Assignment 4 Report

Data Mining

CSE 572

Spring 2018

Submitted to:

Prof. Ayan Banerjee

Ira A. Fulton School of Engineering

Arizona State University

Submitted by:

Lakshmi Sneha Kandukuri (lkanduku@asu.edu**)**
Manasa Pola (mpola@asu.edu**)**
Himaja Tirumalasetti (htirumal@asu.edu**)**
Hitesh Kumar Sannithi (hsannithi@asu.edu**)**
Kushal Reddy Papakannu (kpapakan@asu.edu**)**

May 2, 2018

Table of Contents

1. Introduction

The assignment is to perform **user independent analysis** using classification techniques like decision trees, support vector machines and neural networks on gesture data collected using the sensors. We consider our dataset as highly complex and containing imprecise and uncertain data. The data set is preprocessed for creating a new feature matrix by using feature extraction techniques. Then various classification techniques like neural networks, support vector machines and decision trees are applied on the new feature matrix. The results of Accuracy, Precision, Recall and F1 measure from each classification technique are used to determine the best technique suited for the dataset to identify the gestures.

2. Team Members

Following are the group member of this project

Lakshmi Sneha Kandukuri (lkanduku@asu.edu)

Manasa Pola (mpola@asu.edu)

Himaja Tirumalasetti (htirumal@asu.edu)

Hitesh Kumar Sannithi (hsannithi@asu.edu)

Kushal Reddy Papakannu (kpapakan@asu.edu)

3. Data Preprocessing

The assignment 1 is to collect the sensor data. All the team members should go to Impact lab for collecting the data. Any one of the team member should volunteer for gesturing in front of the screen wearing 2 wristbands, one on right and one on left hand. The gestures are identified by the sensors using the hand movement of the person wearing the wristbands. The gestures are captured using 4 sensors Gyroscope, Accelerator, EMG Sensor, Orientation. According to movement of the hand, each sensor captures the data, for example, if the gesture consists of rotations, orientation sensor data is useful to identify the gesture.

In this task we are considering the raw data from 10 users i.e. group DM29, DM31, DM32, DM33, DM34, DM35, DM02, DM03, DM04, DM05. Each group data consists of 200 CSV files where

each person performed 10 gestures 20 times. Each Action generates data from 34 sensors across 45 time series. This data is processed in the following way:

- Numerical data of one action of one gesture is transposed leading to a matrix of size 34X45. All the remaining actions of this gesture are transposed and appended below one another creating a matrix of size 680X45.
- Similarly, the above task is performed for all the gestures and appending all of them one below the another creating an input matrix for one group of dimensions 6800X45 i.e 680*10X45.
- The above procedure is repeated for 10 groups and 10 csv files created. Data of all the 10 files are merged into a single file “TrainData.xls” with dimensions of 68000X42.
- The data of remaining 27 users is merged into a single file “TestData.xls”
- TrainData.xls and TestData.xls files are merged into a single file task4data.xlsx

Refer DM_Task4.m file for the MATLAB code.

4. Creation of New Feature Matrix

- For each gesture we have identified the dominant features which showed maximum variance. The user specific data for each sensor of all the gestures is extracted, which creates a matrix of dimensions 2000x45.
- We have applied FFT for the above matrix, which gives us a matrix of dimensions 2000X4. We have transposed the above matrix (i.e. 4X2000) and stored in a temporary matrix. Similarly, we have constructed the transposed matrices for all the sensors of one gesture and stacked in temporary matrix creating a matrix dimensions of 20X2000. The temporary matrix is finally re-transposed creating a final matrix of dimensions 2000X20 let’s say it is StackedTransposeMatrix.
- Principle component analysis PCA is applied to StackedTransposeMatrix which returns a coefficient matrix (20x20).
- We have multiplied StackedTransposeMatrix with coefficient matrix of PCA creating a new feature matrix of dimensions 2000x20.

5. Classification Techniques

5.1 Terminology

Confusion Matrix:

A clean and unambiguous way to present the prediction results of a classifier is to use a confusion matrix.

True Positives (TP) - These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes.

True Negatives (TN) - These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no.

False Positives (FP) – When actual class is no and predicted class is yes.

False Negatives (FN) – When actual class is yes but predicted class in no.

Using these four parameters we are calculating Accuracy, Precision, Recall and F1 score.

