ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 1, January 2020

Diabetes Prediction based model using SVM

Gulshan Bleem, Ishpreet Singh Virk

Department of computer Science and Engineering, BBSBEC, Fatehgarh Sahib, Punjab, India

Department of computer Science and Engineering, BBSBEC, Fatehgarh Sahib, Punjab, India

ABSTRACT: With the increase in number of the patients suffering from diabetes, early detection and prediction of diabetes is the major area of concern. In this study, the SVM based model using data mining techniques to analyse the available data to predict the occurrence of diabetes. This model is applied on the large dataset of PIMA. In combination of cluster and class based approach which uses KNN for clustering is used for classification. K-means is a simple and widely used technique but it is highly sensitive towards initial centroid and outliers which further affect the prediction accuracy. The aim is to determine a way to improve the initial centroid selection for KNN and retain maximum original dataset to enhance the performance of SVM. Results show that accuracy of the classification model using SVM and KNN are 77.92% and 64.93%, respectively. Further, using the classification results, this paper analyses the risk associated with diabetic and non-diabetic patients.

KEYWORDS: Data mining K-means clustering Weighted k-means clustering Logistic regression.

I. INTRODUCTION

With the ever-increasing data in every field, specifically in healthcare sector, there is a need to utilize the obtained information in a productive way. Focusing on usage and advantages of various technologies and their applications, they can result in refined

patient data analysis and better overall healthcare predictions. These technologies and platforms can provide provisions for improved solutions, individual health plans and development of clinical support systems.
Nowadays, there has been an extensive increase in the number of diabetic patients. According to WHO report, about 69 million people in India were diagnosed with diabetes in 2015 [2]. It is estimated that this count will rise to 98 million by the end of 2030. Diabetes is caused due to high rise in blood sugar level. The most common types of diabetes are Type 1, Type 2 and gestational diabetes[3]. Type 1 diabetes is caused when the human body is not able to generate sufficient amount of insulin. In type 2 diabetes, the body is unable to make proper use of produced insulin. Gestational diabetes develops in pregnant women and in most cases goes away after the birth of baby but they are highly prone to type 2 later on. In order to identify the high risk patients, advanced technologies can be used. The emerging technologies play a pivotal role in addressing these challenges and Data mining proves to be an appropriate field.
Data mining analyzes the data to obtain connections and nd patterns within the data and using them predictions can be made or the recent trends can be uncovered and analyzed[11]. They are broadly categorized into predictive and descriptive techniques. The predictive techniques have been implemented to ob-tain the results regarding breast cancer, heart disease and diabetes to analyze the current status and make future predictions regarding the occurrence of dis-ease, early detection and preventable patient deaths. This research paper focuses on improving the accuracy of diabetes prediction model. With a median based approach for initial centroid selection, k-means and weighted k-means (WKM) is used to cluster the instances based on the similarity between them and are classified using logistic regression. Based on the results, a comparative study of both the techniques is presented and risk associated with non-diabetic and diabetic patients is analyzed. This paper consists of different sections as mentioned: Section 2 comprises of the work related to disease prediction methods. Section 3 gives an insight of the proposed work. Section 4 presents the results obtained. Section 5 provides a comparative study of the existing work and the proposed model followed by Conclusions in Section 6.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 1, January 2020

II. RELATED WORK

Recently, more efforts have been made to improve the accuracy of prediction. Several techniques have been used to obtain better classification and prediction results. Some of the existing works have been discussed in this section.

Brojo Kishore Mishra et. al author in this paper has proposed a hybrid algorithm for the classification of the diabetes and non diabetes patients. Paper has used k-means algorithm and the gravitation search algorithm. GSA is the optimization based algorithm where the classification is done based on best available centroid. It identifies the true positive and true negative class. K-mean rather than taking the random classifier uses optimized centroid for the classification purpose. The performance given by the hybrid technique is best for the given training and testing set.

Rabindra K. Barik et. al author in this paper has proposed a process of SDI as spatial data infrastructure for the sharing of geospatial big data. This SDI has been integrated to the cloud based architecture to build a combined process of cloud-SDI. This paper has studied various malaria vector-born disease for the state of Maharashtra. They also has used lossless data compression on the data for the easy sharing globally with efficient way.

Archana et al.[17] presented a comparative study of the performance of k-means clustering by varying the distance metrics. Results showed that performance of clustering using k-means is affected by the selection of distance metrics and Euclidean distance metric provides the better result than the others.

