Bio-marker Detection for Type 1 and Type 2 Diabetes using Deep Learning

Abdul Azeez K, Aravindan M, Adhirai Nandhini A, Tejeswinee K

***Abstract*— In developing countries like India, non-communicable diseases such as diabetes have already replaced communicable diseases as the major cause of death. According to data from the International Diabetes Federation(IDF) and 14 cohort studies (representing more than 60 percent of the world population with type 2 diabetes), researchers estimated the burden of type 2 diabetes in 221 countries and territories between 2018 and 2030 and IDF pegs the number of patients with diabetes in India at 65.1 million (it was 50.8 million in 2010) and the number is expected to cross 100 million by 2030 .The number of adults with type 2 diabetes is expected to rise over the next 12 years due to ageing, urbanization, and associated changes in diet and physical activity. In this paper the authors focus on diagnosis of diabetes using the various machine learning techniques of data mining. And, authors have compared various classification techniques such as Naive Bayes, KNN, Adaboost, SVM, Decision tree algorithm J48,Random forest. And three well-performing feature selection algorithms namely, Correlation Feature Subset Selection (CFS), Information Gain(IG) and Gain Ratio (GR) are used to obtain the optimal features contributing to the diabetes disease. Further, Incremental Feature Selection(IFS) techniques are applied to further reduce the feature subset from the optimal feature set.**

***Index Terms* — Incremental Feature Selection, Correlation Feature Subset Selection (CFS), Information Gain(IG) , Gain Ratio (GR)**

I. INTRODUCTION

Diabetes mellitus is a complicated illness poignant, causing each tissue and organ system, with metabolic ramifications extending way on the far side impaired aldohexose metabolism. Biomarkers could replicate the presence and severity of hyperglycemia (i.e. polygenic disorder itself), or the presence and severity of the tube-shaped structure complications of polygenic disorder.

Type 1: Type 1 diabetes can occur at any age, but is most commonly diagnosed from infancy to the late 30s. With this type of diabetes, a person's pancreas produces no insulin. It occurs when the body's own defence system (the immune system) attacks and destroys the insulin-producing cells in the pancreas.

Type 2: Type 2 diabetes is by far the most common type of diabetes - is becoming more common among young people due to lifestyle. People with type 2 diabetes either don't make enough insulin or don't make insulin that the body can use properly. Eventually, the pancreas can wear out from producing extra insulin, and it may start making less and less.

II. LITERATURE SURVEY

Yanqiu Wang; Zhi-Ping Li, proposed a gene coexpression network framework to identify the genes with different coexpression patterns in control and disease. The phenotypic indicators are significantly associated with the outcomes of diabetes and then serve as biomarkers. And they have employed an SVM-based classifier to evaluate the classification of these selected genes for their distinguishing power of classifying different states**[1]**.

Ansam Al-Sabti; Mohamed Zaibi; Sabah Jassim, have used and Integrative Omics approach to identify the sub-network in Diabetes mellitus of Type 2 using a novel network based biomarker identification method to distinguish the disease state from normal state by integrating expression and network datasets. And there proposed approach proves the facility of identifying an accurate biomarker for Type 2 diabetes disorder prognosis due to the including of important topological and network information in scoring the resulting pathways**[2]**.

Azian Azamimi Abdullah ;Nurul Sakinah Fadil ; Wan Khairunizam, have developed a Fuzzy expert system for diagnosis of Diabetes by simple GUI layout design, where user enters his data such as Name, Age, Height, Weight, WC (Waist Circumfernce), WHR (Waist to Hip Ratio) for both men and women**[3]**.

Neeru Lalka; Sushma Jain used same approach as **[3]** and they proposed a method of Insulin Dosage Control (IDC) which enables capturing accurate precison level of probability to recommend the usage of IDC to Type 1 diabetes.And the probability or severity of diabetes in person, lies between 0 and 1**[4]**.

K. Zarkogianni, E. Litsa, K. Mitsis, P. Wu, C. D. Kaddi, C. Cheng, M.D. Wang; and K.S. Nikita have discussed a review of emerging technologies for management of diabetes. They have founded some existing technologies such as Google Smart Lens, iQuickIt Saliva Analyzer, and Abbott developed Freestyle. And they have concluded there review as, Enhanced integration of patient data through the development of multiscale and multilevel physiological models can generate new clinical knowledge and contribute to a more effective personalized diabetes care approach**[5]**.

Abdul Azeez K, Computer Science and Engineering, Rajalakshmi Engineering College, Chennai, India
Aravindan M, Computer Science and Engineering, Rajalakshmi Engineering College, Chennai, India
Adhirai Nandhini A, Computer Science and Engineering, Rajalakshmi Engineering College, Chennai, India
Tejeswinee K, Assistant Professor**,** Dept. of Computer Science and Engineering, Rajalakshmi Engineering College, Chennai, India

Sidong Wei; Xuejiao Zhao; Chunyan Miao in there papers they have used Machine Learning Techniques such as Deep Neural Network, Support Vector Machine (SVM) etc. to identify diabetes and they have used Pima Indian Diabetes data set**[6]**.

Many nice results are created using numerous algorithms. For example, **Asha** and her colleagues used a hybrid model of Genetic algorithmic program and Back Propagation Network to identifiy polygenic disorder **[7]**. They particularly targeted on adopting the rule on some specific input data and reached 84.7% on the known inputs. **Kayaer's** team used GRNN technique **[9]** to spot polygenic disorder. They mentioned a way to build the network and had a similar result as **Gail A. Carpenter** and his cluster has created, which used a really difficult ARTMAP-IC network **[8]**. The technique Kayaer used was abundant simplified compared to Gail's, however it absolutely was still a posh one relevance the dimensions of the data set. From all those researches we will see that all of them explored diabetes identification through one specific methodology, and modified and improved it to its best or approximate best.

