Jade Platform for Diabetes Classifier System

Mrs. Nadia Mahmood Ali[a], Dr.MahaAdham AL-Bayati[b], Dr.TawfeeqF.R.Al-Auqbi[c]

[a]Iraqi commission for computer and informatics, Informatics Institute for Postgraduate Studies, Baghdad/ Iraq.
[b]Computer Science Dept./ College of Science/Mustansiriyah University/ Baghdad, Iraq .
[c] National Diabetes Centre (NDC)/Mustansiriyah University/ Baghdad, Iraq.

Abstract: When designing medical diagnostic programs, disease prediction is considered as a captive task. Machine learning (ML) approaches were successfully used in a variety of applications, which includes the medical diagnoses. Through the development of a classification system, a ML algorithm could greatly aid in solving health-related problems that can help clinicians predict and diagnose diseases and provide patients with treatment at an early stage. This work aims to detect diabetes using ML techniques as decision tree and k-nearest neighbor (KNN) and on the basis of data that is manually collected from Iraqi population society. The work discuss the comparison of such algorithms in terms of accuracy of results..

Keywords: Machine learning ; classification, Diabetes, Decision Tree , Entropy, k-Nearest Neighbor ,Jade, K-fold cross validation , Information Gain.

1. Introduction

ML is an essential domain in the area of the research with an to predict and conduct a systematic overview**(Waren. S and Dubey. N.;2021)**. Diabetes is chronic disease, and became a leading life-style ailment, which is characterized by the prolonged elevated levels of the blood sugar. Failures of some of the organs such as heart, liver, stomach, kidney, and others results from a long run because of diabetes effect. Diabetes mellitus classification may be outlined as **(Sumangali. K. *et al*;2016)**:

- TypeI diabetes: a condition which is dependent upon insulin, it mainly happens in the adolescents and children due to some genetic disorders.
- Type II diabetes: happens in general to the adults about 40 years of age, it is discernible by the high level of blood sugar.
- Gestational diabetes: the type occurring throughout the period of pregnancy.
- Diabetes retinopathy: such disorder results in eye blindness.
- Diabetes neuropathy: it results from the nerve disorder.

Complications Arising as a result of the Diabetes:

The progression of the complications is moderate. Potential complications include arising: **(Choudhury .Aand Gupta D.;2019)**

- Cardio-vascular diseases (CVD): Diabetes results in vividly increasing risks of a variety of the cardiovascular complications;
- Nerve damages (i.e. Neuropathy).
- Kidney damages (i.e. Nephropathy).
- Eye damages (i.e. Retinopathy).
- Damage in foot: the deficient flow of the blood to feet results in increasing risks.
- Acute skin condition: occurrence of fungal and bacterial infections.
- Impairment of hearing: hearing issues are common.
- Alzheimer’s: Increases possibility of Alzheimer’s disease.

Machine learning is frequently used in disease classification and scientists have more interest in the development of those systems for the tracking and diagnoses of the cardio-vascular diseases and diabetes. Based on WHO (i.e. World Health Organization), CVD and diabetes are amongst top 10 death causes over the world

(Alić,et al& Berina;2017). Realizing the development of ML models, those models provide the ability for larger and more complicated data to be analyzed for the purpose of achieving more precise results and present better decisions in the real time with no human interventions.”

This work presents system which is designed to diagnose diabetes through classifications using classes: diabetic, non-diabetic, and probable diabetic. It uses a database for diabeties and subjectthem to two algorithms and compare their results in terms of accuracy in diagnosis and determine the appropriate treatment for each disease condition.”

1. Related work

To detect diabetes, the approach used machine learning classification algorithms where in recent years, data mining technology combined with machine learning technology was utilized with an increase in the frequency for the prediction of the susceptibility of the disease. Numerous rhythms and instrument combinations were produced and investigated by the researchers. Those had highlighted the enormous potentials of this area of study. In this subject, some significant works that are closely associated with the suggested problem have been provided.”

