis of Spacial frequency.
- In Faces-HQ dataset, the dataset is indistinguishable till a frequency level of 500, but then, the models have learned to classify the real and fake images very clearly.
- But in case of FaceForensics++ dataset, the images are only slightly distinguishable on basis of spacial frequencies, that's why the classification results are slightly on the lower side.



Power Spectrum of CelebA dataset between Real and Fake images



Power Spectrum of Faces-HQ dataset between Real and Fake images



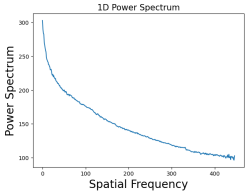
Power Spectrum of FaceForensics++ dataset between Real and Fake images

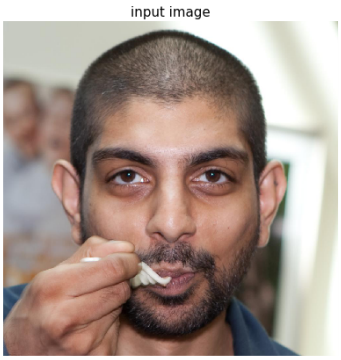
Classification Results

	SVM	LR
CelebA	0.9987499999999999	0.9960000000000001

	SVM	LR
Faces-HQ	1.0	1.0
FaceForensics++	0.857	0.769

Spectral Analysis — Faces-HQ





Conclusion - All results are based on Spacial Frequency Features

- Here in CelebA dataset, the real v/s fake images are very distinguishable on the basis of Spacial frequency.
- In Faces-HQ dataset, the dataset is indistinguishable till a frequency level of 500, but then, the models have learned to classify the real and fake images very clearly.
- But in case of FaceForensics++ dataset, the images are only slightly distinguishable on basis of spacial frequencies, that’s why the classification results are slightly on the lower side.
- SVM outperforms LR in terms of accuracy on the CelebA dataset with a score of 0.99875 compared to 0.996 for LR.
- Both SVM and LR achieve perfect accuracy (1.0) on the Faces-HQ dataset, indicating that both models are able to distinguish between real and fake faces with high precision on this dataset.
- On the FaceForensics++ dataset, SVM again outperforms LR with a score of 0.857 compared to 0.769 for LR. However, it's worth noting that both models perform worse on this dataset than on the other two, indicating that the task of detecting deepfake videos is more challenging than detecting manipulated images. (As expected)

Method 2: Explainability in DeepFake Detection

Pipeline

Preprocessing:

1. Load the dataset (FaceForensics++, Celeb-DF, Deepfake Detection Challenge).
2. Split the video into frames.
3. Crop the face from each frame.
4. Save the face-cropped video.

Model and Train:

1. Load the preprocessed video and labels from a CSV file.
2. Create a PyTorch model using transfer learning with ResNext50 and LSTM.
3. Split the data into train and test data.
4. Train and Test the model, and then save .pt file

Predict:

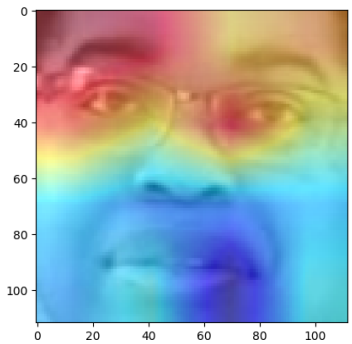
1. Predict the output based on trained weights.

