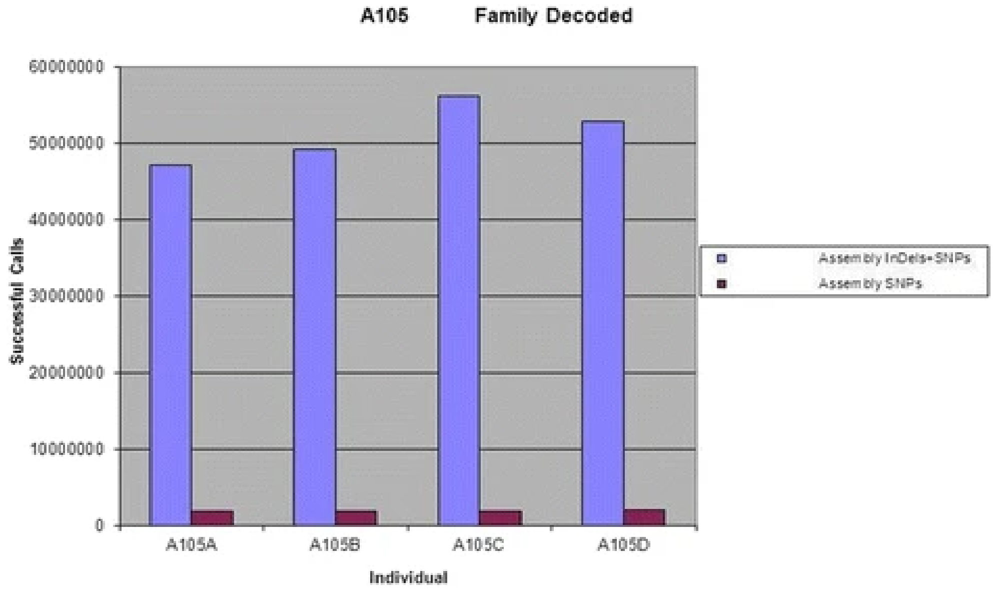
1.



Structural Variations as detected for the whole Genome

**Fig. 1.Here we see the Figure 4. from paper [21] where clearly higher deviation is observed for sum of nucleotides of SNPs plus DIPs (blue) compared to sum of bases for the SNPs (magenta).**

0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39

A,C,G,G,G,A,A,G,G,A,G,C,T,G,G,G,G,A,A,G,T,C,A,T,T,C,G,A,A,T,G,C,T,A,C,G,T,T,A,G

C,A,A,T,A,G,A,C,G,C,A,C,G,G,T,C,TGAGAG,G,T,C,C,A,G,G,T,G,G,GCATTC,C,G,C,G,C,T,T,T,G,G,A,T

G,AGCATATA,T,A,C,CCTGGCCC,G,G,G,T,T,A,G,C,G,A,A,TGACA,A,A,T,G,T,G,G,T,G,A,T,A,C,A,G,G,A,T,A,G,T,A

C,C,C,A,G,C,TTTCG,C,A,T,T,A,C,A,A,A,T,A,A,C,G,GG,A,C,C,A,C,G,A,A,G,T,C,T,ATTAAGTCG,G,G,G,C,A

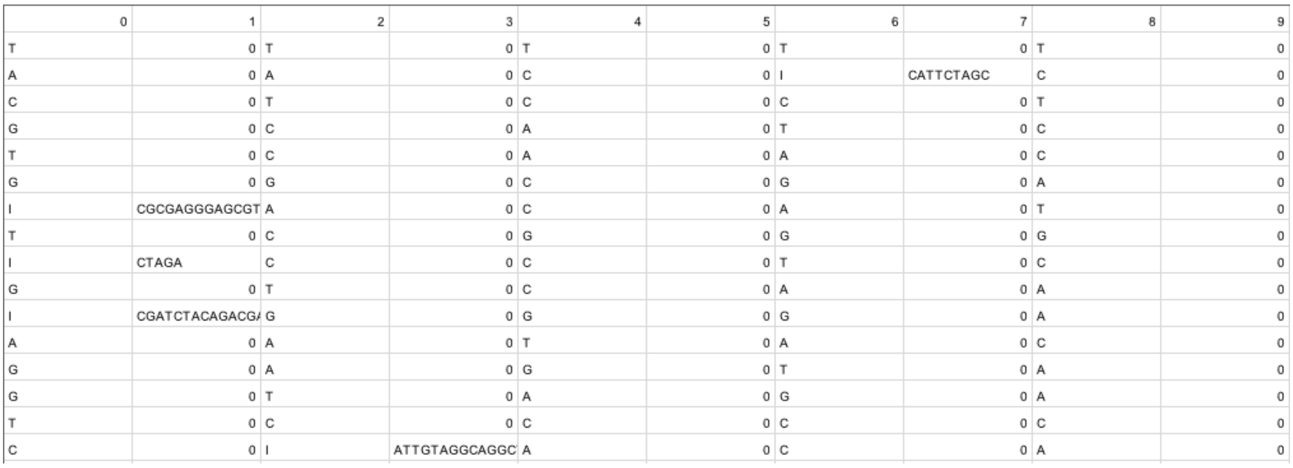
T,C,A,G,T,T,C,A,A,T,T,G,C,A,C,C,T,G,A,A,C,A,A,G,G,G,A,G,A,G,C,A,GGG,C,A,C,G,A,C,GG

C,T,G,G,A,C,T,C,A,A,T,C,A,C,A,G,G,G,C,G,A,GCGGG,T,A,C,T,C,C,T,G,T,C,G,T,G,C,T,G,T,T

A,C,T,G,G,T,A,G,G,T,C,G,G,A,A,A,T,G,T,T,A,G,G,G,C,G,T,G,T,A,G,A,T,A,G,C,C,G,A,A

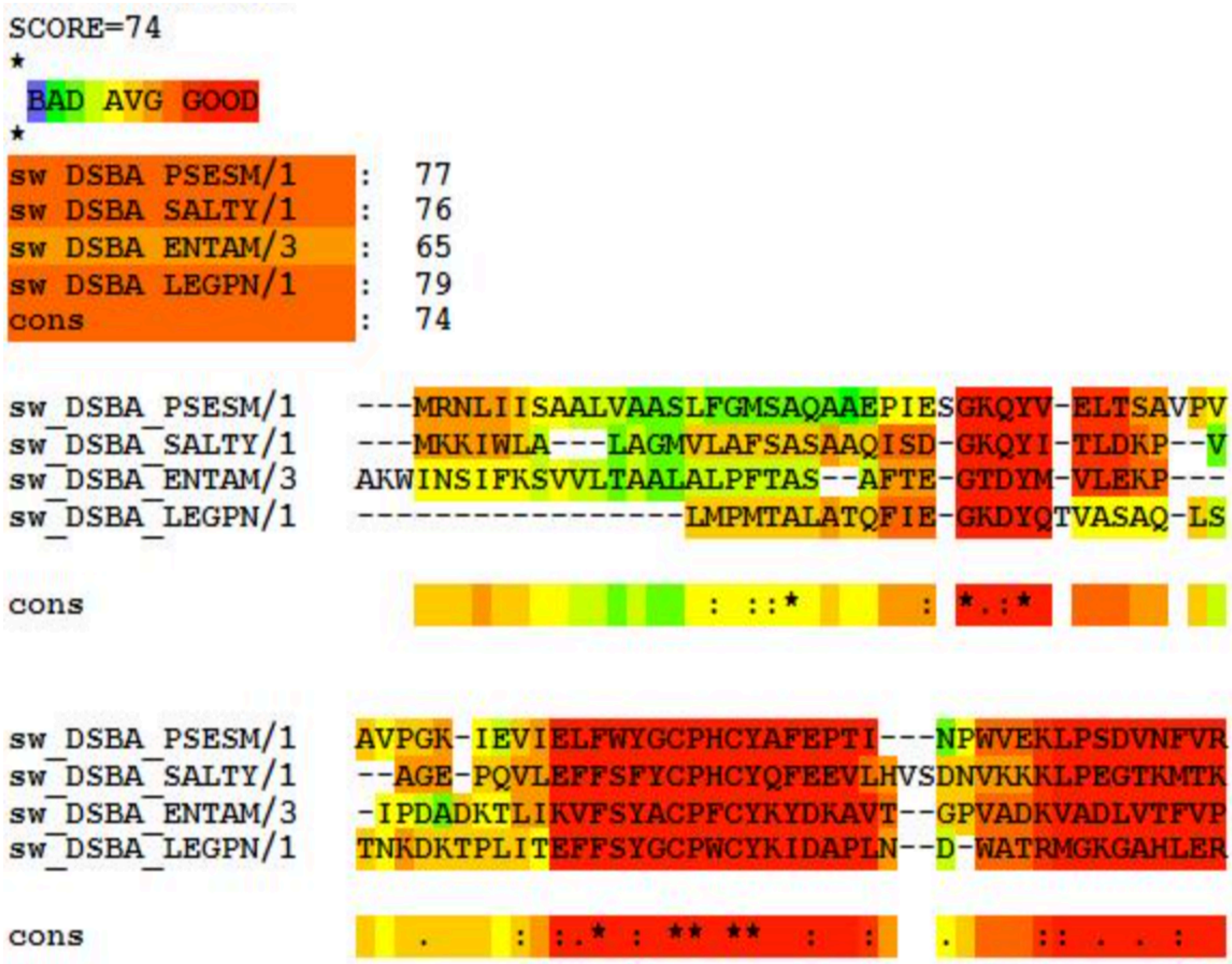
C,A,C,A,T,A,T,A,G,G,T,TAAAAAT,T,A,G,T,T,ACCCATGA,G,A,A,G,T,G,T,A,A,T,G,C,G,C,C,A,A,G,A,G,C,C