According to many researches, many researches have focused on machine learning classification algorithms, H. Wu, et al,**(Wu H., Yang .S., Huang Z., He .J, and Wang .X;2018)** suggest a prediction model for a high-risk T2DM group based upon a new model consisting of two-level algorithms, these are, improved K-mean algorithms and logistic regression algorithms. The results of the suggested model were to avoid omitting too much of the original data. Ensures high quality of the experimental data.,K. Sumangali, H. Ambarkar and B. Geetika, **(Sumangali. K. et al;2016)**: found in their study that the required models in the diagnoses of the diabetes mellitus is obtained by combining RF and CART and this improves and increases the accuracy of the results with an average accuracy of 81.06% as they note that using the single classifier CART (i.e. decision tree) alone with an accuracy of 66.27% for PIMA data set. The increase in estimators will modify the percentage accuracy and the error rate will decrease. D. Gupta and A. Choudhury,**(ChoudhuryA.and Gupta D.;2019)** did a detailed comparative research on a variety of the ML approaches for PIMA Indian Diabetic data set. Performance analysis has been analyzed according to the rate of the accuracy amongst all of the approaches of classification like logistic regression, decision tree, KNN, SVM and Naïve Bayesian. It was found that the logistic regression provides the highest accuracy results for the classification of the diabetic as well as the non-diabetic samples.Y. P. Huang and M. Nashrullah, **(Huang .Y. PandNashrullah. M;2016)**applied SVM and entropy approaches for the evaluation of 3 different data-sets, which include the mammographic mass, colon and spine, and depressin diabetic retinopathy .For the breast and depressin diabetic retinopathy datasets., A. Soofi and A. **Awan(Khalil.R. Mand A. Al-Jumaily;2017)**research had compared 4 models of ML on the classification of data-set for the issue of depression. It has been clear that the SVM classifier outperformed others.

2. JADE (Java Agent Development Framework)

“Jade can be defined as software environment for building the agent systems for managing networked information resources in compliance with specifications of FIPA for the inter-operable multi-agent systems. JADE presents a middleware for developing and executing agent-based applications that have the ability of seamlessly working and interoperating in the wired as well as the wireless environments. In addition to that, JADE supports development of multi-agent systems via pre-defined extensible and programmable agent model and a group of the tools for testing and management **González-briones .A., La Prieta .F. D, Omatu .S, and Corchado .J. M,2018)**. “

3. Machine Learning (ML)

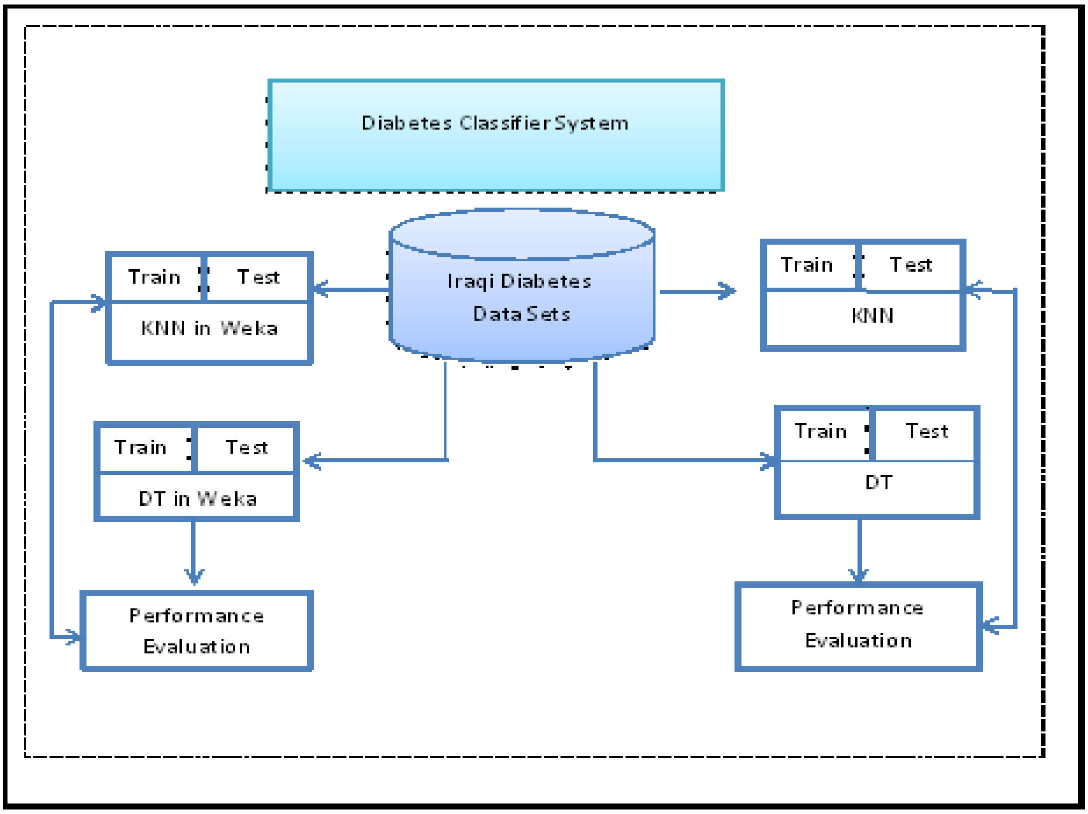
ML can be defined as a field borned from the field of artificial intelligence (AI).where the machine makes the decisions as well as the predictions / forecasts based on the data**(Ray.S;2019)**. There are several ML algorithm types utilized for the classification of the datasets. These techniques fall in one of categories: Unsupervised, Supervised, Reinforcement, Semi-Supervised, Evolutionary Learning, and Deep Learning algorithms**(Waren. S and Dubey. N.;2021)**.”