III. Proposed Computational Framework

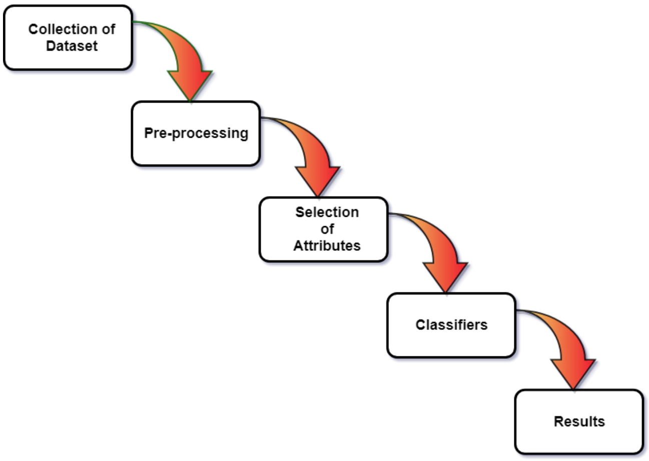
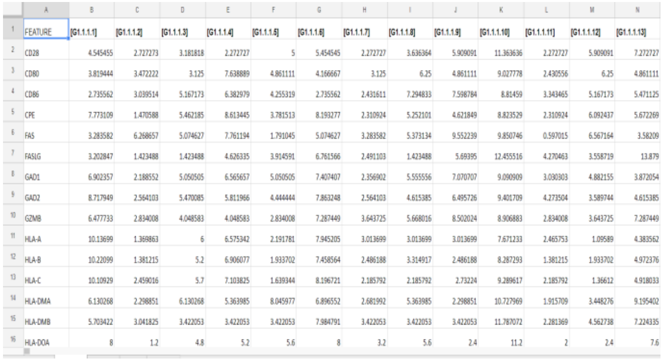


Fig. 1

A. Dataset Generation

The dataset consists of structural and physicochemical properties of proteins related to the genes of type1 and type2 diabetes. It is in the form of CSV(Comma standard value) shown in Fig. 2 .The dataset contains 43 gene sample for Type 1, 47 gene sample for Type 2 Diabetes and 2 gene samples are common to both Type 1 and 2.



FEATURE	[G1.1.1.1]	[G1.1.1.2]	[G1.1.1.3]	[G1.1.1.4]	[G1.1.1.5]	[G1.1.1.6]	[G1.1.1.7]	[G1.1.1.8]	[G1.1.1.9]	[G1.1.1.10]	[G1.1.1.11]	[G1.1.1.12]	[G1.1.1.13]
CD28	4.545455	2.727273	3.181818	2.272727	5	5.454545	2.272727	3.636364	5.909091	11.363636	2.272727	5.909091	7.272727
CD80	3.819444	3.472222	3.125	7.638889	4.861111	4.166667	3.125	6.25	4.861111	9.027778	2.430556	6.25	4.861111
CD86	2.735562	3.039514	5.167173	6.382979	4.255319	2.735562	2.431611	7.294833	7.598784	8.81459	3.343465	5.167173	5.471125
CPE	7.773109	1.470588	5.462185	8.613445	3.781513	8.193277	2.310924	5.252101	4.621849	8.823529	2.310924	6.092437	5.672269
FAS	3.283582	6.268657	5.074627	7.761194	1.791045	5.074627	3.283582	5.373134	9.552239	9.850746	0.597015	6.567164	3.58209
FASLG	3.202847	1.423488	1.423488	4.626335	3.914591	6.761566	2.491103	1.423488	5.69395	12.455516	4.270463	3.558719	13.879
GAD1	6.902357	2.188552	5.050505	6.565657	5.050505	7.407407	2.356902	5.555556	7.070707	9.090909	3.030303	4.882155	3.872054
GAD2	8.717949	2.564103	5.470085	5.811966	4.444444	7.863248	2.564103	4.615385	6.495726	9.401709	4.273504	3.589744	4.615385
GZMB	6.477733	2.834008	4.048583	4.048583	2.834008	7.287449	3.643725	5.668016	8.502024	8.906883	2.834008	3.643725	7.287449
HLA-A	10.13699	1.369863	6	6.575342	2.191781	7.945205	3.013699	3.013699	3.013699	7.671233	2.465753	1.09589	4.383562
HLA-B	10.22099	1.381215	5.2	6.906077	1.933702	7.458564	2.486188	3.314917	2.486188	8.287293	1.381215	1.933702	4.972376
HLA-C	10.10929	2.459016	5.7	7.103825	1.639344	8.196721	2.185792	2.185792	2.73224	9.289617	2.185792	1.36612	4.918033
HLA-DMA	6.130268	2.298851	6.130268	5.363985	8.045977	6.896552	2.681992	5.363985	2.298851	10.727969	1.915709	3.448276	9.195402
HLA-DMB	5.703422	3.041825	3.422053	3.422053	3.422053	7.984791	3.422053	3.422053	3.422053	11.787072	2.281369	4.562738	7.224335
HLA-DOA	8	1.2	4.8	5.2	5.6	8	3.2	5.6	2.4	11.2	2	2.4	7.6

Fig. 2

B. Feature Selection

All the genes have 1437 protein properties each. The properties are broadly classified as G1 to G9 feature descriptors, Fig. 5.1, 5.2 and 5.3 show the parameter for each feature group. To find the specific genes that are most contributing to Type 1 and Type 2 diabetes, feature selection methodologies were investigated. Three mechanisms were applied to find the optimal feature set- Correlation Feature Subset Selection (CFS), Information Gain (IG) and Gain Ratio(GR).

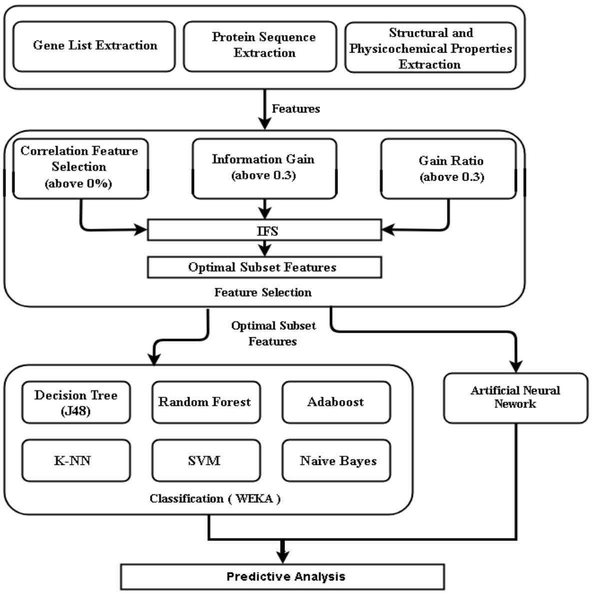


Fig. 3

The feature subsets obtained post feature selection were fed as input to the classification phase wherein classifiers viz, Support Vector Machine (SVM), Random Forest, Decision Tree (J48), Naive Bayes, Adaboost, k-NN were employed and their accuracy in predicting the correct diagnostic class was measured. Fig. 3 shows the proposed framework for feature selection while Fig. 4 depicts the various data mining algorithms that were investigated on the extracted feature subset.