**Fig 2. Randomly generated Genotype data for 8 patients for illustration purpose. The example data chosen for simulation comprises of 40 individuals and 200 genotypic loci.**



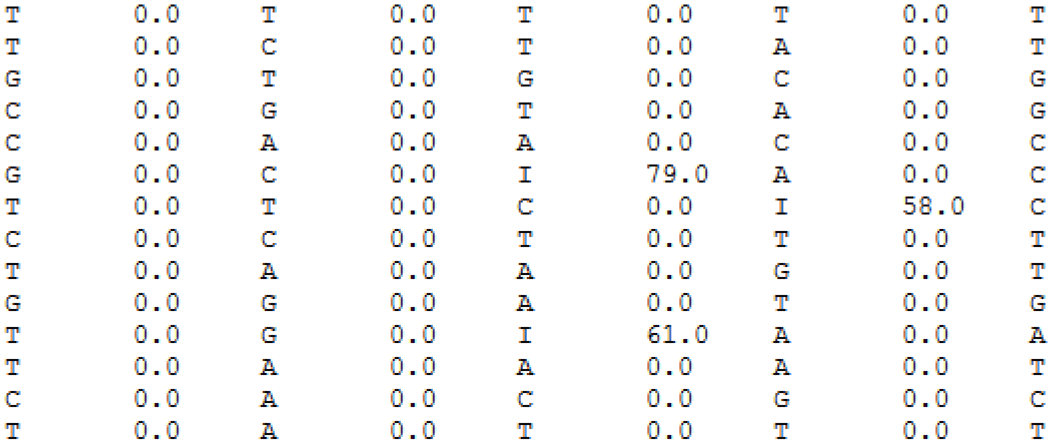
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| T | 0 | T | 0 | T | 0 | T | 0 | T | 0 |
| A | 0 | A | 0 | C | 0 | I | CATTCTAGC | C | 0 |
| C | 0 | T | 0 | C | 0 | C | 0 | T | 0 |
| G | 0 | C | 0 | A | 0 | T | 0 | C | 0 |
| T | 0 | C | 0 | A | 0 | A | 0 | C | 0 |
| G | 0 | G | 0 | C | 0 | G | 0 | A | 0 |
| I | CGCGAGGGAGCGT | A | 0 | C | 0 | A | 0 | T | 0 |
| T | 0 | C | 0 | G | 0 | G | 0 | G | 0 |
| I | CTAGA | C | 0 | C | 0 | T | 0 | C | 0 |
| G | 0 | T | 0 | C | 0 | A | 0 | A | 0 |
| I | CGATCTACAGACG/ | G | 0 | G | 0 | G | 0 | A | 0 |
| A | 0 | A | 0 | T | 0 | A | 0 | C | 0 |
| G | 0 | A | 0 | G | 0 | T | 0 | A | 0 |
| G | 0 | T | 0 | A | 0 | G | 0 | A | 0 |
| T | 0 | C | 0 | C | 0 | C | 0 | C | 0 |
| C | 0 | I | ATTGTAGGCAGGC | A | 0 | C | 0 | A | 0 |

**Figure 3. For illustration purpose, we show how the DIPs columns are generated, by splitting each feature column variable into 2.**



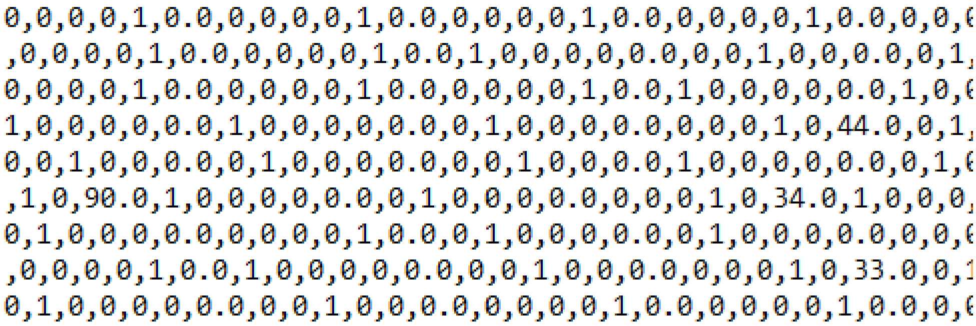
```
SCORE=74
*
 BAD AVG GOOD
*
sw DSBA PSESM/1    :   77
sw DSBA SALTY/1    :   76
sw DSBA ENTAM/3    :   65
sw DSBA LEGPN/1    :   79
cons               :   74

sw_DSBA_PSESM/1     ---MRNLIISAALVAASLFGMSAQAAEPIESGKQYV-ELTSAVPV
sw_DSBA_SALTY/1     ---MKKIWLA---LAGMVLAFSASAAQISD-GKQYI-TLDKP--V
sw_DSBA_ENTAM/3     AKWINSIFKSVVLTAALALPFTAS--AFTE-GTDYM-VLEKP---
sw_DSBA_LEGPN/1     -----------------LMPMTALATQFIE-GKDYQTVASAQ-LS

cons                                  : ::*      : *.:*


sw_DSBA_PSESM/1     AVPGK-IEVIELFWYGCPHCYAFEPTI---NPWVEKLPSDVNFVR
sw_DSBA_SALTY/1     --AGE-PQVLEFFSFYCPHCYQFEEVLHVSDNVKKKLPEGTKMTK
sw_DSBA_ENTAM/3     -IPDADKTLIKVFSYACPFCYKYDKAVT--GPVADKVADLVTFVP
sw_DSBA_LEGPN/1     TNKDKTPLITEFFSYGCPWCYKIDAPLN--D-WATRMGKGAHLER

cons                   .    : :.* : ** **  :  :   .    :: . . :
```

**Figure 4 Example of a sample clustering by multiple sequence alignment with consensus regions and the consensus score with individual scores as well (which we call divergence score in this paper).**



```
T       0.0     T       0.0     T       0.0     T       0.0     T
T       0.0     C       0.0     T       0.0     A       0.0     T
G       0.0     T       0.0     G       0.0     C       0.0     G
C       0.0     G       0.0     T       0.0     A       0.0     G
C       0.0     A       0.0     A       0.0     C       0.0     C
G       0.0     C       0.0     I       79.0    A       0.0     C
T       0.0     T       0.0     C       0.0     I       58.0    C
C       0.0     C       0.0     T       0.0     T       0.0     T
T       0.0     A       0.0     A       0.0     G       0.0     T
G       0.0     G       0.0     A       0.0     T       0.0     G
T       0.0     G       0.0     I       61.0    A       0.0     A
T       0.0     A       0.0     A       0.0     A       0.0     T
C       0.0     A       0.0     C       0.0     G       0.0     C
T       0.0     A       0.0     T       0.0     T       0.0     T
```

