Age, Gender and Ethnicity Detection

Kushal poddar Ritika Jha Daneshwari Kankanwadi Rohan Kumar Akanksha Marskole

***Abstract*—A face of a person plays a very crucial part to identify the person uniquely. Two important traits namely age and gender can enhance the performance of face recognition. Here in this paper we have discussed different Deep Learning architectures to recognize age, gender and ethnicity on the UTK face data[1] provided by Kaggle[2]. Paper contains analysis of different algorithms, data balancing techniques and their influence in obtaining results. This paper also mentions some key co-relations between prediction errors of age, gender and ethnicity. The computing experiment was carried out on Colab, provided by Google.**

I. Introduction

The UTK[1] face dataset as provided by Kaggle[2], in the comma separated value(csv) format, consists of 23705 images, each of 48×48 pixels along with 3 labels that are age, gender and ethnicity.

Since children have immature faces, ethnicity prediction using the UTK face dataset is more challenging than FERET dataset which consists of a less number of images of children[3]. It was now known that methods based on adult faces could not be successfully extrapolated to juvenile faces, particularly as facial identification is highly susceptible to errors when there is an age difference between images of an individual [4]. Deep learning was deployed on UTK face dataset to predict age, where using SVM they got a mean absolute error of 5.25[5]. One other work for Gender prediction on the FERET dataset gives an accuracy of 90.45% and 91.5% for SVM[RBF] and SVM[Linear] respectively[6].

A. Our Contribution:

We have used the UTK face dataset[1], which we first preprocessed to get rid of all kinds of anomalies like empty or repeated data. We are then using this data to train on several different architectures. We first tried the basic Multi layer perceptron followed by SVM and CNN models to predict age, gender and ethnicity separately. Finally we created a Multioutput CNN model which predicts all three of them simultaneously. Looking at the low results in prediction of ethnicity, which might be due to unbalanced ethnicity classes, we augmented our data and again trained the last model for better results. All these models are explained including their algorithms, mathematical preliminaries, architecture diagrams which are discussed in the Methodology section.

WEBSITE:: After saving the model, we have developed a production ready web server using Flask web server for writing the backend APIs. The frontend server was developed using VueJS Framework VueJS compiles the files into static files like html, js, img, css, etc The web application predicts the age ethnicity, and gender of the images in real time For deployment, a paid linux VPS web server is used with NGINX to map the domain name of the application as well.

II. Discussion

(key challenges ,solution to challenges)

The classification of age, gender and ethnicity from images is a difficult task since they depend upon the small, curved and even certain patterns and features. Training a machine to learn these features is a matter of concern for research on this topic.

Also, since the addition of new gender classes worldwide, the dataset has to be improved so that models trained on them can be used for all gender types.

For the UTK dataset, it may be the case that an image with 'Indian' ethnicity may be misclassified as 'Black' due to colour or prominent similar features.

The UTK dataset [1] contains images from age 1 to 5 years also. Since the children within this age range have similar features, which develop as they reach their teenage, gender could be misclassified.

In our project, along with training the model we also provide these analyses that can be of further interest to other researchers.

A. About the UTK dataset[1] -

The provided dataset with 23705 data has redundant data, so we preprocessed and obtained a dataset of 22940 images. The UTK dataset has unbalanced ethnicity classes as shown in Figure 1{pie chart}. To balance them, we augmented the data while training the multi output CNN model. All the dataset visualization is depicted in Figure 1.

Viewpoint was also a challenge, where the face image can be oriented/rotated in multiple dimensions with respect to how the object is photographed and captured in image. No matter the angle in which we capture the image of a person's face, it's still a face. Dugin the augmentation for the multi output CNN model , we rotated the images by certain angles, thus even if we now test for a rotated image, it would give us more accurate results.

III. Methodology

We have used supervised learning techniques - Multi Layer Perceptron, Support Vector Machine and Convolutional networks.

