Identification of Gender and Person using Ear Images

Bachelor of Technology

in

INFORMATION TECHNOLOGY

DEPARTMENT OF INFORMATION TECHNOLOGY

INDIRA GANDHI DELHI TECHNICAL UNIVERSITY FOR WOMEN

May, 2020

Manmeet Kaur-00701032017
Information Technology
Indira Gandhi Delhi Technical University For Women
manmeet0307.kaur@gmail.com

Priyanka Daryani-02201032017
Information Technology
Indira Gandhi Delhi Technical University For Women
priyanka.daryani010@gmail.com

Srishti Kohli-05101032017
Information Technology
Indira Gandhi Delhi Technical University For Women
srishtik94@gmail.com

Nidhi Mahipal-05701032017
Information Technology
Indira Gandhi Delhi Technical University For Women
nidhimahipal21@gmail.com

Megha Varshney -060010302017
Information Technology
Indira Gandhi Delhi Technical University For Women
varshney.megha070@gmail.com

Abstract

Various bio-metric methods exist today to identify either the person or the features unique to him/her. Face recognition, fingerprint scanning, voice recognition are some of the present industry methods implemented for authentication. At bio-forensics level, verification through DNA sequencing is used for identification. In recent studies, it has been found that ear also play a role in providing identity to a person. We approach this problem by leveraging the power of deep learning. The images utilised for the above purpose are captured in an uncontrolled environment. Hence, making it's application possible at industry scalable level. We aim to determine the gender and person from the images of ear by applying Deep Learning Algorithms. We experimented both with scratch learning as well as fine tuning state of the art models.

Contents

Chapter 1

Introduction

The human ear comprises a huge amount of unique and specific features. This gives it the capability to be considered as a strong candidate for Biometrics. Hence, it can be used for human identification. Ear biometrics, therefore has become a hot research topic lately as it is an alternative for biometrics using iris, face or finger. It has been seen that the ear changes only 1.22mm in a year. In early ages, the growth in its size is symmetrical and downward bulging. This is a measurable effect. However, it has been observed under multiple medical studies, that noticeable changes in the ear shape occur after the age of 70 or before the age of 8 years. Furthermore, it undergoes only slight changes in a span of every 10 years.

Also, ear biometrics is more beneficial in cetain scenarios like conduction of Passive Biometrics. Owing to its position on the face (centre of the face profile), ear is one of the best candidates for Passive Biometrics since it can be captured easily from a distance even in absence of the subject's awareness and cooperation. Hence, this serves the requirement of secrecy in authentication systems.

Because of these overwhelming features, ear is a significant modality in forensic science for identification and biometric studies. For example, all biometrics techniques concede some limitations. Face Recognition is tough when facial expressions change, or in case of identical twins or when cosmetics are applied on the face. Similarly, fingerprints of people undergoing Chemotherapy or senior citizens start fading away. Ear recognition can be used as a clear alternative in these situations.

Our research focuses on leveraging the power of Computer Vision to conduct Gender Identification and Person Identification from ear images captured in an unconstrained environment.

Chapter 2

Literature Survey

Following are the interesting questions pertaining to chosen dataset:

- Whether a **person can be recognised** on the basis of the ear?
- Whether the **gender of the person** can be **recognised** on the basis of ear?
- In what domains can this model have **scope of application**?
- Is the dataset **biased towards any gender**? How does gender affect the model performance?
- How is this **better than other biometric models**?

Chapter 3

Dataset Description and Exploration

3.1 Introduction to Data Exploration:

Data exploration is extracting knowledge from data efficiently. Exploratory data analysis aims at providing a set of simple tools to obtain some basic understanding of the data. The results of data exploration help in grasping and understanding of the structure of the data, the distribution of the values, and the presence of extreme values and interrelationships within the data set.

Data visualization is representation of information and data graphically. By making use of visual elements like charts, graphs, and maps, data visualization tools help us understand trends, outliers, and patterns in data.

Data Source: https://data.mendeley.com/datasets/yws3v3mwx3/3

3.2 Dataset Description

The dataset consists of ear images of 164 people totalling to 28,412 images of both males and females. Out of 164 people, there are 98 males and 66 females. The data contains images of both left and right ear.The images cover large variation in pose, scale, illumination, occlusion, resolution, light condition. This dataset could be used on a variety of tasks, e.g., person authentication, classification. We are making use of above data in Gender Recognition and Person Identification.

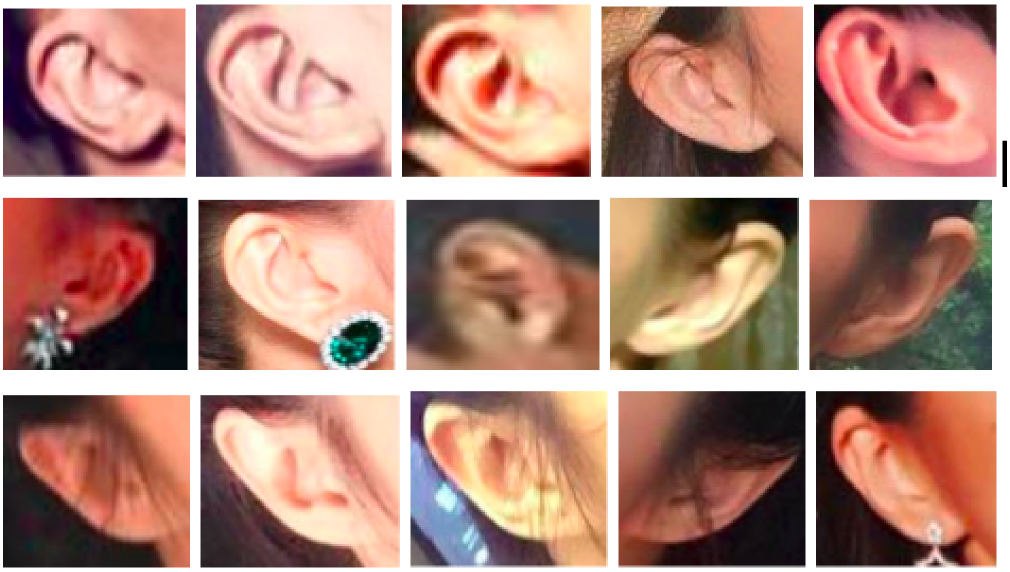
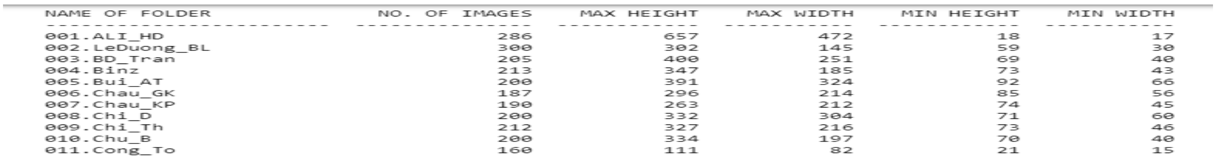


Figure 3.1: Structure of Images that needs preprocessing

3.3 Data Exploration and Inferences

It is observed that some of the ear images are rotated as shown in Fig 3.1 which is non uniform as images are rotated by different angles, this generalisation is made by the fact that ear is elongated in shape with length greater than width,trying keeping a record of images having width greater than length ,which is around 788 images it is concluded that these images require preprocessing.Henceforth we have used Image Augumentation so as to rectify this.

3.4 Data Visualisations



```
NAME OF FOLDER              NO. OF IMAGES   MAX HEIGHT   MAX WIDTH   MIN HEIGHT   MIN WIDTH
-----------------------   ---------------  -----------  -----------  -----------  -----------
001.ALI_HD                            286          657         472           18           17
002.LeDuong_BL                        300          302         145           59           30
003.BD_Tran                           205          400         251           69           40
004.Binz                              213          347         185           73           43
005.Bui_AT                            200          391         324           92           66
006.Chau_GK                           187          296         214           85           56
007.Chau_KP                           190          263         212           74           45
008.Chi_D                             200          332         304           71           60
009.Chi_Th                            212          327         216           73           46
010.Chu_B                             200          334         197           70           40
011.Cong_To                           160          111          82           21           15
```

Figure 3.2

Figure 3.2 represents the data observations in tabulated form:

- Column 1: Name of the folder
- Column 2: No. of images in the corresponding folder or No. of samples per person
- Column 3: Maximum height of image in the corresponding folder
- Column 4: Maximum width of image in the corresponding folder
- Column 5: Minimum height of image in the corresponding folder
- Column 6: Minimum height of image in the corresponding folder