Sandeep et al.[10] presented a hybrid approach for diabetes prediction where k-means is used for dimensionality reduction and Classification is done using support vector machine. The initial centroid for k-means are obtained by partitioning the dataset and choosing a centroid from each partition. Results showed that this hybrid approach achieved an accuracy of 92%. Another hybrid approach used k-means clustering. Further, data was classified using SVM[14]. The results showed that the combination of k-means and SVM produces promising results and use of optimization techniques can further enhance the performance of the model.

Wenqian et al.[5] proposed a hybrid approach implementing k-means for data reduction and classi ed PIMA Indian diabetes dataset using J48 decision tree. The correctly clustered 532 instances obtained from k-means were used to obtain he decision tree. Results showed that the accuracy of prediction model came out to be 90.04%.

Kanika et al.[4] proposed an optimized hierarchical clustering method which aimed at reducing the computation cost. Genetic algorithm and SVM were also used for feature selection and classification respectively. PIMA Indians Diabetes dataset was used and the accuracy was improved by 1.351%. Isaac et al.[8] compared the accuracy obtained by implementing k-means algorithm and decision tree. Results showed that both the techniques resulted in high accuracy but statistical results show that k-means algorithm has higher performance than decision tree.

De Amorim et al. [7] presented a comparative study of different feature weighing methods for k-means clustering. A discussion of the nine most innovative techniques to assign weights to feature for k-means is presented out of which six algorithms were implemented using dataset with and without noise. Further, eight key characteristics were discussed which were used to provide a detailed comparison of the algorithms.

Meng et al.[13] presented a comparison of the performance of three different classifiers by using diabetes dataset consisting of 12 attributes. The three models, logistic regression, ANN and decision tree were evaluated based on accuracy and a few more parameters. Experimental results showed that the decision tree and artificial neural network came out to be the best and worst classifiers among the three models.

Ambika et al.[6] presented a comparative analysis of various machine learning techniques by listing the pros and cons of each and comparing their accuracy and other performance parameters. The techniques, namely, decision tree, ANN,

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 1, January 2020

random forest, KNN, naive bayes, logistic regression and SVM have been implemented using the PIMA diabetes dataset. The experimental results showed that logistic regression provides the best accuracy as compared to other algorithms.

Liu[12] implemented logistic regression to classify the dataset using a combination of features. Varying the set of selected attributes a ected the performance of classifier. Results showed that logistic regression proved to be an easy and efficient way. Sun et al.[19] used a combination of logistic regression, random forest algorithm to identify different genes of breast cancer on micro array dataset. The prediction accuracy rates were analyzed by varying the threshold value. Top 20 genes were recognized that are expected to influence the development of breast cancer and a maximum accuracy obtained was 95.57%.

Han Wu et al.[20] also proposed a model that implemented k-means algorithm, logistic regression using WEKA toolkit to predict type 2 diabetes. The main aim was to enhance the accuracy and implement the model using various datasets. This model produced satisfying results and was less time consuming. Accuracy of the proposed model was 3.04% more than the existing ones.

III. PROPOSED WORK

The proposed model consists of several stages as shown in figure 1. Firstly, the data is preprocessed and used to find the initial cluster centers using median based approach. Then, using the obtained initial centroid, data is clustered into two clusters. The incorrectly clustered data items are removed and the rest are fed to the classifier. In the final stage, results of classifier are used to analyze risk associated with the patients.

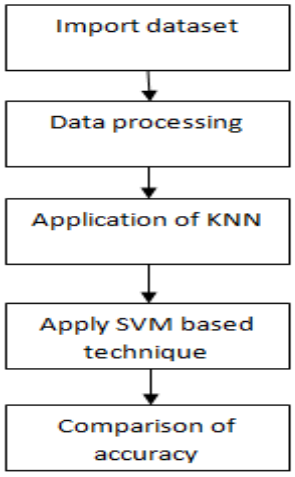

Fig. 1. Design of proposed model

IV. DATASET

The PIMA Indian Diabetes dataset contains data of 768 female patients from Phoenix, Arizona, USA extracted from online repository [1]. It consists of 8 at-tributes and 1 output variable. Out of 768 patients, 500 tested negative whereas 268 tested positive for diabetes. The attributes of the dataset used are shown in figure 2.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 1, January 2020

S.No.	Attributes	Type
1	Number of times pregnant	Numeric
2	Plasma glucose concentration	Numeric
3	Diastolic blood pressure (mm Hg)	Numeric
4	Triceps skin fold thickness (mm)	Numeric
5	2-Hour serum insulin (mu U/ml)	Numeric
6	Body mass index (weight in kg/(height in m)^2)	Numeric
7	Diabetes pedigree function	Numeric
8	Age (years)	Numeric
9	Class variable (0 or 1)	Numeric

Fig. 2. Dataset description

The last attribute indicates whether the patient is non-diabetic(0) or diabetic(1). Sample of dataset used is shown in Figure 3

S.No.	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	PedigreeFunction	Age	Outcome
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1

Fig. 3. Sample of dataset

V. DATA PREPROCESSING

The dataset quality is of utmost importance as it directly a ects the results of prediction. So initially the dataset is preprocessed to replace the incorrect or missing values. The PIMA Indian Diabetes Dataset contains incorrect values for some attributes like blood pressure and BMI. These attribute values can't be zero and were replaced by the column median. Further, each data value (x) was normalized[20] using (1).