Mathematical Functions used are :

1)

$$Sigmoid : S(x) = \frac{1}{1 + \exp^{-1}}$$

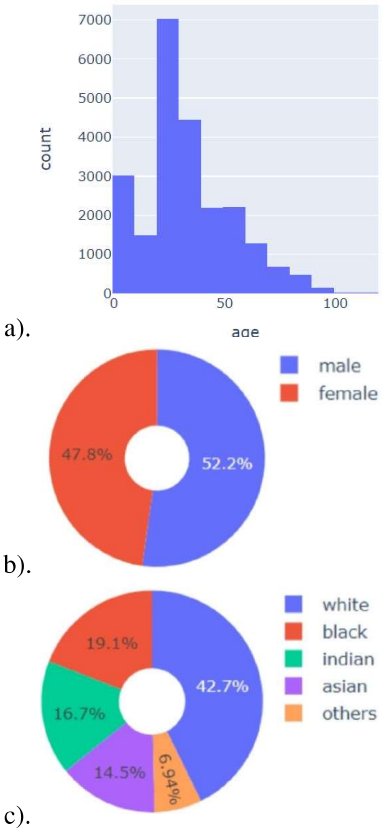


Fig. 1. Data Visualisation a). Age, b). Gnder, c).Ethnicity

where $S(x)$ =sigmoid function and e =Euler's number

2)

$$ReLU : g(z) = max(0, z)$$

3)

$$LeakyReLU : g(z) = max(\in z, z) with \in << 1$$

A. Multi Layer Perceptron (MLP)

MLP is the basic type of Neural Network where apart from an input and output layer, there are multiple hidden layers in between, with each hidden layer having some neurons. The neurons in the hidden layer and output layer have a bias and weights are attached to the input dimensions(or attributes). So the total number of trainable parameters becomes the sum of all biases and weights.

We created an MLP classifier for each of the outputs separately - Gender and Ethnicity, where each of the image pixels is taken as an input feature. The input and hidden layers are the same for both the models, with a change in the output layer. The architecture diagram for the Gender MLP model is as shown in Figure.

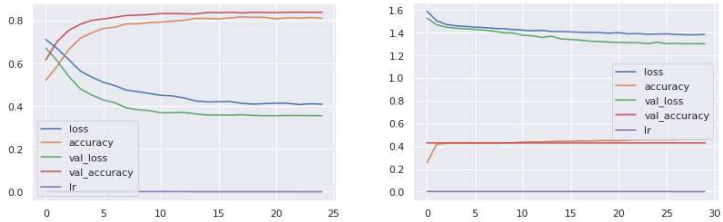
The input layers for both of the models have 2304 (image_height × img_width × number_of_channels) input features. There are three hidden layers, where each layer has a number of neurons in the 4:1 ratio for consecutive layers. Since there is a high possibility that the model may overfit the training data, Dropout layers, with dropout rate of 0.5, have been used after each Dense layer of neurons[9]. The output layers for:

1) Gender MLP model - It has two neurons corresponding to each class (0:Male, 1:Female), that outputs 1 if the class is present otherwise 0. Thus, it is a binary MLP classifier.
2) Ethnicity MLP model - It has five neurons corresponding to each class that outputs 1 if the class is present otherwise 0. Thus, it is a multi class classifier.

The mathematical operation carried out in the hidden and the output layers corresponding to kth neuron is given by the equation :

$$y = w_1x_1 + w_2x_2 + b$$

where, m is the dimension of the input vector, y_i is the predicted output corresponding to j^{th} sample, w_{ij} is the weight corresponding to the input x_i and bk is the bias corresponding to k^{th} neuron.



(a) (b)

Fig. 2. Training history plots for - a). Gender b). Ethnicity

B. Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data, the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimensional space this hyperplane is a line dividing a plane in two parts where class lies on either side.

Since SVMs are prone to overfitting data and aren't useful for large datasets, we spliced the dataset to find unique 2000 images and separated it into testing and training sets.

C. Convolutional Neural Networks (CNN)

Convolutional Neural Networks refer to a sub-category of neural networks: they, therefore, have all the characteristics of neural networks. However, CNN is specifically designed to process images. Their architecture is then more specific: it is composed of two main blocks.

The first block makes the particularity of this type of neural network since it functions as a feature extractor. To do this, it performs template matching by applying convolution filtering operations. The first layer filters the image with several convolution kernels and returns "feature maps", which are then normalized (with an activation function) and / or

resized. This process can be repeated several times: we filter the features maps obtained with new kernels, which gives us new features maps to normalize and resize, and we can filter again, and so on. Finally, the values of the last feature maps are concatenated into a vector. This vector defines the output of the first block and the input of the second.

