International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

An Approach to Predict Early Diabetes Mellitus with An Unsupervised Clustering Technique

Rita Ganguly*[1], Dharmpal Singh*[2]

Submitted: 21/04/2023 **Revised**: 22/06/2023 **Accepted**: 03/07/2023

Abstract: Hyperglycemia which constitutes a considerable imminence to human health. Diabetes may lead to an anomalous rise in glucose levels. Preliminary detection of diabetes reduces the risk of fatality and agony. In our country around 30 million peoples are recognized with this fatal disease. It is tremendously complicated to develop a virtuous and precise diabetes forecasting. The ICMR with diabetic people have taken inventiveness and emerged with various solution but regrettably they endured like leftovers. Clustering is an important technique for the prediction of diabetes. In machine learning the clustering technique contingent on unsupervised learning and classification techniques contingent on supervised learning. In this research work, the factor analysis concept has been solicited to genesis of total effect on the PIMA Indian Diabetic Dataset and designate the prime factors that repercussion on it. K-Means algorithm conviction has been on the total effect data to acquire the cluster in superlative mode and for the quantification of distance the Euclidean distance function has been used. The numbers of clusters have been pronounced on the base of output of the dataset and it causes formation of knowledge based. To predict diabetics various machine learning accession have been solicited on cluster-based dataset. K-Means clustering algorithm used for early diabetic identification containing the data of 165 diabetic patients. The maximum precision, recall and F1-score1.00 obtained by K-Means and accuracy obtained by logistic regression 0.7662, decision tree 0.7269, SVM 0.7835 and random forest 0.7922 respectively. All anticipated outcomes are displayed in a comparison table and pointed out the aspect of research.

Keywords*: *Clustering, Diabetes, K-Means Clustering, Factor Analysis, Disease

1. Introduction

Diabetes scientific name is Diabetes Mellitus, an incurable medical condition brings about accused to irregular sugar level in blood. It causes nerve damage, kidney failure, blindness, and coronary heart diseases. Insulin supervise the diabetes whenever it unconstraint then give rise to diabetes, Li et al[1],Panzarasa[2], Porterand Green[3]. ICMR have shown that Southern state in our country larger proportion of the population affected by diabetes compared to Northern India. Early changes in diabetes can be reconstructible Kuzuya et al[4] describes that pancreas islet cells can be revitalized. In recent days data mining contemplated as statistical confluences in determinations of many kind of fatal illness and disease as mention in Matheus et al[5]. Doctors are also fighting to devote their schedule much on knowledge disclosure in database for anticipating vogue which provide patient better diagnosis scope and metamorphosis in management system of medical databases. The data mining techniques are used to detect and analysis of various fatal diseases. In this research work the clusters are determined in the respect of how data are going to be handled using K-Means algorithm and it provides better accuracy comparing to other classification algorithms. The rest of the paper is organized as follows- section 2 contains related work, section 3 illustrate the proposed system architecture with diversified data exploration, factor analysis, clustering and proposed method with clustering approaches section 4 explains the integrated experimental outcomes and validation, and section 5 explains the conclusion.

2.Related Work

Several researchers design and develop various techniques and algorithms to predict diabetes earlier correctly. Victor Chang et [6] use the Pima Indian Diabetes dataset to test and train Naïve Bayes classifier J48 Decision 3 model and the Random Forest classifier model. The main objective of these works are finding out categorical of values and missing information. Principal Component Analysis (PCA) , K-means clustering were applied on the dataset to accomplish the nomination of feature. Md Kamrul Hassan et al [7] by using Decision Tree, Random Forest, KNN,

[1] *Dr.B.C.Roy Engineering College,West Bengal – 713206, INDIA*
ORCID ID : 0000-0001-6544-941X
[2] *JIS University,West Bengal – 700109, INDIA*
[3] ** Corresponding Author Email: ganguly.rita@gmail.com*

XGBoost, AdaBoost (AB), Naïve Bayes to prognosticate diabetes. In this work they used the PIDD comprising of 768 data with eight attributes. They used pre-processing procedure with filling in missing values rejecting outliers and correlation based on selecting features. J.J Khanam et al [8] use multiple machine learning methods. M A Sarwar et al [9] employed SVM, KNN, NB, RF, LR to aid practitioners and physicians in the early detection of diabetes. L Alturki et al[10] employs machine learning techniques to construct a robust framework and investigate the predictive features to detect risk.