Accuracy:

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same. Therefore, you can look at other parameters to evaluate the performance of your model.

Precision:

Precision is the number of True Positives divided by the sum of number of True Positives and False Positives. Put another way, it is the number of positive predictions divided by the total number of positive class values predicted.

Precision can be thought of as a measure of a classifiers exactness. A low precision can also indicate large number of False Positives.

Recall:

Recall is the number of True Positives divided by the sum of number of True Positives and False Negatives. Put another way it is the number of positive predictions divided by the

number of positive class values in the test data. It is also called Sensitivity or the True Positive Rate.

Recall can be thought of as a measure of a classifiers completeness. A low recall indicates many False Negatives.

F1 Score:

The F1 Score is the 2*((precision*recall)/ (precision + recall)). It is also called the F Score or the F Measure. Put another way, the F1 score conveys the balance between the precision and the recall.

5.2 Decision Trees

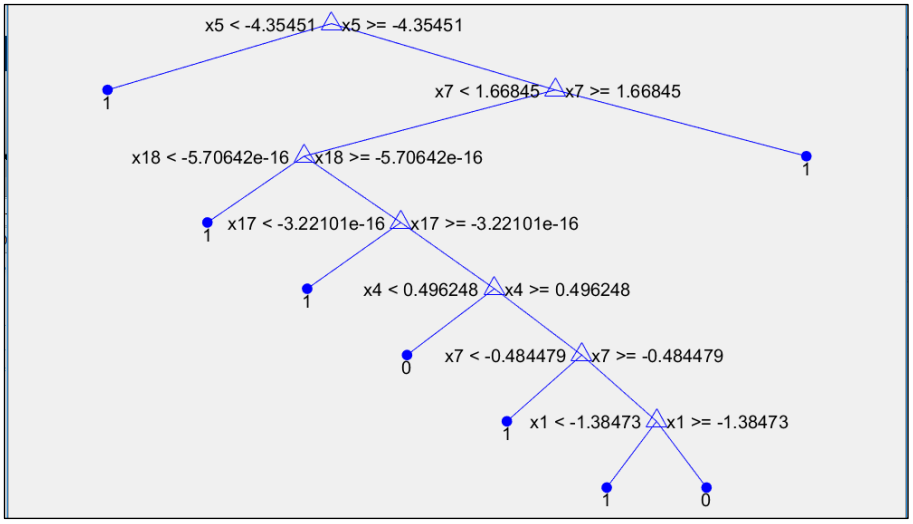
Decision tree methodology is a commonly used data mining method for establishing classification systems based on multiple covariates or for developing prediction algorithms for a target variable. This method classifies data into branch-like segments that construct an inverted tree with a root node, internal nodes, and leaf nodes. The approach is nonparametric and can efficiently deal with large, complicated datasets without imposing a complicated parametric structure. When the sample size is large enough, dataset can be divided into training and test datasets. Using the training dataset to build a decision tree model and a test dataset to decide on the appropriate class label.

For the above dataset we have considered 60% of the data to be training dataset and 40% of the data to be test data.

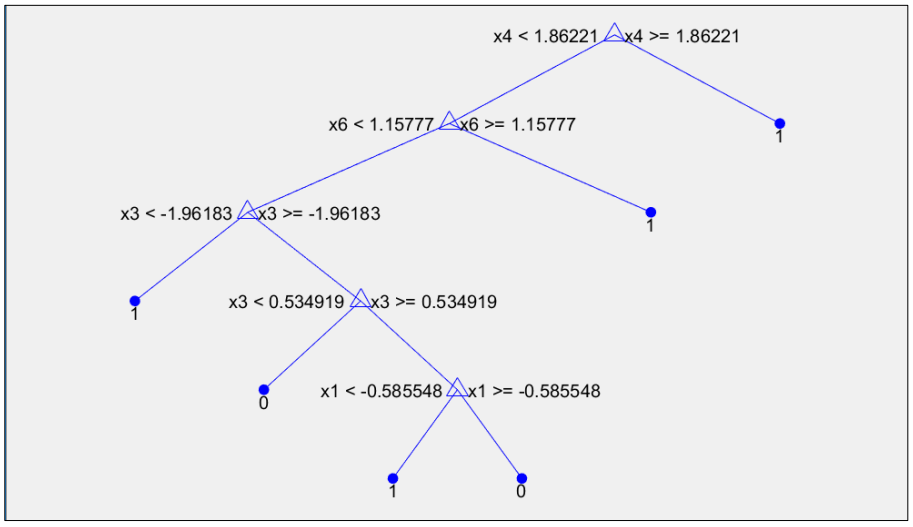
Below tables provides the values of Accuracy, Precision, Recall and F1 Score of different groups obtained using Decision Tree.