The second block is not characteristic of a CNN: it is in fact at the end of all the neural networks used for classification. The input vector values are transformed (with several linear combinations and activation functions) to return a new vector as the output. This last vector contains as many elements as there are classes: element i represents the probability that the image belongs to class i. Each element is therefore between 0 and 1, and the sum of all is worth 1. Some architectures have been already made which are accurate[7], being inspired from them one, we developed different architectures.

The architecture for the CNN model that predicts gender is as shown in Figure 3. Here we have used six convolutional layers. each followed by a batch normalization and max pooling layer. After flattening i.e reducing the dimension of output from convolutional layer to one-dimension, we have obtained the output from the neuron which gives output 0 and 1, where 0:Male and 1:Female class. A similar model was constructed for age prediction and ethnicity classification.

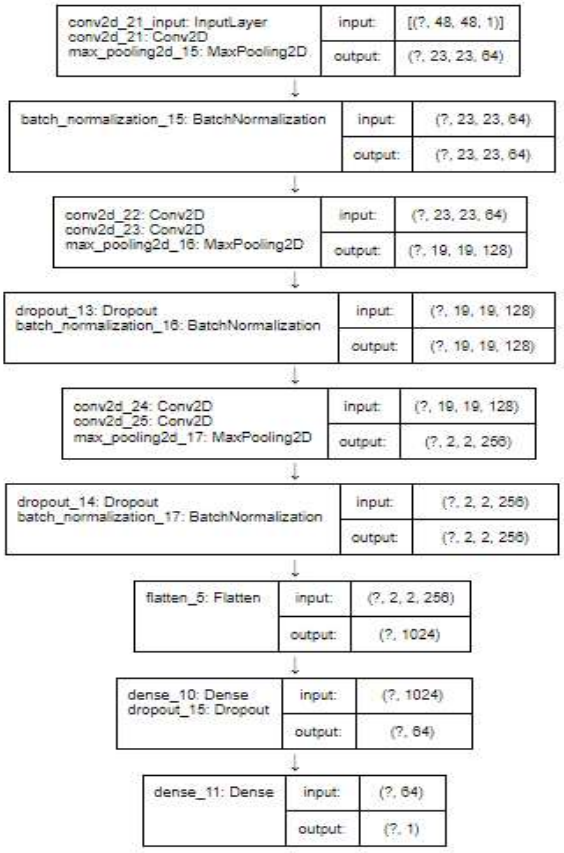
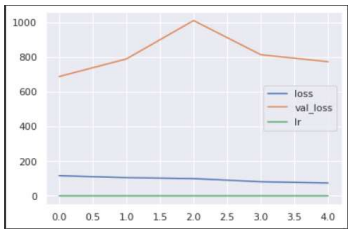


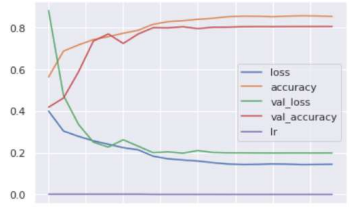
Fig. 3. CNN Gender Model Architecture

Since age is a continuous data, the last activation function used for age was 'Linear' which gave continuous output for age from one neuron. To classify the gender and ethnicity, the activation function used was 'Sigmoid'. Since there are five classes of ethnicity, we one-hot encoded the output label that was obtained by five different neurons. The optimizer used was 'Adam' optimiser and the loss functions were, for age: mean square error(mse), ethnicity: categorical cross entropy, and for gender: binary cross entropy. The training history of all three models is as below Figure 4.

a).



b).



c).

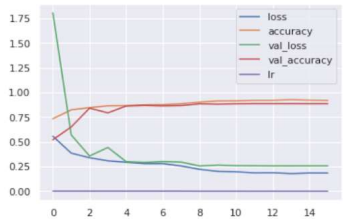


Fig. 4. Plots of training history of CNN models a). Age, b). Ethnicity and c). Gender

D. Multi Output CNN

Neural networks can produce more than one output at once. For example, if we want to predict age, gender and race of a person in an image, we could either train three separate models to predict each of them or train a single model that can produce all three predictions at once. Earlier we made different CNN models for each output and thus the training and testing on the dataset is done three times, separately for each output. But now, in this short experiment, we tried to develop and train a deep CNN that can produce multiple outputs - age, gender and ethnicity, simultaneously. Some attempts to make such a model using a keras tuner have been done[8].

