CN240 DATA SCI

DEPRESSION DATASET ANALYSIS

Department of Electrical and Computer Engineering, Faculty of Engineering
Thammasat School of Engineering | Thammasat University

Table of content

Problem statement

- Depression is a common illness worldwide[1], and rising every year from 2007 to 2018.[2]

Objectives

- Automated depression detect.
- Automated facial emotion detect.
- The model get at least 80% accuracy.

Methods

- Determine what features to use by using filter, wrapper ,embedded method.
- Detect and get landmark from the face using openCV.

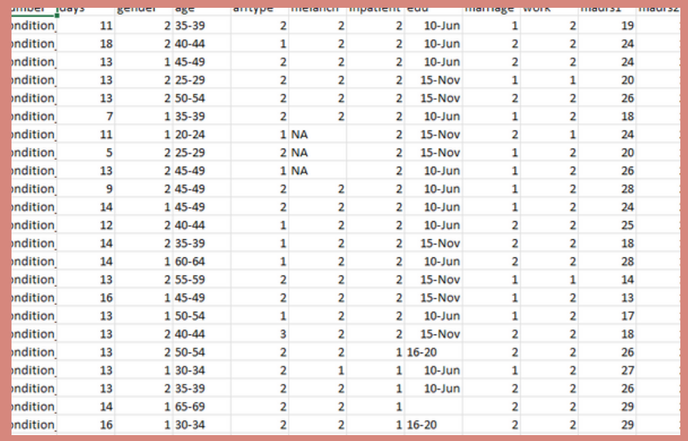
[1] Institute of Health Metrics and Evaluation, Global Health Data Exchange (GHDx), 2019. [Online]. Available:https://www.kaggle.com/arashnic/the-depression-dataset
[2] M. E Duffy and J. M Twenge and T. E Joiner, "Trends in mood and anxiety symptoms and suicide-related outcomes among U.S. undergraduates, Journal of Adolescent Health," United State of America. Accessed: Jul. 3, 2019.[Online]. Available:https://pubmed.ncbi.nlm.nih.gov/31279724/

Pre-processing

We add 18 data to the control and 27 data to the condition. Make a total of 100 data people divided into condition 50 and control 50 people.

The new data we inserted has the following fields:
1.edu 2.marriage 3.age 4.gender 5.depression or non depression.

We use these fields to help filling our old data with null.



ondition	11	2	35-39	2	2	2	10-Jun	1	2	19	
ondition	18	2	40-44	1	2	2	10-Jun	2	2	24	
ondition	13	1	45-49	2	2	2	10-Jun	2	2	24	
ondition	13	2	25-29	2	2	2	15-Nov	1	1	20	
ondition	13	2	50-54	2	2	2	15-Nov	2	2	26	
ondition	7	1	35-39	2	2	2	10-Jun	1	2	18	
ondition	11	1	20-24	1	NA	2	15-Nov	2	1	24	
ondition	5	2	25-29	2	NA	2	15-Nov	1	2	20	
ondition	13	2	45-49	1	NA	2	10-Jun	1	2	26	
ondition	9	2	45-49	2	2	2	10-Jun	1	2	28	
ondition	14	1	45-49	2	2	2	10-Jun	1	2	24	
ondition	12	2	40-44	1	2	2	10-Jun	2	2	25	
ondition	14	2	35-39	1	2	2	15-Nov	2	2	18	
ondition	14	1	60-64	1	2	2	10-Jun	2	2	28	
ondition	13	2	55-59	2	2	2	15-Nov	1	1	14	
ondition	16	1	45-49	2	2	2	15-Nov	1	2	13	
ondition	13	1	50-54	1	2	2	10-Jun	1	2	17	
ondition	13	2	40-44	3	2	2	15-Nov	2	2	18	
ondition	13	2	50-54	2	2	1	16-20	2	2	26	
ondition	13	1	30-34	2	1	1	10-Jun	1	2	27	
ondition	13	2	35-39	2	2	1	10-Jun	2	2	26	
ondition	14	1	65-69	2	2	1		2	2	29	
ondition	16	1	30-34	2	2	1	16-20	2	2	29	

data csv [1]

Not Depressed	Male	73
Not Depressed	Female	18
Not Depressed	Male	19
Not Depressed	Male	21
Not Depressed	Female	85
Not Depressed	Male	79
Not Depressed	Female	59

new data csv [2]

How do we fill missing value?

afftype&inpatient

afftype
1: bipolar II
2: unipolar
3: bipolar I

inpatient
1: inpatient
2: outpatient

NA
NA
NA
NA

from afftype and inpatient
which is a missing value
in both control and condition
We've added the value 0 instead
missing values

0
0
0
0

activity

From the activity values of both control and condition found in the data, the new 50 people we added have no activity values and have a Gaussian distribution
We add the average instead
1:condition_female 2:condition_male
3:control_female 4:control_male

How do we fill missing value?

Madrs

0	4
0	6
0	4
0	6
0	4
0	6
0	3
0	0
0	2
0	6
0	6
0	3
0	0

1	18
1	28
0	14
1	13

We fill madrs score for control by fill numbers in the range 0-6 from paper[1] said about the patient in the normal range and We found paper[2] proof average madrs score in healthy controls to determine the normal range of values has 95% confidence interval and had S.D 3.5-4.5,
So we choose 95 % of control average madrs score is 3.5 and We have 5% of healthy controls but have madrs high score.

We fill madrs score for condition by a random number in the range 7-34 from paper[1] said about the patient is in the depression range

[1] A review of studies of the Montgomery-Asberg Depression Rating Scale in controls: implications for the definition of remission in treatment studies of depression, National Library of Medicine, 2004. [Online]. Available:https://pubmed.ncbi.nlm.nih.gov/15101563/

[2] Aripiprazole (Abilify): Depression, Major Depressive Disorder (MDD), NCIB, 2004. [Online]. Available:https://www.ncbi.nlm.nih.gov/books/NBK409740/#:~:text=The%20MADRS%20scoring%20instructions%20indicate,depression%2C%E2%80%9D%20and%20a%20total%20score

How do we fill missing value?

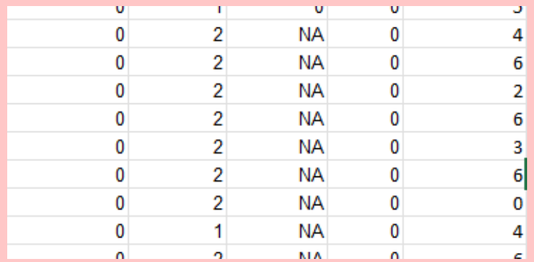
we use **mode** for these columns

EDU

0	..19-24	17
1	25-34	20
2	35-44	19
3	45-49	14
4	50-54	13
5	55-79	17

we collapse the age range and fill missing data with use mode by considering with Age and marriage

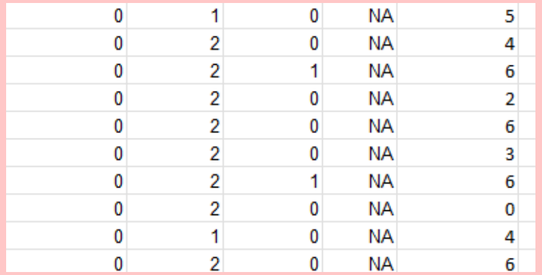
MARRIAGE



0	2	NA	0	4
0	2	NA	0	6
0	2	NA	0	2
0	2	NA	0	6
0	2	NA	0	3
0	2	NA	0	6
0	2	NA	0	0
0	1	NA	0	4

we fill missing data with use mode by considering with Age and work

WORK



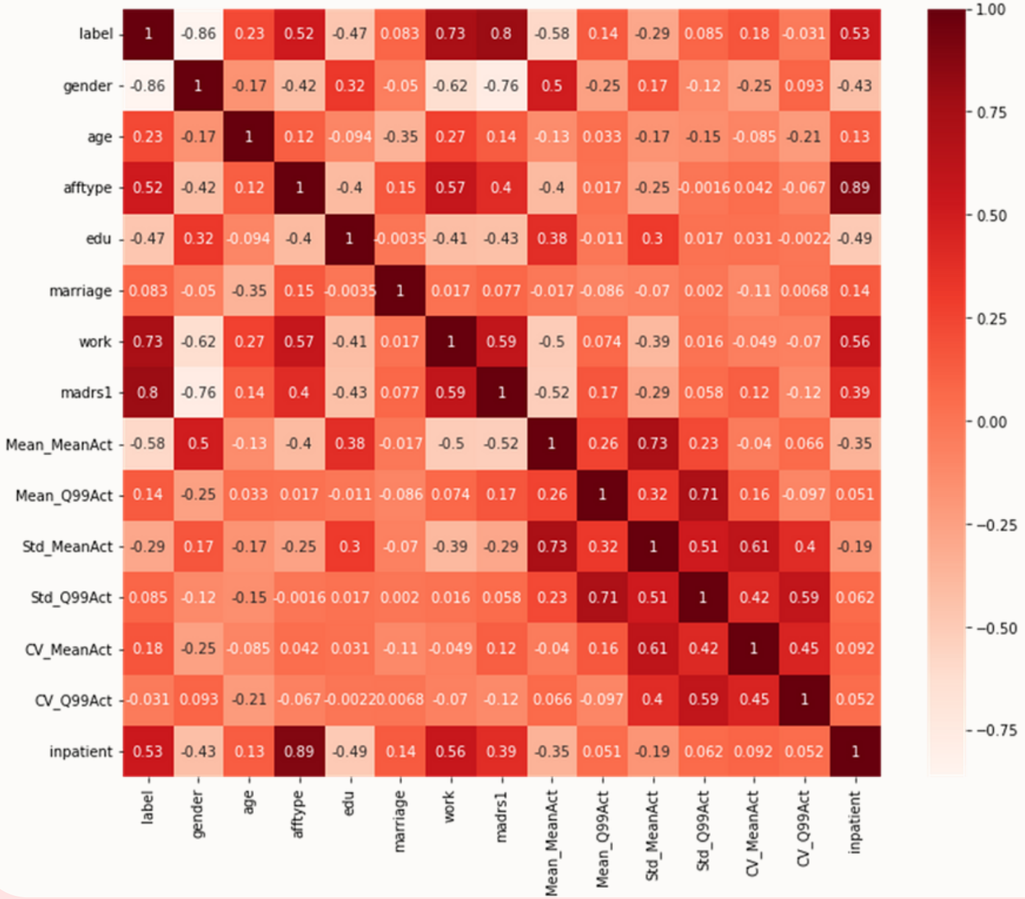
0	1	0	NA	5
0	2	0	NA	4
0	2	1	NA	6
0	2	0	NA	2
0	2	0	NA	6
0	2	0	NA	3
0	2	1	NA	6
0	2	0	NA	0
0	1	0	NA	4
0	2	0	NA	6

we fill missing data with use mode by considering with Age and edu

Feature Selection

Filter Method

Correlation Table



gender	0.864333
afftype	0.516398
work	0.733799
madrs1	0.803343
Mean_MeanAct	0.583698
inpatient	0.528472

We Choose the feature which have absolute of the correlation to the output more than 0.7

If it between 0.5-0.7 we check the correlation between each feature to eliminate some of the features
afftype and inpatient have the correlate of 0.89 so we choose only inpatient because inpatient is more correlate to the output.