Gesture	Accuracy	Precision	Recall	F1 Score
About	0.98	0.32	0.23	0.3421
And	0.98125	0.18182	0.25	0.21053
Can	0.99	0.253	0.34	0.4966
Cop	0.98875	0.4	0.25	0.30769
Deaf	0.985	0.25	0.25	0.25
Decide	0.98	0.1	0.125	0.11111
Father	0.9825	0.321	0.125	0.2187
Find	0.99	0.256	0.32	0.342
Go out	0.985	0.16667	0.125	0.14286
Hearing	0.9825	0.143	0	0.32

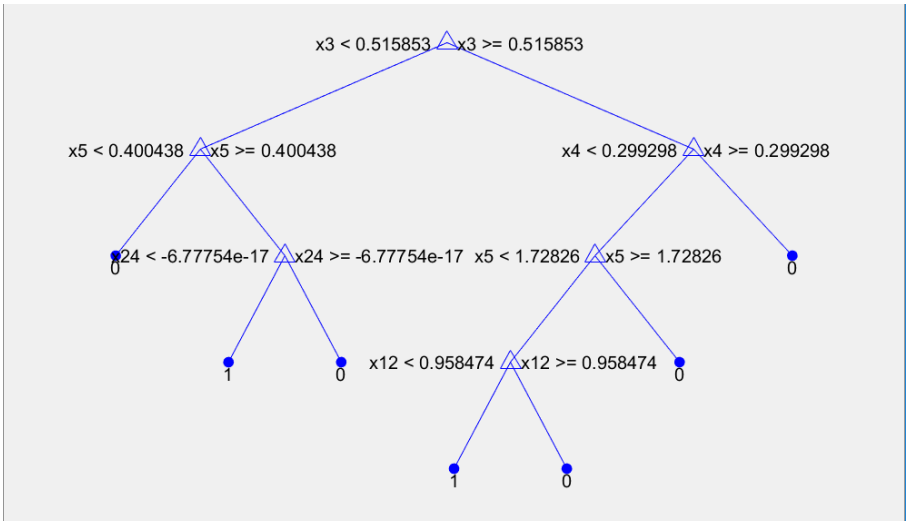
ABOUT



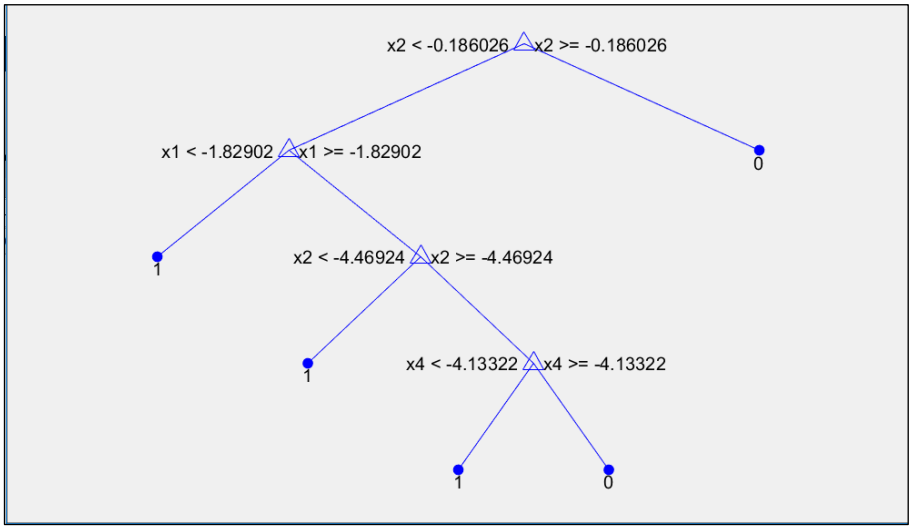