As shown in the architecture diagram in Figure 5, the model is pretty straightforward CNN until the bottleneck returned by the Global Max Pooling layer. Here Global Max Pooling layer is used so that the features extracted could be used in different neural networks for different output categories. Up to the bottleneck layer we have repeated the convolutional blocks with 'Leaky Relu' layers, batch normalization and average or max pooling.

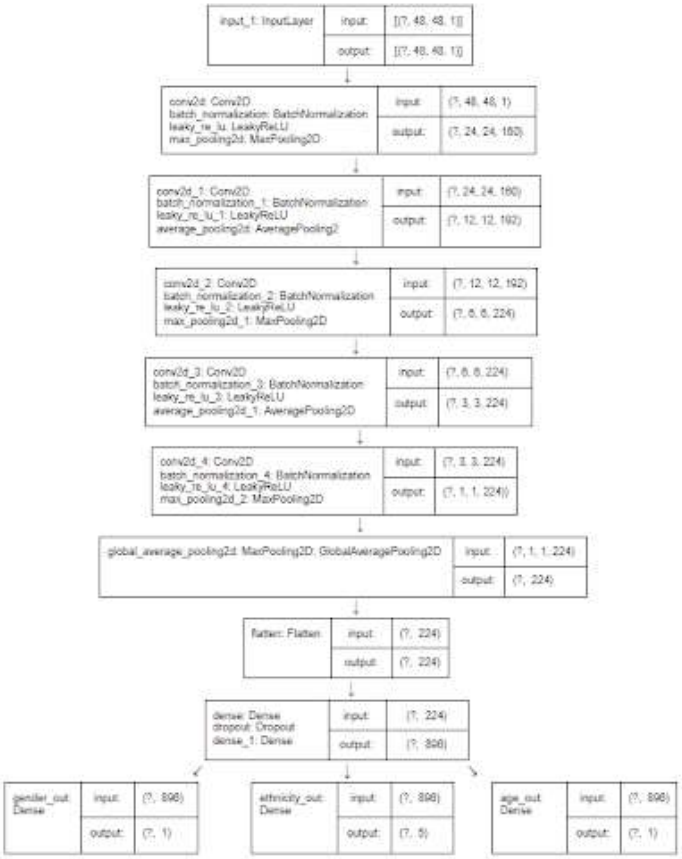


Fig. 5. Multi output CNN architecture Diagram

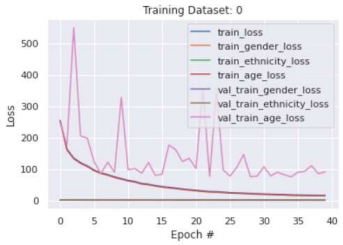
The bottleneck layer's output is branched out into three separate branches to predict three different outputs. For predicting age, the bottleneck layer's output is given as input to the dense layer with linear activation. Thus, a 'linear regressor' is created to predict age.

Similarly, for race and gender the same bottleneck output is given as input to another dense layer with 'Sigmoid' activation function. The optimizer used is root mean square propagation(rmsprop).

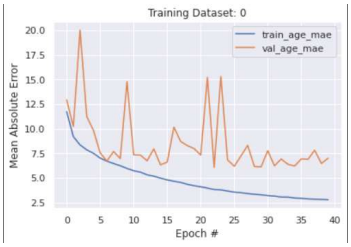
Loss functions used are mean squared error(mse) for age, categorical cross entropy for ethnicity and binary cross entropy for gender output. The training hostory can be seen in Figure 6.

The first training of this model gave us less satisfactory results as discussed in the results section for ethnicity prediction. We analysed that this was due to imbalanced ethnicity classes i.e. ethnicity class 0 had much more images than other ethnicity classes. So we augmented the data of other ethnicity classes which reduced the effect of imbalanced data on the performance of the model. Also the challenge of viewpoint variation is solved as now the model is trained with images which have faces at different angles. We rotated and zoomed in the images slightly to augment the data as we can observe in Figure 7.

a).



b).



c).