VI. CONCEPTUAL DIAGRAM

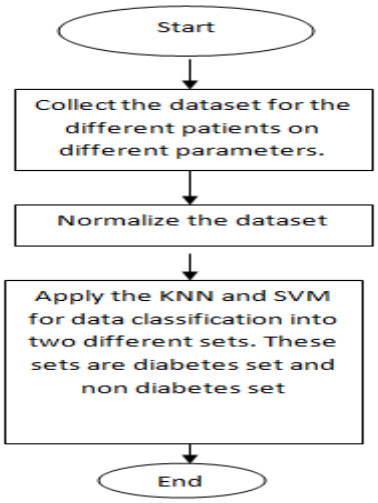

Fig. 4 Conceptual Diagram

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 1, January 2020

VII. ALGORITHM

Step 1 collect the dataset for the 9 different parameters for the patient. Distribute the data in the metric of the rows and columns
$\{d_{i,j}\}i=1\ldots n, j=1\ldots 9$
There are n items and each item has 9 parameters.
Step2 Normalize the dataset for all the empty values to be filled by the average value of the respective parameter.
Empty_field=average($d_{i,k}$)
Step3 calculate the weighted k-mean for the centroid identification.
Weighted_mean=average(Wi,j)/(i*j)
Step 4 classify the data based on centroid value of the dataset.
Class1=diabetes(dij)
Class2=diabetes(dij)
Step 5 identify the performance for classification.

VIII. INITIAL CENTROID CALCULATION

The k-means algorithm starts by choosing random initial centroid and after several iterations it obtains the final cluster centers. These initial centroid play a vital role in k-means clustering. Selection of initial centroid can lead to more precise final cluster centers, thus improving the accuracy of clustering. In the proposed work, initial centroid are selected as shown in figure 4.

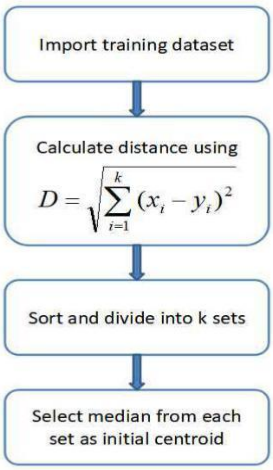

Fig. 5. Initial centroid selection

IX. K-MEANS CLUSTERING

K-means clustering is an unsupervised approach where the observations are partitioned into predefined number of clusters based upon the similarity between them. The clusters formed follow the below mentioned properties: 1. Sk i=1 Ci = N where N = Number of observations and k = Number of clusters 2. Ck ∩ Ck 0 = φ ∀ k 6= k 0 In the second step, k-means clustering is applied on training dataset with specific initial centroids. The correctly clustered data is collected and fed to the logistic regression algorithm and the incorrectly clustered data is discarded so that the model is trained with more accurate data values.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 1, January 2020

X. RISK ANALYSIS

Risk analysis basically refers to analyzing the outcomes of previous observations and based on that predicting the chances or risk probability for new ones. Based on the results of hybrid model, risk analysis is performed for non-diabetic and diabetic patients.

XI. RESULTS

R tool provides a convenient and more user friendly interface to implement and study the experimental results. Using the median based approach, two initial centroids were obtained. Based on these initial centroid, data is clustered into two clusters using k-means clustering. The clusters obtained are shown in figure 3. The cluster having data points marked in green, refer to non-diabetic patients cluster and the one in red indicates the diabetic cluster. Further, the correctly clustered data values were used for classification of instances using SVM.