AND



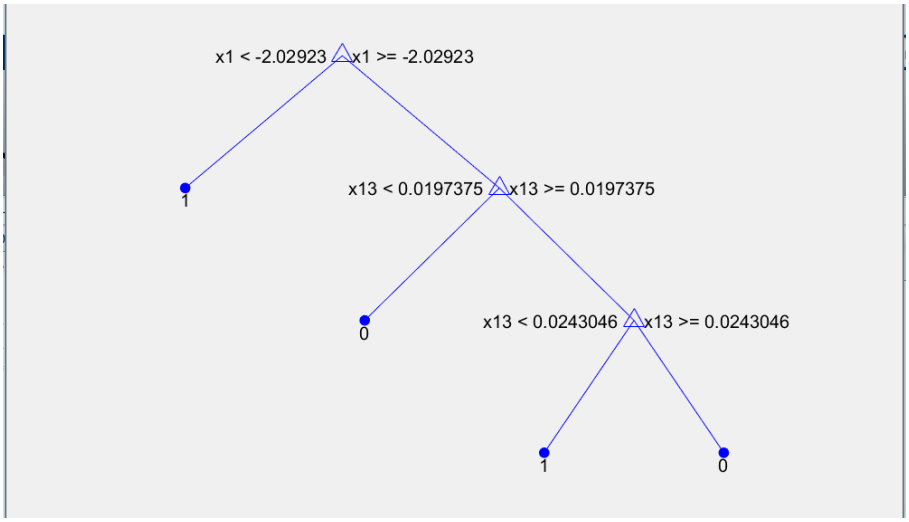
CAN


x3 < 0.515853
x3 >= 0.515853
x5 < 0.400438
x5 >= 0.400438
x4 < 0.299298
x4 >= 0.299298
x24 < -6.77754e-17
x24 >= -6.77754e-17
x5 < 1.72826
x5 >= 1.72826
0
0
1
0
x12 < 0.958474
x12 >= 0.958474
0
1
0

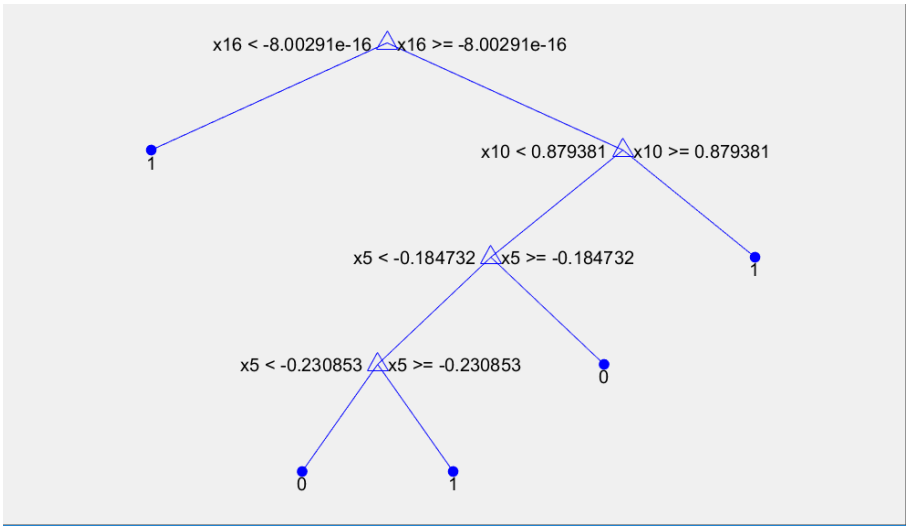
COP


x2 < -0.186026
x2 >= -0.186026
x1 < -1.82902
x1 >= -1.82902
0
1
x2 < -4.46924
x2 >= -4.46924
1
x4 < -4.13322
x4 >= -4.13322
1
0