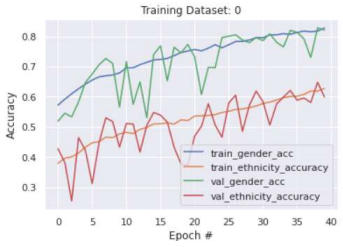


Fig. 6. Multi output CNN model a). Plott of losses of MultiOutput CNN, b). Age mean absolute error plotted against epochs, c). Training history of accuracy of gender and ethnicity

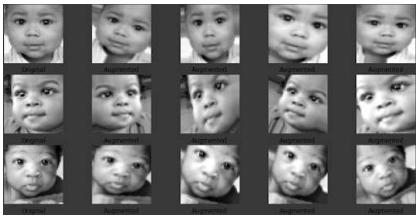


Fig. 7. Augmented Images

IV. Results

A. Multi Layer Perceptron (MLP)

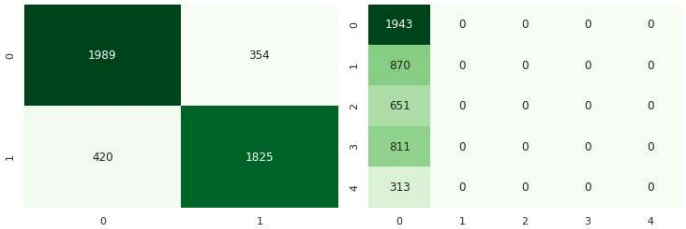


Fig. 8. Confusion Matrix- a). Gender MLP model b). Ethnicity MLP model

When the model was tested against test data, the accuracy obtained for the gender model was: 83.1% and for ethnicity model: 42.3%.

Here, two different models were created to see how models perform for different outputs when the same inputs are given.

Thus a new model which gives all the three outputs simultaneously is required.

B. Support Vector Machine (SVM)

The SVM model for age gave an accuracy of 82% while the ethnicity model gave an accuracy of 52%. Even after changing the parameters including c and gamma, it did not gain much higher accuracy. The mean absolute error was 15.3 years, which was a large difference. Clearly SVM wasn't able to solve the problem in detecting age.

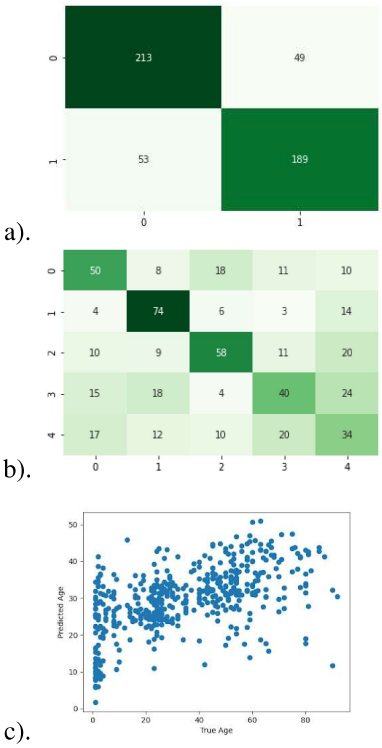


Fig. 9. SVM model a). Confusion matrix of gender model, b). Confusion matrix of ehtnicity model, c). Scatter plot of age model

C. Convolutional Neural Networks (CNN)

The CNN age model gave a mean absolute error of 6.85 whereas the gender model gave 89.42% and the ethnicity gave 81.5% of accuracy on the test dataset, respectively. The scatter plot of predicted and actual age is shown in Figure 9. The confusion matrices for both the ethnicity and gender model are shown in Figure 10.

D. Multi Output CNN

The Multi Output CNN when trained over 40 epochs gave unsatisfactory results with mean absolute error for age as 5.09, which is better than what we got in CNN. The accuracy for gender and ethnicity came out to be 84.3% and 60.5% respectively. We can see that model performed well for age and gender, although it is unable to predict ethnicity significantly.

The results of this model are presented in Figure 11.