Anuja et al[11] introduced a classification model with SVM which acquire accuracy 78% using Radial Basis Function (RBF). Harleen et al[12] introduced a determining technique for early detection of diabetes. It was followed by 3 steps- removal of features, initialization, and estimation of parameters. The classification-based algorithm C4.5 which provided 91% accuracy with training data relevance analysis. Ravi et al[13] introduced six various techniques to achieve discriminant results. Other researchers are used various techniques to achieve the best result with minimization of errors [14].

3. Methodology

From the machine learning repository obtained the dataset by applying the supervised and unsupervised modelling. Next eliminating the labels K-Means clustering unsupervised method is applied in the dataset. For the supervised method, trained dataset with supervised machine learning algorithms logistic regression, decision tree, SVM, random forest and finally analyzed the overall results.

3.1. Diversified Data exploration

When decision involvement engaged with more than single variable, then it is known as diversified data exploration.

Factor Analysis

Step 1: Multiply the whole data with 1000 to make it normalized.

Step 2: Calculate the correlation Matrix.

Step 3: Calculate the Eigen value and Eigen Vector

Step 4: Calculate Percentage Contribution

Step 5: Less Percentage contribution (<10) has been ignored.

Step 6: Contribution of Eigen Vector as corresponding Eigen Value.

Step 7: Calculate the cumulative effect values for all items,

Step 8: A relation has been formed by using the cumulative effect value of all elements to produce total effect values.

3.2. Clustering

Unsupervised learning method of machine learning is basically known as clustering. This is a task to split up the data elements into a number of groups where homogeneous data points are more indistinguishable to other data points in one group and heterogeneous data points in other groups. The podiums of clustering are as follows:

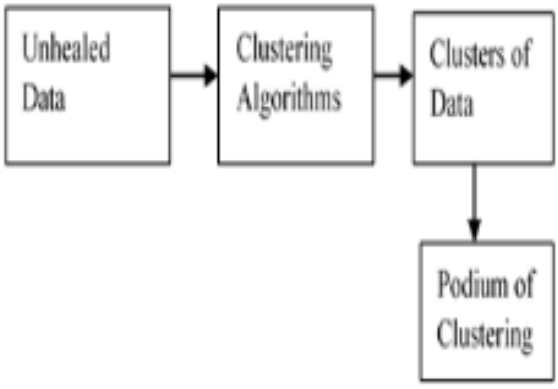


Fig1: Podium of Clustering

3.3. Proposed Method

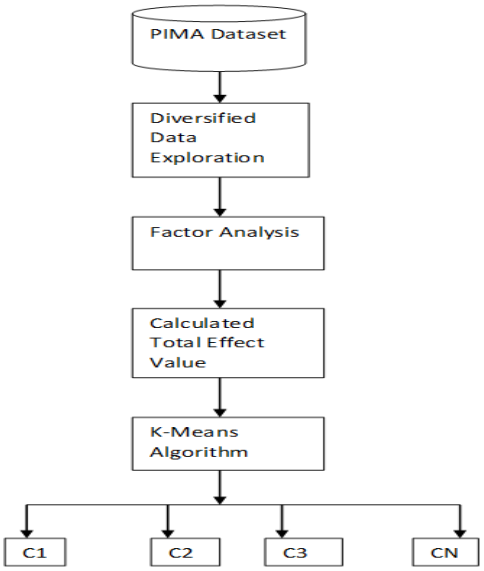


Fig 2: Block diagram of the methodology

In this research work to find the total effect values, the concept of factor analysis has been solicited. Euclidean distance has been estimated by using the concept of K-Means algorithm to fabricate the prime cluster. The prime cluster erected with knowledge based and generated new dataset appraised on it.