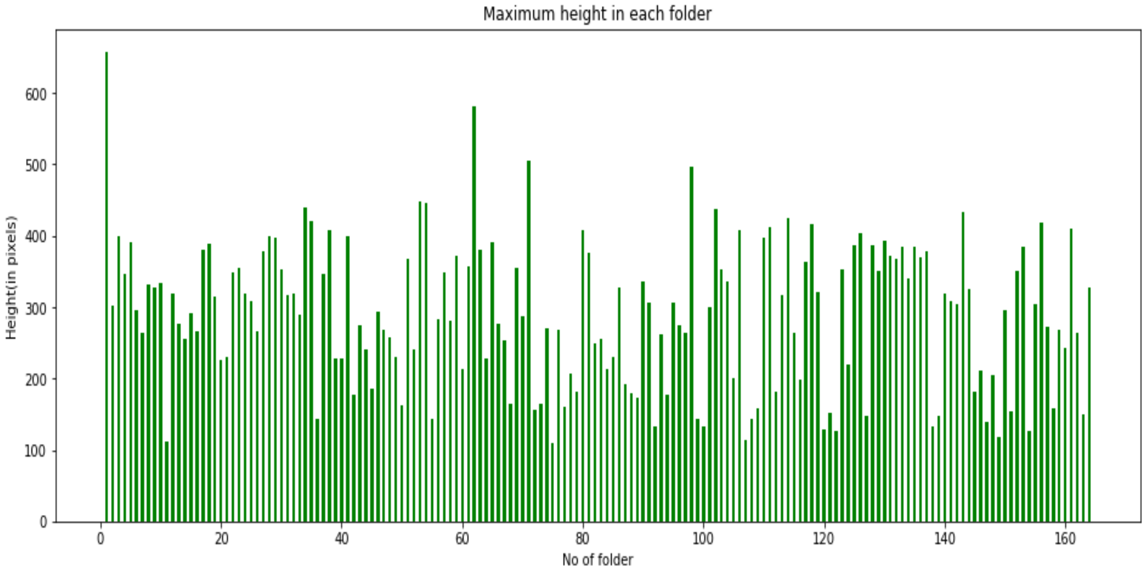


Figure 3.3: Maximum height of image corresponding to each person

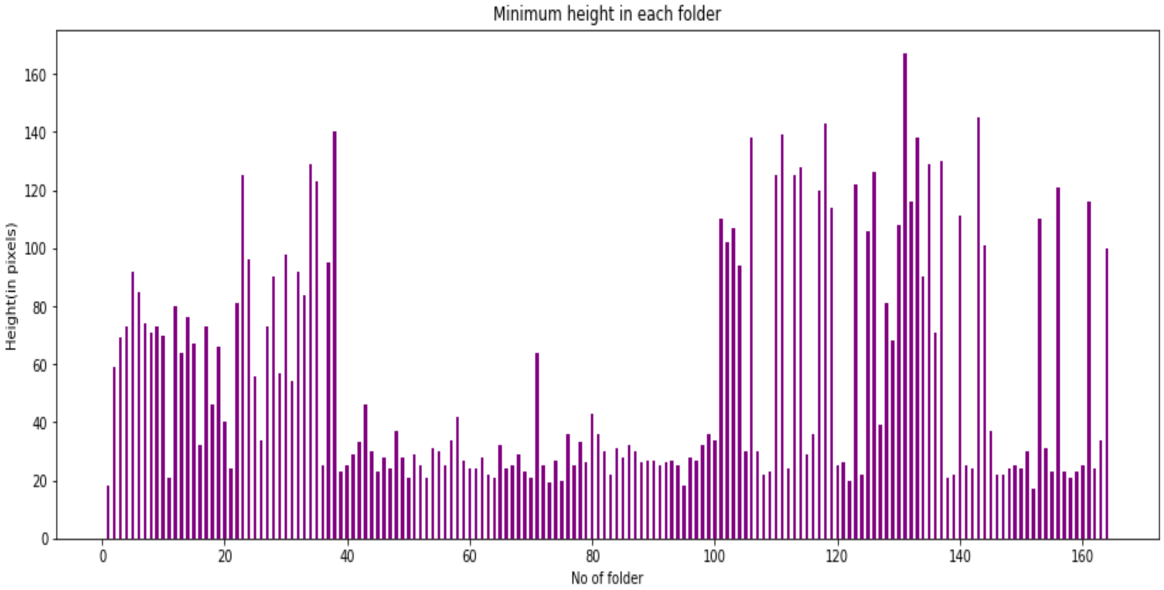


Figure 3.4: Minimum height of image corresponding to each person

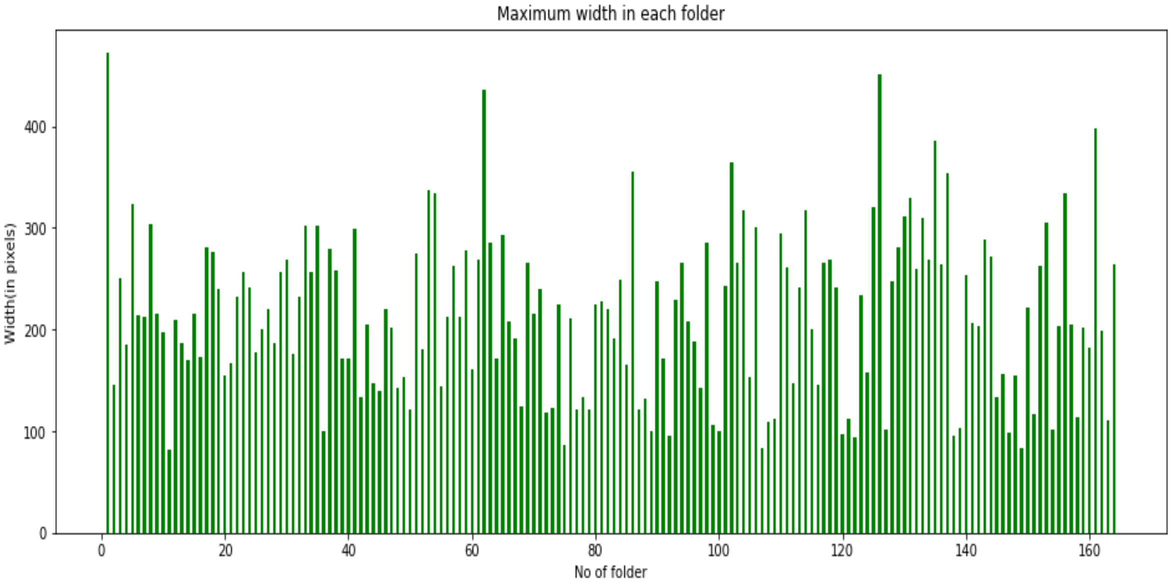


Figure 3.5: Maximum width of image corresponding to each person

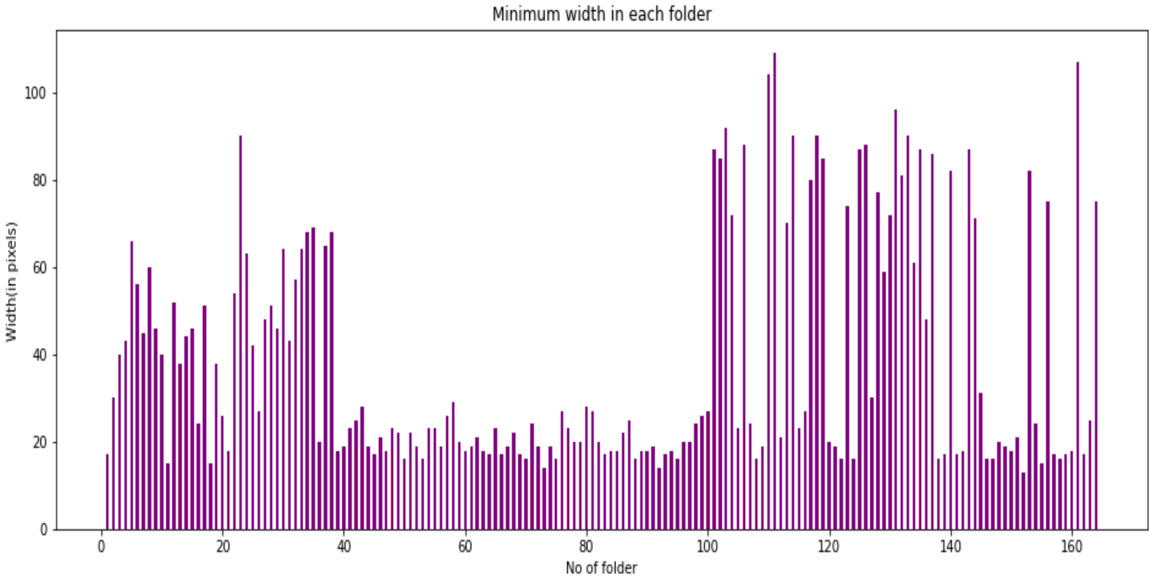


Figure 3.6: Minimum width of image corresponding to each person

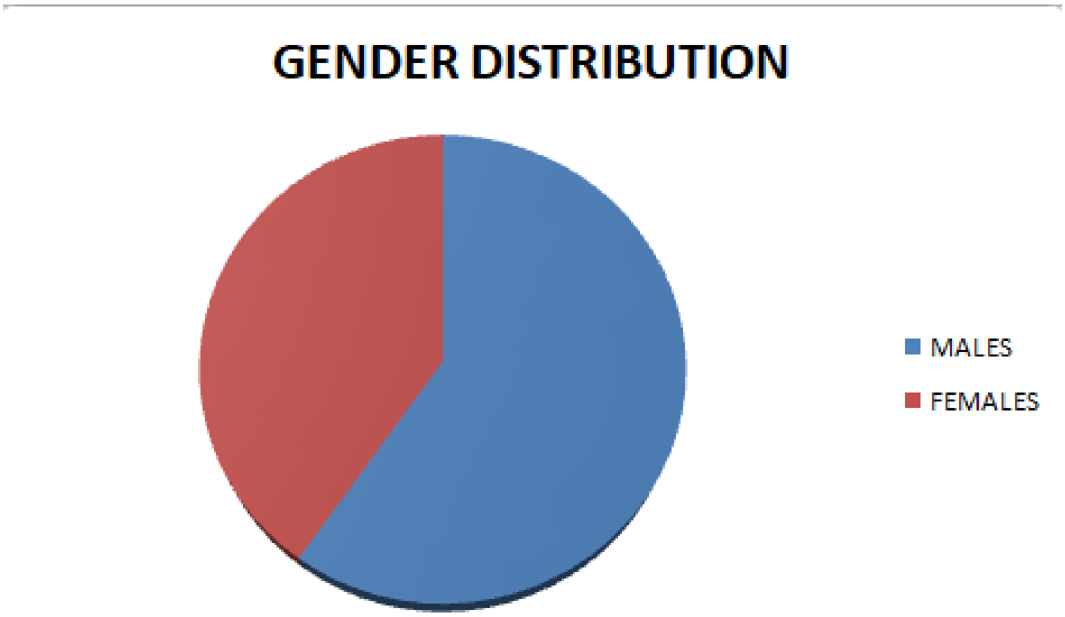


Figure 3.7: Gender distribution in sample

We made use of Gender Distribution so as to do binary classification of our data.

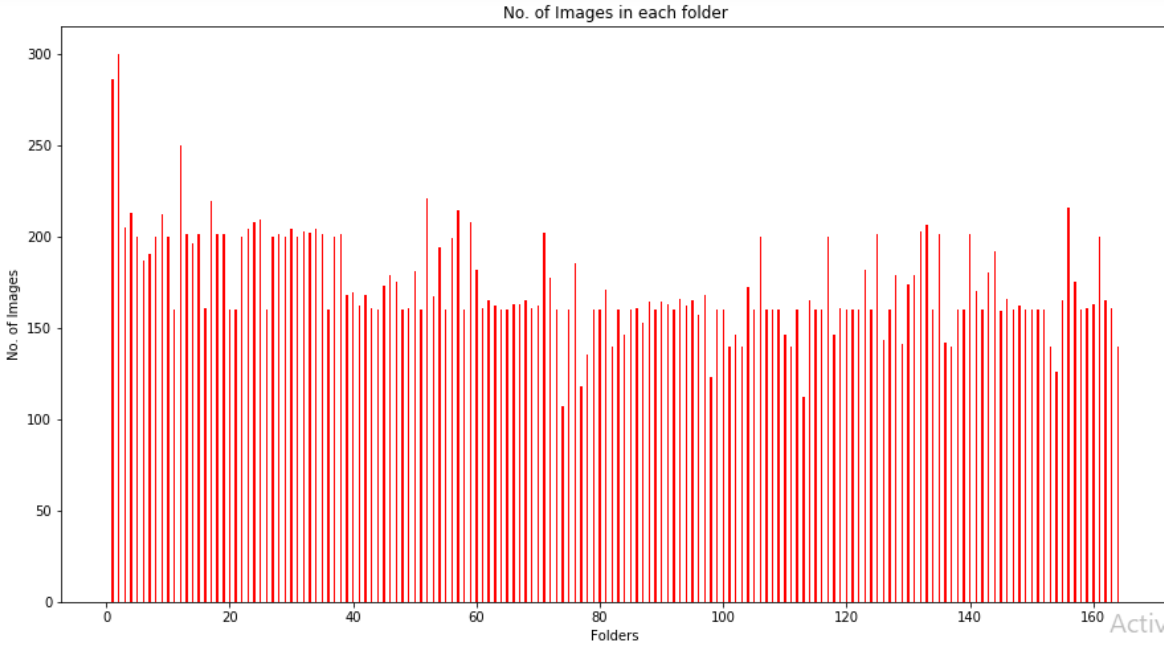


Figure 3.8: No. of samples per person

- There are 164 folders belonging to 164 different people consisting of ear image samples of people.

1. Height and Width of Images

   (All values are in pixels)

   Below graph plots visualizes variations of height and width of images and the number of images corresponding to a particular height or width.