Performance metrics for the multi output gender and ethnicity model in Table I:

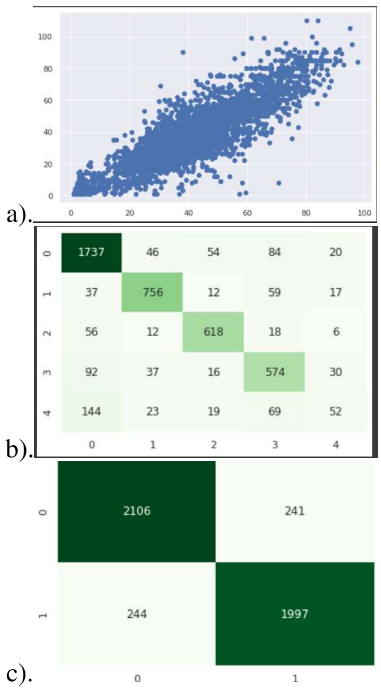


Fig. 10. CNN model a). Scatter plot of age model, b). Confusion matrix of ehtnicity model, c). Confusion matrix of gender matrix

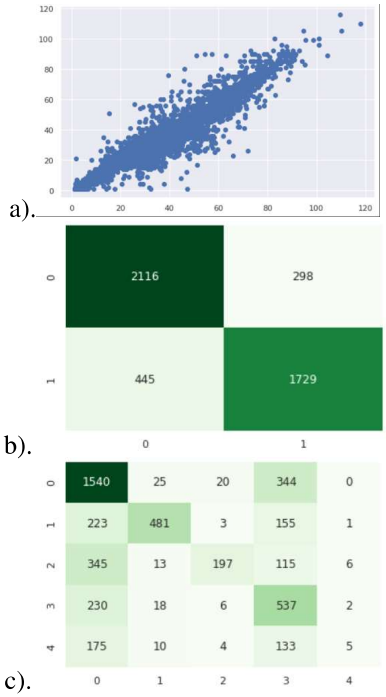


Fig. 11. Multi output CNN odel a). Scatter plot of age model, b). Confusion matrix of gender model, c).confusion matrix of ethnicity model

This model was again trained on augmented images. But due to increased amount of data, computationally we were unable to complete the training process, thus we trained it for only 30 epochs and the accuracy obtained was 82.42% and 62.47% for gender and ethnicity respectively and the mean absolute

Metrics	Class 0 (Male)	Class 1 (Female)
Precision	0.81	0.85
Recall	0.87	0.77
F1 score	0.84	0.81
Support	2384	2204

(a) Gender model's performance table

Metrics	White	1	2	3	Others
Precision	0.61	0.88	0.86	0.42	0.36
Recall	0.80	0.56	0.29	0.68	0.02
F1 score	0.69	0.68	0.43	0.52	0.03
Support	1929	863	676	793	327

(b) Ethnicity model's performance table

TABLE I
PERFORMANCE METRICS OF MULTI OUTPUT CNN MODEL

error for age regressor model reduced up to 2.21. As we can see that a lesser number of epochs gave better results than the previous model, we assume that a training of 40 epochs could further have enhanced the accuracy.

V. CONCLUSION

We have finally compared all our results in Table II . We can infer that CNN models have worked out the best for us, although the results of Multi Output CNN were quite comparative, but due to the inefficiency of ethnicity prediction we can't declare it as the best model.

category↓/ model→	MLP	SVM	CNN	Multi Output CNN
Age(mse)	-	15.3	6.85	5.09
Gender(Accuracy)	83.1%	82%	89.42%	84.3%
Ethnicity(Accuracy)	42.3%	52%	81.5%	60.5%

TABLE II
COMPARISON OF ALL MODELS

We have now tried to pluck out some key inferences from the results obtained by the CNN models.

1) In Gender prediction, out of 4588 samples of testing data, 412 samples were misclassified. Out of these misclassified faces, 200 belong to the age group of 10 years and below, 72 are from age group 10 to 25 years, 64 are from age group of 25 to 40 years and 76 from age group of 40 years and more. Also among the 412 misclassified, 181 belong to ethnicity class 'White' and only 29 belong to class 'Others'.
2) In gender classification, out of 4588 samples of testing data, 588 were misclassified. Out of these misclassifications, 311 belong to class 'Female' and 277 to 'Male'. Also among these 588 misclassified samples, 240 belong to the age group of 25 to 40 years, and only 94 belong to the age group of 10 years and below.