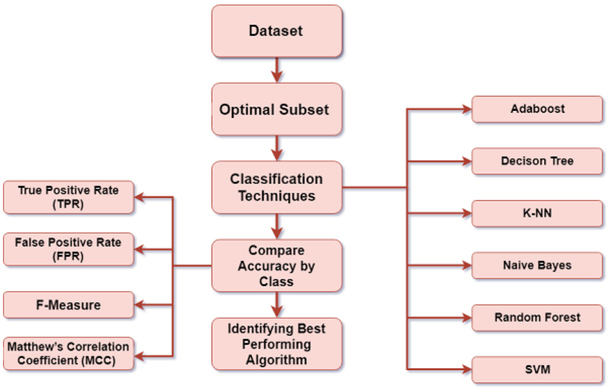
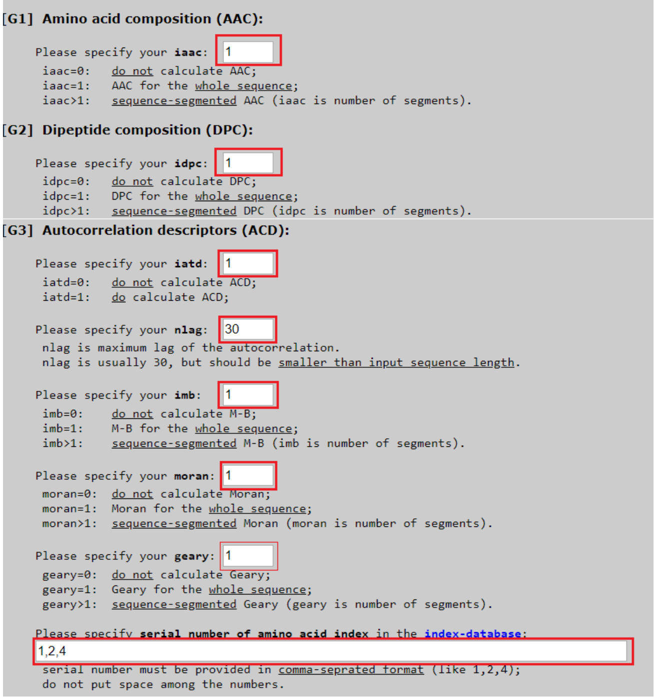

Dataset
Optimal Subset
Classification Techniques
Compare Accuracy by Class
Identifying Best Performing Algorithm
True Positive Rate (TPR)
False Positive Rate (FPR)
F-Measure
Matthew's Correlation Coefficient (MCC)
Adaboost
Decison Tree
K-NN
Naive Bayes
Random Forest
SVM

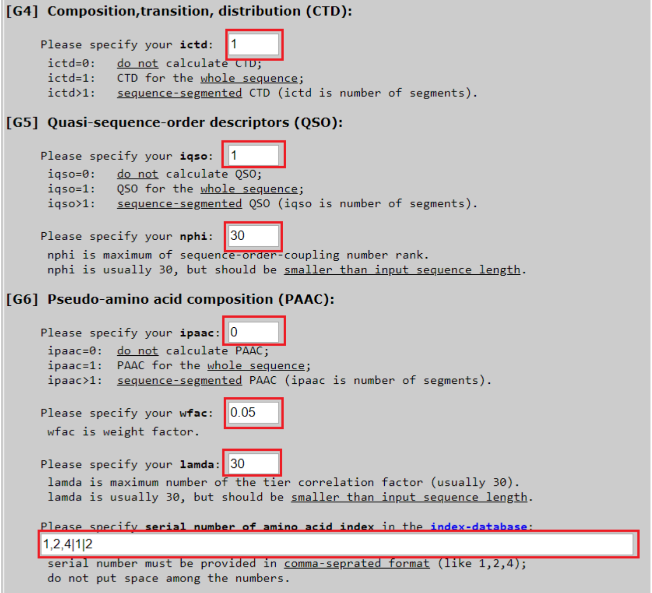
Fig. 4

The accuracy of each class Type 1 and Type 2 diabetes is measured by True Positive Rate (TPR), False Positive Rate (FPR), F-Measure and Mathew's Correlation Coefficient (MCC).



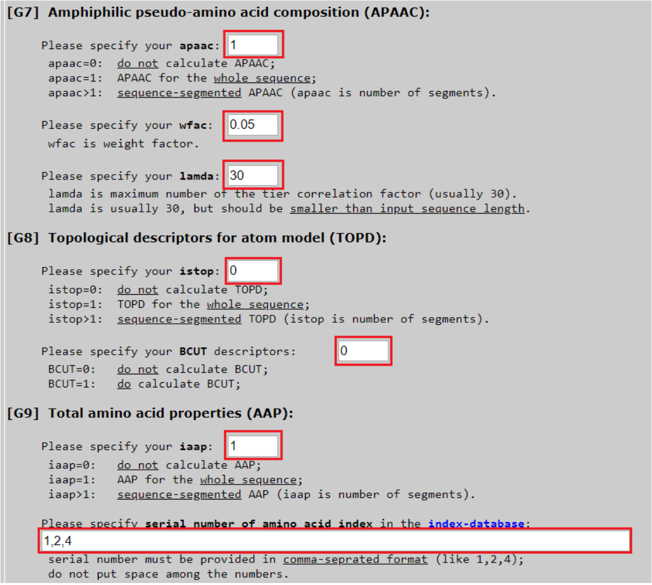
```
[G1] Amino acid composition (AAC):

    Please specify your iaac: 1
      iaac=0:   do not calculate AAC;
      iaac=1:   AAC for the whole sequence;
      iaac>1:   sequence-segmented AAC (iaac is number of segments).

[G2] Dipeptide composition (DPC):

    Please specify your idpc: 1
      idpc=0:   do not calculate DPC;
      idpc=1:   DPC for the whole sequence;
      idpc>1:   sequence-segmented DPC (idpc is number of segments).
[G3] Autocorrelation descriptors (ACD):

    Please specify your iatd: 1
      iatd=0:   do not calculate ACD;
      iatd=1:   do calculate ACD;

    Please specify your nlag: 30
      nlag is maximum lag of the autocorrelation.
      nlag is usually 30, but should be smaller than input sequence length.

    Please specify your imb:  1
      imb=0:    do not calculate M-B;
      imb=1:    M-B for the whole sequence;
      imb>1:    sequence-segmented M-B (imb is number of segments).

    Please specify your moran: 1
      moran=0:  do not calculate Moran;
      moran=1:  Moran for the whole sequence;
      moran>1:  sequence-segmented Moran (moran is number of segments).

    Please specify your geary: 1
      geary=0:  do not calculate Geary;
      geary=1:  Geary for the whole sequence;
      geary>1:  sequence-segmented Geary (geary is number of segments).

    Please specify serial number of amino acid index in the index-database:
    1,2,4
      serial number must be provided in comma-seprated format (like 1,2,4);
      do not put space among the numbers.
```