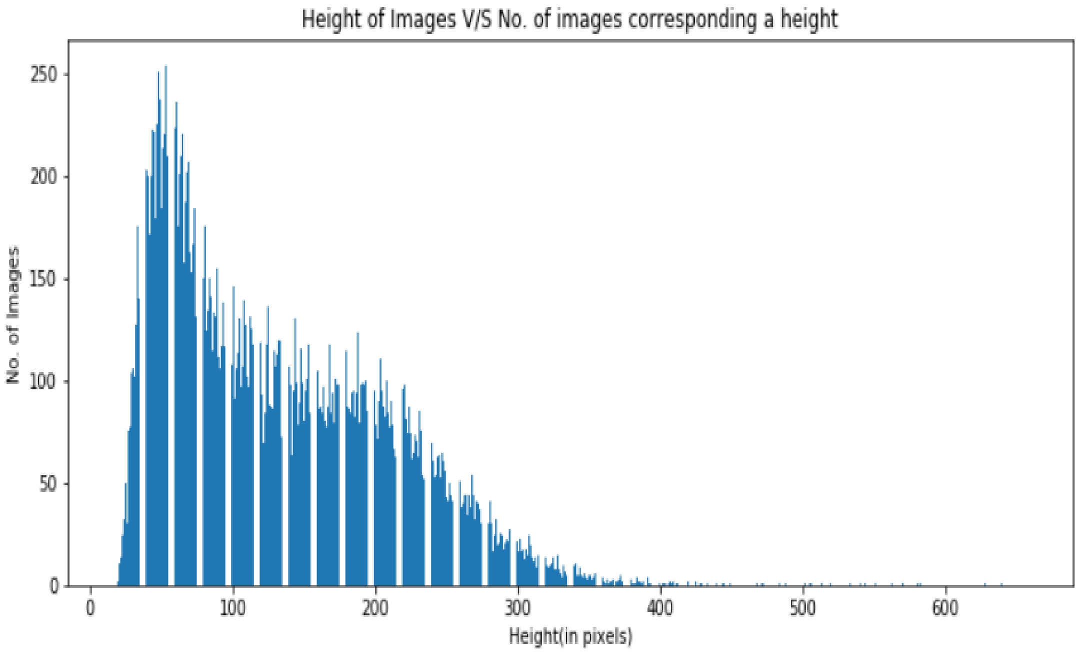


Figure 3.9: Number of images corresponding to height

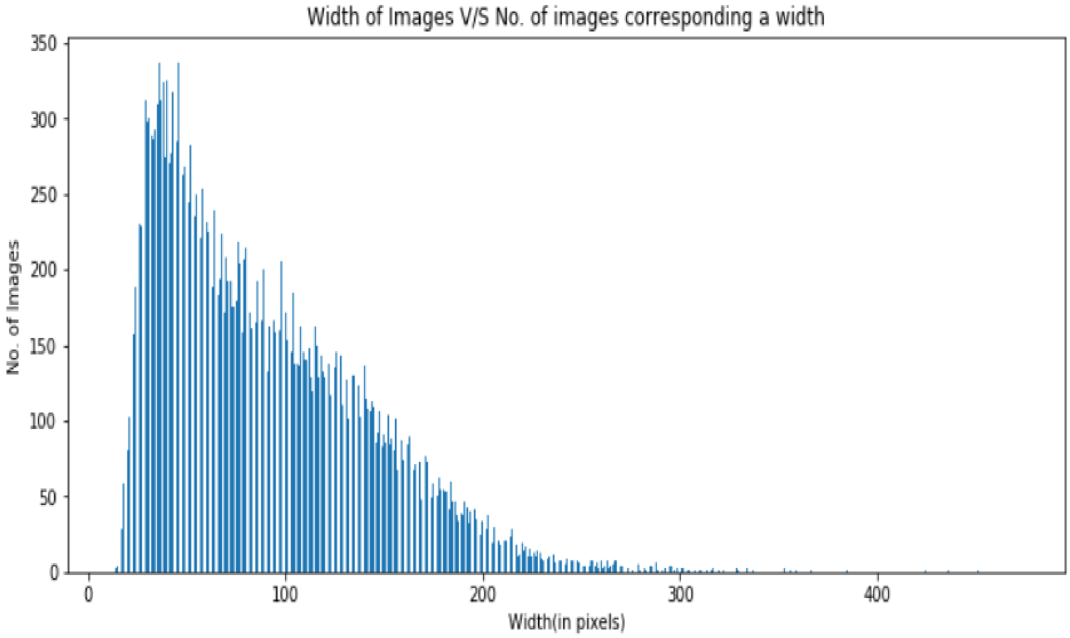


Figure 3.10: Number of images corresponding to width

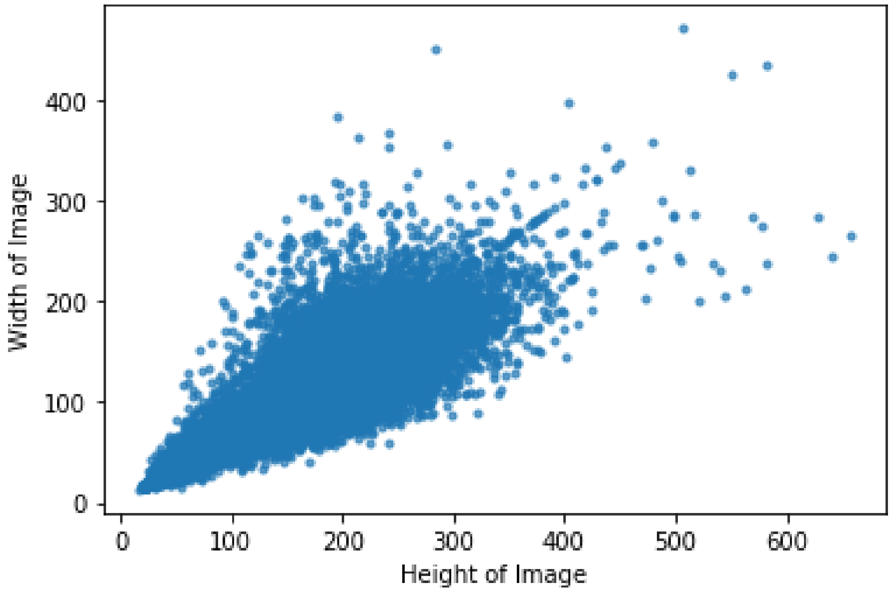


Figure 3.11: Relationship between height and width of images

3.5 Observations

- Figure 3.9 depicts height of all images v/s no. of images corresponding to a particular height, and we observed the following.

  The minimum height is observed as 17. The maximum height is observed as 675. The no. of images corresponding maximum height is observed as 53. The no. of images corresponding minimum height is observed as 498. The range of height is observed as [17-675]. The mean height is observed as 131.524. The median height is observed as 115.

- Figure 3.10 depicts width of all images v/s no. of images corresponding to a particular width, and we observed the following.

  The minimum width is observed as 13. The maximum width is observed as 472. The no. of images corresponding maximum width is observed as 46. The no. of images corresponding minimum width is observed as 330. The range of width is observed as [13-472]. The mean width is observed as 89.09. The median width is observed as 77.

- Figure 3.11 depicts the relationship between height and width of all images and we observed the following:

  Correlation between height and width is observed as 0.89.