4. Classification

The term classification is defined as a machine learning problem**(Sen.P. C, Hajra. M, and Ghosh M.;2020)**,Classification can be defined as an approach the categorization or assignment of the class labels to pattern set under a teacher's supervision. The decision boundaries are produced for discriminating between the patterns that belong to separate classes. Data-sets are partitioned at first to *training* and *testing* datasets, and the classifier, which is a construct (algorithm) that discriminates between classes of patterns, and it is trained on the training dataset to create a model. The testing dataset is utilized for the evaluation of the classifier's generalization capability **(Khalil.R. Mand A. Al-Jumaily;2017)**]."

5. Design of Proposed System

In an aim to detect diabetes ,this study presents the system `which is developed whose design is show in figure (1) and presented in the next following section.



Figure(1) Flowchart of the General Structure of the proposed DCS

7.1 Dataset

As mention, the data is collected from the Iraqi society, It is obtained from the lab of the Medical City Hospital and (Specialized Iraqi Society Centre for the Endocrinology and Diabetes at Al-Kindy Teaching Hospital). Patient files are consider to extract the data from , and to *arrange* it in database, haveto get the diabetes dataset available for forthcoming. The data includes the following attributed of the Patients' personal and medical information as well as laboratory analysis. This information is listed below:

- Patients No
- Name

- Age
- Gender
- Sugar Level Blood
- Body Mass Index (BMI)
- Creatinine ratio(Cr)
- Cholesterol (Chol)
- Fasting lipid profile, which includes the total, LDL, VLDL, HDL Cholesterol and Triglycerides(TG).
- Urea
- Diabetes Class (patient's diabetes disease class could be Diabetic, Predict-Diabetic or Non-Diabetic)."
- HBA1C

In this senses a total of 1000 patient data sample are collected each of which consists of 16 attributes, One of these attributes is designated as the class attribute; the set of possible which hold the value of cornpones class that represent the tgra of diabetes type. It gives a judgment about the person's states of being diabetic (Y), not -diabetic(N) or probable-diabetic(P).

7.2 Classification Methods

There are many methods that are utilized to detect diabetes. Data samples have been partitioned to target class in the classification approach and this same has been forecasted for every one of the data points. For example, a patient can be classified as "diabetes", or "non-diabetes", and "probably" on a premise of their illnesses with the use of data classification approaches. Some of routinely utilized approaches have been discussed below:"

7.2.1 k-Nearest Neighbors (k-NN)

KNN can be described as a lazy algorithm of learning for the instance based learning. It has been utilized for the classification of the objects according to their closest examples of training in feature space. The object is categorized in a class to which its k-nearest neighbors belongs. In kNN, classification of a new test feature vector has been specified by its k-nearest neighbor classes. K-NN is carried out with the use of the metrics of Euclidean distance for locating nearest neighbor. The metrics of the Euclidean distance d(X1,Y1) between a pair of the points x1 and x2 has been computed with the use of the following equation:"

$x1=(x_{11}, x_{12},\ldots, x_{1n})$

$x2=(x_{21}, x_{22},\ldots, x_{2n})$

$$\text{dis}(x1 + x2) = \sqrt{\sum_{i=1}^{n}(x_{1i-}x_{2i})^2} \qquad (1)$$

In this work, and based on Equation No. (1), algorithm (1) of KNN is applied.