Fig. 5.1



```
[G4] Composition,transition, distribution (CTD):

    Please specify your ictd: 1
      ictd=0:   do not calculate CTD;
      ictd=1:   CTD for the whole sequence;
      ictd>1:   sequence-segmented CTD (ictd is number of segments).

[G5] Quasi-sequence-order descriptors (QSO):

    Please specify your iqso: 1
      iqso=0:   do not calculate QSO;
      iqso=1:   QSO for the whole sequence;
      iqso>1:   sequence-segmented QSO (iqso is number of segments).

    Please specify your nphi: 30
      nphi is maximum of sequence-order-coupling number rank.
      nphi is usually 30, but should be smaller than input sequence length.

[G6] Pseudo-amino acid composition (PAAC):

    Please specify your ipaac: 0
      ipaac=0:  do not calculate PAAC;
      ipaac=1:  PAAC for the whole sequence;
      ipaac>1:  sequence-segmented PAAC (ipaac is number of segments).

    Please specify your wfac: 0.05
      wfac is weight factor.

    Please specify your lamda: 30
      lamda is maximum number of the tier correlation factor (usually 30).
      lamda is usually 30, but should be smaller than input sequence length.

    Please specify serial number of amino acid index in the index-database:
    1,2,4|1|2
      serial number must be provided in comma-seprated format (like 1,2,4);
      do not put space among the numbers.
```

Fig 5.2



```
[G7] Amphiphilic pseudo-amino acid composition (APAAC):

    Please specify your apaac: 1
      apaac=0:  do not calculate APAAC;
      apaac=1:  APAAC for the whole sequence;
      apaac>1:  sequence-segmented APAAC (apaac is number of segments).

    Please specify your wfac: 0.05
      wfac is weight factor.

    Please specify your lamda: 30
      lamda is maximum number of the tier correlation factor (usually 30).
      lamda is usually 30, but should be smaller than input sequence length.

[G8] Topological descriptors for atom model (TOPD):

    Please specify your istop: 0
      istop=0:  do not calculate TOPD;
      istop=1:  TOPD for the whole sequence;
      istop>1:  sequence-segmented TOPD (istop is number of segments).

    Please specify your BCUT descriptors:     0
      BCUT=0:   do not calculate BCUT;
      BCUT=1:   do calculate BCUT;

[G9] Total amino acid properties (AAP):

    Please specify your iaap: 1
      iaap=0:   do not calculate AAP;
      iaap=1:   AAP for the whole sequence;
      iaap>1:   sequence-segmented AAP (iaap is number of segments).

    Please specify serial number of amino acid index in the index-database:
    1,2,4
      serial number must be provided in comma-seprated format (like 1,2,4);
      do not put space among the numbers.
```

Fig. 5.3

IV. EXPERIMENTAL RESULTS

The investigation of existing techniques revealed the importance of selecting important features for classification. 10-fold cross validation was employed to measure the performance of the data mining algorithms. Two performance parameters were identified to rank the algorithms.

A. Accuracy

The degree to which the result of a measurement, calculation, or specification conforms to the correct value or a standard.

$$\text{Recall} = \frac{tp}{tp+fn}$$

$$\text{Precision} = \frac{tp}{tp+fp}$$

$$\text{Accuracy} = \frac{tp+tn}{tp+tn+fp+fn}$$

F-measure

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

This is also known as the F_1 measure, because recall and precision are evenly weighted.

B. Matthews' Correlation Co-efficient (MCC)

It's the measure of the quality of binary (two-class) classifications. It takes into account true and false positives and negatives .

$$MCC = \frac{(TPXTN)-(FPXFN)}{\sqrt{(TP+FN)X(TN+FP)X(TP+FP)X(TN+FN)}}$$

TP = true positives: number of examples predicted positive that are actually positive; FP = false positives: number of examples predicted positive that are actually negative; TN = true negatives: number of examples predicted negative that are actually negative; FN = false negatives: number of examples predicted negative that are actually positive.

Mean absolute error (MAE) : MAE measures the average magnitude of the errors in a set of forecasts, without considering their direction. It measures accuracy for continuous variables. The MAE is the average over the verification sample of the absolute values of the differences between forecast and the corresponding observation. The MAE is a linear score which means that all the individual differences are weighted equally in the average;

Root mean squared error (RMSE) : RMSE is a quadractic scoring rule which measures the average magnitude of the error. The difference between forecast and corresponding observed values are each squared and then averaged over the sample. Finally, the square root of the average is taken. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. This means the RMSE is most useful when large errors are particularly undesirable.

Kappa statistic 0.9403 <- agreement of prediction with true class. Mean absolute error 0.0309 <- not squared before averaging. Root mean squared error 0.1493 <- squared before averaging, so large errors have more influence. Relative absolute error 6.9047 % <- Relative values are ratios, and have no units.

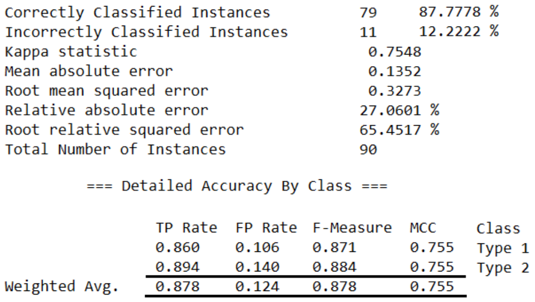
Before feature selection, the dataset consisted of 1437 attributes and by extracting the gene set using GSEC we get 90 attributes of datasets. (43gene sample of type1 and 47 gene sample of type2).

Investigation was carried out using WEKA 3.4 tool open source data mining suite. Once the dataset was pre-processed, feature selection techniques were explored.

A threshold greater than or equal to 0.3 was chosen for Information Gain (IG) and greater than or equal to 0.3 was chosen for Gain Ratio (GR) and 0% above for Correlation Feature Subset Se. Correlated Feature Subset Selection (CFS) is an automated method that uses Best-First Search strategy to identify and narrow down to the optimal feature subset. The CFS subset evaluation algorithm extracted a subset containing 153 attributes. The output generated three subsets, one for each of the above mentioned mechanisms.