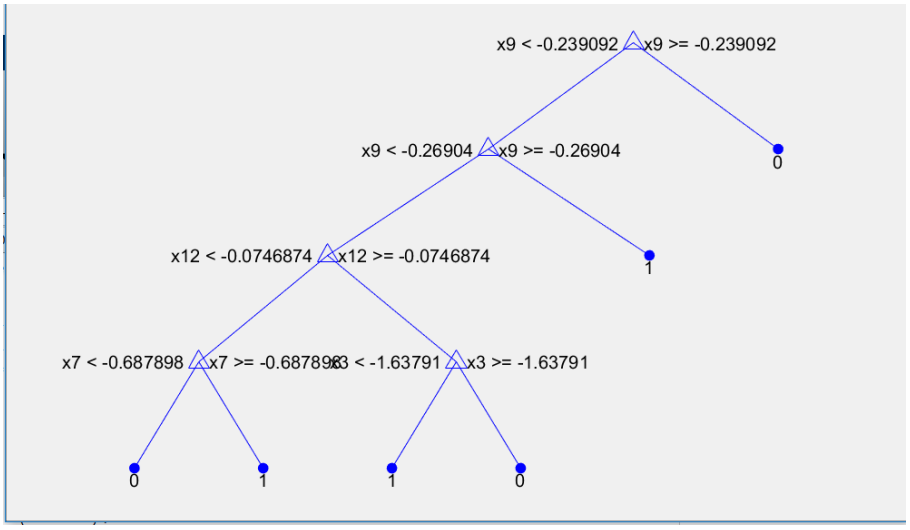
DEAF



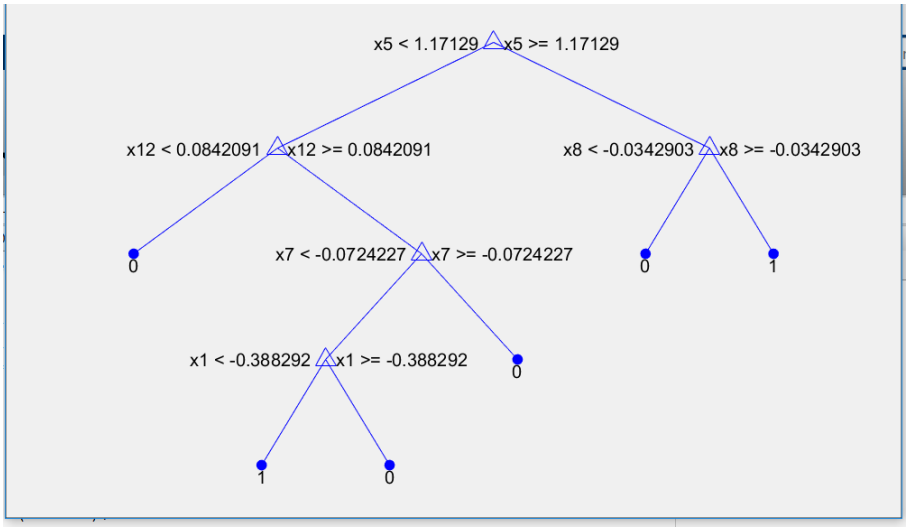
DECIDE



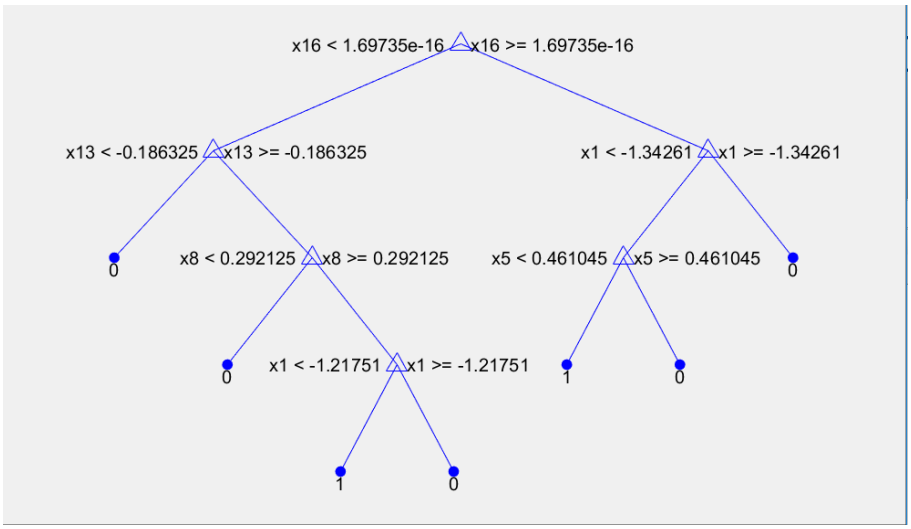
FATHER



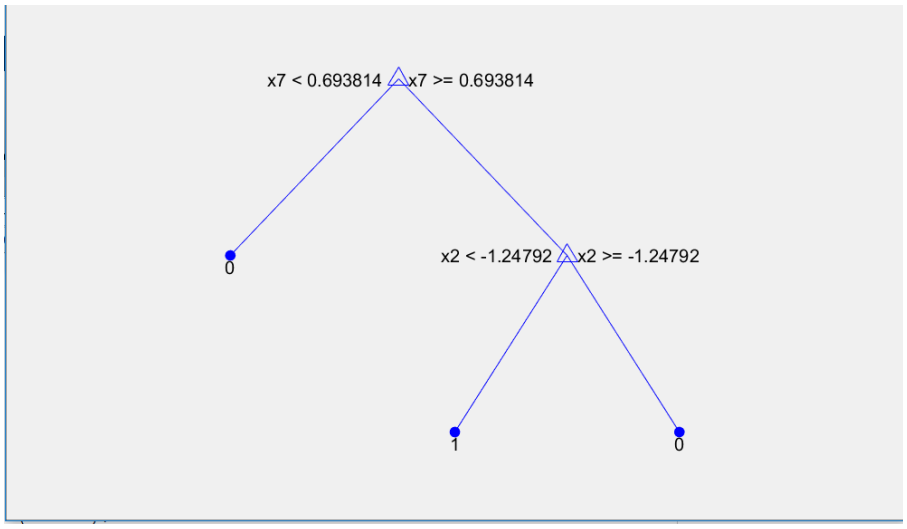
FIND



GO OUT



HEARING



Support Vector machines(SVM) are supervised learning models with associated learning algorithms that analyze data used for classification analysis. Given a set of training examples an SVM training algorithm builds a model that assigns new examples, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

Below tables provides the values of Accuracy, Precision, Recall and F1 Score of different groups obtained using SVM.

Gesture	Accuracy	Precision	Recall	F1 Score
About	0.9	0.46	0.24	0.45
And	0.8875	0.12	0.16	0.57
Can	0.95	0.83333	0.625	0.71429
Cop	0.9	0.471	0.125	0.1975
Deaf	0.9	0.18	0.8	0.2938
Decide	0.9	0.654	0.678	0.6657