**Figure 5. The DIPs are replaced by the corresponding divergence from consensus score, lying between 0 and 100.**



```
0,0,0,0,1,0.0,0,0,0,0,1,0.0,0,0,0,0,1,0.0,0,0,0,0,1,0.0,0,0,0
,0,0,0,0,1,0.0,0,0,0,0,1,0.0,1,0,0,0,0,0.0,0,0,1,0,0,0.0,0,1,
0,0,0,0,1,0.0,0,0,0,0,1,0.0,0,0,0,0,1,0.0,1,0,0,0,0,0.0,1,0,0
1,0,0,0,0,0.0,1,0,0,0,0,0.0,0,1,0,0,0,0.0,0,0,0,1,0,44.0,0,1,
0,0,1,0,0,0.0,0,1,0,0,0,0.0,0,0,1,0,0,0.0,1,0,0,0,0,0.0,0,1,0
,1,0,90.0,1,0,0,0,0,0.0,0,1,0,0,0,0.0,0,0,0,1,0,34.0,1,0,0,0,
0,1,0,0,0,0.0,0,0,0,0,1,0.0,0,1,0,0,0,0.0,0,1,0,0,0,0.0,0,0,0
,0,0,0,0,1,0.0,1,0,0,0,0,0.0,0,0,1,0,0,0.0,0,0,0,1,0,33.0,0,1
0,1,0,0,0,0,0.0,0,0,1,0,0,0.0,0,0,0,0,1,0.0,0,0,0,0,1,0.0,0,0
```

**Figure 6. Now, the single nucleotide variations, SNVs or SNPs, are also one-hot encoded**

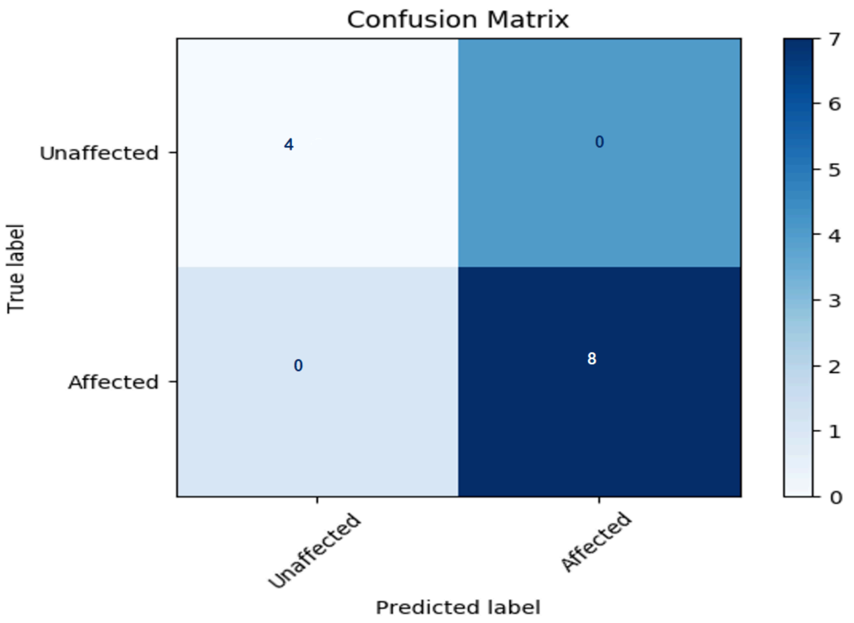
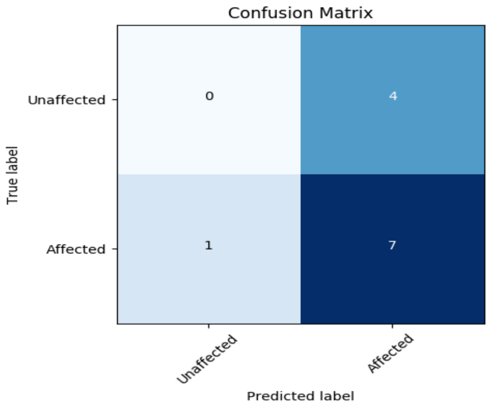
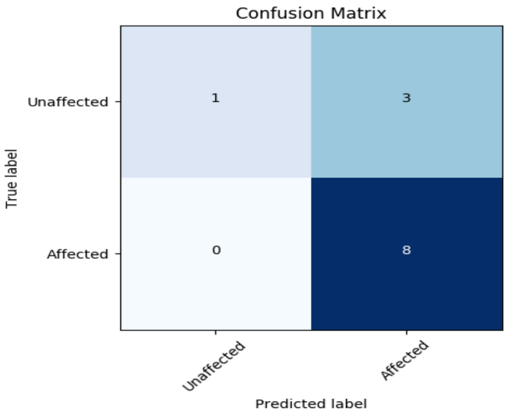


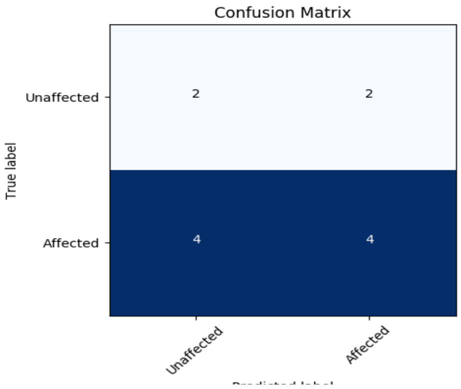
**Figure 7. Confusion MATRIX for ExhaustiveDNN model for simulated dataset.**

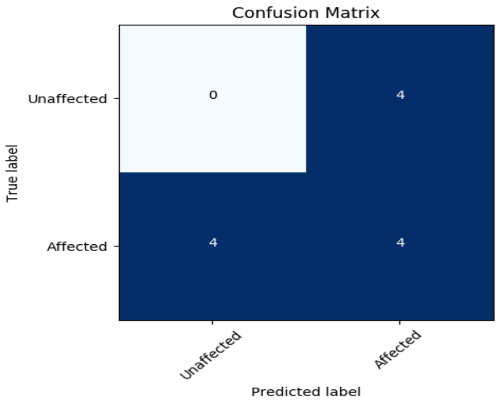
| Algorithm | Accuracy % |
|---|---|
| Exhaustive Deep Neural Network | 100 |
| Logistic Regression | 58.33 |
| AdaBoost | 66.67 |
| GradientBoost | 50 |
| Naïve Bayes | 75 |
| Bagging | 33.33 |
| Support Vector | 66.67 |
| Random Forest | 50 |
| Extra Tree Classifier | 66.67 |

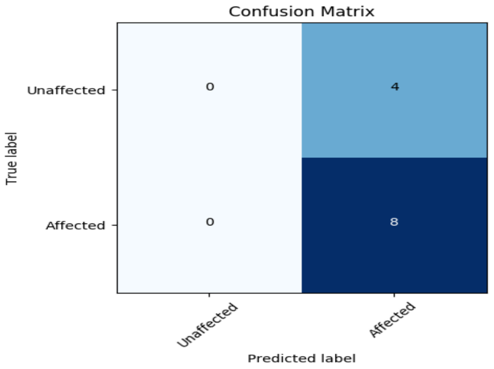
**Table 1. ExhaustiveDNN outperforming some of the popular Machine Learning methods for simulated dataset**

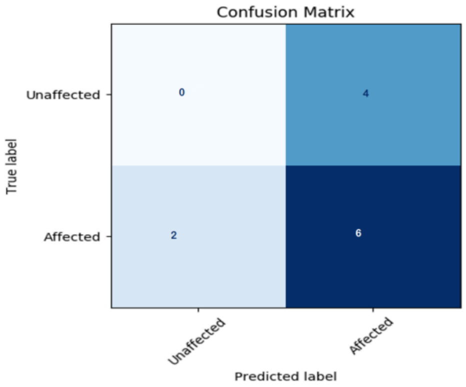

Confusion Matrix
Unaffected
Affected
True label
0
4
1
7
Unaffected
Affected
Predicted label


Confusion Matrix
Unaffected
Affected
True label
1
3
0
8
Unaffected
Affected
Predicted label


Confusion Matrix
Unaffected
Affected
True label
2
2
4
4
Unaffected
Affected
Predicted label


Confusion Matrix
Unaffected
Affected
True label
0
4
4
4
Unaffected
Affected
Predicted label


Confusion Matrix
Unaffected
Affected
True label
0
4
0
8
Unaffected
Affected
Predicted label


Confusion Matrix
Unaffected
Affected
True label
0
4
2
6
Unaffected
Affected
Predicted label