4. Integrated Experimental Outcomes and Validation

4.1. Explanation of Dataset

PIMA Indian Diabetic Dataset consists of 768 instances having 8 attributes. All instances represent the patient is diabetic or non diabetic. The set of attributes like preg, plas, pres, skin, insu, mass, pedi, age indicates the significant amount of disorders results diabetic or non diabetic.

Step 1: Multiply the whole dataset with 1000 to make it normalized.

preg	plas	pres	skin	insu	mass	Pedi	age
6000	148000	72000	35000	0	33600	627	50000
1000	85000	66000	29000	0	26600	351	31000
8000	183000	64000	0	0	23300	672	32000
1000	89000	66000	23000	94000	28100	167	21000
0	137000	40000	35000	168000	43100	2288	33000
5000	116000	74000	0	0	25600	201	30000
3000	78000	50000	32000	88000	31000	248	26000
10000	115000	0	0	0	35300	134	29000
2000	197000	70000	45000	543000	30500	158	53000
8000	125000	96000	0	0	0	232	54000

4000	110000	92000	0	0	37600	191	30000
10000	168000	74000	0	0	38000	537	34000
10000	139000	80000	0	0	27100	1441	57000
1000	189000	60000	23000	846000	30100	398	59000
5000	166000	72000	19000	175000	25800	587	51000
7000	100000	0	0	0	30000	484	32000
0	118000	84000	47000	230000	45800	551	31000
7000	107000	74000	0	0	29600	254	31000
1000	103000	30000	38000	83000	43300	183	33000
1000	115000	70000	30000	96000	34600	529	32000
3000	126000	88000	41000	235000	39300	704	27000
8000	99000	84000	0	0	35400	388	50000
7000	196000	90000	0	0	39800	451	41000

Step 2: Find the correlation matrix of the dataset.

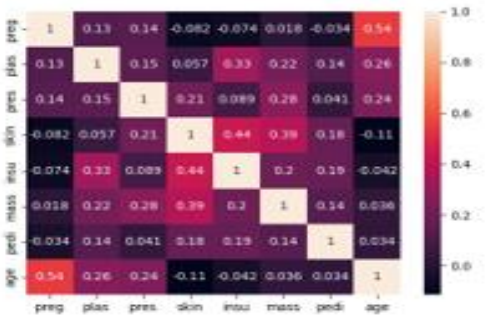


Fig 3 : Correlation Matrix

Step 3: Generate the Eigen Values and Eigen Vectors.

E_Value: [2.09269285 1.72575856 0.42871538 0.40463767 0.67231641 1.03269431 0.76943678 0.87374804]

E_Vector: [[0.10161586 0.59761523 0.59723558 -0.12286461 -0.23524407 -0.00225154 -0.42464923 0.06589433]

[0.38824326 0.17466077 0.22524852 0.39700586 -0.03134486 -0.45302574 0.38670801 -0.41713094]

[0.35460826 0.17762979 0.15658447 -0.07419779 0.66023607 0.54572248 0.21113145 0.09477771]

[0.44837492 -0.33327545 -0.05782493 0.62313328 -0.08528794 0.23087459 -0.47875956 0.03343396]

[0.44410043 -0.25077819 -0.07458731 -0.56291592 0.22085382 -0.33197474 -0.4224835 -0.35642736]

[0.4506338 -0.12451008 -0.03026941 -0.33501329 -0.65380952 0.35676736 0.33446368 0.05529702]

[0.28068857 -0.07149033 -0.00707364 -0.01488583 0.15493672 -0.45013663 0.31881366 0.82250707]

Step 4: Eigen Value and Percentage Contribution Calculation

Table 1: Eigen Value and Percentage

Data Attribute	***Eigen Value***	***Percentage Contribution***
Preg	2.09269285	26
Plas	1.72575856	21.95
Pres	0.42871538	5.36
Skin	0.40463767	6.74
Insu	0.67231641	8.40
Mass	1.03269431	12.90
Pedi	0.76943678	9.62
Age	0.87374804	10.92