Father	0.9	0.49	0.12	0.42
Find	0.9	0.57	0.25	0.51
Go out	0.9	0.43	0.17	0.53
Hearing	0.9375	0.61538	0.13	0.7619

5.4 Neural Networks

Pattern recognition is the process of training a neural network to assign the correct target classes to a set of input patterns. Once trained the network can be used to classify patterns it has not seen before. This dataset can be used to design a neural network that classifies the action as one of the recognized Gesture.

The standard network that is used for pattern recognition is a two-layer feedforward network, with a sigmoid transfer function in the hidden layer, and a softmax transfer function in the output layer.

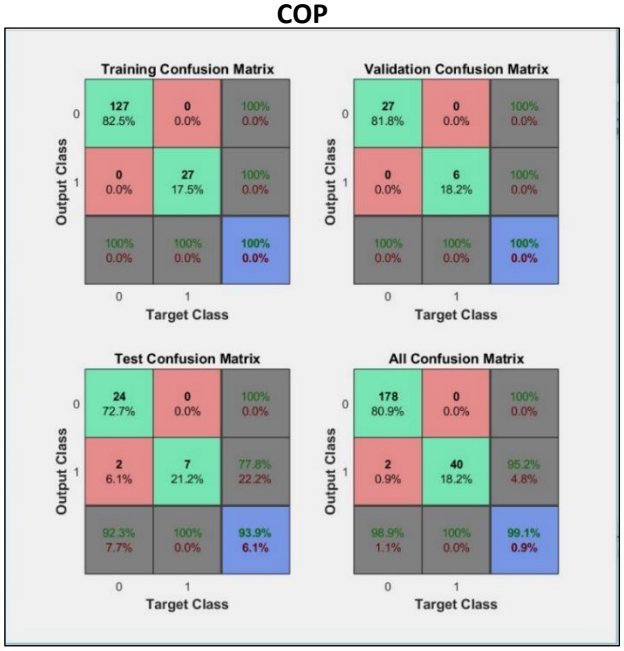
Below sample images gives the values of Confusion Matrix, Accuracy, Precision, Recall and F1 Score of one group obtained using neural networks.