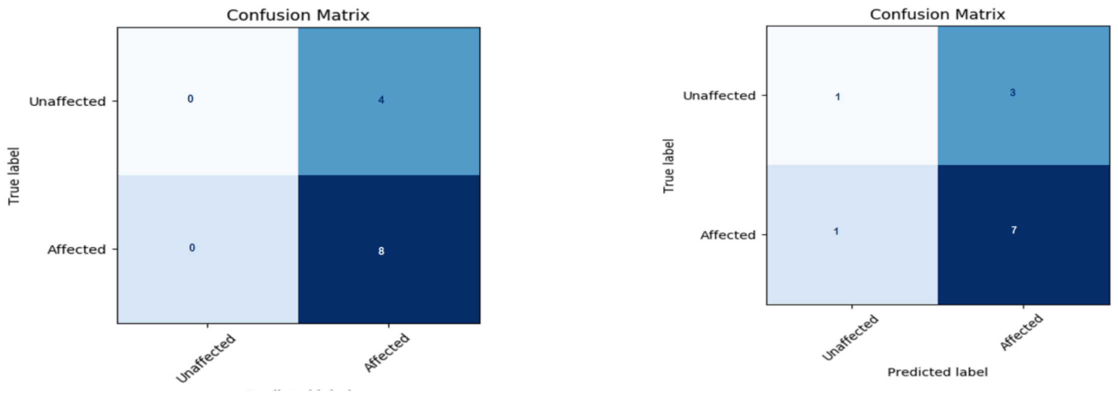


**Fig 8.Top left to bottom right: Confusion MATRIX for logistic regression, Naïve Bayes, Gradient Boost, Bagging approach, AdaBoost, RandomForest, Support Vector & Extratree Classifier respectively for simulated dataset**

| # Hidden Layers | # Hidden Units in Each Layer | Average Score of K-fold (k = 10) |
|---|---|---|
| 2 | 8 | 0.9600000023841858 |
| 3 | 8 | 0.9400000005960465 |
| 4 | 8 | 0.9600000023841858 |
| 5 | 8 | 0.9400000035762787 |
| 6 | 8 | 0.9600000023841858 |
| 7 | 8 | 0.9600000023841858 |
| 2 | 9 | 0.9800000011920929 |
| 3 | 9 | 0.9800000011920929 |
| 4 | 9 | 0.9600000023841858 |
| 5 | 9 | 0.9200000047683716 |
| 6 | 9 | 0.6333333551883698 |
| 7 | 9 | 0.9800000011920929 |
| 2 | 10 | 0.9400000005960465 |
| 3 | 10 | 0.9600000023841858 |

| | | |
|---|---|---|
| 4 | 10 | 0.6333333551883698 |
| 5 | 10 | 0.9600000023841858 |
| 6 | 10 | 0.9600000023841858 |
| 7 | 10 | 0.6333333551883698 |
| 2 | 11 | 0.9600000023841858 |
| 3 | 11 | 0.9400000005960465 |
| 4 | 11 | 0.9400000005960465 |
| 5 | 11 | 0.9600000023841858 |
| 6 | 11 | 0.9400000035762787 |
| 7 | 11 | 0.9400000005960465 |

**Table 2**. ExhaustiveDNN leading to several different average accuracy score for various combinations of hidden layers and hidden units.

| **Algorithm** | **Accuracy %** |
|---|---|
| Exhaustive Deep Neural Network | 78.5 |
| Logistic Regression | 94.64 |
| AdaBoost | 76.78 |
| GradientBoost | 76.78 |
| Naïve Bayes | 76.78 |
| Bagging | 78.5 |
| Support Vector | 94.64 |
| Random Forest | 78.5 |
| Extra Tree Classifier | 78.5 |