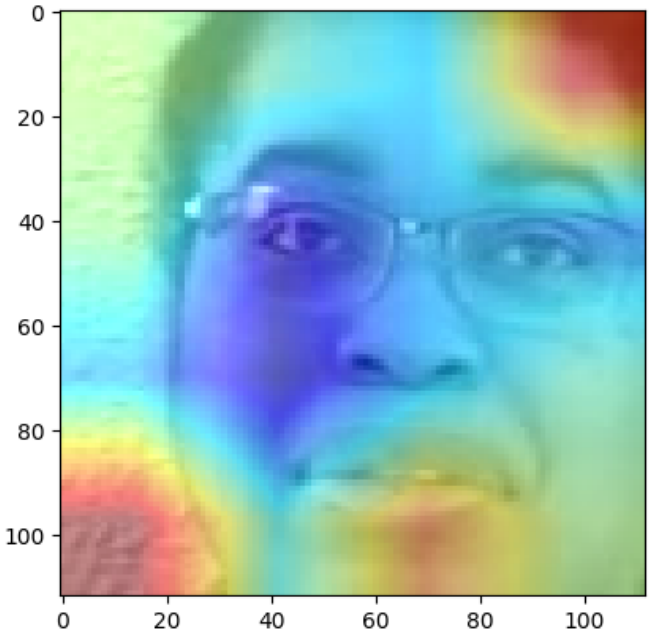
Results:

Classified both correctly as real images.

`confidence of prediction: 99.89238381385803`

`confidence of prediction: 99.98657703399658`





Conclusions

- We tried to implement Grad-CAM here and above were the results.
- We Observe that the predictions being correct on deepfake detection, but the Grad-CAM is not able to correctly identify the real reasons behind the explainability of the predictions in some cases
- That means the model is not learning the correct features for classification between real and fake.

Method 3 : Deepfake Detection in Video

Dataset

We used the DFDC dataset (Deepfake detection challenge)

- The data is comprised of .mp4 files, split into compressed sets of ~10GB apiece. A metadata.json accompanies each set of .mp4 files, and contains filename, label (REAL/FAKE), original and split columns, listed below under Columns.
- The full training set is just over 470 GB.

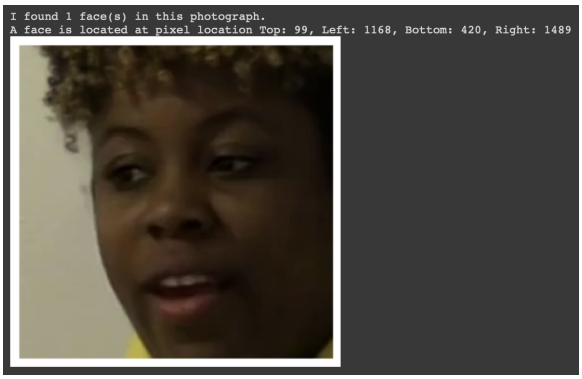
	label	split	original
aagfhgtpmv.mp4	FAKE	train	vudstovrck.mp4
aapnvogymq.mp4	FAKE	train	jdubbvfswz.mp4
abarnvbtwb.mp4	REAL	train	None
abofeumbvv.mp4	FAKE	train	atvmxvwyns.mp4
abqwwspghj.mp4	FAKE	train	qzimuostzz.mp4

- This plot shows that there is class imbalance in the dataset with 80% samples being fake and just 20% being real.

Face recognition

- We used the face recognition package to capture an image frame from the video.