Height and width data is used to feed average height and width to the input tensor thereby helping in gender Identification and Person Identification.

Chapter 4

Proposed Methodology

In this research, we try to identify the gender of person from their ear image. This a binary classification problem.

4.1 Gender Identification

4.1.1 Custom CNN

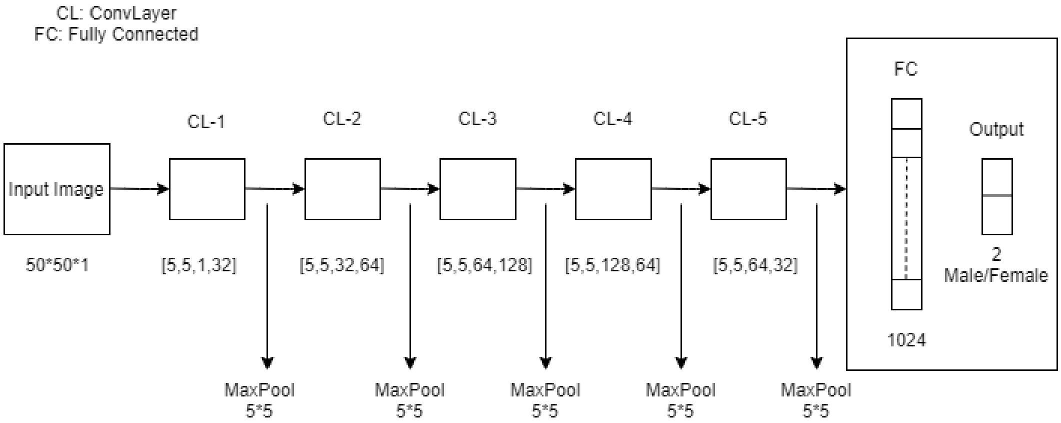
4.1.1.1 Description

Conventional Convolutional Neural Network can be described as the sequence of various steps, first one being the Convolution step which comprises of conversion of the input image to feature map or matrix using feature detector. In our model, the Rectified Linear Unit activation function(relu) is applied to increase non-linearity. Max-pooling is used to make the pooled feature map, which in turn also prevents over-fitting. The pooled feature map is flattened into column, in order to feed into our model. Following this, is the fully connected layer and the output layer.

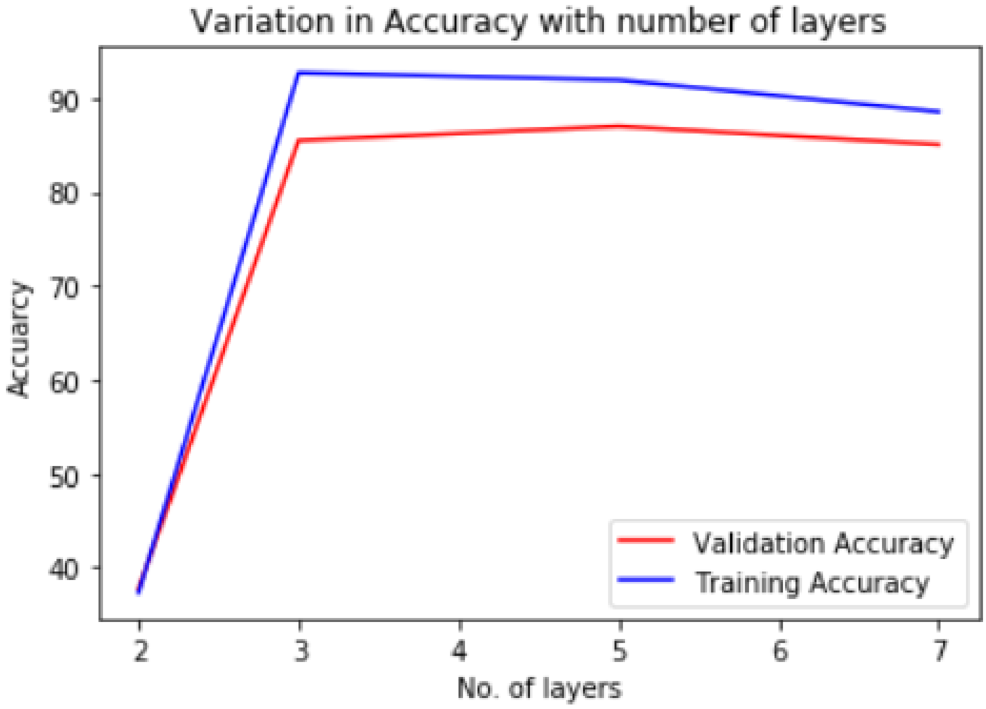
In the model built for this research, input images of 50*50 dimensions are fed after pre-processing(resizing and gray-scaling), followed by 5 hidden layers and a fully-connected layer. The final output layer has dimension equal to the number of classes to be predicted in our case(Gender) which is 2 i.e Male/Female.

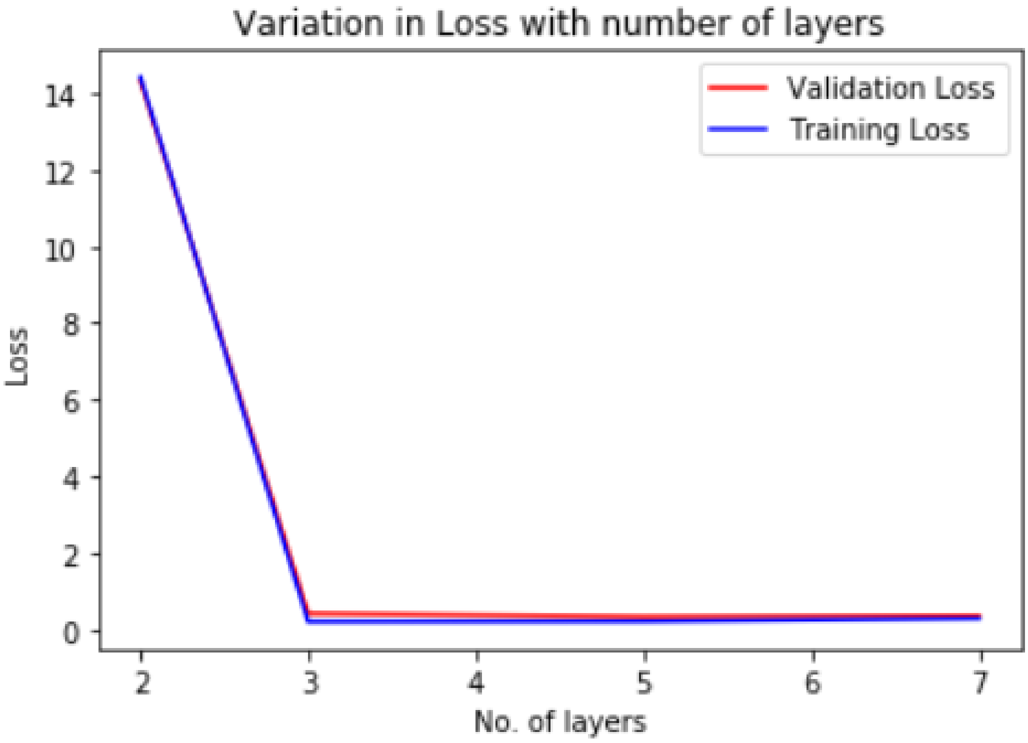
The data is divided in train and test set with ratio 75:25.