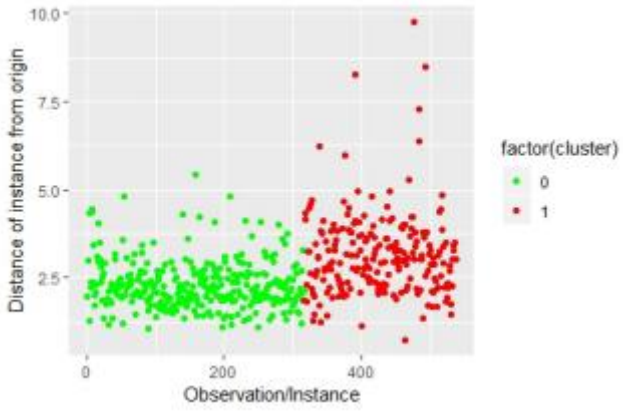

Fig. 6: K-means clustering

Performance parameters The confusion matrix contains four different results, namely, true positive(TP), true negative(TN), false positive(FP) and false negative(FN) based on which several performance parameter are calculated. Accuracy is the ratio of correctly predicted instances to total number of instances. The accuracy obtained for the proposed model is 77.67%

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN}$$

The value of recall obtained using equation for the model is 0.9630.

$$\text{Recall}=\frac{TP}{TP+TN}$$

Precision is calculated using equation and the value obtained is 0.9623.

$$\text{Precision}=\frac{TP}{TP+FP}$$

F-Score is a measure of accuracy that takes into account both precision and recall by calculating their harmonic average using equation. The value obtained using proposed model is 0.9629.

$$\text{F-score}=2*\frac{precision*recall}{precision+recall}$$

Specificity, also known as true negative rate is calculated using equation and value obtained is 0.9855.

$$\text{Specificity}=\frac{TN}{TN+FP}$$

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 1, January 2020

XII. KAPPA COEFFICIENT

Kappa coefficient is a performance parameter to measure the homogeneity or consistency of the given model. Its value lies between 0 and 1. The value closer to 0 means that there is less agreement between the actual and predicted outcomes whereas a value closer to 1 represents that there is more resemblance between the actual and predicted values

$$K=\frac{P0-Pe}{1-Pe}$$

$$\text{Where } P_0=\frac{TP+TN}{N}$$

$$\text{and } P_e=\frac{[(TP+FN)*(TP+FP)+(TN+FN)]}{N2}$$

The obtained value of kappa coefficient for proposed model is 0.9291.

Confusion Matrix and Statistics

Actualvalue	Predictedvalue 0	1
0	156	1
1	6	68

Accuracy : 0.9697
95% CI : (0.9386, 0.9877)
No Information Rate : 0.7013
P-Value [ACC > NIR] : <2e-16

Kappa : 0.9291
Mcnemar's Test P-value : 0.1306

Sensitivity : 0.9630
Specificity : 0.9855
Pos Pred Value : 0.9936
Neg Pred Value : 0.9189
Prevalence : 0.7013
Detection rate : 0.6753
Detection Prevalence : 0.6797
Balanced Accuracy : 0.9742

'Positive' class : 0

XIII. RISK ANALYSIS

Based on the results obtained from classification, we performed risk analysis for both non-diabetic and diabetic patients shown in figures 5 and 6. In case of non-diabetic patients, the data values that lie close to zero have a high probability of having diabetes in the future whereas the ones that lie far away have low probability of having diabetes. This analysis can be helpful to patients having high risk, as they can take necessary precautions to avoid the disease. In case of diabetic patients, the ones that lie near zero can take necessary precautions to control their blood sugar levels and the ones that lie far away are at a high risk and need to take immediate measures to control the blood sugar level, since it can prove to be fatal.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 1, January 2020

Fig. 7: Risk analysis of non-diabetic patients

Fig. 8: Risk analysis of diabetic patients

XIV. DISCUSSION

The experimental results show that using the proposed model, the number of correctly clustered instances increased as compared to existing work. A comparison of the results obtained using k-means clustering is illustrated in table 1.

Author	**Correctly clustered instances**
Wenqian et al.[5]	69.27%
Mustafa et al.[9]	74.21%
Han Wu et al.[20]	76.69%
Our proposed model (Using SVM)	77.67%

Table 1: Comparison of K-means clustering results

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 1, January 2020

Further, the quality measures such as accuracy, recall, precision, F score, specificity and kappa coefficient for previous and proposed work are depicted in Table 2 and also shown in Figure 7. The comparative analysis shows that our model performs better than existing model in terms of all parameters used to measure the performance. Table 2: Performance measure indices of previous work and proposed work

Performance metric	**Previous work[12]**	**Proposed work**
Accuracy	64.93%	77.92%
F score	0.954	0.9629
Kappa	0.8975	0.9291
Precision	0.9547	0.9629
Recall	0.954	0.9630
Specificity	0.9024	0.9855