- Next, we moved to detect the face from the captured frame



- We then tried to identify the **face landmarks** within an image

- Some more examples

Frame by frame detection

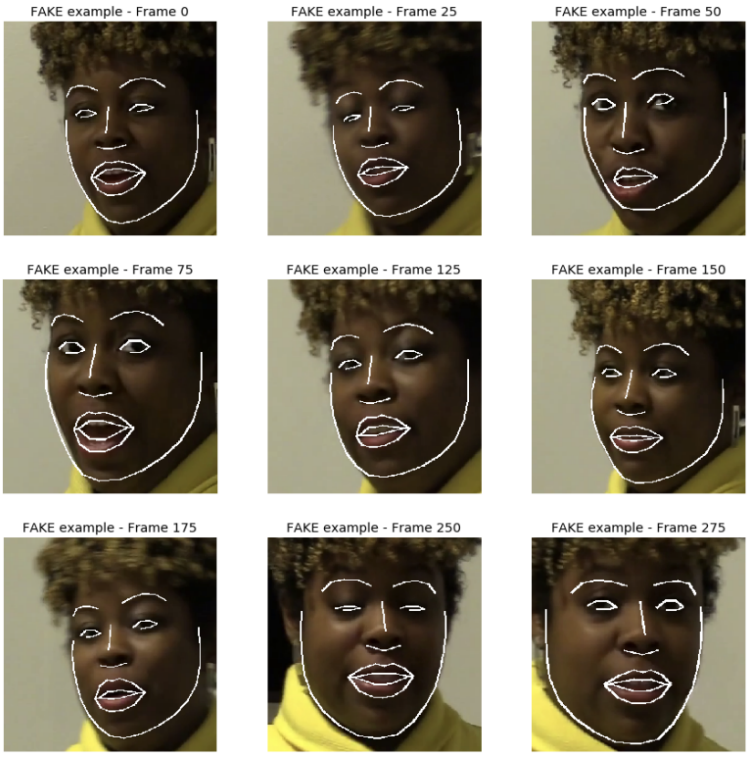
- The real power may come from looking at how the "face" changes or doesn't change as the video progresses

- We will take the FAKE example video `akxoopqjqz.mp4`
- First we looped through the frames of the video file and appended them to a list called `frames`

```
The number of frames saved:  300
```

Following this we extracted the frames from each one of these and drew the facial landmarks as well

```
Count find face in frame 100
```



Bias in DeepFake Detection

Our Goal was to Detect racial and gender bias in state-of-the-art deepfake detectors on videos of humans.

1. **Dataset**: FakeAVCeleb, which contains videos of different racial and gender groups.
2. **Models**: XceptionNet, EfficientNetB4, EfficientNetB4ST, EfficientNetAutoAttnB4, and EfficientNetAutoAttnB4ST. These models use ensemble CNN models, specifically EfficientNetB4 CNN architectures, with an attention mechanism to see which local part of the image contributes the most to the predictions.
3. **Approach on Videos**:
   a. Take 30 frames from each video.
   b. Use Blazeface face extractor to extract a face from each frame.
   c. Pass each frame to the deepfake detector to get a score between 0 and 1.
   d. Average the scores of all 30 frames to get a score for the video.

4. **Predictions**: The predictions of each model on the test sets are in the directory Predictions/pred_ensembles and Predictions/preds_cross_efficient_transformer.txt file. The notebook video_predictions_colab.ipynb was used to get predictions from the ensemble models.
5. **Metrics**:
   a. Accuracy
   b. Precision, Recall, F1
   c. ROC curves (TPR vs FPR)
   d. AUC of ROC curves
6. References:

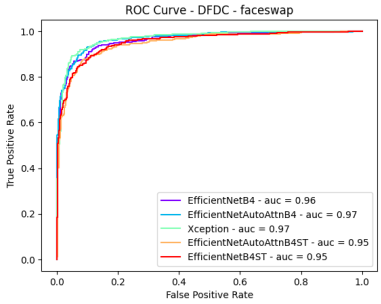
   https://github.com/polimi-ispl/icpr2020dfdc

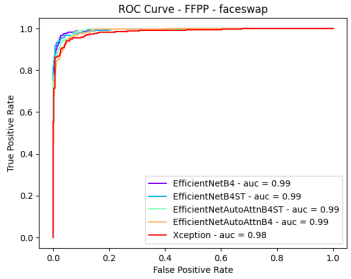
   https://github.com/davide-coccomini/Combining-EfficientNet-and-Vision-Transformers-for-Video-Deepfake-Detection