Finally we get 5 features which are Mean_MeanAct, madrs1, inpatient, gender, work

Feature Selection with sklearn and Pandas, Towards Data Science, 2019. [Online]. Available:https://towardsdatascience.com/feature-selection-with-pandas-e3690ad8504b

Feature Selection

Wrapper Method

We use the **RFE (Recursive Feature Elimination)** in this method

We use **linear Regration model** on this method

Here are the top 7 features

Rank 1 is CV_Q99Act
Rank 2 is CV_MeanAct
Rank 3 is gender
Rank 4 is work
Rank 5 is afftype
Rank 6 is inpatient
Rank 7 is marriage

After We rank the features, we find the optimum number of features check by the score of each model from 1 to 14 features. Here are the optimum numbers of features and score

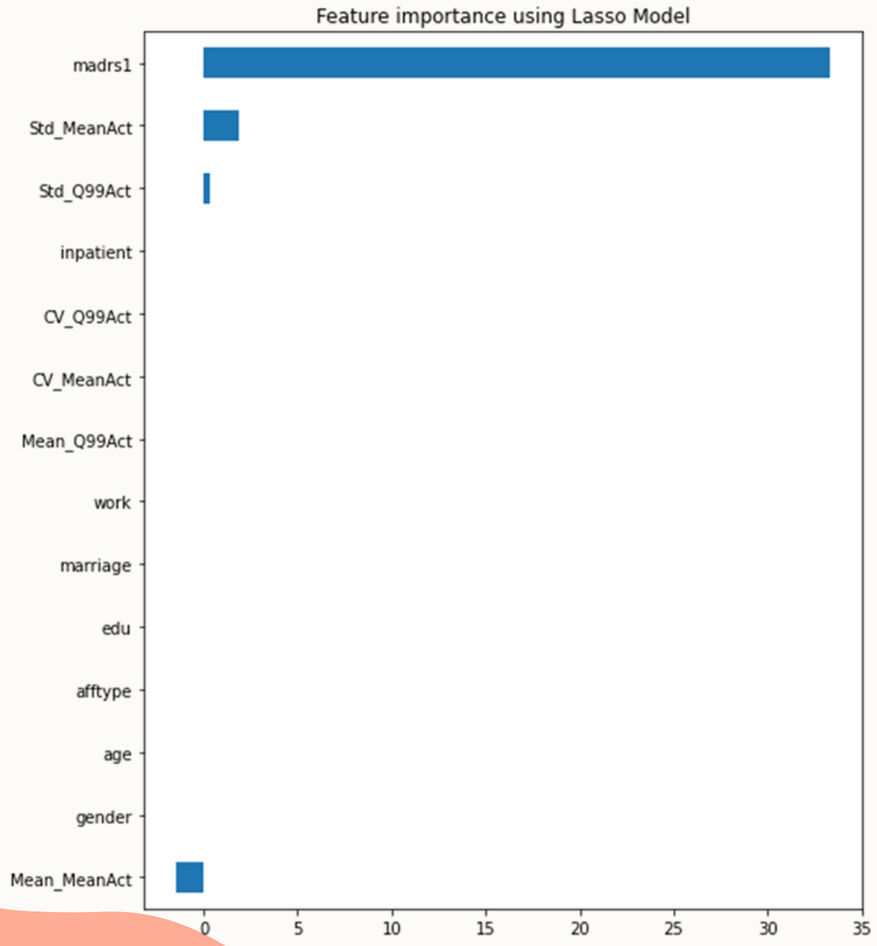
Optimum number of features: 4
Score with 4 features: 0.543370

Finally we got 4 features which are CV_MeanAct, work, CV_Q99Act, gender

Feature Selection with sklearn and Pandas, Towards Data Science, 2019. [Online]. Available:https://towardsdatascience.com/feature-selection-with-pandas-e3690ad8504b

Feature Selection

Embedded



We use LassoCV to pick and eliminate features

The model pick 4 features and eliminate other 9 features

Finally We got 4 features which are madrs1, Std_MeanAct, Std_Q99Act, Mean_MeanAct

Feature Selection with sklearn and Pandas, Towards Data Science, 2019. [Online]. Available:https://towardsdatascience.com/feature-selection-with-pandas-e3690ad8504b

Feature Selection

Days	Melanch	Madrs2

Based on our research, We found that melanch [1] had no effect on depression, and madrs2 was the post-treatment assessment. In the days part, our calculated activity values were averages. An average has been made using days to help determine the value. With that said, we have eliminated the information in these 3 parts.

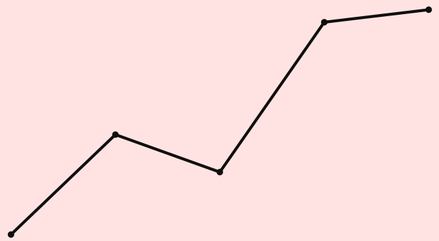
[1] Impaired implicit learning and reduced pre-supplementary motor cortex size in early-onset major depression with melancholic features, sciencedirect, 2009. [Online]. Available:https://www.sciencedirect.com/science/article/abs/pii/S0165032709001281

CSV DATA

Training : 64%	Validate : 16%	Test : 20 %

Preprocessing data

1st	FOLD 1	FOLD 1	FOLD 1	FOLD 1	FOLD 1
2nd	FOLD 2	FOLD 2	FOLD 2	FOLD 2	FOLD 2
3rd	FOLD 3	FOLD 3	FOLD 3	FOLD 3	FOLD 3
4th	FOLD 4	FOLD 4	FOLD 4	FOLD 4	FOLD 4
5th	FOLD 5	FOLD 5	FOLD 5	FOLD 5	FOLD 5



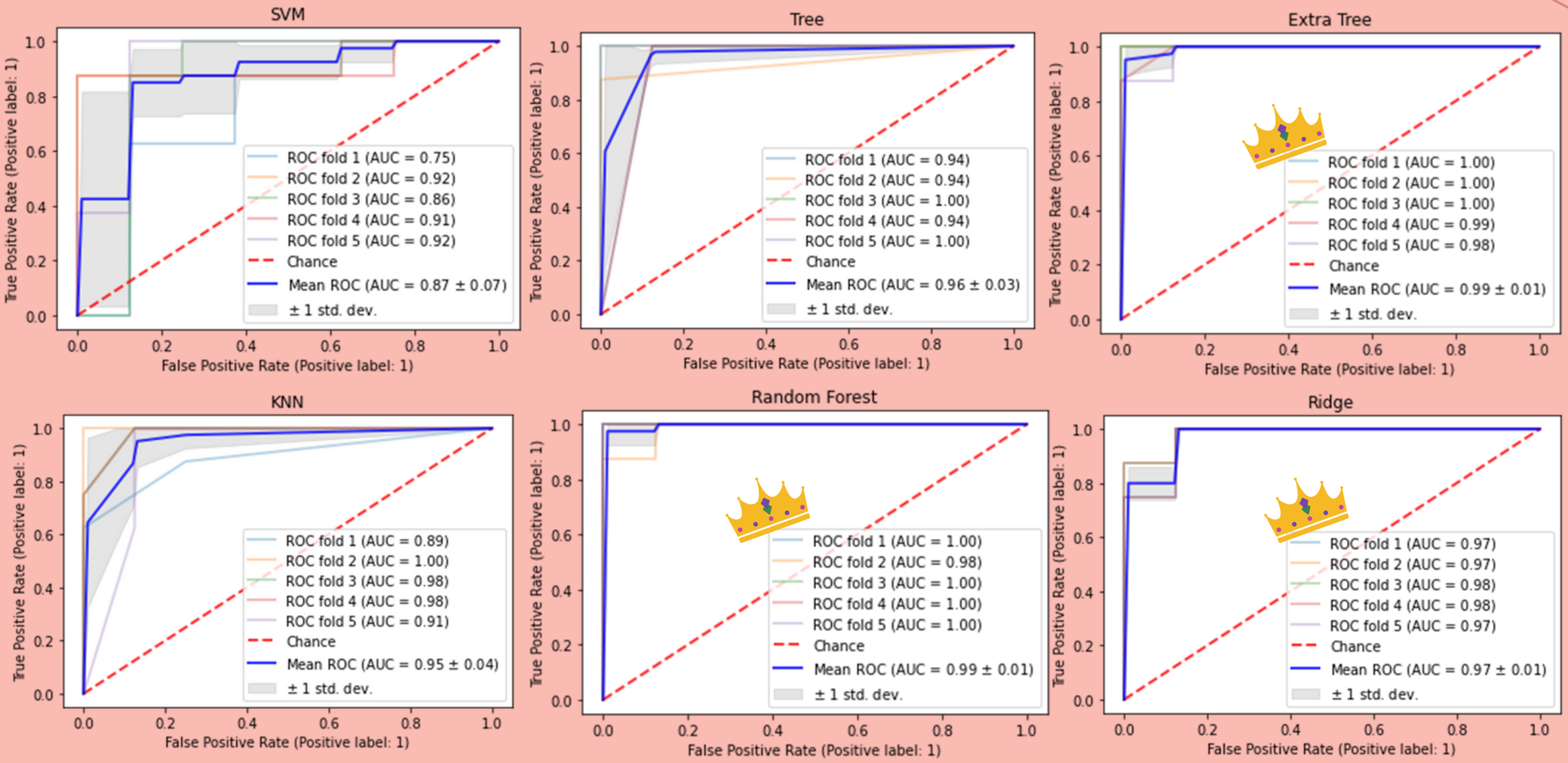
Plot ROC graph

The ratio between control and condition is the same for all FOLDs.