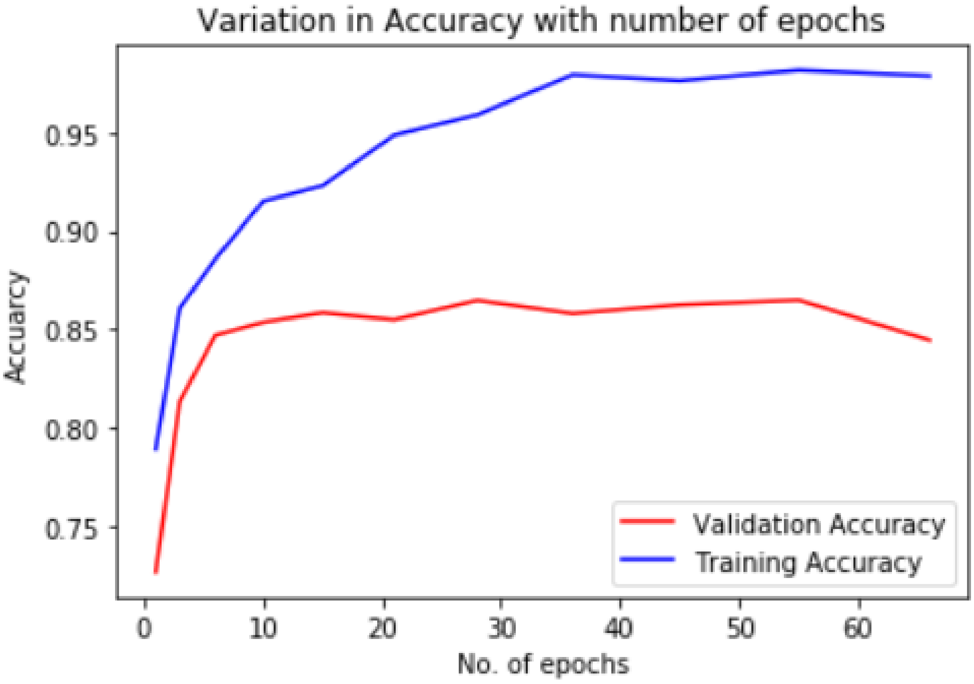
4.1.1.2 Architecture of Implemented CNN model

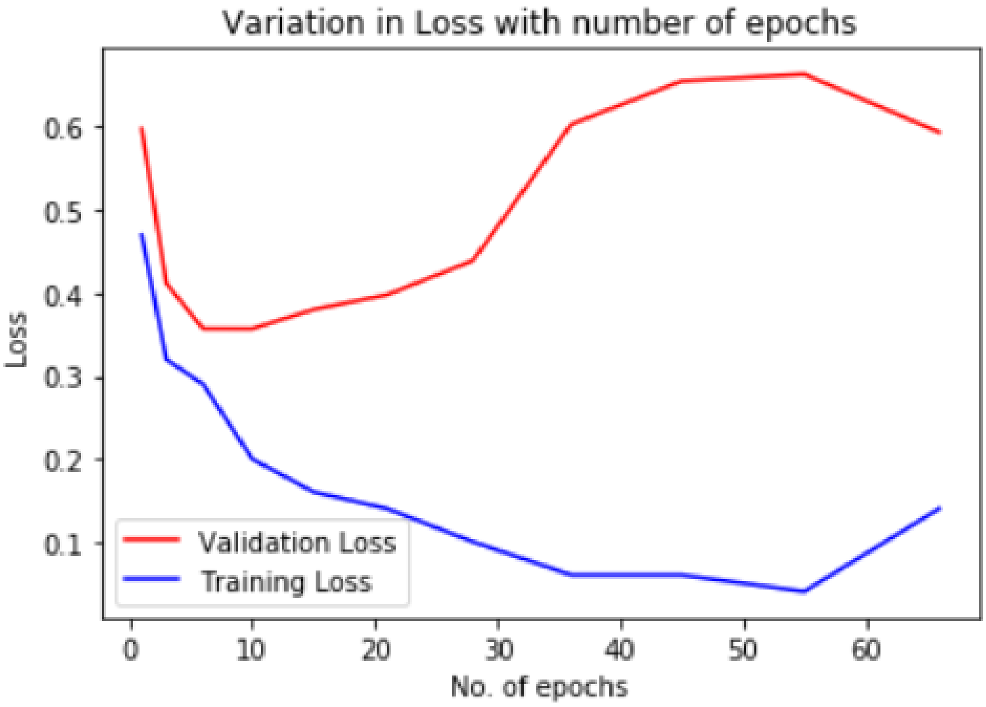


4.1.1.3 Visualization of Results and Parameter Tuning




Variation in Loss with number of layers
Validation Loss
Training Loss
Loss
14
12
10
8
6
4
2
0
2
3
4
5
6
7
No. of layers


Variation in Accuracy with number of epochs
Accuarcy
0.95
0.90
0.85
0.80
0.75
Validation Accuracy
Training Accuracy
0
10
20
30
40
50
60
No. of epochs



4.1.1.4 Conclusion

After visualizing the results of parameter tuning i.e changing number of layers and epochs, a custom CNN model of 5 hidden layers was built and was trained with 15 epochs (validation accuracy was observed to maximum and validation loss minimum) after which validation accuracy of 87% was achieved.

The below image shows the Actual vs Predicted labels by the model, where M is for Male and F is for Female. All the 12 images are taken from the shuffled test set.

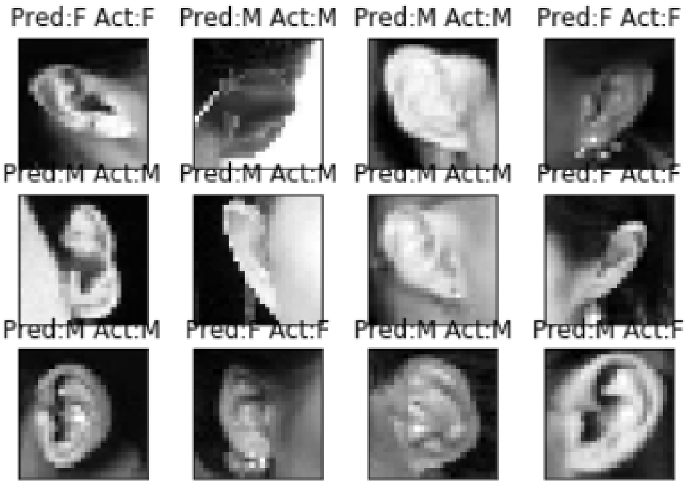


Figure 4.1: Actual vs Predicted Labels

4.1.2 Transfer Learning

4.1.2.1 Description

Transfer Learning is a technique where knowledge gained by model while it is being trained on one of the previous tasks is stored and is utilized as a starting point to solve a new similar task. The knowledge gained can be reused partly or wholly to accelerate the training and improve the performance of a model on the problem of interest.

For Gender identification, the pre-trained network has been fine-tuned, in order to adapt the weights entirely during the training of the model.The data was read from data folder which further contains 164 sub-folders.

Based on sub-folder name two new folders were created inside train and test set i.e Male and Female.First 98 sub-folders belong to male and remaining belongs to female. The data is divided in train and test set with ratio 80:20.For best results Image Augmentation technique has been adopted in which each test image has been re-scaled to 1./255,rotated within range of 20 ,zoomed 30 % , translated in both vertical and horizontal direction by 20 %.

4.1.2.2 Architecture

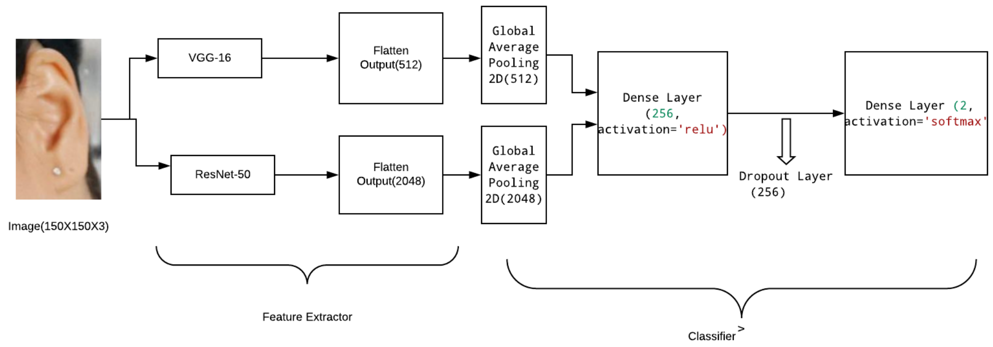

Image(150X150X3)
VGG-16
ResNet-50
Flatten Output(512)
Flatten Output(2048)
Global Average Pooling 2D(512)
Global Average Pooling 2D(2048)
Dense Layer (256, activation='relu')
Dropout Layer (256)
Dense Layer (2, activation='softmax'
Feature Extractor
Classifier

Figure 4.2: Architecture

4.1.2.3 VGG 16

VGG16 is a pre-trained convolutional neural network with a depth of 16 layers and 14,846,530 parameters. This deep learning model is used for image classification and has weights trained on ImageNet.