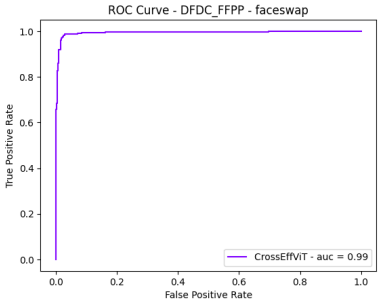
Additional Details

```
ETHNICITIES= ['African','Asian (East)', 'Asian (South)','Caucasian (American)','Caucasian
 (European)']
GENDERS= ['men','women']
METHODs = 'faceswap'
MODELS = ['EfficientNetB4_DFDC', 'EfficientNetB4_FFPP', 'EfficientNetAutoAttnB4_DFDC', 'Ef
ficientNetB4ST_FFPP', 'Xception_DFDC', 'EfficientNetAutoAttnB4ST_FFPP', 'EfficientNetAutoA
ttnB4ST_DFDC', 'EfficientNetAutoAttnB4_FFPP', 'EfficientNetB4ST_DFDC', 'Xception_FFPP', 'C
rossEffViT_DFDC_FFPP']
```

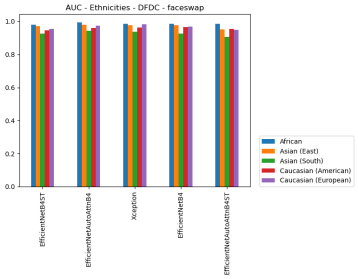
Since we have an imbalanced dataset, we choose only one fake generated video against each real video for each generation method. We will evaluate the results on the three different generation methods separately.