Step 5: Find the major factors where pres,skin,insu and pedi discarded due to the less (<10) percentage contribution. The Eigen vectors of the four attributes have been furnished below:

Table 2: Eigen Vectors of Major Attributes

Preg	***Plas***	***Mass***	***Age***
0.10161586	0.59761523	-0.00225154	0.06589433
0.38824326	0.17466077	-0.45302574	-0.41713094
0.35460826	0.17762979	0.54572248	0.09477771
0.44837492	-0.33327545	0.23087459	0.03343396
0.44410043	-0.25077819	-0.33197474	-0.35642736
0.4506338	-0.12451008	0.35676736	0.05529702
0.28068857	-0.07149033	-0.45013663	0.82250707
0.18185766	0.62146718	-0.05944197	0.0703063

Step 6: Estimation of substantial factors using principles ($\sqrt{(Eigen\ Value\ X\ square(Eigen\ Vector))}$) using the selected Eigen values as embellished in table 1 and Eigen vectors embellished in table 2. The major factors have been embellished in below:

Table 3: Eigen Vector Corresponding Eigen Value

Data Attribute /Eigen Value	*2.09269285*	*1.72575856*	*1.03269431*	*0.87374804*

Preg	0.0149374251	0.469173346	0.000005152	0,004058721
plas	0.2180522615	0.040075667	0.208560303	0.162643951
pres	0.0456440687	0.041449722	0.302642266	0.008396640
skin	0.2908274328	0.145913899	0.054167422	0.001044886
insu	0.2853087773	0.082617024	0.111994248	0.118750426
mass	0.2937651488	0.020365686	0.129346929	0.002858226
pedi	0.1139730581	0.266182027	0.205908655	0.632371643
age	0.0478427288	0.507371905	0.003590643	0.004620422

Step 7: The cumulative effect value for all data attributes have been calculated in above table and summed up according to row wise. The cumulative values for all attributes have been embellished below:

Table 4: Cumulative Effect Value of Items

Data Attribute	***Cumulative Effect***
Preg	0.488175
plas	0.629332
pres	0.398133
skin	0.491954
insu	0.59867
mass	0.446336
pedi	1.218435
age	0.563426

Step 8: To generate the total effect value a relation has been fabricated by using cumulative effect value .Totaleffectvalue= Preg*(0.488175)+plas*(0.629332)+pres*(0.398133)+skin*(0.491954)+insu*(0.59867)+mass*(0.446336)+pedi*(1.218435)+age*(0.563426). Using the relation, a resultant total effect value has been furnished in table in sorted order:

Preg	plas	pres	skin	insu	mass	pedi	age	Tota Effect Value
1000	0	48000	20000	0	24700	140	22000	53028.09
2000	74000	0	0	0	0	102	22000	60066.57
3000	80000	0	0	0	0	174	22000	64418.46
2000	84000	0	0	0	0	304	21000	66042.59
0	73000	0	0	0	21100	342	25000	69861.28
1000	0	68000	35000	0	32000	389	22000	71931.7
0	94000	0	0	0	0	256	25000	73554.78
1000	0	74000	20000	23000	27700	299	21000	78118.27
7000	105000	0	0	0	0	305	24000	83390.93
0	99000	0	0	0	25000	253	22000	86165.9
2000	99000	0	0	0	22200	108	23000	86279.27
5000	44000	62000	0	0	25000	587	36000	86972.69
4000	90000	0	0	0	28000	610	31000	89299.44
6000	114000	0	0	0	0	189	26000	89552.26
5000	0	80000	32000	0	41000	346	37000	89602.16
6000	96000	0	0	0	23700	190	28000	89930.52
6000	0	68000	41000	0	39000	727	41000	91565.58
6000	91000	0	0	0	29800	501	31000	91575.72
1000	80000	55000	0	0	19100	258	21000	93403.37
1000	73000	50000	10000	0	23000	248	21000	93655.45
10000	115000	0	0	0	0	261	30000	94475.72
1000	71000	62000	0	0	21800	416	26000	94741.06

Total effect value is estimated using factor analysis. Now K-Mean algorithm is used on the total effect value, after that different numbers of cluster points are assumed and the different results are embellished.