The input layer is followed by a 16 layer deep network, VGG16.The network takes 150 X 150 X 3 images as input and yields a flattened output of length 512. Weights of last 10 layers has been trained to adapt the weights entirely during the training of the model. The result from VGG16 is fed into 4 layers deep artificial neural network.First layer of following ANN is Pooling layer which is followed by Dense layer of size 256. The dense layer is followed by a dropout layer which is further followed up by Dense layer of size 2.

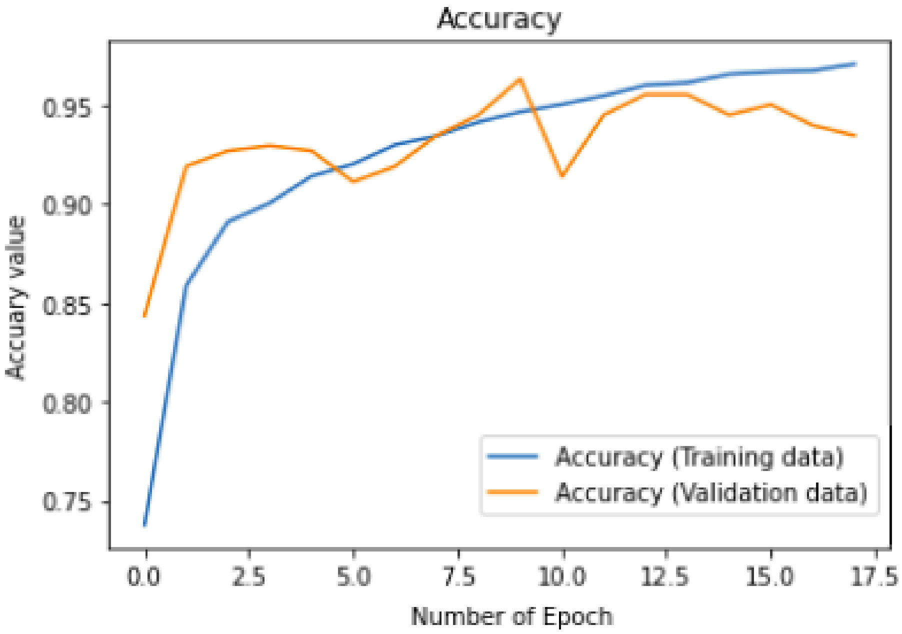


Figure 4.3: VGG 16: Accuracy and validation Accuracy:(0.000001=Adam LR)

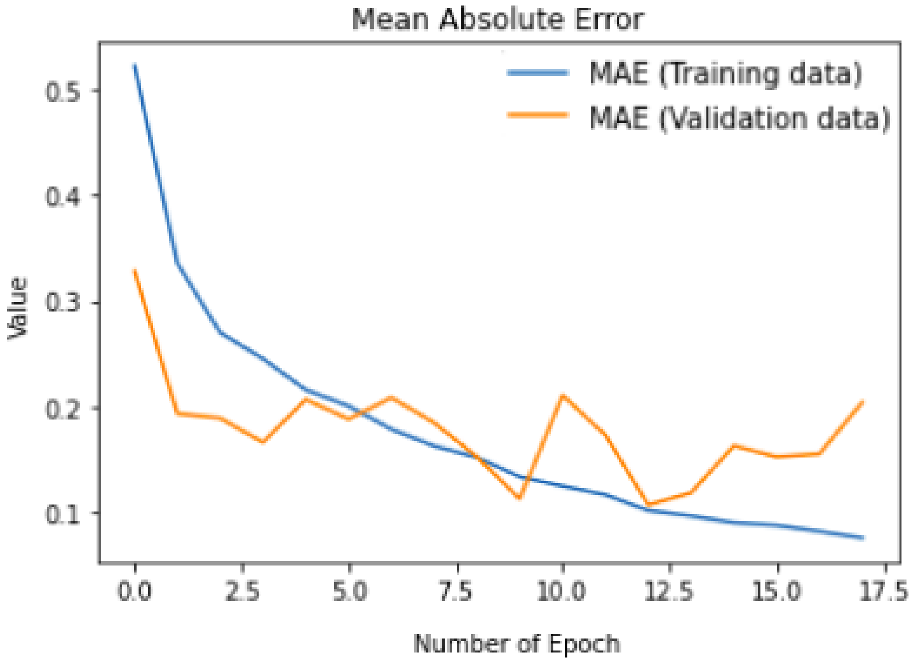


Figure 4.4: VGG 16: MAE and validation MAE:(0.000001=Adam LR)

4.1.2.3.1 Conclusion It has been observed that after tenth epoch,the validation accuracy was decreasing and training accuracy was increasing. On other hand after tenth epoch, the validation loss starts increasing and training loss is further reducing. Hence we can conclude that after tenth epoch, the model is moving towards over-fitting.

Metric	Value
Testing Accuracy	0.9469
Validation Accuracy	0.9635
Testing MAE	0.1339
Validation MAE	0.1132

4.1.2.4 ResNet-50

ResNet-50 is a pre-trained convolutional neural network with a depth of 50 layers and 25,636,712 parameters. This deep learning model is used for image classification and has weights trained on ImageNet.The input layer is followed by a 50 layer deep network, ResNet-50.

The network takes 150 X 150 X 3 images as input and yields a flattened output of length 2048. Weights of last 2 convolutional layers has been trained to adapt the weights entirely during the training of the model. The result from ResNet-50 is fed into 4 layers deep artificial neural network.First layer of following ANN is Pooling layer which is followed by Dense layer of size 256. The dense layer is followed by a dropout layer which is further followed up by Dense layer of size 2.

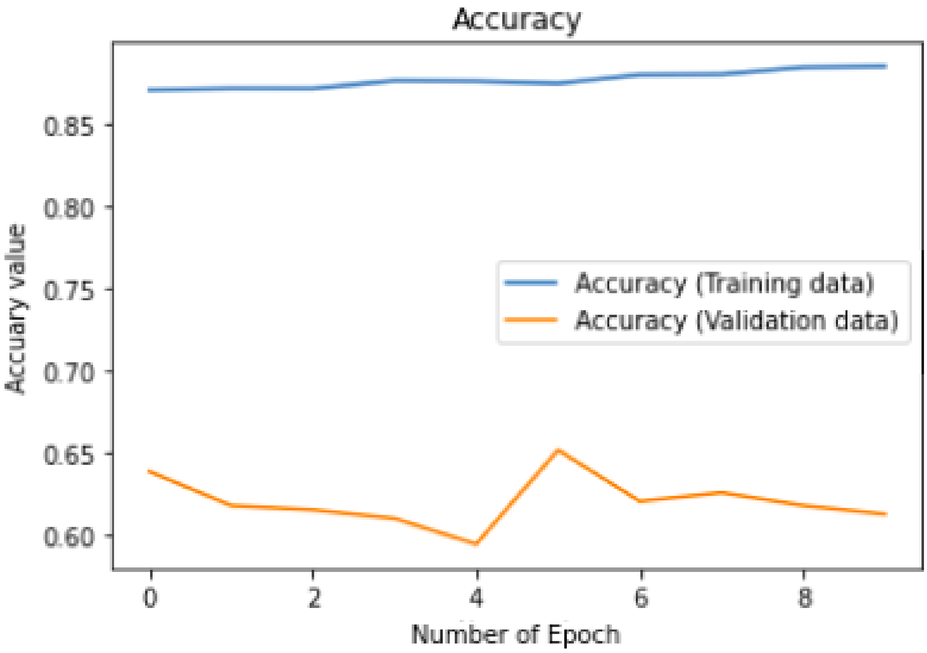


Figure 4.5: ResNet 50: Accuracy and validation Accuracy:(0.000001=Adam LR)