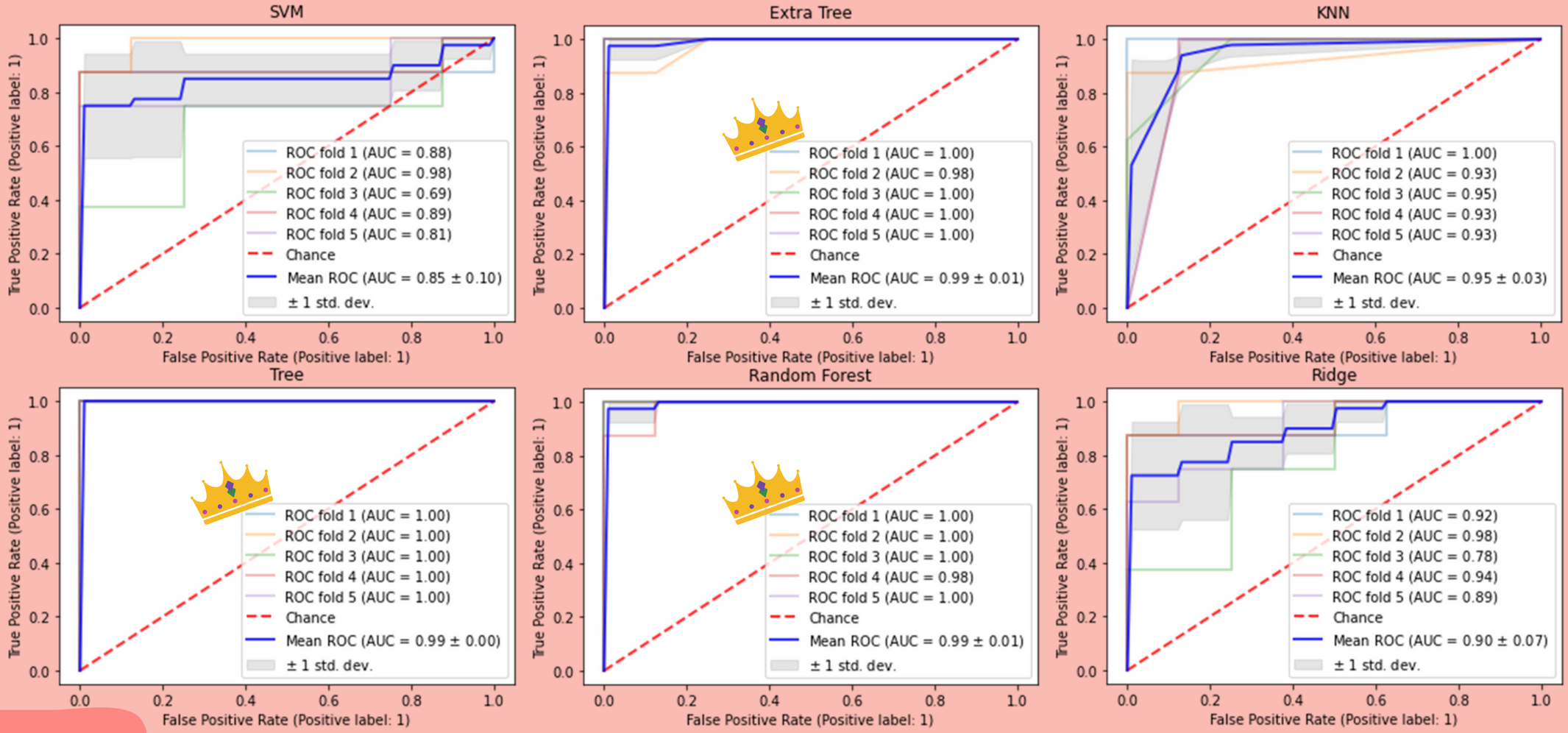
Performance

Accuracy
Precision
Sensitivity (Recall)
F1_Score
Specificity

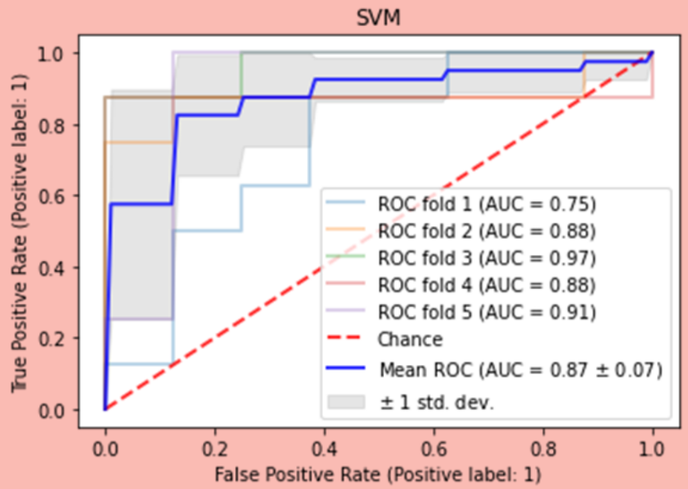
Choosing the best algorithm (Filter)

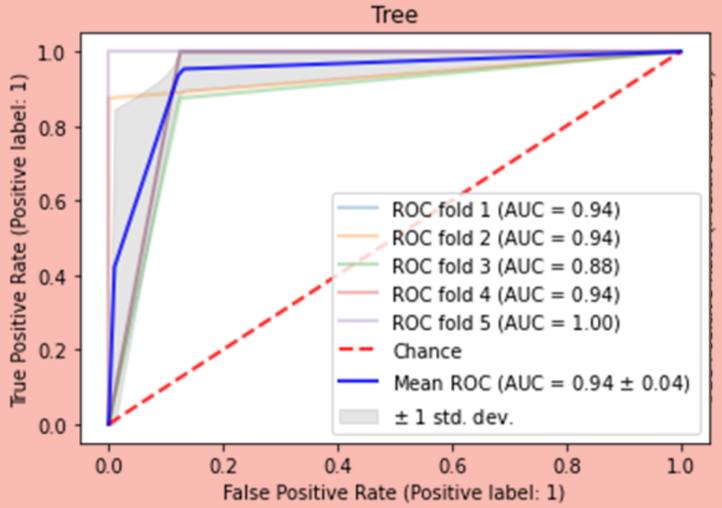


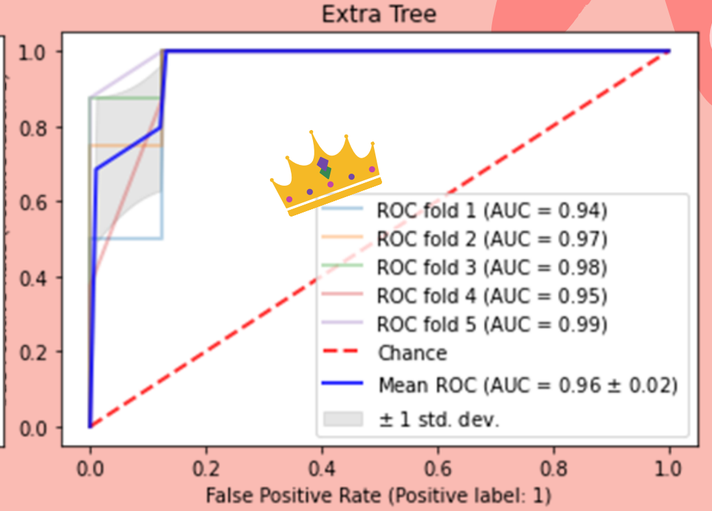
Choosing the best algorithm (Wrapper)

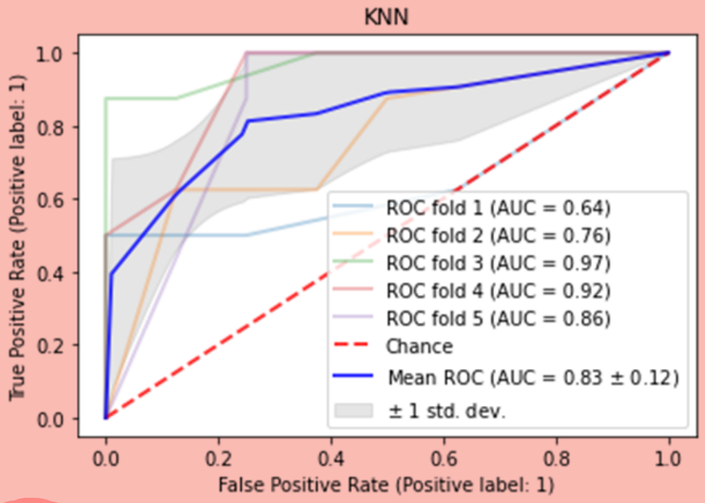


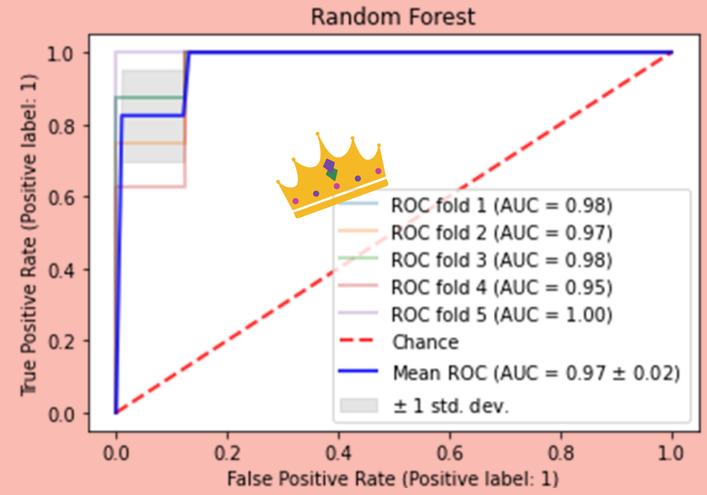
Choosing the best algorithm (Embeded)