The calculated Euclidean Distances regarding different cluster points are as follows:

Table 5: Euclidean Distances with Cluster Points

Cluster Points	***Euclidean Distances***
2	67802.4010
4	46233.7363
124	13849.1589
248	8971.9382
596	1765.52897
696	566.8348

The calculated distance to desired cluster point with calculated cluster centre of each attribute when the cluster point k=2:

Table 6: Cluster Centre Value

Cluster Feature's	***Cluster Centre***	***Distance to desired Cluster Point***
preg	3883.913765	21370.18354
plas	115266.9983	21370.18354
pres	68097.844113	21370.18354
skin	17618.573798	21370.18354
insu	32212.271973	21370.18354
mass	31173.631841	21370.18354
pedi	437.570481	21370.18354
age	33114.427861	21370.18354

The calculated distance to desired cluster point with calculated cluster centre of each attribute when the cluster point k=4:

Table 7: Cluster Centre Value

Cluster Feature's	*Cluster Centre*	*Distance to desired Cluster Point*
preg	4238.6635	8574.266625
plas	116272.0764	8574.266625
pres	67513.1265	8574.266625
skin	12892.6014	8574.266625
insu	4241.0501	8574.266625
mass	30782.8162	8574.266625
pedi	417.539379	8574.266625
age	34715.789976	8574.266625

The PIDD embraces of 768 samples with 8 input attributes and one output attributes. The PIDD swallow 9 attributes which represents the symptoms Pregnancies (preg), Plasma Glucose (plas), Diastolic Blood Pressure (pres), Triceps Skin Fold Thickness (skin), Insulin(2 hrs serum insulin)(insu), Body Mass Index (mass), Diabetes Pedigree Function (pedi), Person's Age (age) and after that Class Label only output. Among the 768 data there are some cases like 28 patients had a diastolic blood pressure of 0, 192 others had a skinfold thickness value as 0, 140 others had serum insulin levels at 0, and 11 more had body mass index record as 0.Accuracy, sensitivity, precision, specificity, F-measure, and error rate metrics are calculated according to the performance assessment.

Accuracy: A ratio of all precisely forecasted samples to the total number of samples.

$$\text{Accuracy}=\frac{TP+TN}{Total}$$

Sensitivity: Categorizes positive samples.

$$\text{Sensitivity}=\frac{TP}{p}$$

Precision: A ratio of the number of precisely forecasted samples to the total positive samples.

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

Specificity: Categorizes negative samples.

$$\text{Specificity}=\frac{TN}{N}$$

F-measure: Harmonic mean of sensitivity and precision

Confusion matrix

It defines the accurate and non-accurate classified samples. The representation of confusion matrix 2 X 2 in below

True positive (TP): The set of correctly classified positive samples.

True Negative (TN): The set of correctly classified negative samples.

False Positive (FP): The set of negative samples that are disarranged as positive samples.

False Negative (FN): The set of positive samples that are disarranged as negative samples.

Table 8: Comparison Table

Method Used	*Accuracy*
Logistic Regression	76.62%
Decision Tree	72.29%
SVM	78.35%