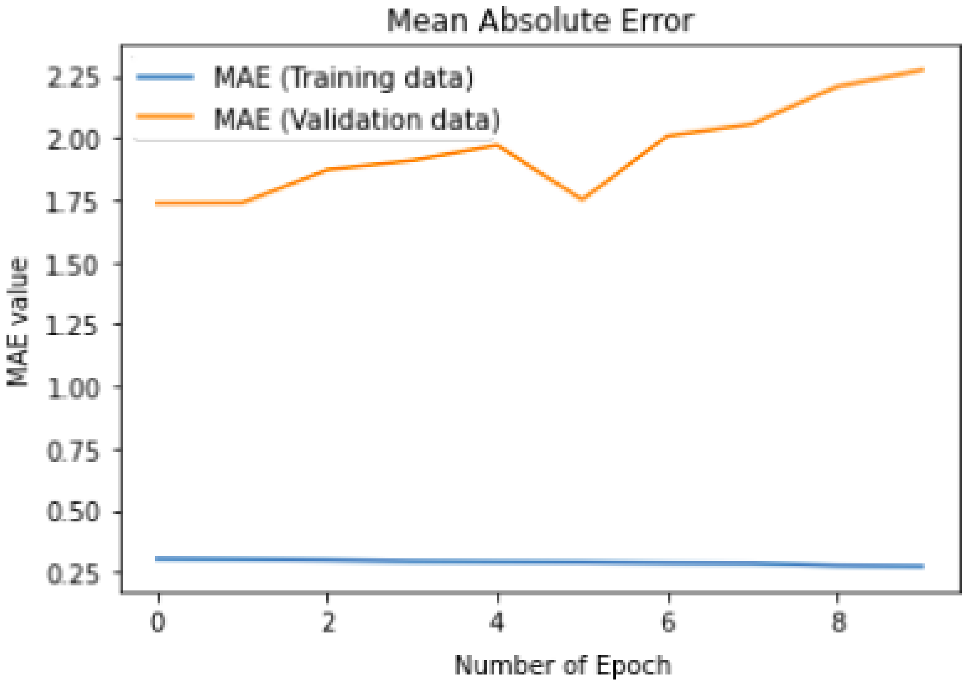


Figure 4.6: ResNet 50: MAE and validation MAE:(0.000001=Adam LR)

4.1.2.4.1 Conclusion It has been observed that after sixth epoch,the validation accuracy was decreasing and training accuracy was increasing. On other hand after sixth epoch, the validation loss starts increasing and training loss is further reducing. Hence we can conclude that after sixth epoch, the model is moving towards overfitting.

Metric	Value
Testing Accuracy	0.8528
Validation Accuracy	0.6510
Testing MAE	0.3354
Validation MAE	1.4740

4.1.2.5 Comparative Study

Model	Number of Epochs	Accuracy (%)
Custom CNN	15	87
VGG 16	10	96.35
ResNet 50	6	65.10

4.2 Person Identification

In this problem, we try to identify the person from their ear image and also study the relationship between number of classes(persons) and accuracy. **We also identify the person from ear images of Males only and Females only to study the role of gender in model performance and whether dataset is biased towards any particular gender.**

4.2.1 Convolutional Neural Network

Convolutional Neural Network comprises of a network of nodes/neurons, to learn the weights and biases, employing convolution as the linear operation. It can be understood as replacing matrix multiplication in a neural network with convolution in atleast one of its layers. The neural network receives pixel points as the input. The input layer provides the attribute values (pixel points) to the subsequent layers. Each node in the subsequent layer learns the weights based on the input from the previous layer and gives the output as the activated result of the learned weights. The network tries to extract features of the image frame to develop embeddings for the same. These embeddings are then pushed to a classifier to assign them the target class labels.

The Custom CNN used in this research uses 70:30 train-test data. 64X64X3 (RGB) images were given as the input. The first layer after the input layer(first hidden layer) consists of 32 filters each of size 3 X 3 X 3, keeping stride as 1 and padding same. We get 28 parameters per filter. The next layer, Convolution layer consists of 64 filters each of size 3 X 3 X 3, keeping stride as 1 and padding same. The next layer is Max Pooling with kernel size 3 giving a convolved image of size 21 X 21 X 3. The next 2 layers, Convolution layers consists of 64 filters each of size 3 X 3

X 3, keeping stride as 1 and padding same. This is followed by Max Pooling with kernel size 3 giving convolved image of 7 X 7 X 3. The succeeding 2 layers are again, Convolution layers with 64 filters and 128 filters respectively, each of size 3 X 3 X 3, keeping stride as 1 and padding same. Max Pooling layer with kernel size 3 giving convolved output of size 2 X 2 X 3 follows. The next layer consists of 128 filters, each of size 3 X 3 X 3, keeping stride as 1 and padding same. The flattened output of 512 dimensions is fed into the next layer i.e. Dense layer having output dimension as 256 followed by Dense Layer with output dimension 128. The final layer is Dense Layer having output dimension equal to the number of classes(Persons).

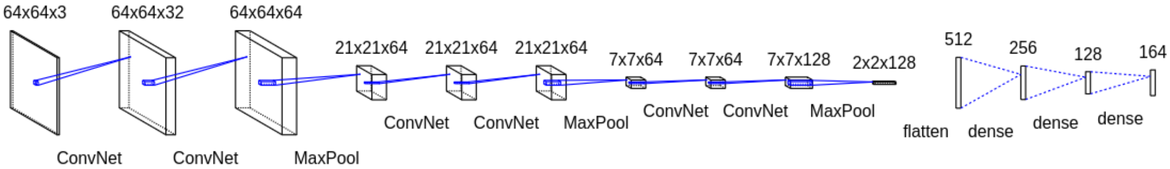


Fig: Proposed CNN Architecture for person identification

No. of Persons	Accuracy (%)
10	71.36
20	61.12
40	54.61
80	51.20
164	50.42

Table : Custom CNN Number of Persons v/s Accuracy

No. of Persons	No. of Epochs	Batch Size	Steps	Accuracy (%)
10	20	64	200	71.36
20	20	64	200	61.12
40	30	64	200	54.61
80	50	64	500	51.20
164	60	64	712	50.42

Table : Custom CNN Results

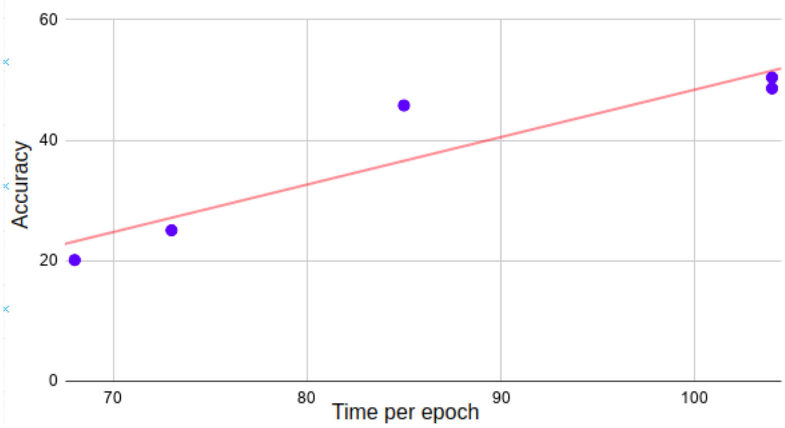


Fig: Accuracy and Time/epoch for CNN

Accuracy and time per epoch are observed to be almost linaerly related

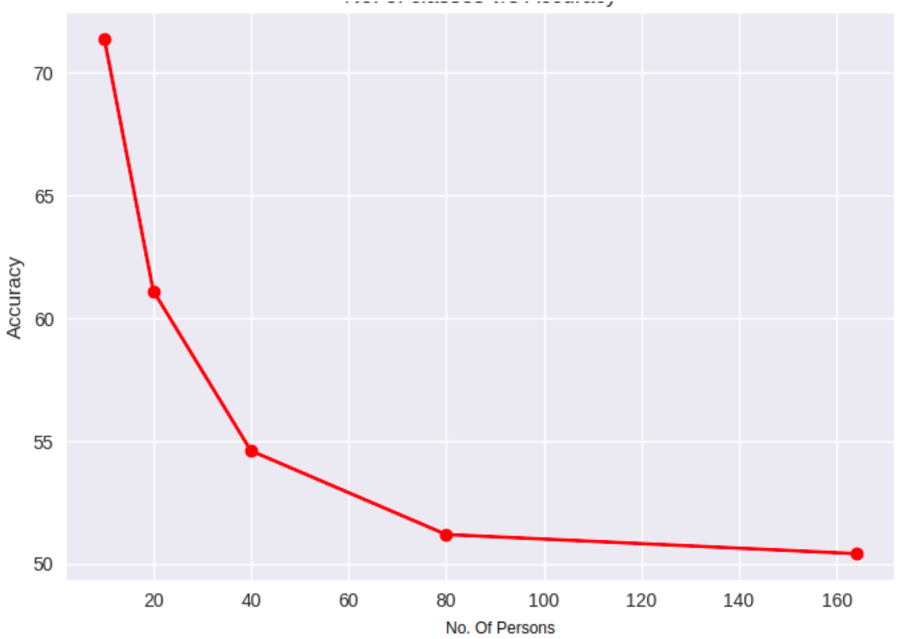


Fig: No. of Persons v/s Accuracy for CNN

It is observed that as the number of classes(Persons) increase, the model performance(Accuracy) decreases exponentially

No. of Persons	No. of Epochs	Batch Size	Steps	Males only Accuracy	Females only Accuracy(%)
10	20	64	200	69.8	72.4
20	20	64	200	60.16	62.18
40	30	64	200	53.84	5 4.92
80	50	64	500	50.78	-

