Journal of Theoretical and Applied Information Technology
15th July 2022. Vol.100. No 13

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

# DIAGNOSIS OF DIABETES MELLITUS BASED COMBINED OF FEATURE SELECTION METHODS

**SHAIMAA HAMEED SHAKER[1],AL-KHALIDI FARAH Q. [2],AMMARFAKHRI MAHDI[3]**

[1,3]Department of Computer Sciences, University of Technology , Iraq, Baghdad

[2]Department of Computer Sciences, Mustansiriyah University, Iraq, Baghdad

E-mail:[1]shaimaa.h.shaker@uotechnology.edu.iq,
[2]farahqaa@uomustansiriyah.edu.iq,[3]ammar.f.mahdi@uotechnology.edu.iq

**ABSTRACT**

Diagnosis of Diabetes-Mellitus in human given a chance to know a person that may be at risk of lengthened tricky situation. This work proposed to diagnosis diabetes mellitus based on a hybrid of carefully selected features method to reduce the features by combining Chi square test and Recursive feature elimination methods, so this was the first stage. Then the output of first stage is the input to the classification stage to obtain the true accuracy as the second stage. Logistic Regression(LR), K-Nearest-Neighbor(KNN), and Naïve-Bayes(NB) methods are used to classify nonappearance or appearance of diabetes mellitus disease. So the contribution of this research was to determine the optimal number of extracted and approved characteristics for rapid accurate diagnosis of disease by using two methods of information reduction and to benefit from the data collected in real in addition to standard data.This work has deal with PIDD from UCI and another dataset that collected from some patients. All the results is evaluated using accuracy(ACC), precision(PR), recall(RE), and f-score(FScore) measures. Algorithms LR, KNN, and NB achieved accuracy about (95% -98%) with combined feature selection method. So LR and NB achieved maximum accuracy(98%)with proposed method based on five and six features of 400 patient records, while KNN algorithm achieved accuracy (87%)with Chi-square test and Recursive feature elimination on diabetes-dataset.

**Keywords:** *Diabetes Mellitus, classification, LogisticRegression, K-Nearest-Neighbor, Naïve-Bayes, Feature Selection.*

## 1. INTRODUCTION:

One of the most dangerous diseases is Diabetes Mellitus (DM) ever since too many of human are affected[1]. DM happens when the human body cells become resistant to insulin or when not enough insulin is produced via the pancreas. The energy that exists in food can't be used effectively via humans due to diabetes [2]. Pre-diabetes is a stage when a doctor isn't able to diagnose the disease despite the high blood sugar level more than it should be. As well as, one of the common types of this disease which often appears from childhood, is insulin-dependent diabetes which is an autoimmune disease caused by attacking antibodies in the body. A patient with a pancreas that will not be able to secrete insulin, and this type is accompanied by other diseases such as damage to the small blood vessels in the eyes of the patient, neuropathy or nephropathy, and the risk of heart disease and stroke increases. There is a type of diabetes that not dependent on insulin, so it is common among children and adolescents due to obesity in young people also means that the pancreas produces an insufficient amount of insulin or that the body does not use it as it should. Moreover there is another type is gestational diabetes which is often discovered in mid or late pregnancy and it is imperative to protect the growth and development of the fetus and then the newborn and some surgeries, medications and infections cause diabetes as another type of infection [3][4]. Table1 shows the normal blood sugar levels.

*Table1: Typical Blood Sugar Levels*

| # | Abstaining (mg/dl) | | Post-Prandial-blood sugor (mg/md) |
|---|---|---|---|
| | Range value | Post-eating | 2 Hr. after utilizing glucose |
| Normal-level | 70-100 | 170-200 | <140 |
| Early -level | 101-126 | 190-230 | 140- 200 |
| Establised-level | > 126 | 230-300 | > 300 |

Classification of data is a process of data mining to find significant patterns that aid in medical diagnosis[5]. To get effective features (not