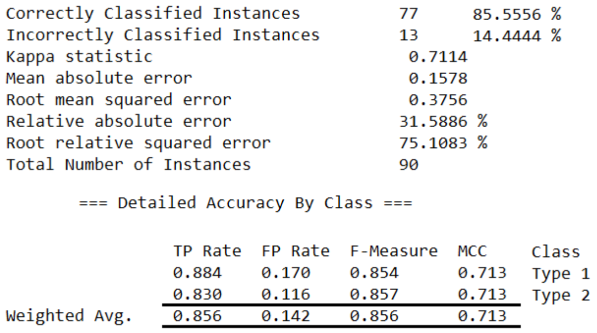
All the six classification algorithms were implemented and their accuracy was measured. Their results are shown below.

Adaboost



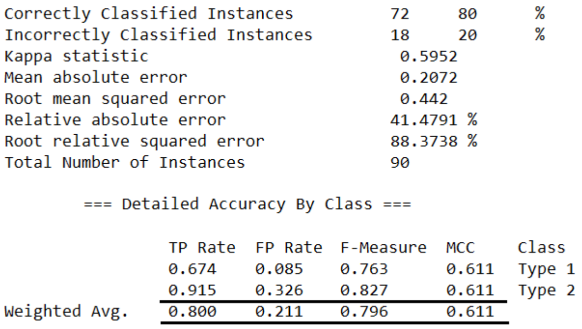
```
Correctly Classified Instances          79         87.7778 %
Incorrectly Classified Instances        11         12.2222 %
Kappa statistic                          0.7548
Mean absolute error                      0.1352
Root mean squared error                  0.3273
Relative absolute error                 27.0601 %
Root relative squared error             65.4517 %
Total Number of Instances               90

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  F-Measure  MCC     Class
                 0.860    0.106    0.871      0.755   Type 1
                 0.894    0.140    0.884      0.755   Type 2
Weighted Avg.    0.878    0.124    0.878      0.755
```

Decision Tree (J48)



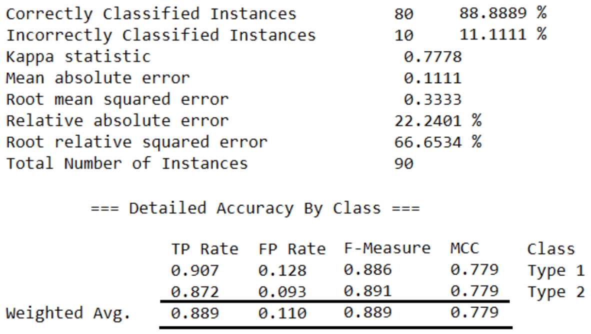
```
Correctly Classified Instances          77         85.5556 %
Incorrectly Classified Instances        13         14.4444 %
Kappa statistic                          0.7114
Mean absolute error                      0.1578
Root mean squared error                  0.3756
Relative absolute error                 31.5886 %
Root relative squared error             75.1083 %
Total Number of Instances               90

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  F-Measure  MCC     Class
                 0.884    0.170    0.854      0.713   Type 1
                 0.830    0.116    0.857      0.713   Type 2
Weighted Avg.    0.856    0.142    0.856      0.713
```

K-NN



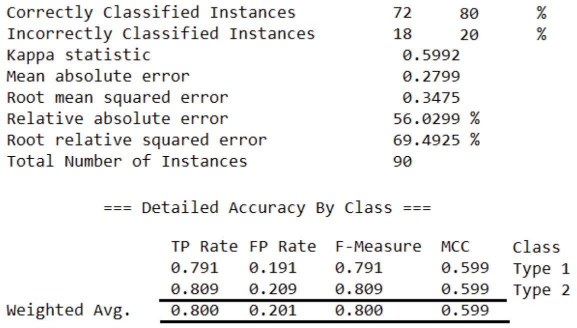
```
Correctly Classified Instances          72         80      %
Incorrectly Classified Instances        18         20      %
Kappa statistic                          0.5952
Mean absolute error                      0.2072
Root mean squared error                  0.442
Relative absolute error                 41.4791 %
Root relative squared error             88.3738 %
Total Number of Instances               90

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  F-Measure  MCC     Class
                 0.674    0.085    0.763      0.611   Type 1
                 0.915    0.326    0.827      0.611   Type 2
Weighted Avg.    0.800    0.211    0.796      0.611
```

Naive Bayes



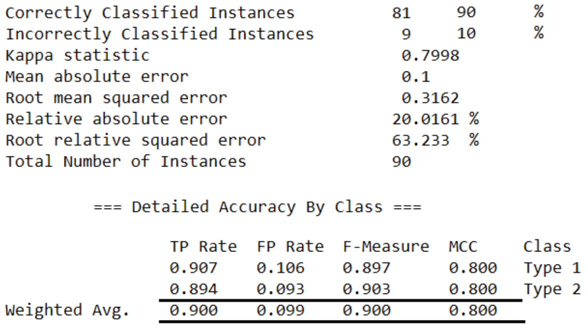
```
Correctly Classified Instances          80         88.8889 %
Incorrectly Classified Instances        10         11.1111 %
Kappa statistic                          0.7778
Mean absolute error                      0.1111
Root mean squared error                  0.3333
Relative absolute error                 22.2401 %
Root relative squared error             66.6534 %
Total Number of Instances               90

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  F-Measure  MCC     Class
                 0.907    0.128    0.886      0.779   Type 1
                 0.872    0.093    0.891      0.779   Type 2
Weighted Avg.    0.889    0.110    0.889      0.779
```

Random Forest



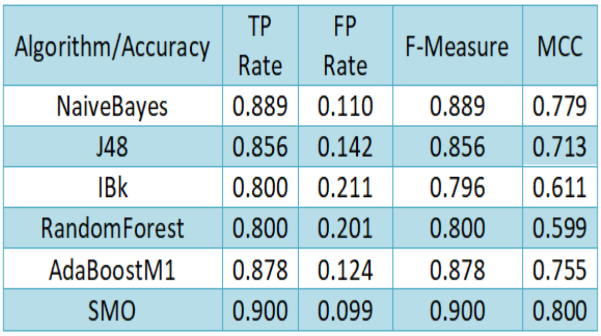
```
Correctly Classified Instances          72         80      %
Incorrectly Classified Instances        18         20      %
Kappa statistic                          0.5992
Mean absolute error                      0.2799
Root mean squared error                  0.3475
Relative absolute error                 56.0299 %
Root relative squared error             69.4925 %
Total Number of Instances               90

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  F-Measure  MCC     Class
                 0.791    0.191    0.791      0.599   Type 1
                 0.809    0.209    0.809      0.599   Type 2
Weighted Avg.    0.800    0.201    0.800      0.599
```