Table 2: Performance measure indices of previous work and proposed work

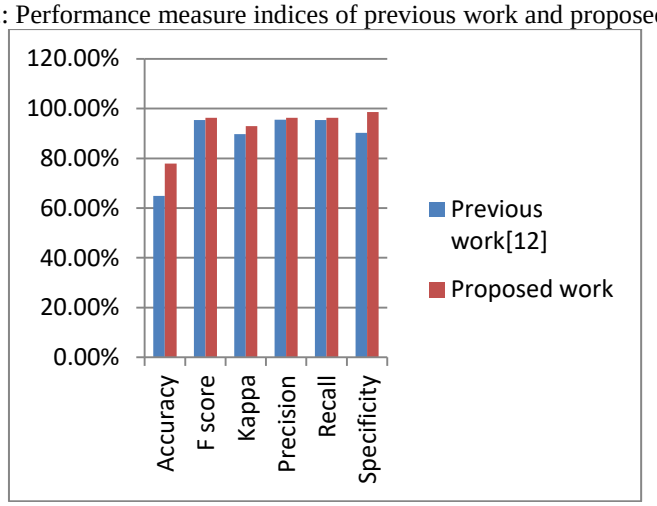

Fig. 9 Comparison

XV. CONCLUSION

With the increasing demand of predictive analysis in the medical field, the aim of this paper is to propose an efficient model for diagnosing and predicting the occurrence of diabetes based on several parameters. The proposed model is a combination of cluster and class based techniques, k-means clustering and logistic regression, respectively. Though, k-means clustering is an efficient technique to obtain clusters based on similarity between the instances but it is sensitive towards the selection of initial centroids. In this paper, we proposed a median based approach for the selection of initial centers to reduce the effect of outliers and further enhance the performance of the classifier. Experimental results showed that about 80% of the original data is retained after clustering, which is fed to the classifier. Accuracy obtained for classification using the proposed model is 77.34% with precision and recall as 96.29% and 96.30%, respectively. Further, the risk associated with diabetic and non-diabetic patients is analyzed using the results of the classifier.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 9, Issue 1, January 2020

REFERENCES

[1] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," Computational and structural biotechnology journal, vol. 15, pp. 104–116, 2017.
[2] A. Singh, A. Yadav, and A. Rana, "K-means with three different distance metrics," International Journal of Computer Applications, vol. 67, no. 10, 2013.
[3] S. Kaur and S. Kalra, "Disease prediction using hybrid k-means and support vector machine," in 2016 1st India International Conference on Information Processing (IICIP), pp. 1–6, IEEE, 2016.
[4] A. H. Osman and H. M. Aljahdali, "Diabetes disease diagnosis method based on feature extraction using k-svm," Int J Adv Comput Sci Appl, vol. 8, no. 1, 2017.
[5] W. Chen, S. Chen, H. Zhang, and T. Wu, "A hybrid prediction model for type 2 diabetes using kmeans and decision tree," in Software Engineering and Service Science (ICSESS), 2017 8th IEEE International Conference on, pp. 386–390, IEEE, 2017.
[6] K. Bhatia and R. Syal, "Predictive analysis using hybrid clustering in diabetes diagnosis," in Control, Automation & Power Engineering (RDCAPE), 2017 Recent Developments in, pp. 447– 452, IEEE, 2017.
[7] L. D. Isaac and C. Sureshkumar, "Diagnosis prognosis and prevention of breast cancer based on present scenario of human life," in 2018 International Conference on Communication information and Computing Technology (ICCICT), pp. 1– 7, IEEE, 2018.
[8] X.-H. Meng, Y.-X. Huang, D.-P. Rao, Q. Zhang, and Q. Liu, "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors," The Kaohsiung journal of medical sciences, vol. 29, no. 2, pp. 93–99, 2013.
[9] A. Choudhury and D. Gupta, "A survey on medical diagnosis of diabetes using machine learning techniques," in Recent Developments in Machine Learning and Data Analytics, pp. 67–78, Springer, 2019.
[10] L. Liu, "Research on logistic regression algorithm of breast cancer diagnose data by machine learning," in 2018 International Conference on Robots & Intelligent System (ICRIS), pp. 157–160, IEEE, 2018.
[11] M. Sun, T. Ding, X.-Q. Tang, and K. Yu, "An efficient mixed-model for screening differentially expressed genes of breast cancer based on lr-rf," IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2018.
[12] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, "Type 2 diabetes mellitus prediction model based on data mining," Informatics in Medicine Unlocked, vol. 10, pp. 100–107, 2018.
[13] M. S. Kadhm, I. W. Ghindawi, and D. E. Mhawi, "An accurate diabetes prediction system based on k-means clustering and proposed classification approach," International Journal of Applied Engineering Research, vol. 13, no. 6, pp. 4038–4041, 2018.