**Table 3. List of various machine and deep learning algorithms with the score of their accuracy.**

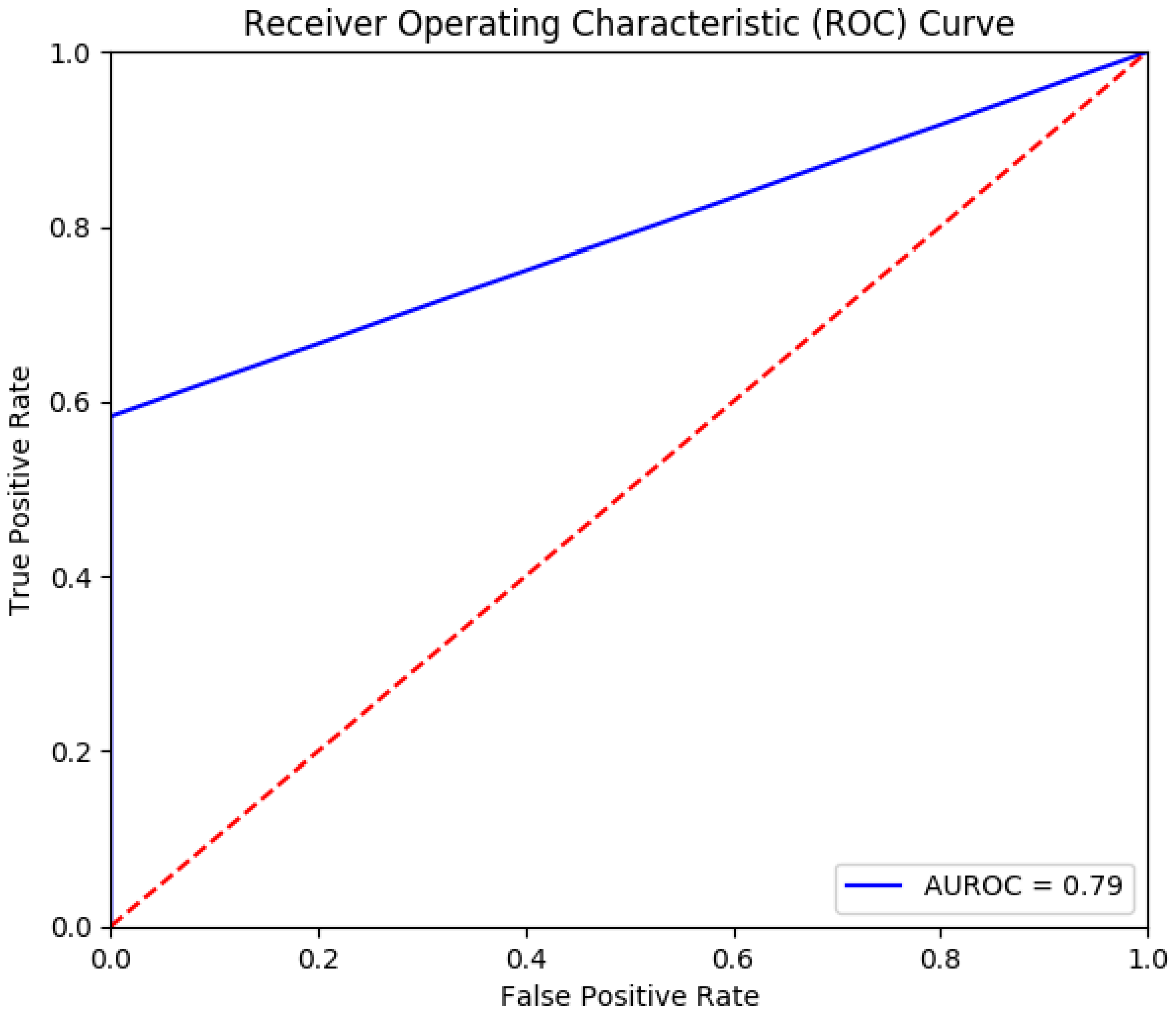


**Figure 9. ROC curve for Logistic regression in DMWAS suite MHHRTATT trait for GTEx V7 pilot dataset**

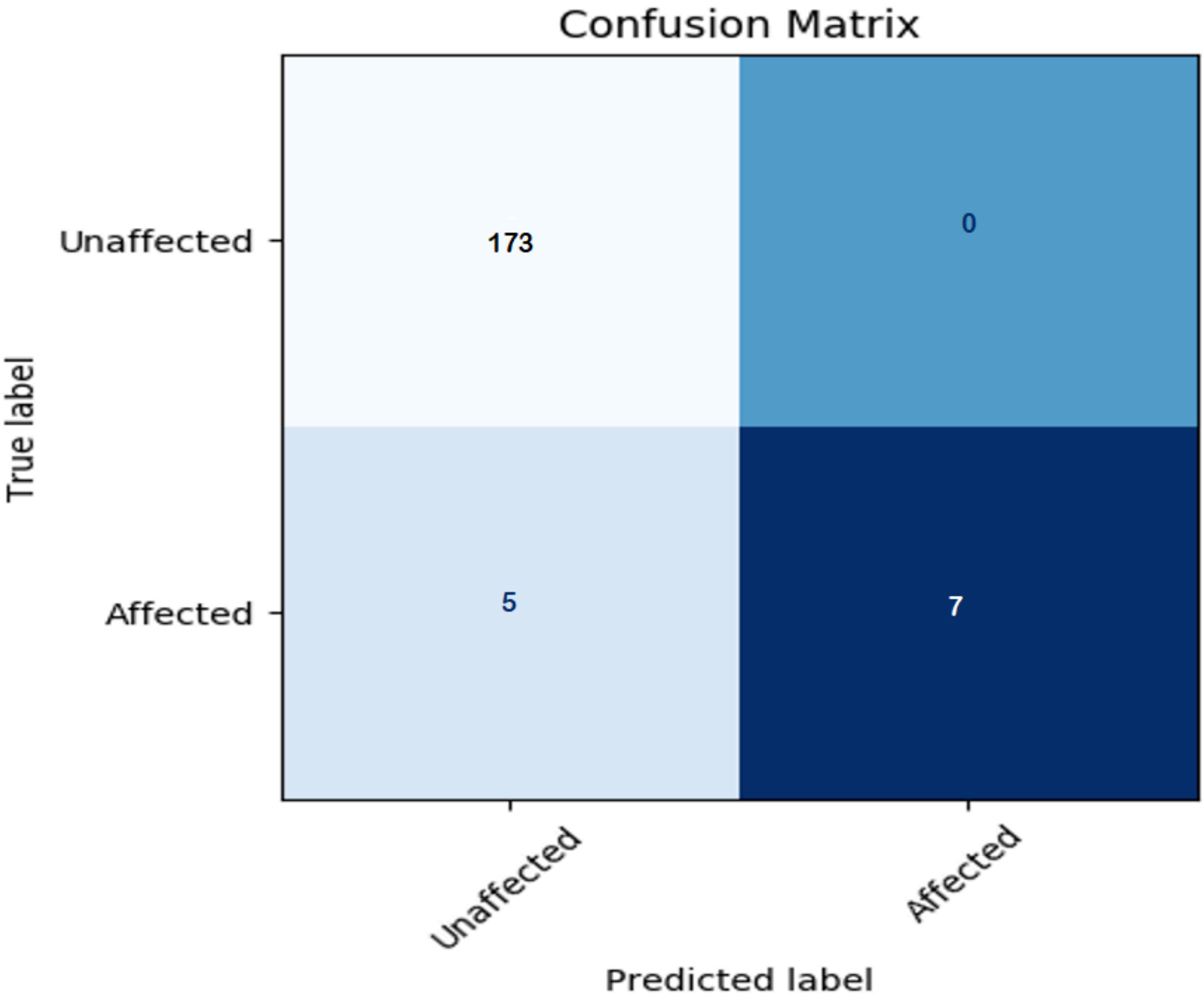
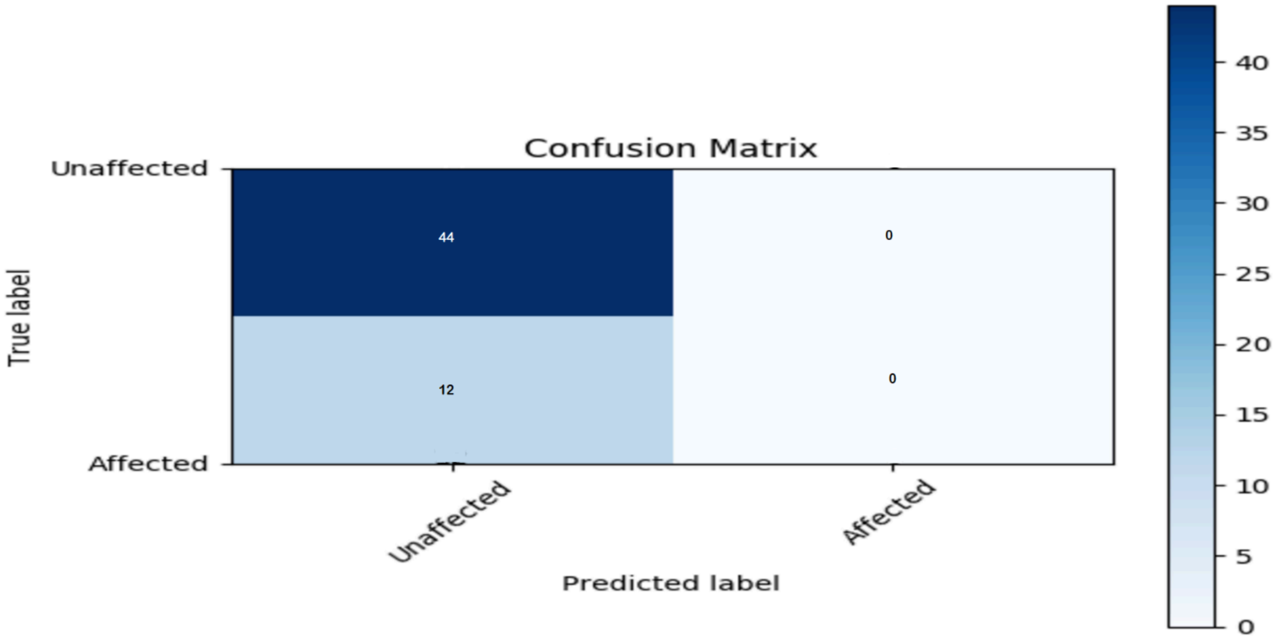
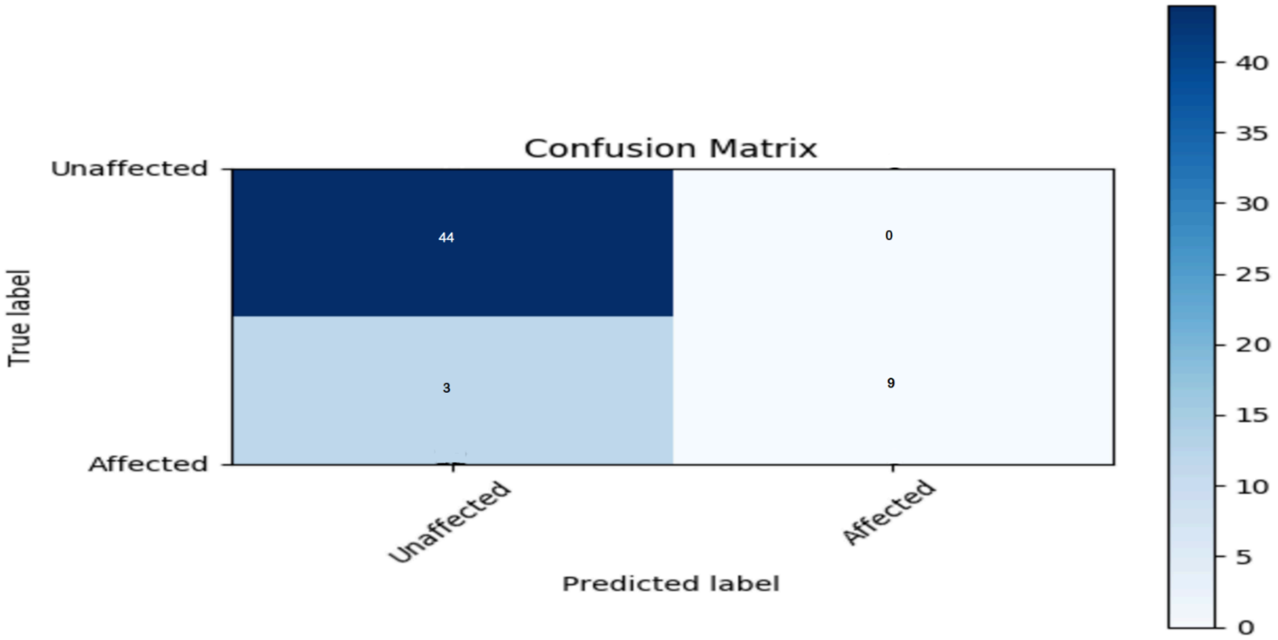
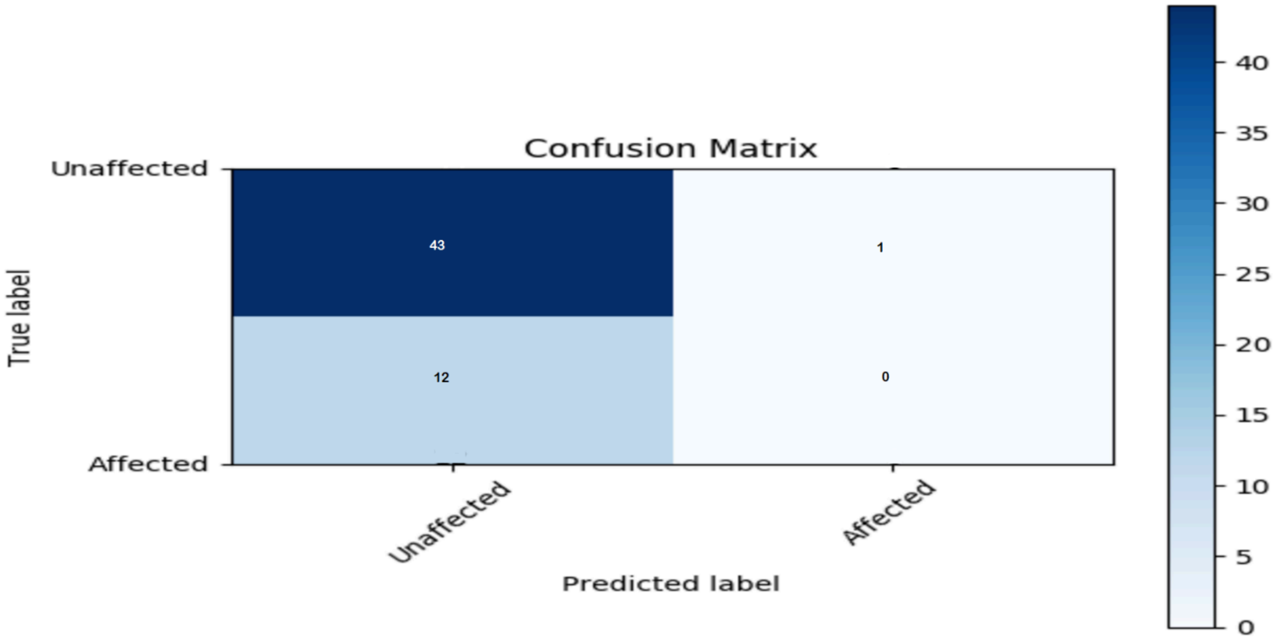
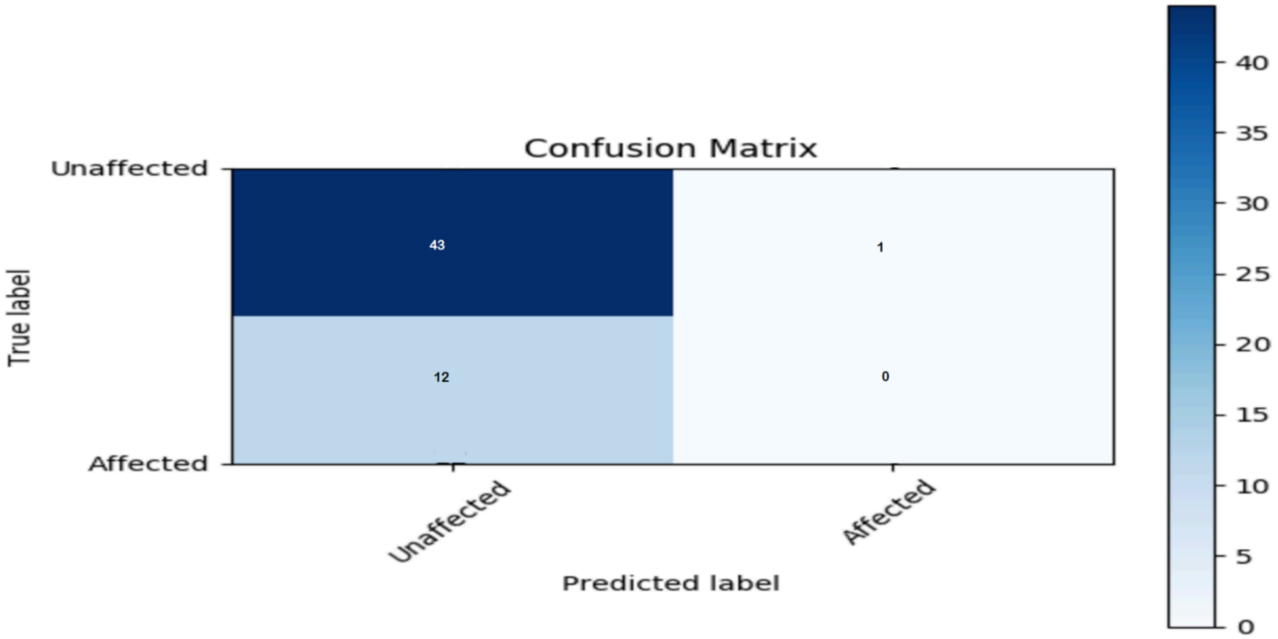