Support Vector Machine (SVM)



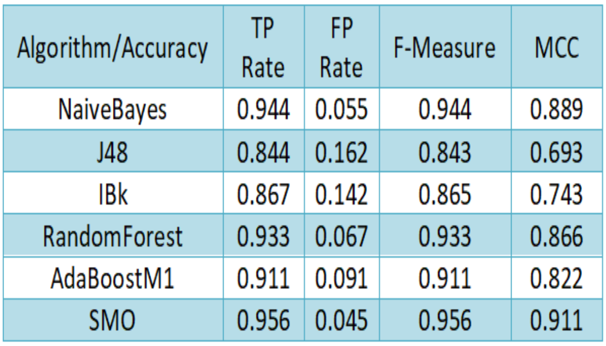
```
Correctly Classified Instances          81         90      %
Incorrectly Classified Instances         9         10      %
Kappa statistic                          0.7998
Mean absolute error                      0.1
Root mean squared error                  0.3162
Relative absolute error                 20.0161 %
Root relative squared error             63.233  %
Total Number of Instances               90

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  F-Measure  MCC     Class
                 0.907    0.106    0.897      0.800   Type 1
                 0.894    0.093    0.903      0.800   Type 2
Weighted Avg.    0.900    0.099    0.900      0.800
```

The result of each algorithm is tabulated in Table 1 for full Dataset and the subset for each CFS, IG, GR are tabulated Table 2, Table 3 and Table 4 respectively. And it is visually represented in the form of Bar Chart as Chart 1 for Pre-feature selection and Chart 2 for Post-feature selection (CFS), Chart 3 for Post-feature Selection(IG) and Chart 4 for Post-feature selection(GR).



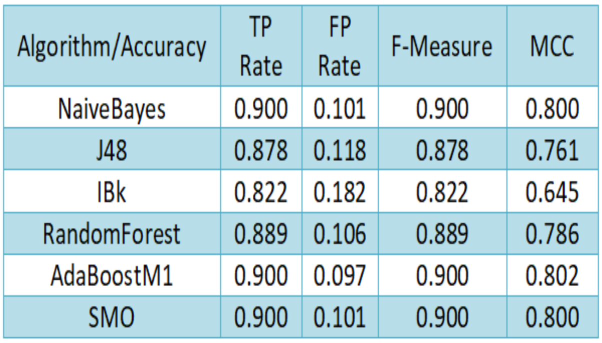
Algorithm/Accuracy	TP Rate	FP Rate	F-Measure	MCC
NaiveBayes	0.889	0.110	0.889	0.779
J48	0.856	0.142	0.856	0.713
IBk	0.800	0.211	0.796	0.611
RandomForest	0.800	0.201	0.800	0.599
AdaBoostM1	0.878	0.124	0.878	0.755
SMO	0.900	0.099	0.900	0.800

Table 1



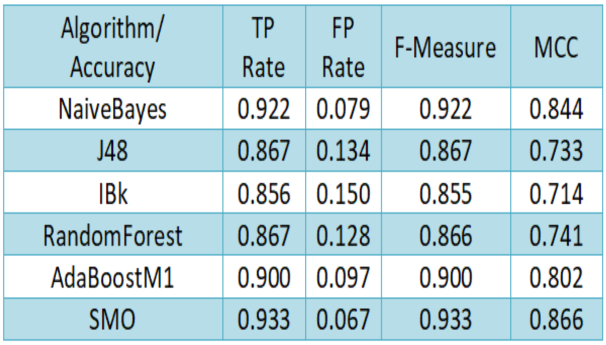
Algorithm/Accuracy	TP Rate	FP Rate	F-Measure	MCC
NaiveBayes	0.944	0.055	0.944	0.889
J48	0.844	0.162	0.843	0.693
IBk	0.867	0.142	0.865	0.743
RandomForest	0.933	0.067	0.933	0.866
AdaBoostM1	0.911	0.091	0.911	0.822
SMO	0.956	0.045	0.956	0.911

Table 2



Algorithm/Accuracy	TP Rate	FP Rate	F-Measure	MCC
NaiveBayes	0.900	0.101	0.900	0.800
J48	0.878	0.118	0.878	0.761
IBk	0.822	0.182	0.822	0.645
RandomForest	0.889	0.106	0.889	0.786
AdaBoostM1	0.900	0.097	0.900	0.802
SMO	0.900	0.101	0.900	0.800

Table 3



Algorithm/ Accuracy	TP Rate	FP Rate	F-Measure	MCC
NaiveBayes	0.922	0.079	0.922	0.844
J48	0.867	0.134	0.867	0.733
IBk	0.856	0.150	0.855	0.714
RandomForest	0.867	0.128	0.866	0.741
AdaBoostM1	0.900	0.097	0.900	0.802
SMO	0.933	0.067	0.933	0.866