Algorithm (1) KNN

```
    K ← represents*ID3*number of the nearest neighbours.
    For every object Z d.
        Compute distance between each object x and z in training set d(x,z)
        Neighbourhood ← the k neighbours, nearest to z in training dataset
        Zclass ← selector class (based on the neighborhood)
    End For
```

using the foresaid dataset and upon implementation this algorithms gain accuracy of 89.33 % compared to that reached via weka tool of 88.8 % .

7.2.2 Decision Tree

A decision tree can be defined as flowchart-like tree structure, in which every one of the internal nodes denotes a test on attribute, every one of the branches represents a test outcome, and class label is denoted by every one of the leaf nodes (i.e. terminal node). It's easy converting the decision trees to rules of classification. Decision tree learning utilizes a decision tree as the predictive model which is used for mapping the observations concerning an item to the conclusions concerning target value of the item. It's a predictive modelling approach that is utilized in the statistics, machine learning and data mining."

In this work, algorithm (2) of Decision Tree is applied.

```
Algorithm(2) ID3(Examples, Attributes, Target_attribute,)
Inpust:   Target_attribute,    // represent attribute whose value will bepredicted  by tree //
          Examples,            // represent training data samples//
          Attributes// represents a list of other attributes which could be tested by learned decision tree//
Outputs:decision tree correctly classifying givenExamples
Begin
CreateRoot node for the tree
If all of the Examples were negative, Return single node tree Root, labelled as Neg.
If all of the Examples were positive, Return single node tree Root, labelled as Pos
IfAttributes is empty, Return single node the tree Root, labelled as most common Target _attribute value in Examples.
  Else Begin
A = attribute from Attributes best classifying the Examples
    Decision attribute for Root = A
For every one of the possible values, vi, of A do
```

```
Add a new tree branch below Root, which corresponds to test A = vi
Let  Examplesvi be a sub-set of Examples which have value vi for A
Add sub tree ID-3(Examples_vi , Target_attribute, Attributes – {A}))
        below this new branch
EndFor
End
    Return the Root
 End
```

Select Optimal Attribute

Central choice in ID-3 algorithm is selecting the attribute to be tested at every one of the nodes in a tree. The attribute which is most beneficial for the classification of the examples was chosen. What is the sufficient quantitative measure of an attribute worth? A statistical characteristic will be characterized, which is referred to as information gain, measuring how well a certain attribute can separate training"examples based on their target classifications. ID-3 utilizes this measure of Information Gain for selecting amongst candidate attributes at every one of the steps while growing tree that is computed with the use of Entropy measure."

i. The ID-3"algorithm utilizes the entropy for the calculation of homogeneity of the sample or characterization of the inclusion of a random set of data. As in the following equation no(2):

$$\text{Entropy(S)} = -P_{pos} \log_2 P_{pos} - P_{neg} \log_2 P_{neg} \qquad (2)$$

- P_{pos} => represents proportion of the positive samples in S.
- P_{neg} => represents proportion of the negative samples in S."

ii. More generally, in the case where target attribute might take on c different values, then entropy of S relative to that c-wise classification has been characterized based on equation (3):

$$\text{Entropy(S)} = \sum_{i=1}^{c} \quad -pi.log2pi \ (3)$$

- P_i represents proportion of S that belongs to the class i.
- if target attribute might take on c possible values,
  Gain the entropy may be as large as log2 c.

iii. The Measures of the Information Gain of Expected Reduction in Entropy equation (4):

$$Gain\ (S, A) = Entropy(s) - \sum_{V \in Values\ (A)} \left(\frac{|Sv|}{|S|} \right) Entropy\ (Sv) \quad (4)$$

- | |: represents the symbol of Cardinality.
- *Values (A)* : represents set of all of the potential values of attribute *A*.
- *Sv:* represents sub-set of S for which attribute *A* has value *v*.
- *Entropy(S):* represents entropy of original collection *S*.

$\sum_{V \in Values\ (A)} \left(\frac{|Sv|}{|S|} \right) Entropy\ (Sv)$: represents the expected entropy value after *S* has been partitioned with the use of attribute *A*.

“”The expected entropy that has been described by this 2nd term simply represents summation of entropy values of every sub-set S_v that has been weighted by fraction of the examples $|S_v| / |S|$ belonging to S_v. Gain (S, A) thus represents expected entropy reduction, which results from knowing attribute *A’s* value.”

Again ,comparing the results gained applying equations (2),(3) and (4) to that got using Weka tool, show that the former perform at accuracy of 99.9 % and hence outperform the latter which perform at accuracy of 98.4 % .