We can therefore say that the chances of predicting a wrong gender is highest among the children that are aged 10 years and below and people that belong to the ethnicity class 'White'.

While classifying the ethnicity, the chances of wrong prediction are almost equivalent for class 'Male' and 'Female'.

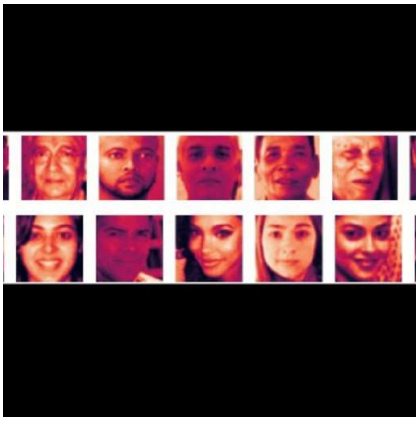


Fig. 12. Wrong Ethnicity

Fig. 13. Predicted Female but Actually Male

Fig. 14. Predicted Male but Actually Female

Also the people between the age group of 25 to 40 years have the highest probability of being misclassified for ethnicity.

Thus from the above analysis we can say that,

1) The ethnicity is based on colour and sharp features of face. Since the images we used were single channelled, some ethnicity values were misclassified.
2) Gender of children with age 10 years or below were misclassified, since the features of both the genders are somewhat similar in this age group.
3) Data augmentation is required to remove the biases among classes. Also, as we can see from the results above, the model trained on augmented models gave better results (for a lesser number of epochs) than that that was trained on the non-augmented images.

FURTHER WORK

In our project, we created Multi Layer Perceptron, Convolutional Neural Networks and Support Vector Machine models for each output separately and a multi output Convolutional Neural Network.

The multi output CNN can be improved by using a different architecture of CNN model. Also, these models can be trained on coloured images from the original UTK dataset to obtain better results. The model that gives the best performance can be used in applications such as crime investigation, cosmetic industries etc.

Furthermore, classification can be done based on each specific part of the face like eyebrows, eyes, lips, etc. Since the final classification would be based on all these, thus the chances of misclassification will reduce and we can be more precise for the classification of ethnicity.

ACKNOWLEDGEMENT

We extend our sincere gratitude to our professors - Prof. T.V. Vijay Kumar and Prof. Prerna Mukherjee for their constant support and guidance.

REFERENCES

[1] UTK Face website. [Online]. Available: https://susanqq.github.io/UTKFace/

[2] The kaggle website. [Online]. Available: https://www.kaggle.com/nipunarora8/age-gender-and-ethnicity-face-data-csv

[3] Ferguson, Eilidh Louise. "Facial identification of children: a test of automated facial recognition and manual facial comparison techniques on juvenile face images." PhD diss., University of Dundee, 2015.

[4] Guo, Guodong, Charles R. Dyer, Yun Fu, and Thomas S. Huang. "Is gender recognition affected by age?." In 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, pp. 2032-2039. IEEE, 2009.

[5] X. Wang, R. Guo and C. Kambhamettu, "Deeply-Learned Feature for Age Estimation," 2015 IEEE Winter Conference on Applications of Computer Vision, Waikoloa, HI, 2015, pp. 534-541, doi: 10.1109/WACV.2015.77.

[6] S. K. Gupta and N. Nain, "Gender Recognition for Juvenile Unconstrained Faces Using Gabor-MeanPool-DCT Feature Model and SVM-Kernel Optimization," 2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), Sorrento, Italy, 2019, pp. 499-504, doi: 10.1109/SITIS.2019.00085.

[7] The kaggle website. [Online]. Available: https://www.kaggle.com/marktsvirko/face-recognition-with-keras

[8] The kaggle website. [Online]. Available: https://www.kaggle.com/dijiswiki/keras-tuner-multi-output-cnn

[9] C. Ranjan "Rules-of-thumb for building a Neural Network" in towards data science, 2019. [Online]. Available: https://towardsdatascience.com/17-rules-of-thumb-for-building-a-neural-network-93356f9930af