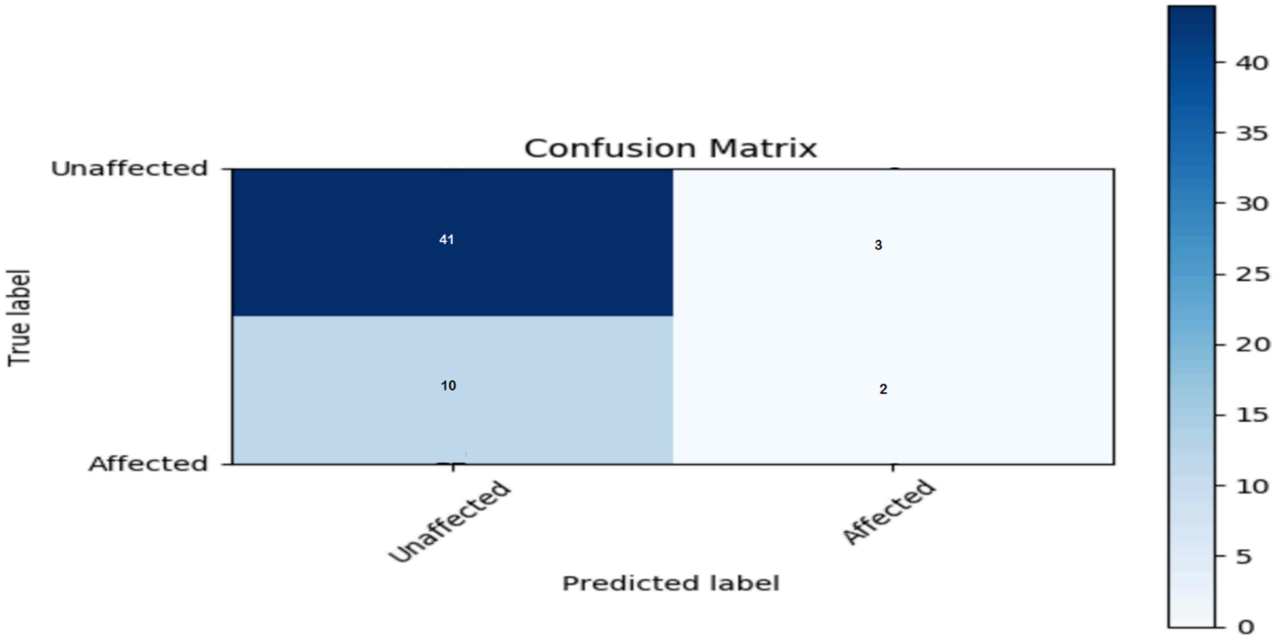
**Figure 10. Confusion Matrix of Logistic Regression of DMWAS on GTEx V7 Pilot data for MHHRTATT trait giving accuracy of 97.3%**