7.2.3 K-fold cross validation

“k-fold cross-validation can be defined as a testing system for the algorithms of machine learning. In this approach, data is divided to “k” parts, one of which will be separated for training before testing as illustrate in Figure (2). The benefit of this approach is that it decreased bias that is related to random sampling approach .

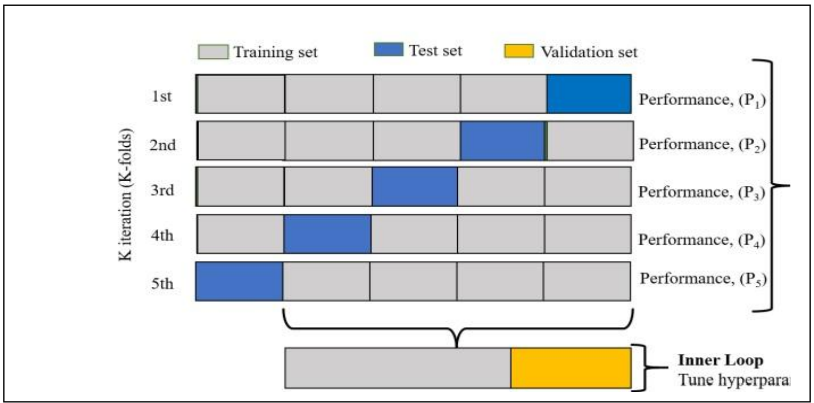


Figure (2) K-fold cross validation

As stated before ,this work uses the diabetes data set which is collected from the Iraqi society , and the purpose of the diagnostic data set is to predict diabetes .As depicted in the table (1), these data samples fall into three classes ;and are divided using different partitioning percentages table (1) presents the accuracy levels gained using KNN algorithms with table (2) presents the accuracy levels gained using DT algorithms.

Table (1) KNN Accuracy

Classes		No. of samples/ Class	60%	70%	80%
Class Y	Training	844	506	591	675
	Testing		338	253	169
Class N	Training	103	62	72	82
	Testing		41	31	21
Class P	Training	53	32	37	42
	Testing		21	16	32
Accuracy		1000	84.75	89.33	92.45

‘

Classes		No. of samples/ Class	60%	70%	80%
Class Y	Training	844	506	591	675
	Testing		338	253	169
Class N	Training	103	62	72	82
	Testing		41	31	21
Class P	Training	53	32	37	42
	Testing		21	16	32
Accuracy		1000	99.5	99.66	99.75

Table (2) Decision Tree Accuracy

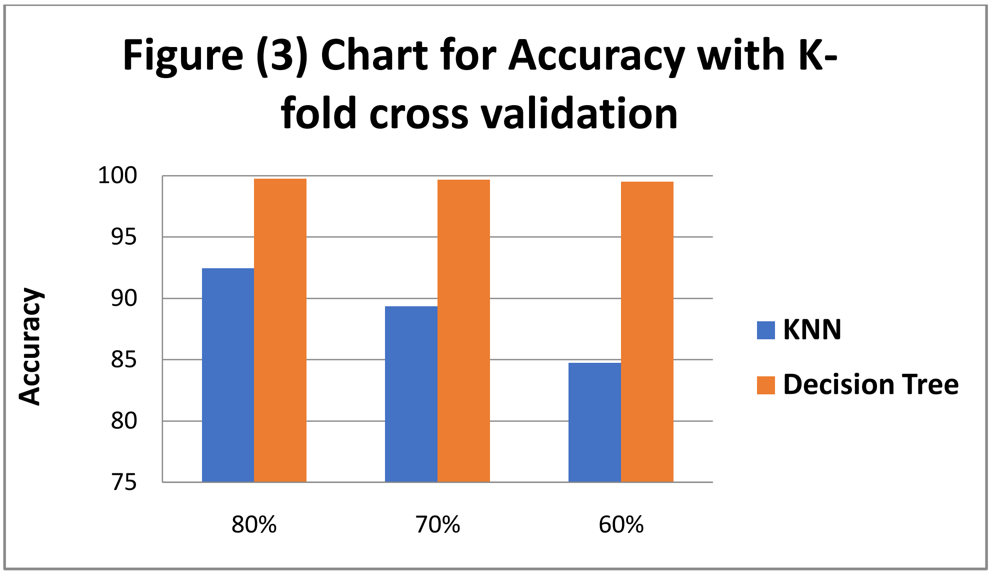
8. Result Discussion

As presented, accuracy increases with an increase in percentage of machine learning training level .It is worth mentioning that cross validation has been mainly utilized in the applied machine learning for the estimation of the skill of a model of machine learning on the unseen data. Which means that using a limited sample for the estimation of the way the model will be expected to perform in general when it’s utilized for making predictions about unused data during model training.” table (3) illustrate a comparison between KNN an Decision Tree