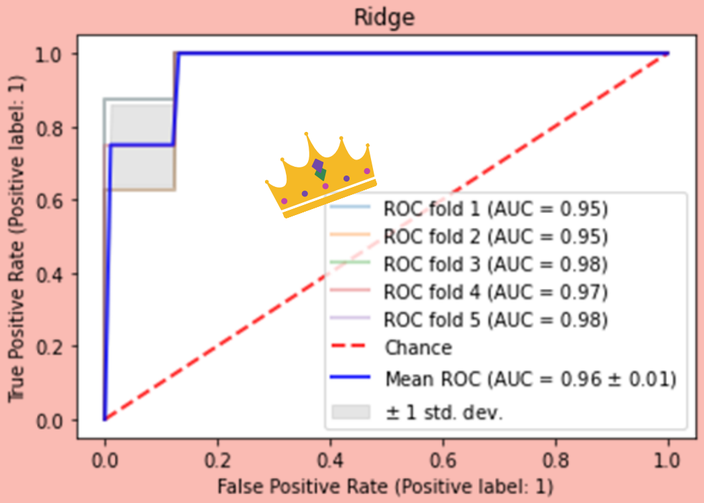






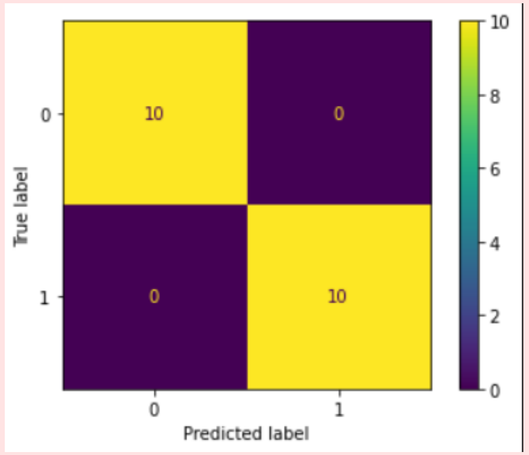






Filter

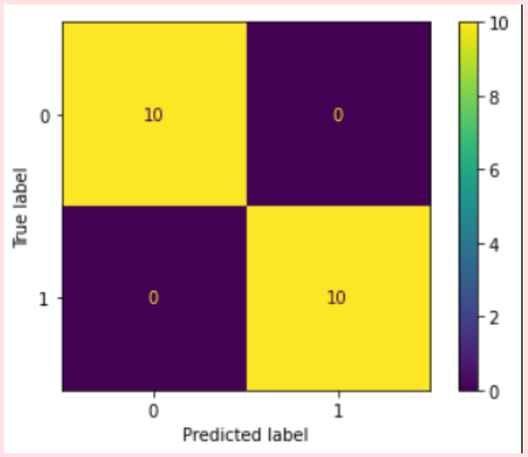
Extra Tree



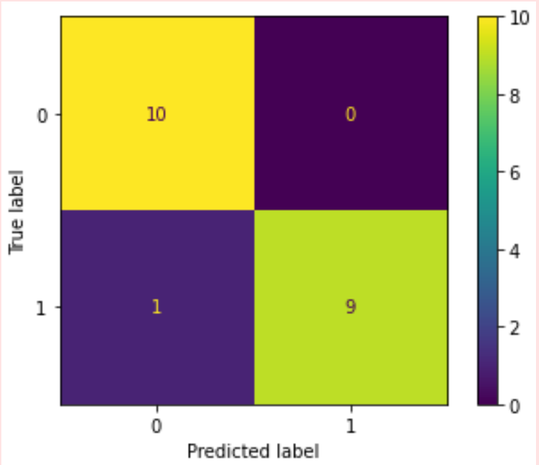
Accuracy : 100%
Precision : 100%
Recall : 100%
F1_Score : 100%
Specificity : 100%

Random Forest

Accuracy : 100%
Precision : 100%
Recall : 100%
F1_Score : 100%
Specificity : 100%



Ridge



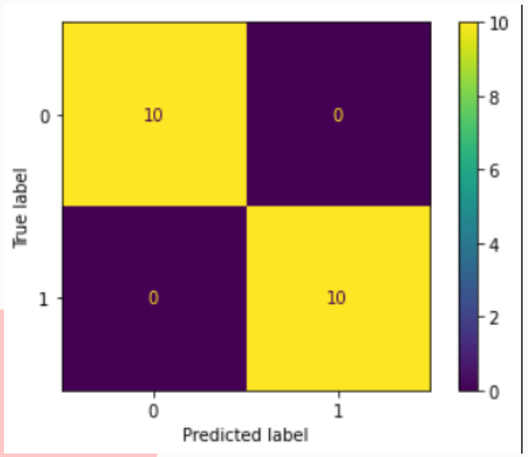
Accuracy : 95%
Precision : 100%
Recall : 90%
F1_Score : 94.73%
Specificity : 100%

1 is false negative.(similar to control)

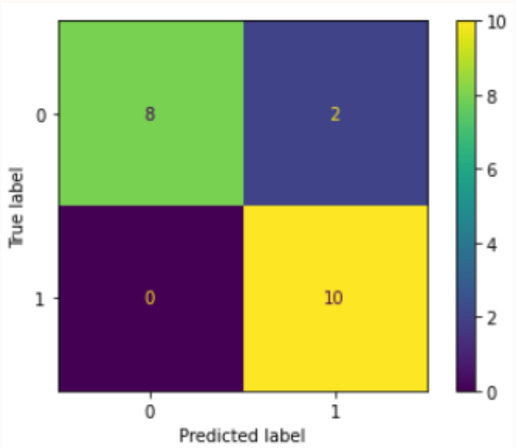
Wrapper

Extra Tree

Accuracy : 100%
Precision : 100%
Recall : 100%
F1_Score : 100%
Specificity : 100%



Random Forest



Accuracy : 90%
Precision : 83.3%
Recall : 100%
F1_Score : 90.91%
Specificity : 80%

4 is false positive.(similar to condition)

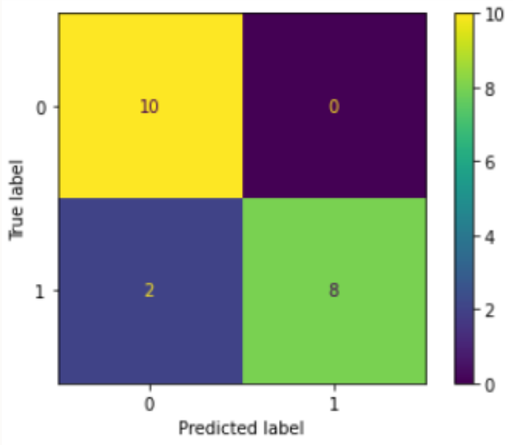
9 is false positive.(similar to condition)

Ridge

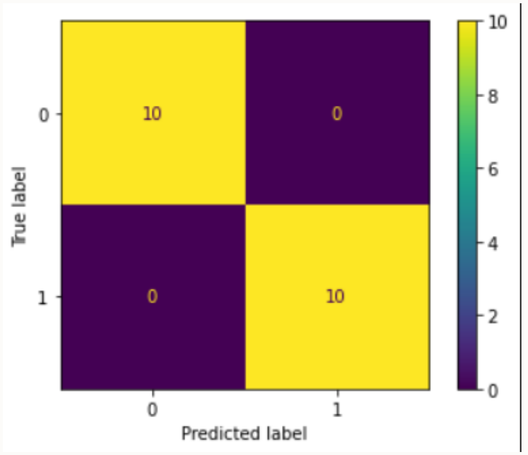
Accuracy : 90%
Precision : 100%
Recall : 80%
F1_Score : 88.89%
Specificity : 100%

1 is false negative.(similar to control)

16 is false negative.(similar to control)

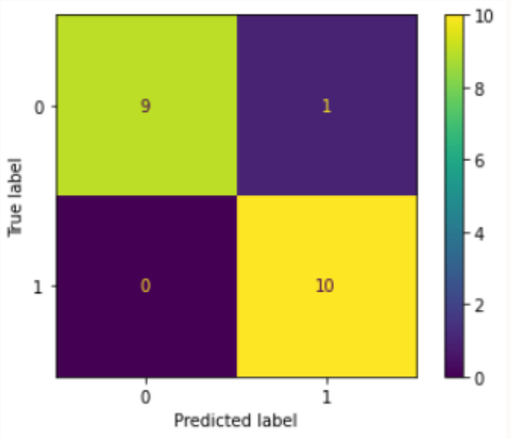


Embedded



Extra Tree

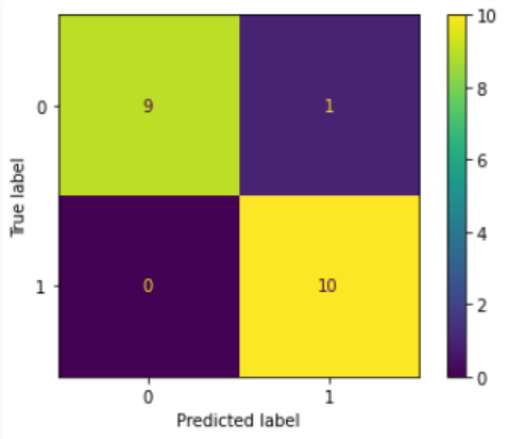
Accuracy : 100%
Precision : 100%
Recall : 100%
F1_Score : 100%
Specificity : 100%