Random Forest	79.22%
K-Means	100%

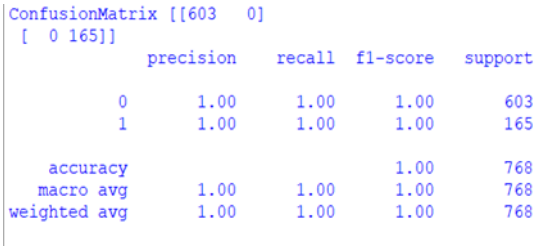
Confusion Matrix for cluster point =2

TN=603	FP=0
FN=0	TP=165

K-Means Clustering Results

Diabetic	165	21.48%
Non-Diabetic	603	78.52%

Confusion Matrix for cluster point =2

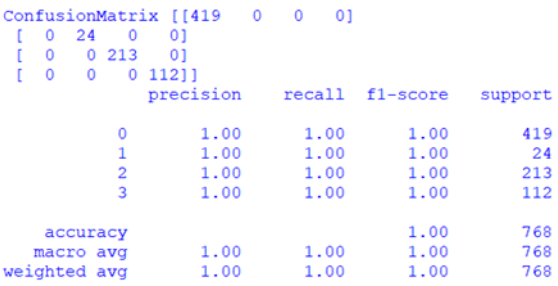


```
ConfusionMatrix [[603   0]
 [  0 165]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       603
           1       1.00      1.00      1.00       165

    accuracy                           1.00       768
   macro avg       1.00      1.00      1.00       768
weighted avg       1.00      1.00      1.00       768
```

Confusion Matrix for cluster point =4



```
ConfusionMatrix [[419   0   0   0]
 [  0  24   0   0]
 [  0   0 213   0]
 [  0   0   0 112]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       419
           1       1.00      1.00      1.00        24
           2       1.00      1.00      1.00       213
           3       1.00      1.00      1.00       112

    accuracy                           1.00       768
   macro avg       1.00      1.00      1.00       768
weighted avg       1.00      1.00      1.00       768
```

5. Conclusion

Diabetes disorder is required for proper early detection and prevention. In this research paper the factor analysis diversified data exploration technique is used for better and more accurate results. The K means clustering algorithm is used on PIDD and factor analysis is implemented to form a knowledge based. Using K-means clustering algorithm clustered the diabetes dataset to generate groups of the dataset based on classes. After that assembling a model of diabetes mellitus using different supervised ML algorithms, exhibiting the actualness of supervised learning on an unsupervised clustering dataset is the primary contributions of this research. The clustering accuracy provides better result comparing to classification techniques. Prediction of disorganization figures out to construct approaches to get gripes with it in beforehand. Imminent accentuate will be placed on the development of solicitations with user-friendly functions that will assist in the prognosis of this illness and provide recommendations for treatment and daily activities.

Author contributions

Rita Ganguly: is the inventor of this study, she provides system architecture for this model, she performed the final validation. She wrote, corrected, and restructured the entire menuscripted to complete this project. **Dharmpal Singh:** helped the writing on this document. All authors consent to submitting their manuscript to this journal. Conceptualization, Methodology, Software, Field study.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Li, L., Tang, H., Wu, Z., Gong, J., Gruidl, M., Zou, J., Tockman, M., Clark, R.A. (2004). Data mining techniques for cancer detection using serum proteomic profiling. Artificial Intelligence in Medicine, 32(2): 71- 83. doi.org/10.1016/j.artmed.2004.03.006

[2] Panzarasa, S. (2010). Data mining techniques for analyzing stroke care processes. Proceedings of the 13th World Congress on Medical Informatics, pp. 939-943. https://doi.org/10.3233/978-1-60750-588-4-939

[3] Porter, T., Green, B. (2009). Identifying diabetic patients: a data mining approach. AMCIS 2009 Proceedings, 500.

[4] Kuzuya, T., Nakagawa, S., Satoh, J., Kanazawa, Y., Iwamoto, Y., Kobayashi, M., Kashiwagi, A., Araki, E., Ito, C., Inagaki, N., Iwamoto, Y., Kasuga, M., Hanafusa, T., Haneda, M., Ueki, K., Committee of the Japan Diabetes Society on the Diagnostic Criteria of Diabetes Mellitus. (2002). Report of the Committee on the classification and diagnostic criteria of diabetes mellitus. Diabetes Research and Clinical Practice, 55(1): 65-85. https://doi.org/10.1016/s01688227(01)00365-5

[5] Matheus, C.J., Piatetsky-Shapiro, G., McNeill, D. (1996). 20 selecting and reporting what is interesting: The kefir application to healthcare data. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10. 1. 1.200.7107.