Algorithm	60%	70%	80%
KNN	84.75	89.33	92.45
Decision Tree	99.5	99.66	99.75

Table (3) Compare accuracy between KNN and Decision tree

Table (3) summarizes the accuracy levels gained ,due to different partitioning percentages ,by the two ML algorithms.



As the Chart in Figure (3), shows DT scores a higher accuracy than the KNN algorithm, so we rely on it to build rules for diagnosing diabetes and train them on these rules to obtain high accuracy results.

9. Conclusion and Future Scope

This ”paper seeks to provide an approach to mobile learning systems a suitable prediction model for diagnosing diabetes with high accuracy can be generated. According to several researchers' experiences, we proposed learning that is successively is to be considered for mobile diagnosis where a model using Jade platform due to develop mobile learning agency . research consists of a high-precision machine learning algorithm, the DT algorithm. For the purpose of making a valid comparison with the machine learning results of KNN and Decision Tree, it was necessary to implement this model using a diabetic dataset. The K-fold was a cross-validation to enhance the data set’s validity and rationality

The proposed model ensures acceptable prediction accuracy that allows for realistic data set is entered by diabetics for rapid prediction of their condition.

The suggested model had proven to be suitable for prediction with high speed and accuracy. An advantage of the suggested model is that it avoids the deletion of much of the original data. Ensures experimental data high quality. Another advantage is that the proposed model is built using Java, Java applications typically run in secure automated environments for providing many native application features by embedding them in HTML pages. It is possible, in the future to develop other models that serves to build robust prediction system for one of the most vital health issues diabetes . use other models of the algorithms of machine learning.”

References

Alić.et al and Berina, “Classification of metabolic syndrome patients using implemented expert system,” CMBEBIH 2017. Springer, Singapore, pp. 601–607, 2017.

Choudhury.A and Gupta. D, A Survey on Medical Diagnosis of Diabetes Using Machine Learning Techniques, vol. 740.Springer Singapore, 2019.

González-briones .A., La Prieta .F. D, Omatu .S, and Corchado .J. M, “Multi-Agent Systems Applications in Energy Optimization Problems : A State-of-the-Art Review,” Energies, pp. 1–28, 2018, doi: 10.3390/en11081928.

Huang Y. PandNashrullah. M, “SVM-based Decision Tree for medical knowledge representation,” 2016 International Conference on Fuzzy Theory and Its Applications, iFuzzy 2016, 2017, doi: 10.1109/iFUZZY.2016.8004949.

Khalil .R. M and A. Al-Jumaily, “Machine learning based prediction of depression among type 2 diabetic patients,” Proceedings of the 2017 12th International Conference on Intelligent Systems and Knowledge

Engineering, ISKE 2017, vol. 2018-Janua, pp. 1–5, 2017, doi: 10.1109/ISKE.2017.8258766.

Ray. S, “A Quick Review of Machine Learning Algorithms,” Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing: Trends, Prespectives and Prospects, COMIT Con 2019, pp. 35–39, 2019, doi: 10.1109/COMITCon.2019.8862451.

Sen. P. C., Hajra. M., and Ghosh .M., Emerging Technology in Modelling and Graphics, vol. 937. Springer Singapore, 2020.

Sumangali .K, B. Geetika. S. R, and H. Ambarkar, “A classifier based approach for early detection of diabetes mellitus,” 2016 International Conference on Control Instrumentation Communication and Computational Technologies, ICCICCT 2016, pp. 389–392, 2017, doi: 10.1109/ICCICCT.2016.7987979.

Waren .S and Dubey .N, “Detection of Diabetes Using ML Algorithms : A Survey,” vol. 1, no. 02, pp. 287–291, 2021.

Wu H., Yang .S., Huang Z., He .J, and Wang .X, “Type 2 diabetes mellitus prediction model based on data mining,” Informatics in Medicine Unlocked, vol. 10, no. August 2017, pp. 100–107, 2018, doi: 10.1016/j.imu.2017.12.006.

[