Table 4

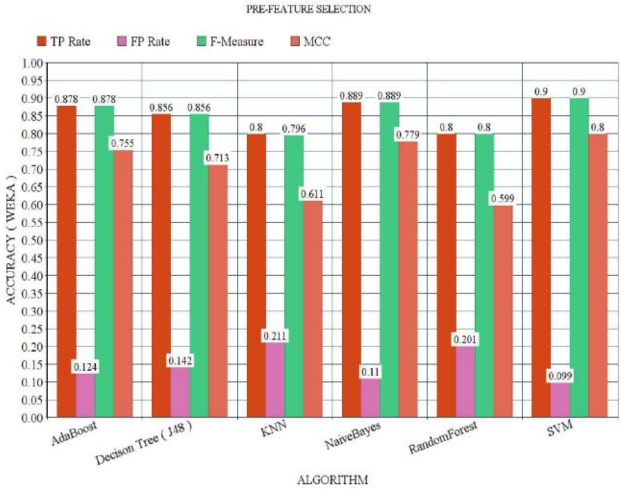


Chart 1

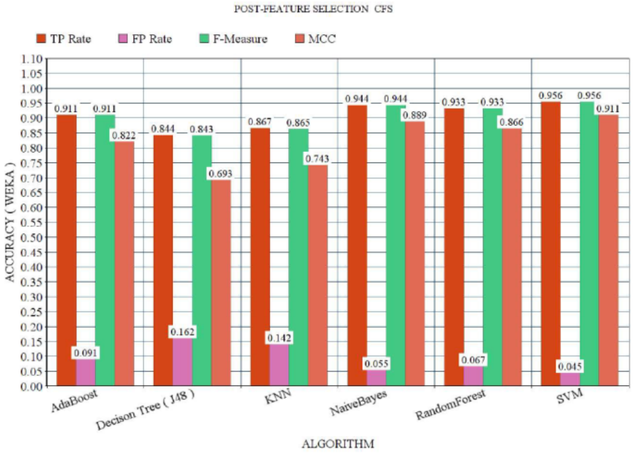


Chart 2

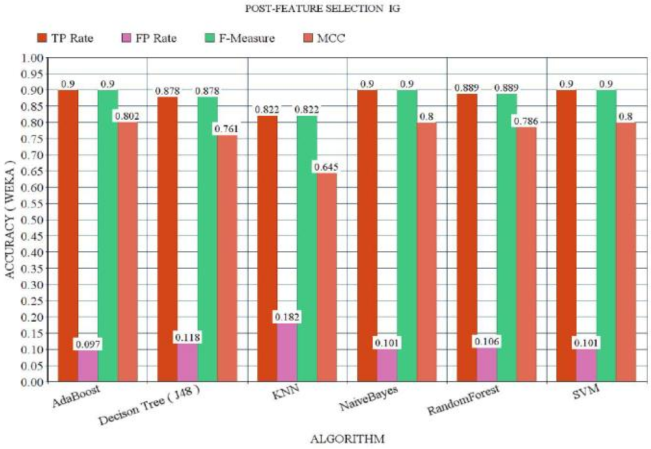


Chart 3

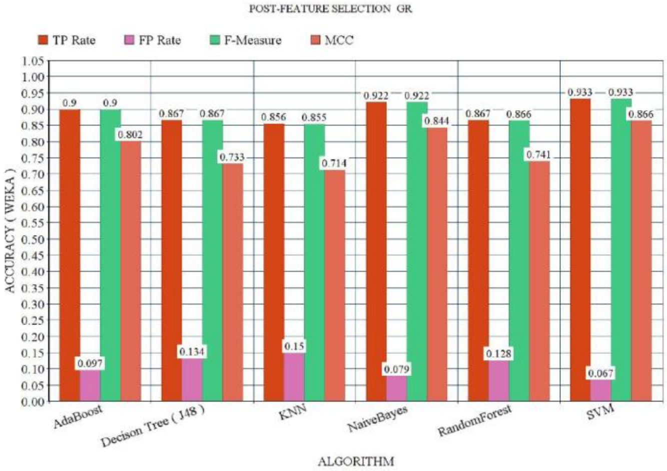
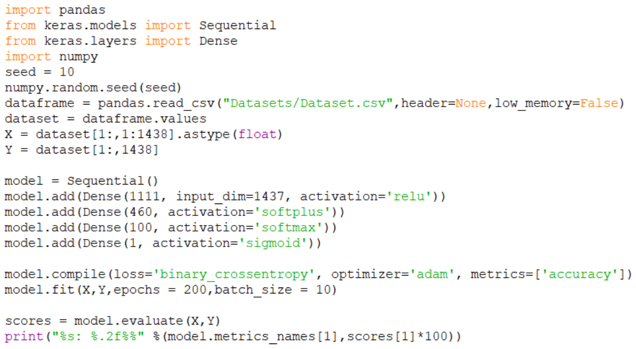


Chart 4

V. Results

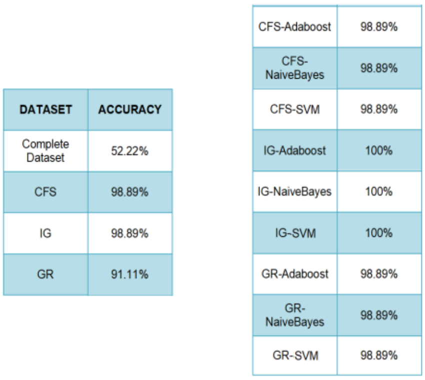
The below mention code is used to evaluate the full dataset and the subset data of CFS , IG , GR and another subset of CFS-Naive Bayes, CFS- Adaboost and CFS-SMO, IG-Naive Bayes, IG- Adaboost and IG-SMO and GR-Naive Bayes, GR-Adaboost and GR-SMO.(Code differs for every subset)

,



```
import pandas
from keras.models import Sequential
from keras.layers import Dense
import numpy
seed = 10
numpy.random.seed(seed)
dataframe = pandas.read_csv("Datasets/Dataset.csv",header=None,low_memory=False)
dataset = dataframe.values
X = dataset[1:,1:1438].astype(float)
Y = dataset[1:,1438]

model = Sequential()
model.add(Dense(1111, input_dim=1437, activation='relu'))
model.add(Dense(460, activation='softplus'))
model.add(Dense(100, activation='softmax'))
model.add(Dense(1, activation='sigmoid'))

model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
model.fit(X,Y,epochs = 200,batch_size = 10)

scores = model.evaluate(X,Y)
print("%s: %.2f%%" %(model.metrics_names[1],scores[1]*100))
```

The evaluate method in the above mentioned code gives an accuracy as shown in below Table 5. And the accuracy is pictorically visualized in the Chart 5 for full dataset and pre-feature selection, and Chart 6 for post-feature selection.



DATASET	ACCURACY
Complete Dataset	52.22%
CFS	98.89%
IG	98.89%
GR	91.11%

CFS-Adaboost	98.89%
CFS-NaiveBayes	98.89%
CFS-SVM	98.89%
IG-Adaboost	100%
IG-NaiveBayes	100%
IG-SVM	100%
GR-Adaboost	98.89%
GR-NaiveBayes	98.89%
GR-SVM	98.89%