redundant relevant features) from dataset via features selection process after cleaning the dataset(non redundant, not missing) , each record in dataset is classified according the effective features. In 2016, K. Thangadurai, and N.Nandhini [6] suggested a system that used classification techniques, including Genetic algorithm, SVM, C4.5, EM, and K means to classify diabetes data. The good organization of the model was evaluated by a data from the UCI-PIMA and results of accuracy with EM was 70%, whereas, the C4.5 algorithm achieved 71.2% accuracy, K means achieved 77% accuracy, SVM reached 68% accuracy, and the Genetic algorithm achieved 78.1% accuracy. K. Saravanapriya and J. Bagyamani in 2017[7] analyzed the achievement of the categorization techniques in the DM dataset. Achieved accuracy range between 74% to 80% of various classification techniques such as NB, J48, Random Forest, Multilayer Perceptron, JRIP, KNN, SVM, and RBF Network. K. Akyol and B. Şen in 2018[8] introduced a system based the weighting approaches as feature-selection and Gradient Boosted Trees, AdaBoost, and RF ensemble learning algorithms with accuracy of prediction upon a combination of AdaBoost and Stability Selection approach is somewhat better compared to other algorithms with accuracy of 73.88%. M. Faisal, U. Ahmed, and I. Sarker in 2019[9] suggested a method of to predict diabetes mellitus, which used NaïveBayes, C4.5, SVM, KNN algorithms. The results of classification-accuracy was 68%, 73%, 70%,and 71% respectively. R.Kumarmangalnd, and et al. in 2021[10] suggest an enhanced ML-Framework for type2-Diabetes classification based on imbalanced-data and missing values, so this approach provided the best accuracy of 98% on the test data. In 2022, Y. Meiz, Y. Blma, and T.Özgür[11] present an evaluation of MLA with different CAD datasets, such that the Random-Forest displayed results for its classification functioning, while KNN-ML was less perform at all intends and others such as logistic-regression has changeable categorization working at every point

The aim of this paper is to diagnosis of diabetes mellitus with optimal cost and efficient performance by using collection of dataset from PIDD and some patients records accumulated from medical-center in Baghdad then depending on some features extraction by Chi-square test and Recursive feature elimination. Then MLA as classifier (LR, KNN, and NB). Section2 of this paper introduce some theoretical background of typical diagnosis diabetes , techniques of feature selection and some classification based MLA. While section3 presents the suggested method, and section4 resents the results and some discussion. Section5 is the conclusions.

## 2. TYPICAL DIAGNOSIS OF DIABETES TECHNIQUES:

Automatic diagnosis of DM systems may be done by machine-learning(ML) approaches which have some advantages and some limitations. The effectiveness of automated diagnostics depends mainly on the efficiency of the approved typical model[12]. Fig.1 illustrates the typical model of DM diagnosis. Datasets divided into two sets: training set and testing set, in some cases, researchers created the database as per their need or accessible databases can be regarded as standard databases, since a variety of research and tests are performed on them. Best classification performance is dependent on efficient data preparation and data preprocessing methods. Handling missing values of attributes through a meaningful approach can substantially improve a machine learning model's performance [13].

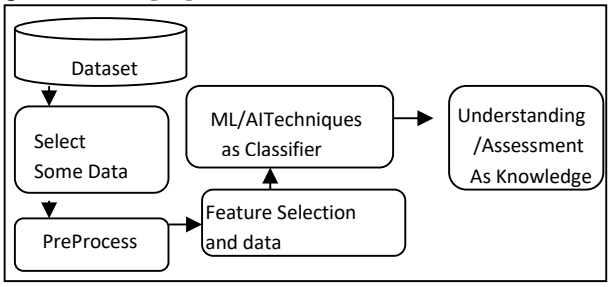


*Figure1: The main Steps of the Diabetes diagnosis typical model*

Feature Selection(FS) is significant that the data set is pre-processed before mining process is used so that repeated data can be removed or the unstructured data can be counted by transformation of the dataset. Theoretical strategies for selecting proper features differ for a different challenge to another. Employing feature selection is the important step to simplifying the learning part of the mining stages and enhancing the performance without altering the primary structure of data mining methods [14] and [15]. Some classification based MLA such as K-Nearest Neighbors (KNN) which makes a clustering process based on the proximity relations between objects. It works in the coordinate plane with the linear decomposition method that obtains neighbor data using the Euclidean distance between data points [16][17]. Linear regression(LR) is an arithmetical/statistical method that is used for analytical task. LR makes forecast for incessant/numericvalues[18][19].

Journal of Theoretical and Applied Information Technology
15th July 2022. Vol.100. No 13

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Naïve-Bayes (NB) re-scans the entire dataset for each new classification process which might cause it to operate relatively slowly [20]. There are some other ML approaches like SVM,RF and J48 [21][22].

## 3. PROPOSED DIAGNOSIS METHOD DESIGN

The proposed method of DM diagnosis is consists of 5 stages as shown in fig2. This core of this work is the feature selection-stage which is based on the Chi-square test and Recursive feature elimination that have the ability for choosing the best subset of attributes for classification so this stage is capable on diabetes disease prediction. The diagnosis-part of this work received some attributes as an input features from dataset. The number of these features is ten from dataset1 and eight features from dataset2. So the result of diagnosis-part as an output is the finding mellitus or no mellitus diabetic disease.

### 3.1 Data Preparation Part

This work deals with data of patients as dataset1 that consists of 988 persons as male and females, bigger than or equal 6 years old, and 747 are mellitus diabetic disease ,241 are no-mellitus diabetic disease. Dataset1 is accumulate for 14 months from patients of medical center in Baghdad-city depend on some questionnaire about the conditions.