Random Forest

Accuracy : 95.00%
Precision : 90.91%
Recall : 100 %
F1_Score : 95.24%
Specificity : 90.00%

2 is false positive.(similar to condition)

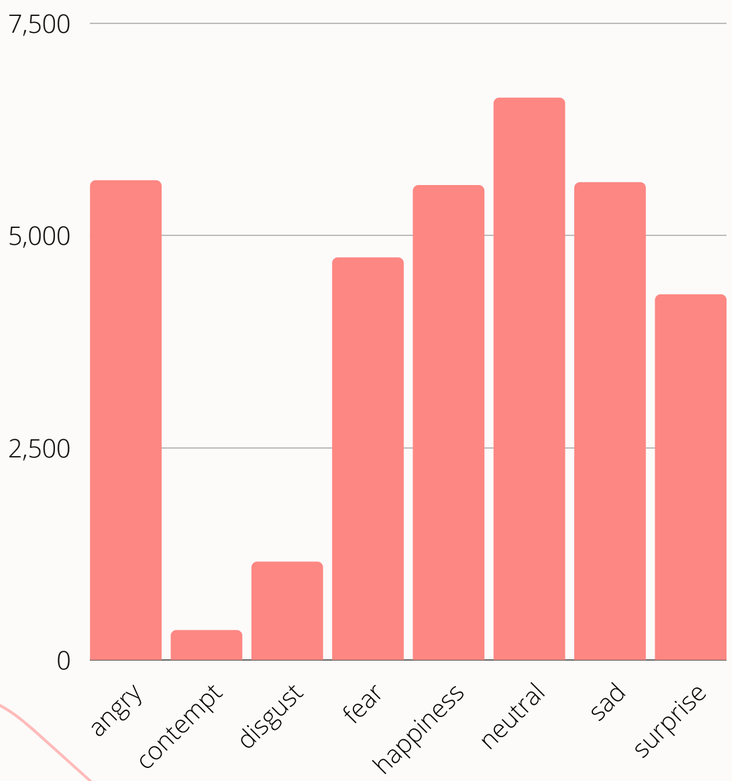


Ridge

Accuracy : 95.00%
Precision : 90.91%
Recall : 100 %
F1_Score : 95.24%
Specificity : 90.00%

2 is false positive.(similar to condition)

Image Data



Find more data from 2 databases

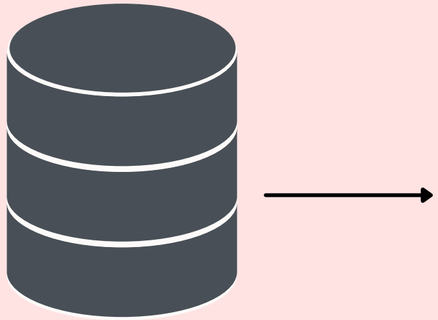
After we found the data we change their file extension to jpg and resize it to 256*256 pixel and relabel them.

We use algorithm to separate male and female, and algorithm that detect the same picture by using openCV

We separate each person to only be in the same set of data, and we find the ratio of male and female.

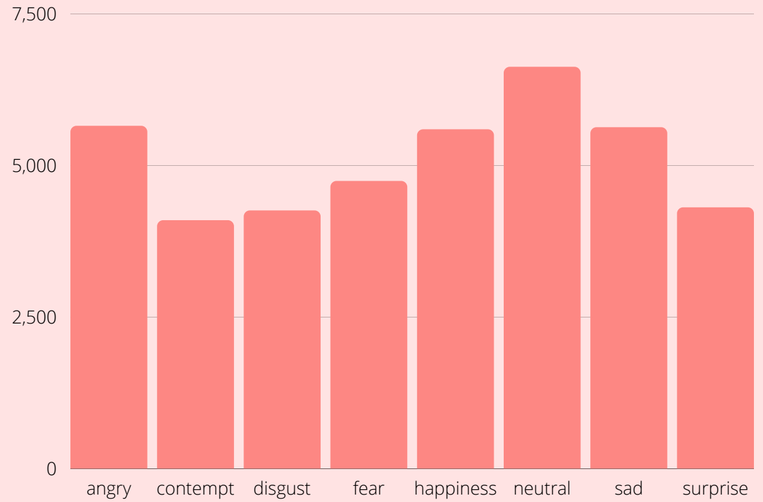
Image Data

Training : 64%	Validate : 16%	Test : 20 %



Preprocessing data

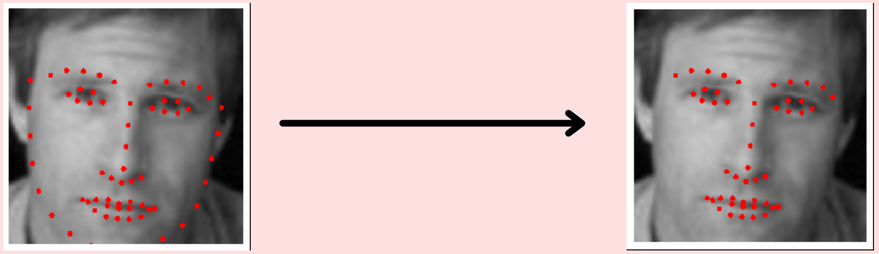
1st	FOLD 1	FOLD 1	FOLD 1	FOLD 1	FOLD 1
2nd	FOLD 2	FOLD 2	FOLD 2	FOLD 2	FOLD 2
3rd	FOLD 3	FOLD 3	FOLD 3	FOLD 3	FOLD 3
4th	FOLD 4	FOLD 4	FOLD 4	FOLD 4	FOLD 4
5th	FOLD 5	FOLD 5	FOLD 5	FOLD 5	FOLD 5



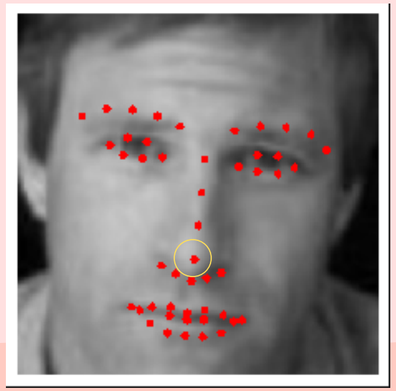
In training set we augment the disgust and contempt to match the remaining class by superpixel, invert image,hue image , add light, saturation, flip vertical, flip horizontal, and rotation randomly

Feature Selection

- Face Detecting and landmark using dlib

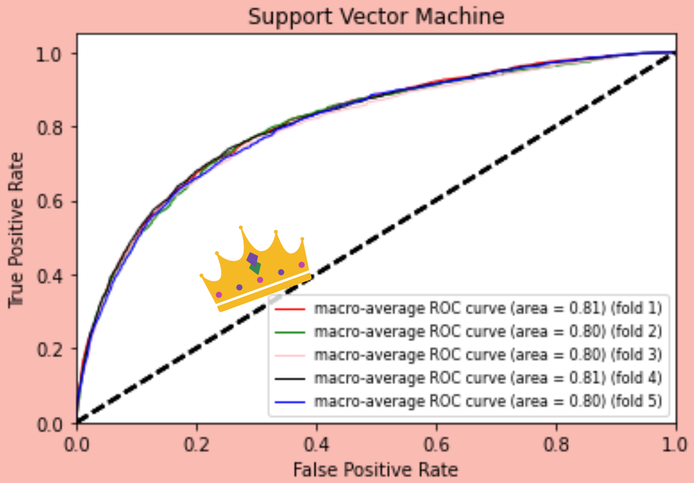


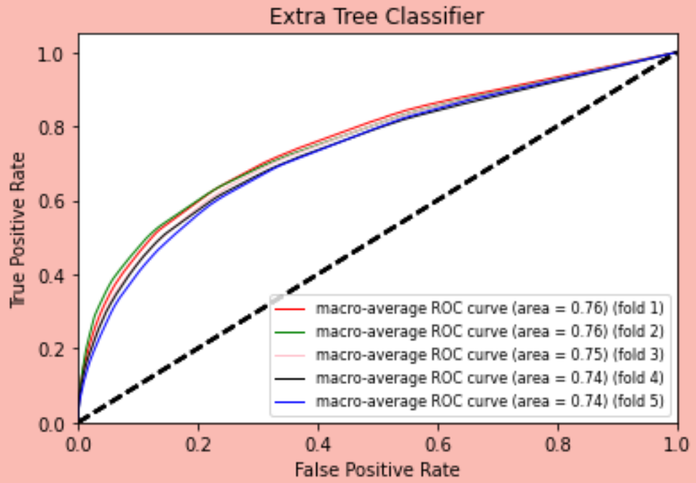
- mark only used point. (eyes , eyebrows, nose and mouth).
- label each point and group them. (ex. left-eye: 36-41 etc.)
- We get the total of 51 points.