ABOUT

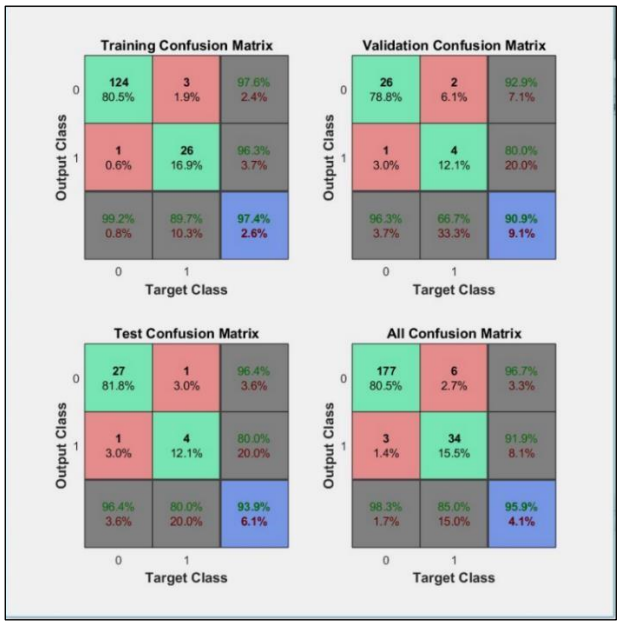
AND

CAN

COP

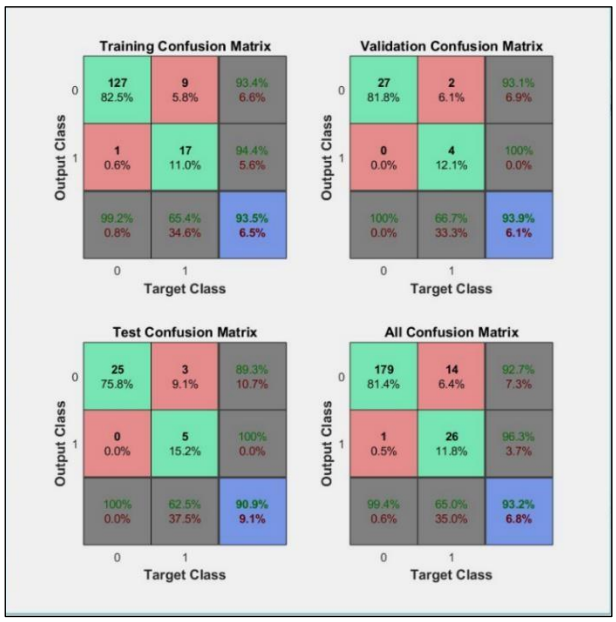


DEAF



DECIDE

FATHER



FIND

GOOUT

HEARING

Below tables provides the values of Accuracy, Precision, Recall and F1 Score of different groups obtained using Neural Networks.

Gesture	Accuracy	Precision	Recall	F1 Score
About	0.643	0.49	0.67	0.566034
And	0.628	0.73	0.64	0.682044
Can	0.654	0.67	0.55	0.604098
Cop	0.767	0.35	0.43	0.385897
Deaf	0.723	0.56	0.73	0.633798
Decide	0.769	0.44	0.36	0.396
Father	0.878	0.69	0.61	0.647538
Find	0.634	0.56	0.42	0.48
Go out	0.843	0.79	0.68	0.730884
Hearing	0.712	0.85	0.82	0.834731

ROC CURVE:

A Receiver Operating Characteristic (ROC) Curve is a way to compare diagnostic tests. It is a plot of the true positive rate against the false positive rate.

A ROC plot shows:

- The relationship between sensitivity and specificity. For example, a decrease in sensitivity results in an increase in specificity.
- Test accuracy; the closer the graph is to the top and left-hand borders, the more accurate the test. Likewise, the closer the graph to the diagonal, the less accurate the test. A perfect test would go straight from zero up the top-left corner and then straight across the horizontal.
- The likelihood ratio; given by the derivative at any cutpoint.

Test accuracy is also shown as the area under the curve (which you can calculate using integral calculus). The greater the area under the curve, the more accurate the test. A perfect test has an area under the ROC curve (AUROCC) of 1. The diagonal line in a ROC curve represents perfect chance.

6. Summary

We can observe from the above data that accuracies for User Independent analysis were significantly higher than the accuracies for User Dependent Analysis. When a user performs wrong gesture in user dependent analysis it leads to discrepancies in the test data. But if a user performs wrong gesture, then it would be counterbalanced by the data from other users in User Independent Analysis. This leads to no influence of training model over the test data causing the accuracies to improve.