[6] V.Chang, J.Bailey, Q.A.Xu. Z.Sun, Pima Indians diabetes Mellitus Classification Based on Machine Learning (ML) Algorithms, Neural Computer Applications,(2022) 1-17.

[7] M.K.Hasan, M.A.Alam, D.Das, E.Hossain, M.Hasan, Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers, IEEE Access 8(2020) 76516-76531.

[8] J.J. Khanam , S.Y.Foo, A Comparison of Machine Learning Algorithms for Diabetes Prediction, ICT Express (2021).

[9] M.A.Sarwar, N.Kamal, W.Hamid, M.A.Shah, Prediction of Diabetes Using Machine learning Algorithms in Healthcare in:2018 24th International Conference on Automation and Computing, ICAC,IEEE,2018,pp1-6.

[10] L.Alturki, K.Aloraini, A.Aldughayshim, S.Albahli, Predictions of Readmissions and Length of Stay of Diabetes Related Patients, in: 2019 IEEE/ACS 16th International Conference on Computer Systems and Applicattions, AICCSA,IEEE,2019,pp.1-8

[11] V. Anuja and R.Chitra., "Classification Of Diabetes Disease Using Support Vector Machine", International Journal of Engineering Research and Applications (IJERA), vol.3,Issue 2, pp. 1797-1801, 2013.

[12] Aiswarya I., S. Jeyalatha and Ronak S., "Diagnosis Of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP), vol.5, ,No. 1, pp. 1-14, 2015.

[13] K.Rajesh and V.Sangeetha,"Application of Data Mining Methods and Techniques for Diabetes Diagnosis," in proceedings of International journal of Engineering and Innovative Technology, vol.2, Issue 3, pp. 43-46, 2012.

[14] Harleen and Dr. Pankaj B.,"A Prediction Technique in Data Mining for Diabetes Mellitus," Journal of Management Sciences and Technology, vol. 4, Issue 1, pp. 1-12, 2016.

[15] Ravi S. and Smt T., "Prognosis of Diabetes Using Data mining Approach-Fuzzy C Means Clustering and Support Vector Machine," International Journal of Computer Trends and Technology (IJCTT), vol. 11, No. 2, pp. 94-98, 2014.

[16] G. Krishnaveni*, T. Sudha," A Novel Technique To Predict Diabetic Disease Using Data Mining Classification Techniques" in International Conference on Innovative Applications in Engineering and Information Technology (ICIAEIT2017), vol. 3, Issue 1, pp. 5-11, 2017.

[17] Raj A., Vishnu P., and Kavita B.,"K-Fold Cross Validation and Classification Accuracy of PIMA Indian Diabetes Data Set Using Higher Order Neural Network and PCA", International Journal of Soft

Computing and Engineering (IJSCE), Volume-2, Issue-6, pp. 436-438, January 2013.

[18] Vrushali B., and Rakhi W., "Review on Prediction of Diabetes using Data Mining Technique", International Journal of Research and Scientific Innovation (IJRSI), Volume IV, Issue IA, pp. 43-46, January 2017.

[19] Thirumal P., and Nagarajan N.," Utilization of Data Mining Techniques for Diagnosis of Diabetes Mellitus - A Case Study", ARPN Journal of Engineering and Applied Sciences, Vol. 10, No. 1, pp. 8-13, January 2015.

[20] Huamaní, E. L. ., Leon-Ayala, R. ., Alva-Mantari, A. ., & Meneses-Claudio, B. . (2023). Prototype of a Mobile Application for the Detection of Car Accidents on the Roads of Peru . International Journal on Recent and Innovation Trends in Computing and Communication, 11(3), 37–42. https://doi.org/10.17762/ijritcc.v11i3.6198

[21] Jang Bahadur Saini, D. . (2022). Pre-Processing Based Wavelets Neural Network for Removing Artifacts in EEG Data. Research Journal of Computer Systems and Engineering, 3(1), 43–47. Retrieved from https://technicaljournals.org/RJCSE/index.php/journal/article/view/40