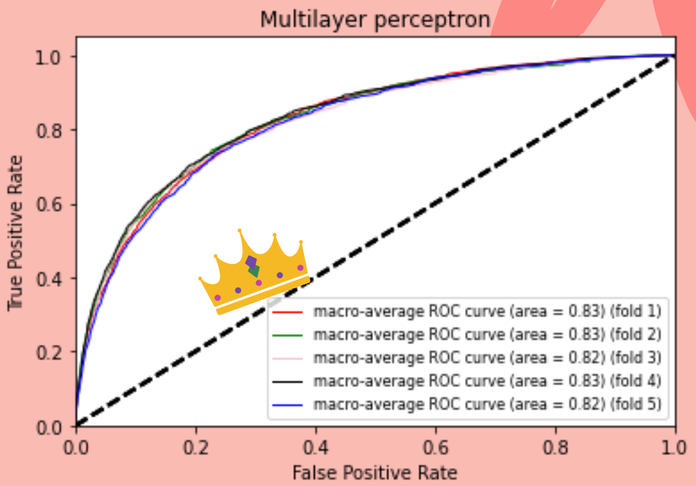


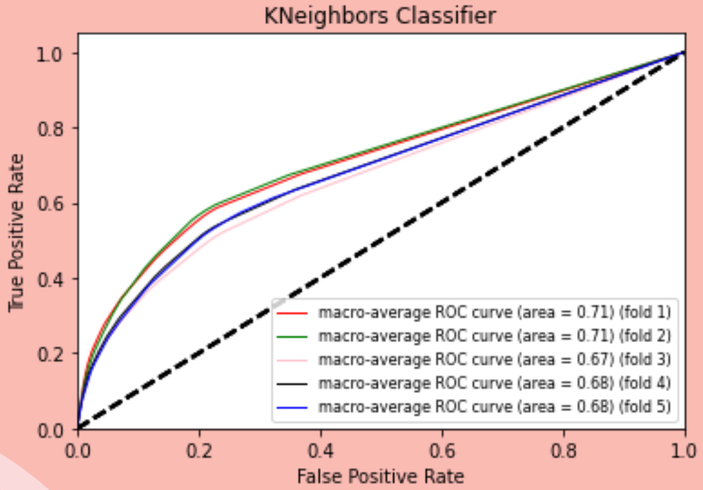
- We use the centre point of the nose and find the cosine distance between all the points and the centre point to be the features and we find the length and height of the eyes and mouth to be the features, which get the total of 67 features,
- write all the output to the csv file.

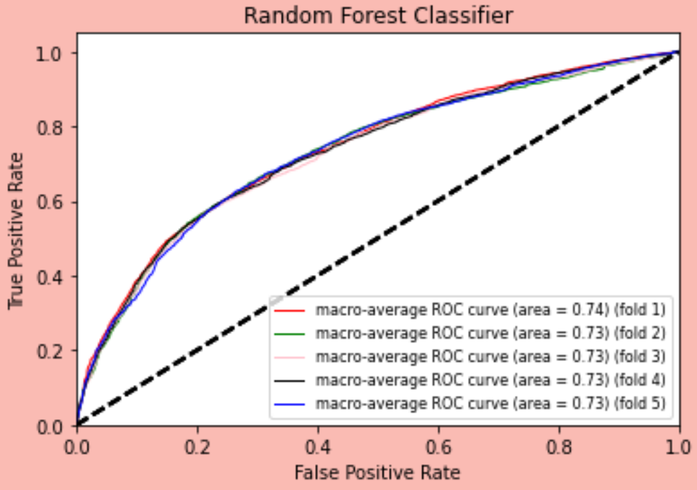
Choosing the best algorithm

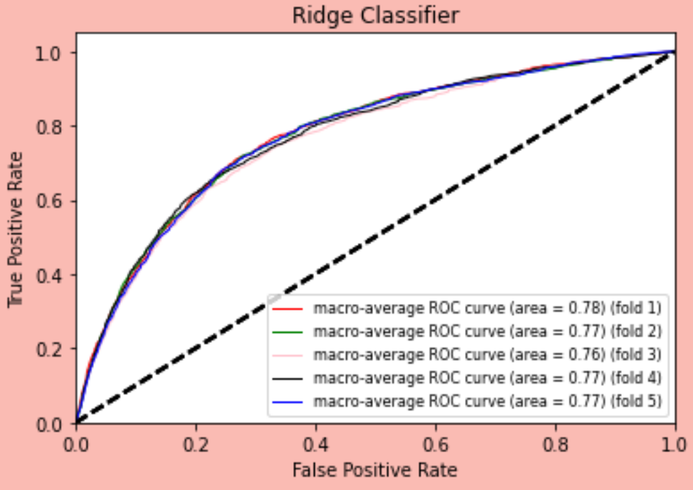








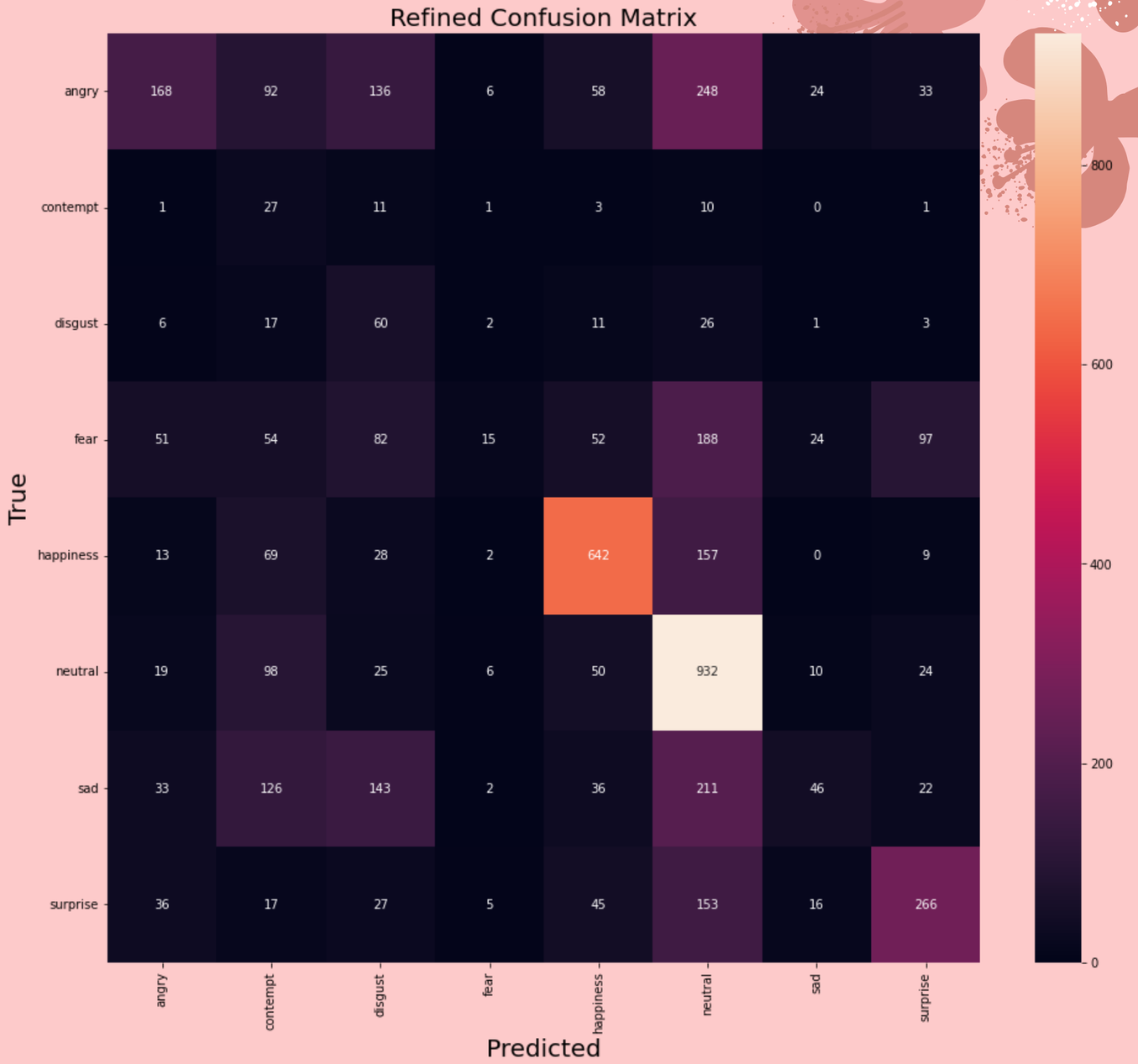




Support-vector machine

	precision	recall	f1-score
angry	51%	22%	31%
contempt	5%	50%	10%
disgust	12%	48%	19%
fear	38%	3%	5%
happiness	72%	70%	71%
neutral	48%	80%	60%
sad	38%	7%	12%
surprise	58%	47%	52%