Table :Custom CNN Number of Persons v/s Accuracy for individual Genders

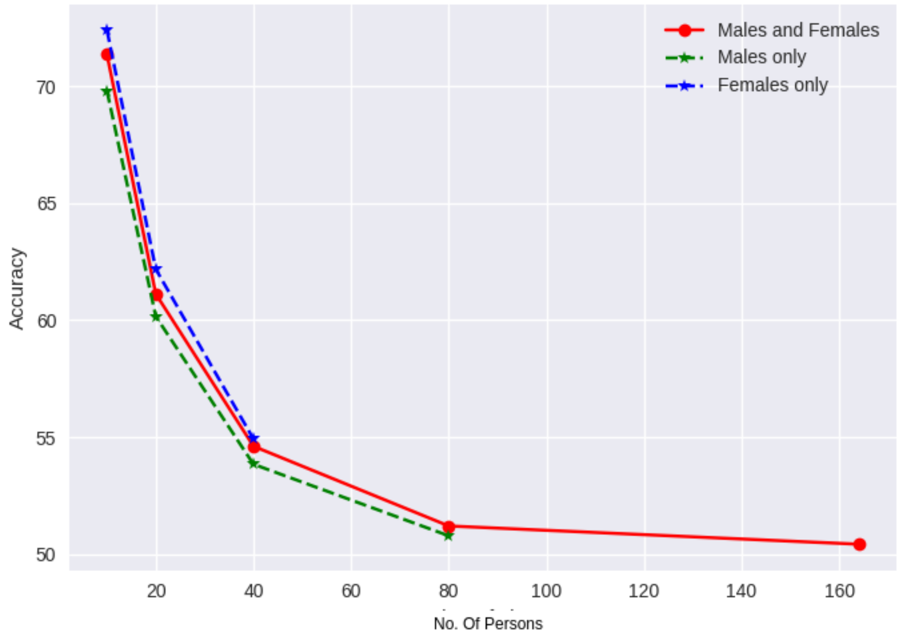


Fig: No. of Persons v/s Accuracy for CNN (Males Only and Females Only)

The accuracy for males and females individually is comparable to both males and females together. Hence data is not observed to be biased.

The accuracy for males and females individually is comparable to both males and females together. Hence, no role of gender in model performance is observed.

4.2.2 Transfer Learning

Transfer Learning is a technique where knowledge gained by model while it is being trained on one of the previous task is stored and is utilized as a starting point to

solve a new similar task.The knowledge gained can be reused partly or wholly to accelerate the training and improve the performance of a model on the problem of interest.
For person identification, we use pre-trained weights and only draw inference from Transfer Learning.

4.2.2.1 ResNet50

ResNet-50 is a pre-trained convolutional neural network with a depth of 50 layers and 25,636,712 parameters. This deep learning model is used for image classification and has weights trained on ImageNet. We incorporate the concept of transfer learning, wherein the weights obtained by training on one dataset are used for subsequent learning.

The input layer is followed by 50 layer deep network, ResNet-50. Since we use ResNet for feature extraction, the last layer(classifier) of the network is excluded. This network takes 64 X 64 X 3 images as input and yields a flattened output of length 2048. This is fed to our 4 layers deep artificial neural network consisting of Dense layers with output dimensions as: 512,256,128,No. of classes.

No. of Persons	No. of Epochs	Batch Size	Steps	Accuracy (%)
10	10	64	100	58.97
20	10	64	100	47.87
40	10	64	100	38.01
80	10	64	161	27.92
164	20	64	500	25.06

Table : ResNet50 Results

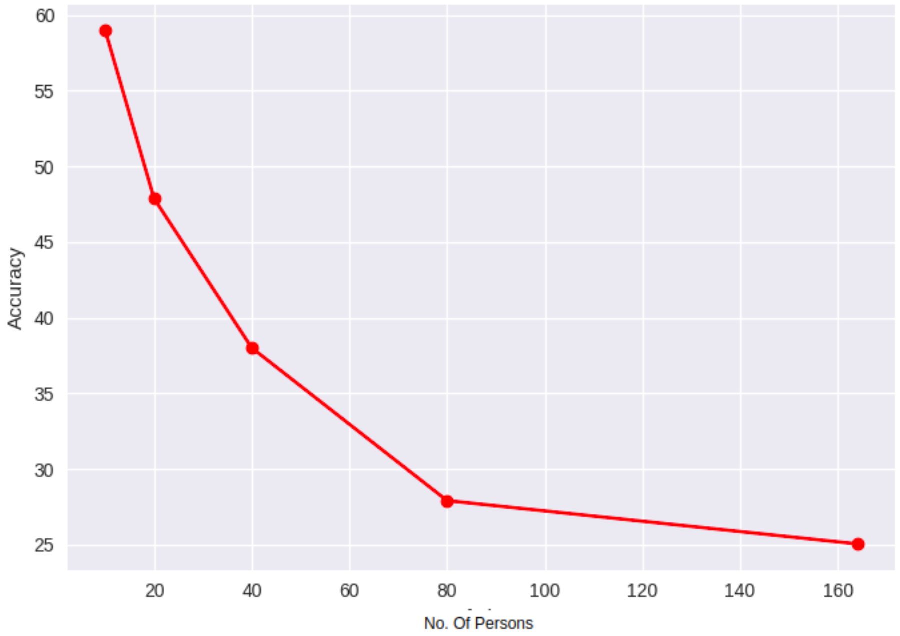


Fig: No. of Persons v/s Accuracy for ResNet50

It is observed that as the number of classes(Persons) increase, the model performance(Accuracy) decreases exponentially

4.2.2.2 VGG 19

VGG19 is a pre-trained convolutional neural network with a depth of 26 layers and 143,667,240 parameters. This deep learning model is used for image classification and has weights trained on ImageNet. We incorporate the concept of transfer learning, wherein the weights obtained by training on one dataset are used for subsequent learning. The input layer is followed by 26 layer deep network, VGG19. Since we use VGG19 for feature extraction, the last layer(classifier) of the network is excluded. This network takes 64 X 64 X 3 images as input and yields a flattened output of length 512. This is fed to our 4 layers deep artificial neural network consisting of Dense layers with output dimensions as: 512,256,128, No. of classes.

No. of Persons	Accuracy (%)
10	71.36
20	61.12
40	54.61
80	51.20
164	50.42

Table : VGG19 Results

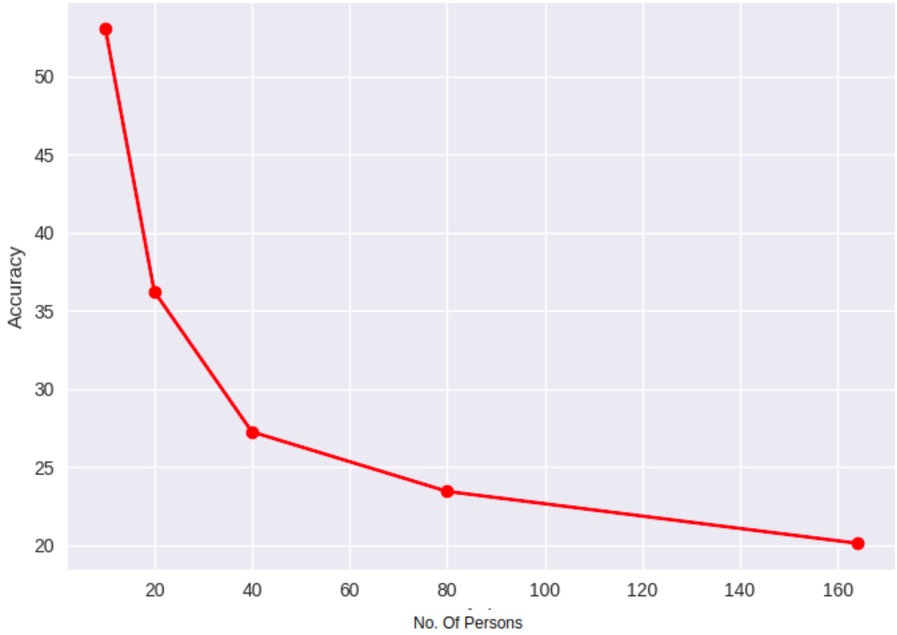


Fig: No. of Persons v/s Accuracy for VGG19

It is observed that as the number of classes(Persons) increase, the model performance(Accuracy) decreases exponentially

4.2.2.3 Conclusion

We have successfully conducted Person Identification from Ear Images. The relationship between accuracy and the number of classes(Persons) for ear recognition/ person identification has been studied successfully - Accuracy decreases as the number of classes increases. The dataset is not observed to be biased towards any particular gender as the model performance is observed to be almost same for both genders individually and when both considered together.

Chapter 5

Future Work Scope

Currently, in our research, we have conducted supervised analysis based on the provided labels in the dataset. We carried out two major tasks namely, Gender and Person Identification.

Furthermore, as a future work, we propose experimentation on right/left ear detection or recognition, which might be used to solve the problem of whether left-ear image can be matched with the right-ear image for person identification. This can be considered under the area of semi-supervised/unsupervised learning.

Chapter 6

Bibliography

1. Data-set:https://data.mendeley.com/datasets/yws3v3mwx3/4

2. https://www.sciencedirect.com/science/article/pii/S2352340919309850