Table 5

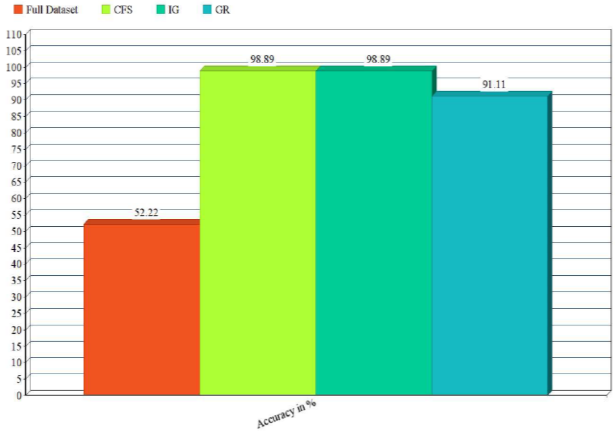


Chart 5

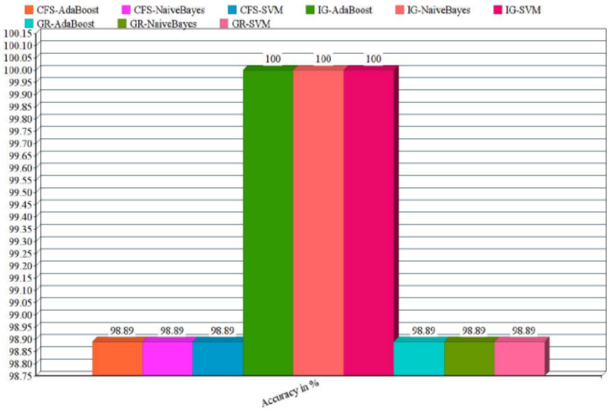


Chart 6

VI. Conclusion

Our case study is able to identify the gene causing Type 1 and Type 2 diabetes. Although in this paper, we have generated a new dataset that consists of genetic information that pertains to theType 1 and Type 2 disease. The dataset, initially extracted from the KEGG database, consisted of 1437 structural and physicochemical protein properties that were extracted from the PROFEAT server. It consisted of 43 gene sample for Type 1, 47 gene sample for Type 2 Diabetes and 2 gene samples are common to both Type 1 and 2.
(1 Missing gene sample in Type 1)

KEGG database, in future has the storage of gene information of all homo-sapiens, which enables to classify the infants having the possibility of diabets in earlier stage itself.

References

[1] Y. Wang and Z. Liu, "Identifying biomarkers of diabetes with gene coexpression networks," 2017 Chinese Automation Congress (CAC), Jinan, 2017, pp. 5283-5286. doi: 10.1109/CAC.2017.8243719

[2] A. Al-Sabti, M. Zaibi and S. Jassim, "An Integrative Omics Approach to Identify Sub-Network Biomarker in Type 2 Diabetes Mellitus," 2017 European Modelling Symposium (EMS), Manchester, 2017, pp. 53-58. doi: 10.1109/EMS.2017.20

[3] A. A. Abdullah, N. S. Fadil and W. Khairunizam, "Development of Fuzzy Expert System for Diagnosis of Diabetes," 2018 International Conference on Computational Approach in Smart Systems Design and Applications (ICASSDA), Kuching, 2018, pp. 1-8. doi: 10.1109/ICASSDA.2018.8477635

[4] N. Lalka and S. Jain, "Fuzzy based expert system for diabetes diagnosis and insulin dosage control," International Conference on Computing, Communication & Automation, Noida, 2015, pp. 262-267. doi: 10.1109/CCAA.2015.7148385

[5] K. Zarkogianni et al., "A Review of Emerging Technologies for the Management of Diabetes Mellitus," in IEEE Transactions on Biomedical Engineering, vol. 62, no. 12, pp. 2735-2749, Dec. 2015. doi: 10.1109/TBME.2015.2470521

[6] S. Wei, X. Zhao and C. Miao, "A comprehensive exploration to the machine learning techniques for diabetes identification," 2018 IEEE 4th World Forum on Internet of Things (WF-IoT), Singapore, 2018, pp. 291-295. doi: 10.1109/WF-IoT.2018.8355130

[7] Karegowda A. G., Manjunath A. S. and Jayaram M. A., "Application of genetic algorithm optimized neural network connection weights for medical diagnosis of pima Indians diabetes," International Journal on Soft Computing, vol. 2. 2, 2011, pp. 15-23.

[8] Carpenter G. A. and Markuzon N., "ARTMAP-IC and medical diagnosis: Instance counting and inconsistent cases," Neural Networks, vol. 11. 2, 1998, pp. 323-336.

[9] Kayaer K and Yildirim T, "Medical diagnosis on Pima Indian diabetes using general regression neural networks," Proceedings of the international conference on artificial neural networks and neural information processing, 2003, pp. 181-184.

[10] https://spu.fem.uniag.sk/cvicenia/ksov/fuskova-ulicna/Data%20mining/cvicenie8_classification/Cv8_classification_some%20interpretations.pdf

[11] https://www.onlinecharttool.com

[12] https://en.wikipedia.org/

[13] https://timesofindia.indiatimes.com/india/90-of-diabetics-are-unaware-of-their-condition/articleshow/46206555.cms

[14] https://www.indiatoday.in/education-today/gk-current-affairs/story/98-million-indians-diabetes-2030-prevention-1394158-2018-11-22

[15] https://www.genome.jp/kegg/

[16] http://bidd2.nus.edu.sg/cgi-bin/profeat2016/main.cgi

[17] https://www.mayoclinic.org/diseases-conditions/diabetes/diagnosis-treatment/drc-20371451

[18] K. Tejeswinee, Gracia Jacob Shomona, R. Athilakshmi,Feature Selection Techniques for Prediction of Neuro-Degenerative Disorders: A Case-Study with Alzheimer's And Parkinson's Disease,Procedia Computer Science,Volume 115,2017,Pages 188-194,ISSN 1877-0509, https://doi.org/10.1016/j.procs.2017.09.125

[19] https://www.google.com/

[20] https://www.draw.io/

[21] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3339236/

Abdul Azeez K, is a final year student at Rajalakshmi Engineering College, Chennai, Thandalam, India. He is pursuing his Bachelor of Engineering in Computer Science and Engineering. And, he is currently doing his internship in Hitachi Solutions India, Chennai. He has a wide range of interests in C# coding and also in Microsoft technologies (Azure, Dynamics 365, ASP.NET MVC etc.,).

Aravindan M, is a final year student at Rajalakshmi Engineering College, Chennai, Thandalam, India. He is pursuing his Bachelor of Engineering in Computer Science and Engineering. And he is a Certified Senior System Architect (**CSSA**) and Certified System Architect (**CSA**) which validates his ability to use Pega to design and build for reusability across multiple lines of business.

Adhirai Nandhini A, is a final year student at Rajalakshmi Engineering College, Chennai, Thandalam, India. She is pursuing her Bachelor of Engineering in Computer Science and Engineering. And, she is passionate about learning new technologies and their real time application includes exploring the new technological devices.

Tejeswinee K M.E., is an Assistant professor in the Department of Computer Science and Engineering at Rajalakshmi Engineering College, Chennai, Thandalam, India. She guided many Bachelor Degree level projects and published a variety of papers in many conferences and journals.