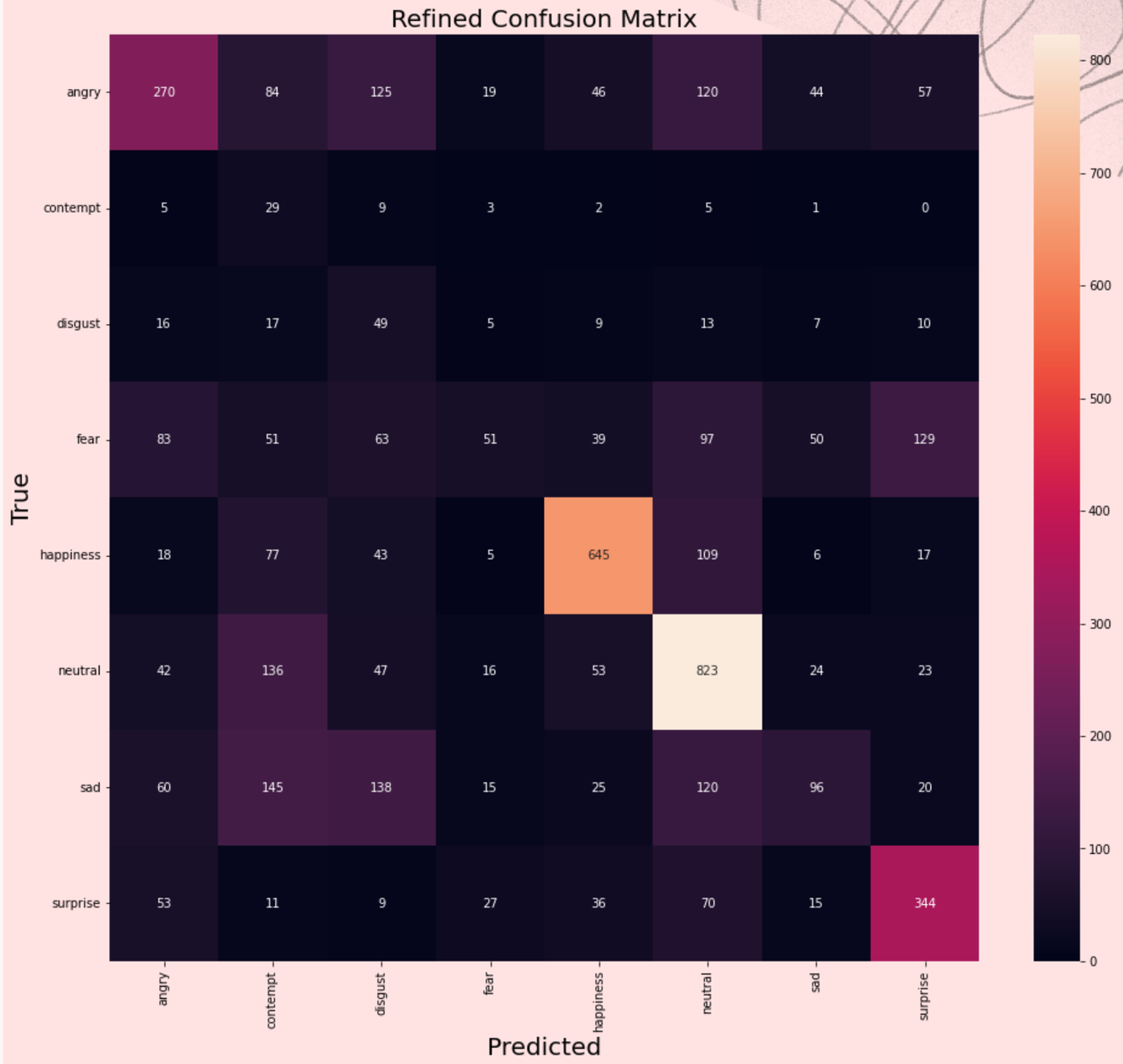
average recall	40.88%
average sensitivity	40.83%
average precision	40.43%
average f1-score	42.22%
accuracy	45.51%



Multilayer perceptron

	precision	recall	f1-score
angry	49%	35%	41%
contempt	5%	54%	10%
disgust	10%	39%	16%
fear	36%	9%	14%
happiness	75%	70%	73%
neutral	61%	71%	65%
sad	40%	16%	22%
surprise	57%	61%	59%

average recall	44.38%
average sensitivity	44.27%
average precision	41.73%
average f1-score	48.62%
accuracy	48.30%



COMPARE BETWEEN 2 TYPE OF FACE

angry

disgust

neutral

angry

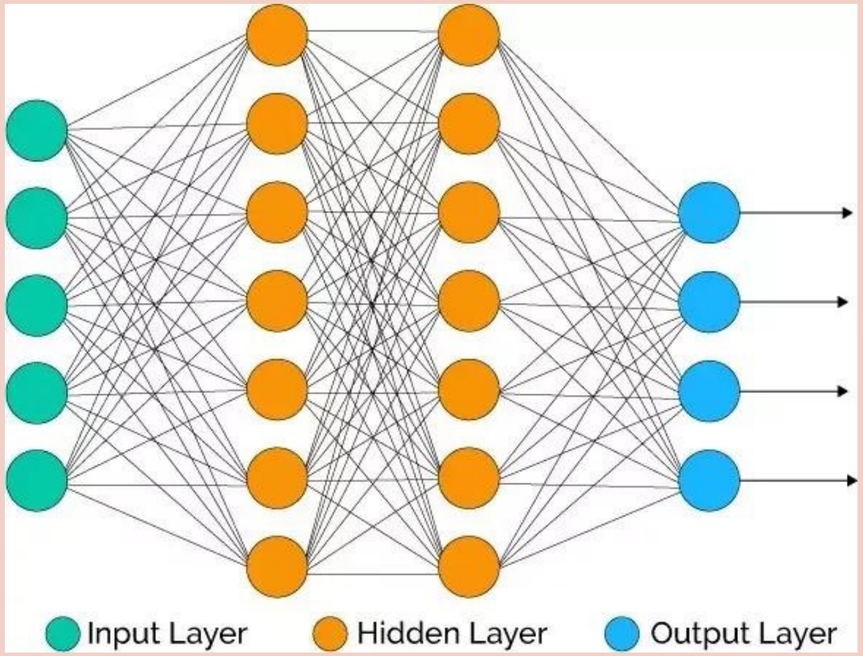
neutral

sad

neutral

surpise

ANN



The function of Neural Networks is when inputs to a network are multiplied by the weight of each leg. The resulting inputs on all neuron legs are added together and compared to a predetermined threshold. is greater than the threshold and the neuron will output its output.

We will use 2 hidden layers.

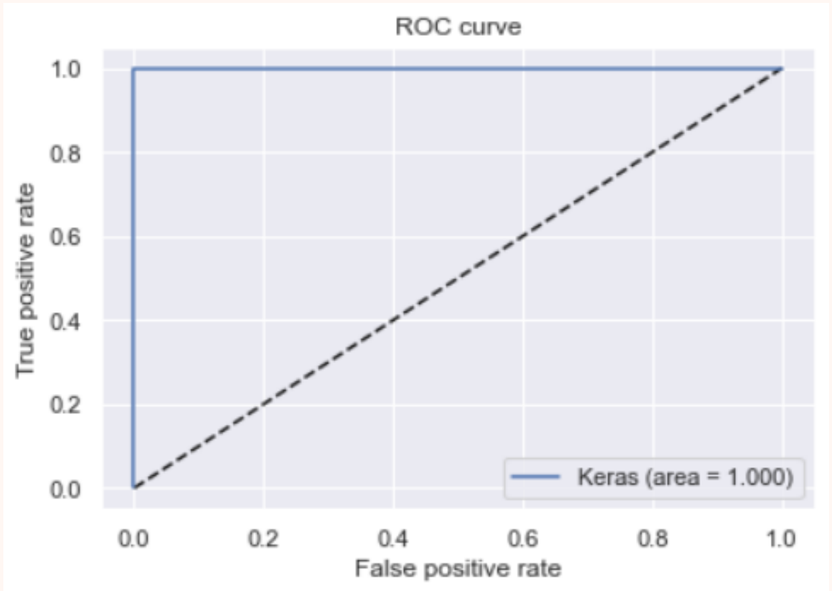
ANN CSV

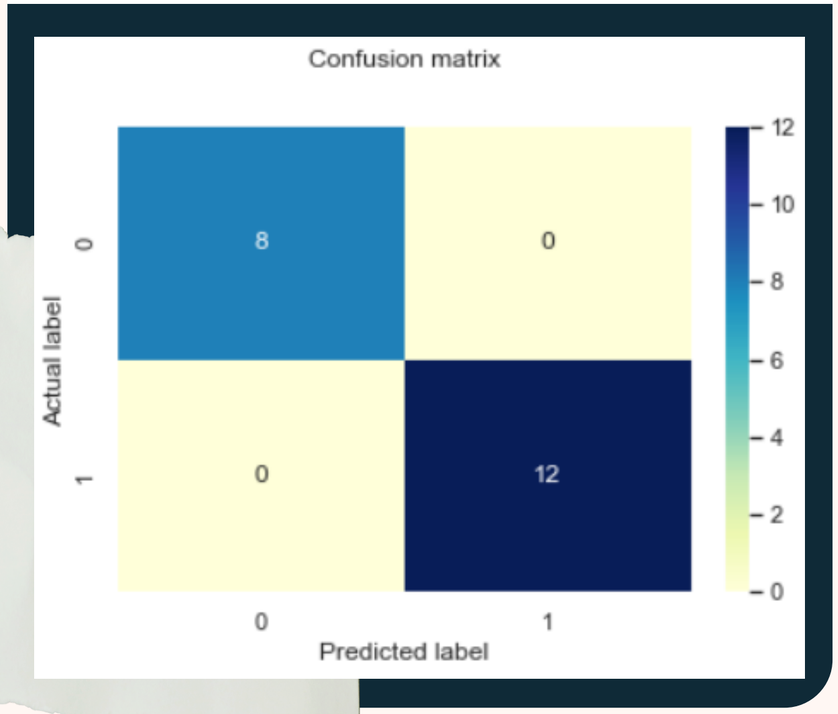
Hyperparameter

2 hidden layer

input	layer 32 units		activation	fn = relu
hidden	layer 16 units		activation	fn = relu
hidden	layer 8	units	activation	fn = relu
output	layer 1	units	activation	fn = softmax

using adam optimizer to compile ann
batch = 16
epochs = 100





Accuracy : 100%
Precision : 100%
Recall : 100%
F1_Score : 100%
Specificity : 100%