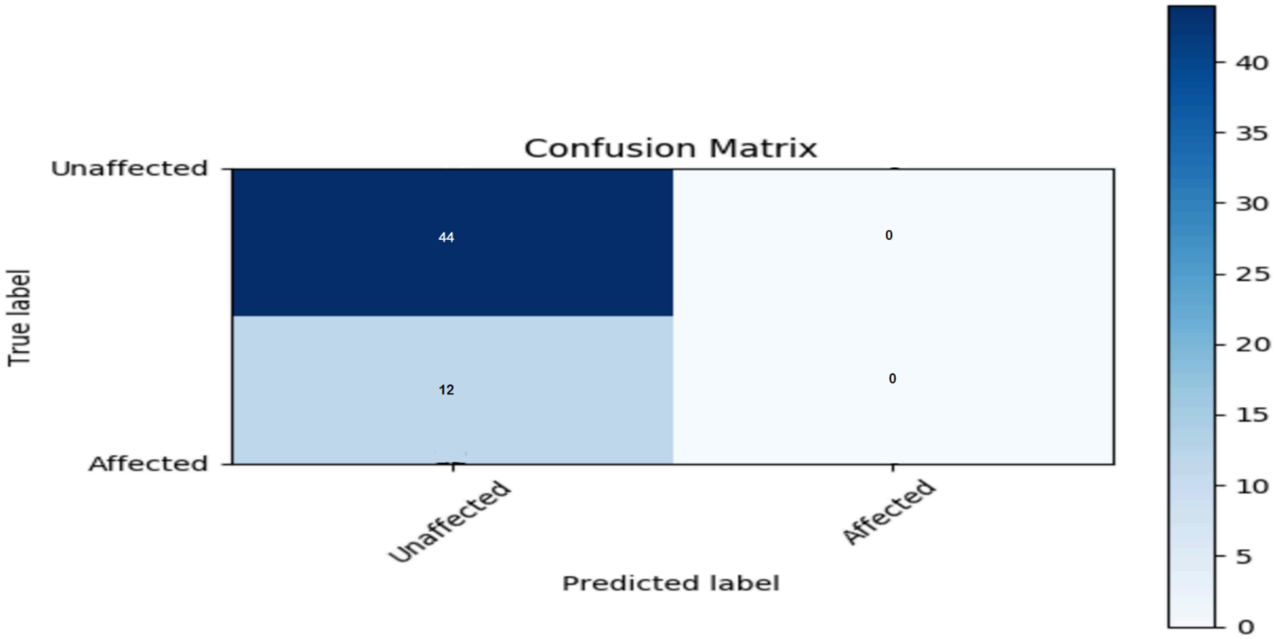

Confusion Matrix
True label
Unaffected
Affected
44
0
12
0
Unaffected
Affected
Predicted label
40
35
30
25
20
15
10
5
0

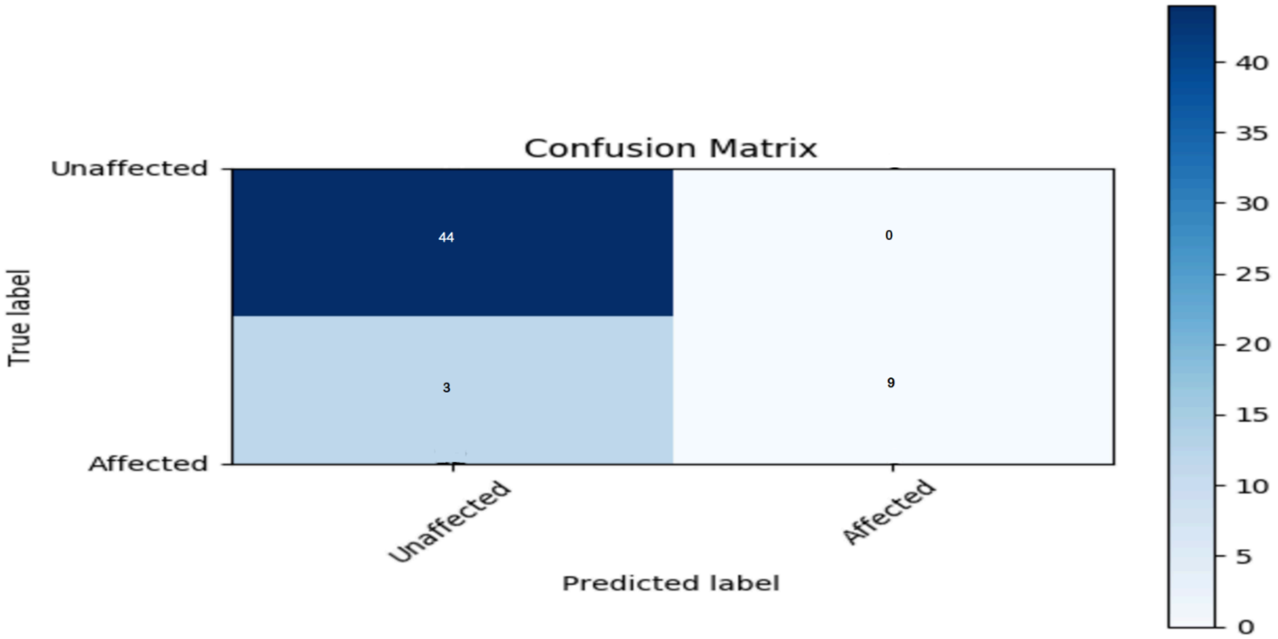

Confusion Matrix
True label
Unaffected
Affected
44
0
3
9
Unaffected
Affected
Predicted label
40
35
30
25
20
15
10
5
0

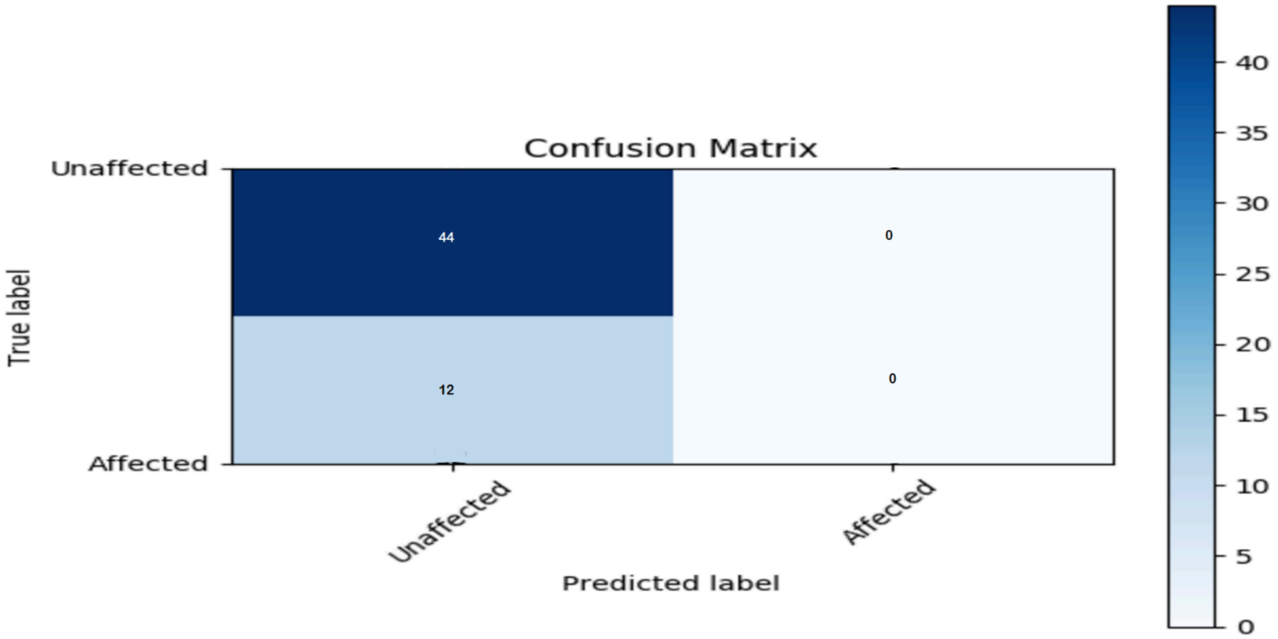

Confusion Matrix
True label
Unaffected
Affected
Predicted label
Unaffected
Affected
43
1
12
0
40
35
30
25
20
15
10
5
0


Confusion Matrix
True label
Unaffected
Affected
Predicted label
Unaffected
Affected
43
1
12
0
40
35
30
25
20
15
10
5
0


Confusion Matrix
True label
Unaffected
Affected
41
3
10
2
Unaffected
Affected
Predicted label
40
35
30
25
20
15
10
5
0


Confusion Matrix
True label
Unaffected
Affected
44
0
12
0
Unaffected
Affected
Predicted label
40
35
30
25
20
15
10
5
0


Confusion Matrix
Unaffected
Affected
True label
44
0
3
9
Unaffected
Affected
Predicted label
40
35
30
25
20
15
10
5
0


Confusion Matrix
Unaffected
Affected
True label
44
0
12
0
Unaffected
Affected
Predicted label
40
35
30
25
20
15
10
5
0