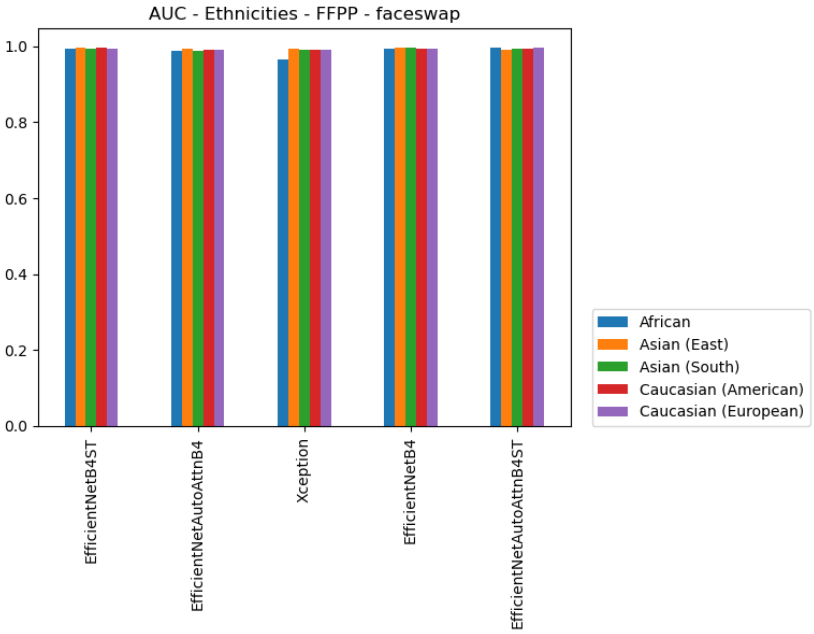




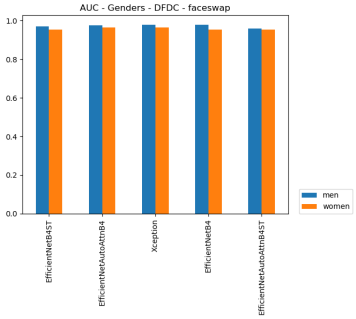


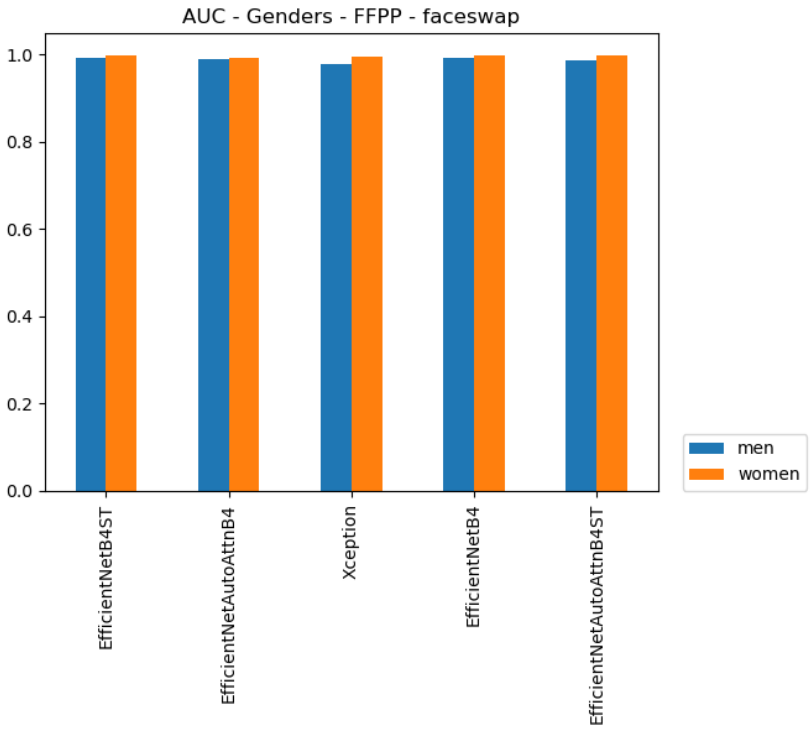
Ethnicities



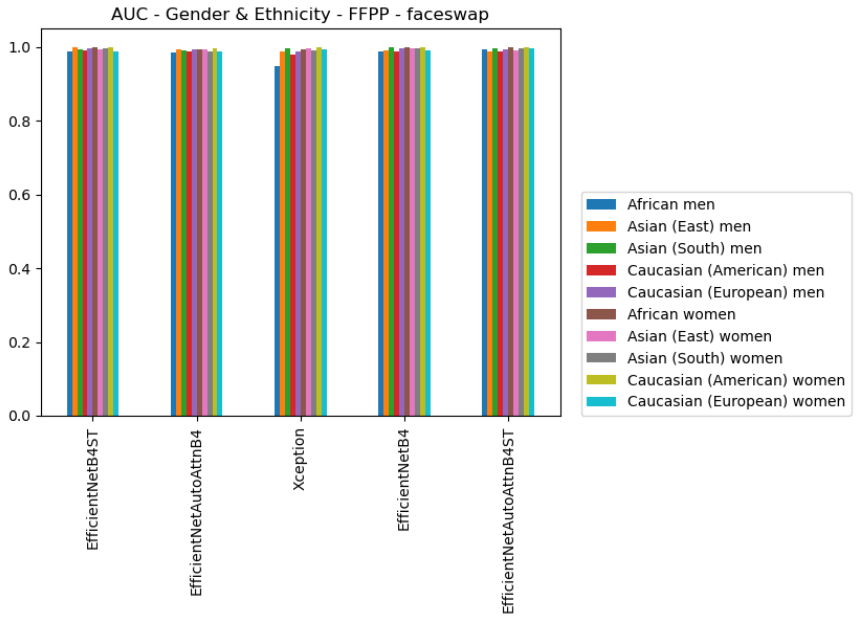


Genders





Combining Gender and Ethnicity



Conclusion

- All the models considered by us, give a very high AUC score, which suggests that our bias detection task is performing very well. It means that our classifier is able to accurately distinguish between biased and unbiased instances with a high degree of confidence, and is making very few mistakes.
- In the varied ethnicities, roughly all show similar results with Asian(south) being slightly lower than others.
- In Male and Female also, both achieved almost similar AUC's.