ANN image

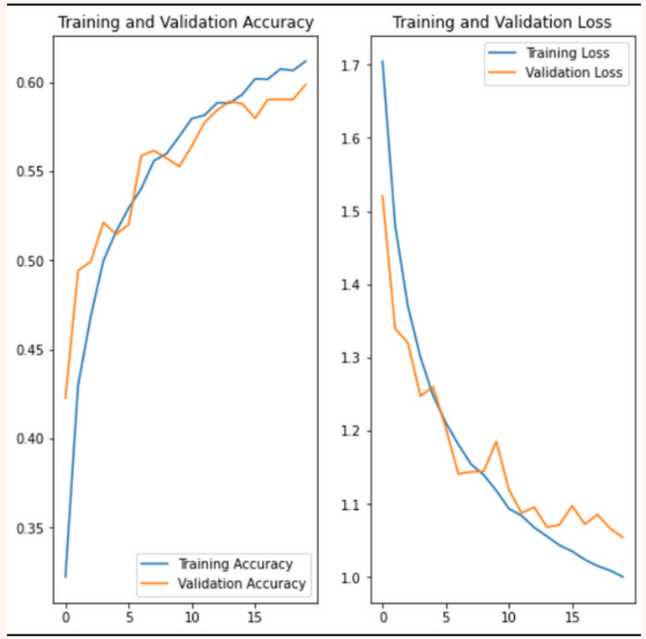
Hyperparameter

2 hidden layer

input layer 64 units activation fn = relu
hidden layer 32 units activation fn = relu
hidden layer 16 units activation fn = relu
output layer 8 units activation fn = softmax

using adam optimizer to compile
ann
batch = 16
epochs = 100

test_acc = 0.4935
test_loss = 1.2752

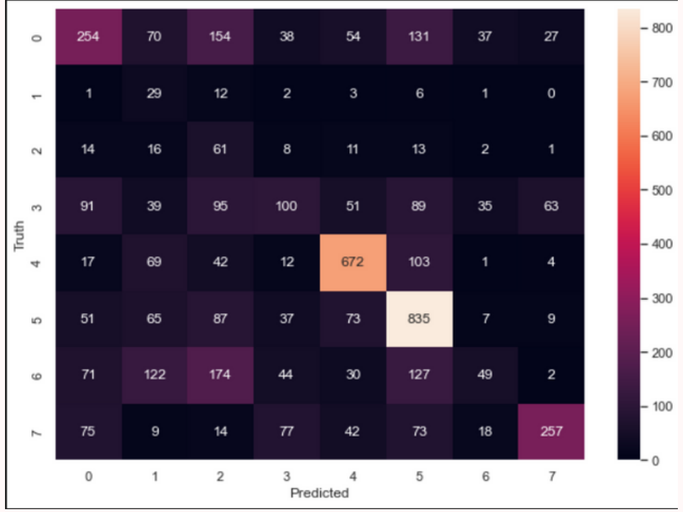


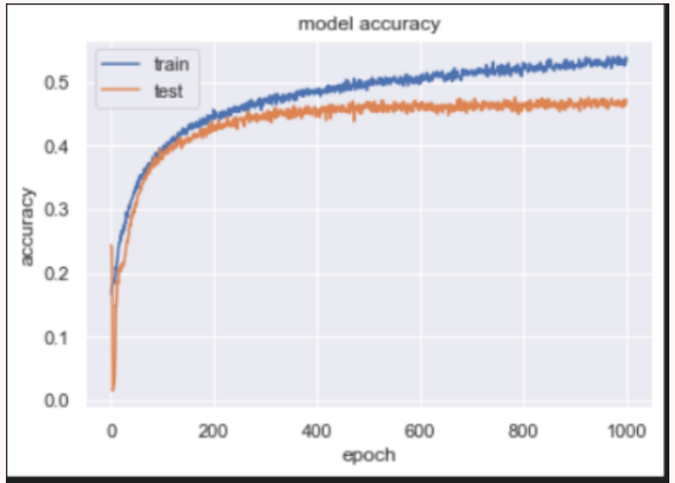
	Accuracy	Loss	after Fix Overfitting	Accuracy	Loss
train	98.35%	0.0532	===>	61.15%	1.0008
validate	56.81%	3.2850	===>	59.85%	1.0547

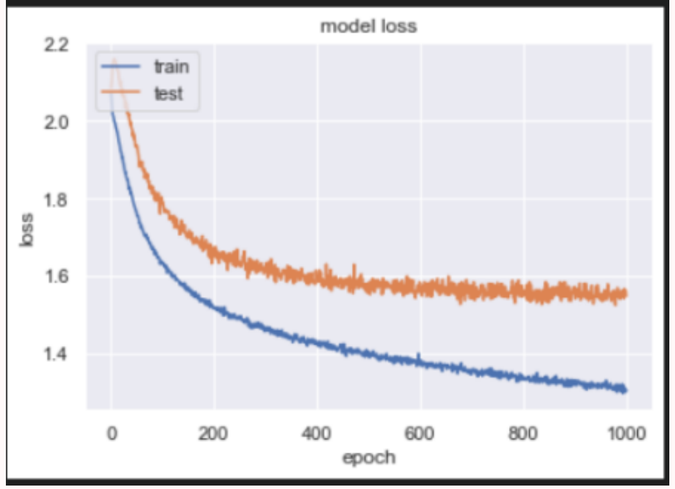
We use dropout and augmentation to fix overfitting.

ANN image

	precision	recall	f1-score	support
0	0.44	0.33	0.38	765
1	0.07	0.54	0.12	54
2	0.10	0.48	0.16	126
3	0.31	0.18	0.23	563
4	0.72	0.73	0.72	920
5	0.61	0.72	0.66	1164
6	0.33	0.08	0.13	619
7	0.71	0.45	0.55	565
accuracy			0.47	4776
macro avg	0.41	0.44	0.37	4776
weighted avg	0.52	0.47	0.47	4776







CNN image

acc = 0.9039
val_acc = 0.6273
loss = 0.2640
val loss = 1.5699

test_acc = 0.6105
test_loss = 1.1003

Discussion and Conclusion

- We get the data from the Kaggle including CSV, image data
- We visualize the data to get the features from the CSV file and research which feature is effect the result, then we fill the find more data and fill in the missing value.
- We use 3 methods to select the features.
- **First**, we use the filter method which selects features from their correlation.
- **Second** we use the wrapper method which uses the RFE method to rank the features and find the optimum features.
- **Lastly**, we use an embedded method that uses LassoCV to select and eliminate features.
- Then We train the model using a feature from each method and compare them, the best method
- While we test the model we found that person the get predicted wrong had some variable close to the other condition.
- After we visualize the image data we found that some expression was very low compared to other expressions, then we find more of that expression data.
- then we use a face-landmark to extract the features.
- then we train and test the model then we see what the model was confused

Discussion and Conclusion (Cont.)

- We use ANN sequential classifier, Both csv and image file we use the same amount of layers, optimizer and activation funnction, but we use different node due to their output.
- then we found that there was a overfitting so we fix it.
- We use efficientnetB0 classifier on image and test it performance.

REFERENCES

[1] Institute of Health Metrics and Evaluation, Global Health Data Exchange (GHDx), 2019. [Online]. Available:https://www.kaggle.com/arashnic/the-depression-dataset

[2] M. E Duffy and J. M Twenge and T. E Joiner, "Trends in mood and anxiety symptoms and suicide-related outcomes among U.S. undergraduates, Journal of Adolescent Health," United State of America. Accessed: Jul. 3, 2019.[Online]. Available:https://pubmed.ncbi.nlm.nih.gov/31279724/

[3] The Depression Dataset, Kaggle, Feb. 2021. [Online]. Available: https://www.kaggle.com/arashnic/the-depression-dataset

[4] Depression and Motor Activity, Kaggle, Apr. 2021. [Online]. Available: https://www.kaggle.com/docxian/depression-and-motor-activity#Activity-Data---Exploration

[5] Emotion Detection From Facial Expressions, Kaggle, Dec. 2016. [Online]. Available: https://www.kaggle.com/c/emotion-detection-from-facial-expressions/overview

[6] Institute of Health Metrics and Evaluation. Global Health Data Exchange (GHDx), 2019. [Online]. Available:http://ghdx.healthdata.org/gbd-results-tool?params=gbd-api-2019-permalink/d780dffbe8a381b25e1416884959e88b

[7] World Health Organization, Depression, Sep. 13, 2021. [Online]. Available:https://www.who.int/news-room/fact-sheets/detail/depression

REFERENCES (CONT.)

[8] What is the suicide rate among persons with depressive disorder, 2022. [Online]. Available:https://www.medscape.com/answers/286759-14675/what-is-the-suicide-rate-among-persons-with-depressive-disorder-clinical-depression
[9] Predicting Depression Data, 2022. [Online]. Available:https://www.kaggle.com/datasets/jeremyteo/predicting-depression-data
[10] Impaired implicit learning and reduced pre-supplementary motor cortex size in early-onset major depression with melancholic features, sciencedirect, 2009. [Online]. Available:https://www.sciencedirect.com/science/article/abs/pii/S0165032709001281
[11] Feature Selection with sklearn and Pandas, Towards Data Science, 2019. [Online]. Available:https://towardsdatascience.com/feature-selection-with-pandas-e3690ad8504b
[12] A review of studies of the Montgomery-Asberg Depression Rating Scale in controls: implications for the definition of remission in treatment studies of depression, National Library of Medicine, 2004. [Online]. Available:https://pubmed.ncbi.nlm.nih.gov/15101563/
[13] Aripiprazole (Abilify): Depression, Major Depressive Disorder (MDD), NCIB, 2004. [Online]. Available:https://www.ncbi.nlm.nih.gov/books/NBK409740/#:~:text=The%20MADRS%20scoring%20instructions%20indicate,depression%2C%E2%80%9D%20and%20a%20total%20score

Member

นภทีป์	ละปะชัย	6210612609
ณิชชารีย์	คัคนางจนันท์	6210612617
อนุรดี	บรรดาศักดิ์	6210612625
กชกร	สุรินทร์	6210612674
จิรพนธ์	กันภัย	6210612690
ปารินทร์	โสภิตลาภธนา	6210612740
ปฤญจ์กาณฑ์	ดิษฐ์สมบูรณ์	6210612781
ปุณยวัจน์	เอี่ยมทวีสิน	6210612799
ธาม	พลศรีสุทธิกุล	6210612807
สิทธิโชค	งามพิพัฒน์โชคชัย	6210612815

Thank You!

Do you have any questions for me before we go?