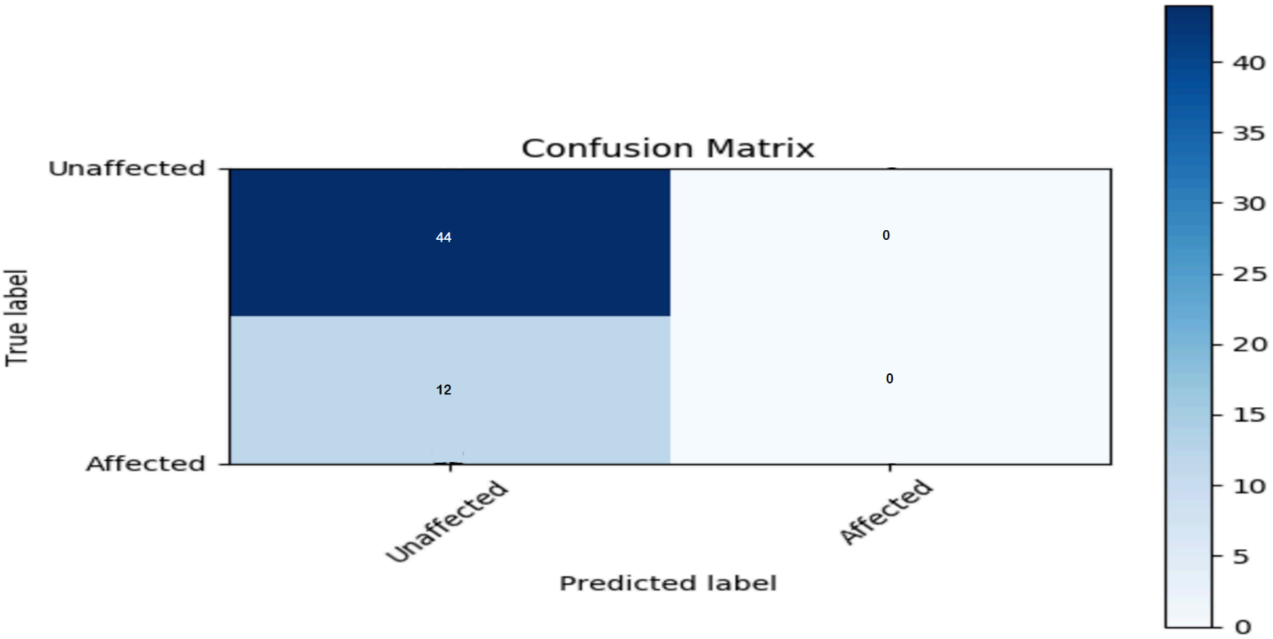


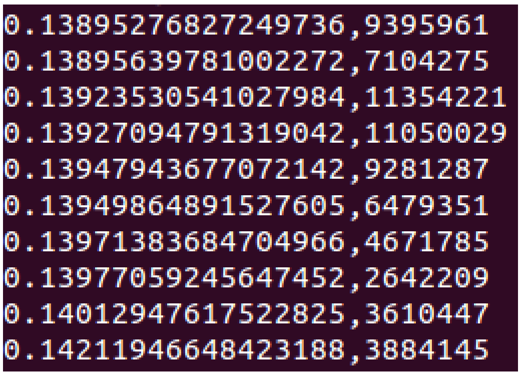
**Figure 11: Confusion Matrices top to bottom for ExhaustiveDNN, Logistic Regression, AdaBoost, GradientBoost, Naïve Bayes, Bagging, Support Vector, Random Forest, ExtraTreesClassifier for MHHRTATT phenotype GTEx V7 Pilot dataset.**

| PDValues | ColumnName |
|---|---|
| 0.690258855 | 289 |
| 0.690258855 | 232_I |
| 0.690258855 | 53 |
| 0.690258855 | 9 |
| 0.690258855 | 3 |
| 0.690258855 | 288_I |
| 0.690258855 | 233 |
| 0.690258855 | 377 |
| 0.690258855 | 267 |
| 0.690258855 | 259 |
| 0.690258855 | 145 |
| 0.690258855 | 392_I |
| 0.690258855 | 214_C |
| 0.690258855 | 356_I |
| 0.690258856 | 226_I |

| | |
|---|---|
| 0.690258856 | 234_I |
| 0.690258857 | 196_C |
| 0.690258857 | 380_T |
| 0.690258857 | 68_I |
| 0.690258858 | 296_I |
| 0.690258858 | 396_T |
| 0.690258858 | 110_A |
| 0.690258858 | 206_G |
| …<br>…<br>…<br>0.711782557 | 395 |
| 0.712538851 | 137 |
| 0.712730046 | 81 |
| 0.712760458 | 399 |
| 0.713352305 | 121 |
| 0.713557698 | 183 |
| 0.714217146 | 197 |
| 0.714282215 | 349 |
| 0.714400629 | 389 |
| 0.716420812 | 149 |
| 0.719423753 | 329 |
| 0.724220414 | 113 |
| 0.72980916 | 187 |

**Table 4**: **List of partial score and the corresponding column explanatory genomic variant variable**

| PDValues | ColumnName |
|---|---|



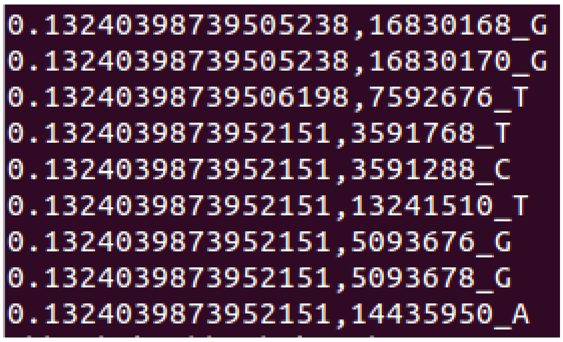
```
0.13895276827249736,9395961
0.13895639781002272,7104275
0.13923530541027984,11354221
0.13927094791319042,11050029
0.13947943677072142,9281287
0.13949864891527605,6479351
0.13971383684704966,4671785
0.13977059245647452,2642209
0.14012947617522825,3610447
0.14211946648423188,3884145
```

| **GTEx Pilot 5M.PED.MAP File ROW Number** | **Genotype** |
|---|---|
| 2348991 | Chromosome 9 position 95811874 and variant Id P1_M_061510_9_203_M |
| 1776069 | Chromosome 6 variant Id P1_M_061510_6_987_P position 162112867 |

| | |
|---|---|
| 2838556 | Chromosome 12 variant Id P1_M_061510_12_59_P genomic position 5223453 |
| 2762508 | Chromosome 11 variant Id P1_M_061510_11_420_M genomic position  93911243 |
| 2320322 | Chromosome 9 variant Id P1_M_061510_9_163_M genomic position 78004294 |
| 1619838 | Chromosome 6 variant Id P1_M_061510_6_181_P genomic position 48930947 |
| 1167947 | Chromosome 4 variant Id P2_M_061510_4_715_M genomic position 137617593 |
| 660553 | Chromosome 2 variant Id P1_M_061510_2_509_P genomic position 233364549 |
| 902612 | Chromosome 3 variant Id P1_M_061510_3_309_M genomic position 145931899 |
| 971037 | Chromosome 3 variant Id P1_M_061510_3_402_P genomic position 192063195 |

**Table 5**: **List of top 10 partial score as per the logistic regression and the corresponding column explanatory genomic variant variable column number as per the GTEx V7 pilot data numbering. The corresponding genomic co-ordinates can be found using the .MAP and .PED file information from GTEx dataset as described in 'Optimized Feature set for MHHRTATT biomarkers' section and are also shown in the table**

| **PDValues** | **ColumnName** |
|---|---|



```
0.13240398739505238,16830168_G
0.13240398739505238,16830170_G
0.13240398739506198,7592676_T
0.1324039873952151,3591768_T
0.1324039873952151,3591288_C
0.1324039873952151,13241510_T
0.1324039873952151,5093676_G
0.1324039873952151,5093678_G
0.1324039873952151,14435950_A
```

| GTEx Pilot 5M.PED.MAP File ROW Number | Genotype |
|---|---|
| 4207542 | Chromosome 23 variant Id kgp30994055 genomic position 52587347 |
| 4207543 | Chromosome 23 variant Id kgp31134917 genomic position 52588392 |
| 1898169 | Chromosome 7 Variant Id kgp11290556 genomic position 70226068 |
| 897942 | Chromosome 3 Variant Id kgp5923265 genomic position 142797398 |
| 897822 | Chromosome 3 Variant Id kgp18185020 genomic position 142711709 |
| 3310378 | Chromosome 14 Variant Id kgp28093020 genomic position 97615238 |
| 1273419 | Chromosome 5 Variant Id kgp22643217 genomic position 13809129 |
| 1273420 | Chromosome 5 Variant Id kgp22679345 genomic position 13809146 |
| 3608988 | Chromosome 17 Variant Id kgp5104948 genomic position 4991686 |

**Table 6**: **List of bottom 10 partial score as per the logistic regression and the corresponding column explanatory genomic variant variable column number as per the GTEx V7 pilot data numbering. The corresponding genomic co-ordinates can be found using the .MAP and .PED file from GTEx dataset information as described in 'Optimized Feature set for MHHRTATT biomarkers' section and are also shown in the table.**