Some analyzing and preprocessing on the dataset was done to remove the redundant and to complete a missing values, so some encoding of attributes of each record of information using correlation matrix .Table2 describes this collected dataset1. Dataset2 is PIDD from the UCI-Data World[23][24]. Table3 shows the main features of dataset2 that included a total of 521 samples, 201 of them were no-diabetic, and 320 cases are diabetic.

#### 3.1.1 Data Preprocessing

Some of preprocessing steps are applied to organize the raw data for using it in analytic process, so missing-data and duplicate-data are preprocessed to obtain efficient data analysis. Algorithm1 shows the preprocessing of missing data in dataset. Algorithm2 shows the steps of cleaning dataset from duplicate the data-record.

#### 3.1.2 Attribute Data Normalization

Normalization is a proactive step for the classification step and its importance appears when the data values appear in a disparate and inconsistent manner, which affects the classification results.

*Table2: Main features of Dataset1*

| # | Feature | Feature-Value | Type |
|---|---|---|---|
| 1 | Gender | 1-Male/2- Female | Nominal |
| 2 | Age | 6-80 years-old | Numerical |
| 3 | Hb-A1C | 4.80-5.90 % | Numerical |
| 4 | Glucose | 4.11-6.05 mmol/L | Numerical |
| 5 | HDL Cholesterol | 0.78-1.6 mmol/L | Numerical |
| 6 | LDL Cholesterol | 0-2.6 mmol/L | Numerical |
| 7 | Total Cholesterol | <5.2 | Numerical |
| 8 | Triglyceride | 0.86-1.9 mmol/L | Numerical |
| 9 | Creatinine | 62-106 µmol/L | Numerical |
| 10 | Case | 1-Pve/2-Nve | Nominal |

*Table3:Main features of Dtaset2*

| # | Feature | Feature-Value | Type |
|---|---|---|---|
| 1 | Glucose | Plasma-glucose after 2hrs in the glucose- tolerance test | Numerical |
| 2 | Blood pressure | mmHg | Numerical |
| 3 | Skin thickness | mm | Numerical |
| 4 | Insulin | 2h-sinsulin (muU/ml) | Numerical |
| 5 | BMI | Mass-index of Body (kg/m) : Sqr(weight/height) | Numerical |
| 6 | DPF | Diabetes pedigree function | Numerical |
| 7 | Age | =>25 to <85 | Numerical |
| 8 | Result | 1-Negative/2-Positive | Nominal |

**Algorithm1: Data filling Algorithm**

Input: dataset
Output: filling dataset

Begin
Step 1: Check if no-field-value then replace with "0"
Step 2: Remove all records with "0" value
Step 3:Return filling dataset
End

**Algorithm2: Non-duplicate data-record Algorithm**

Input: Filling dataset
Output:Cleaning dataset

Begin
Step 1: for each data-record
Step1.1 Begin
Step1.2: for each attribute in data-record
Step1.2.1: Begin
Check if values are unique and not duplicate then go step1.2.2
Else remove One of the double record
Step1.2.2: End
Step2: End
Step3: Return cleaning Dataset
End

Eq1 shows the Z-score approach of normalization, A values are normalized on a basis of the mean as well as the standard deviation of A .

$$V' = \frac{(v - \bar{A})}{\sigma_A} \dots\dots (1)$$

In which $\sigma_A$ and $\bar{A}$ are mean and standard deviation of the attribute (A), respectively.

Journal of Theoretical and Applied Information Technology
15th July 2022. Vol.100. No 13

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

### 3.2 Feature Selection Part

Feature selection stage, might be utilized for increasing the performance of classification i.e. improves the quality of data in dataset by excluding the redundant and irrelevant values, reduces complexity, help in understanding the classification with accurate, and assist in updating the model chi-square test, Recursive feature elimination are filter category of feature selection.

#### .3.2.1 Chi(χ2)-test as Feature Optimal-Extraction.

The Chi-(χ2)-test used for determining the best characteristics for a certain dataset via specifying these characteristics that affecting the output class label. The test was act on 2 groups for determining the likeness level concerning their variances. Eq2 shows the χ2 , algorithm3 shows the χ2 test.

$$\chi 2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \dots \dots (2)$$

where $O_{ij}$ $is$ $an$ observed value, and $E_{ij}$ $is$ $an$ expected value. The features are selected and classified by the lowest chi-square test score.

#### 3.2.2 Recursive Feature Elimination for best feature Selection.

RFE method goals to detect the finest achieving of some features. At the beginning the training on the primary features then weighting each of them. After that selected only lowest one of weight among primary features then re-find the weight of the remain features. These steps are repeated awaiting the preferred number of features is achieved that means is the end criterion. Algorithm4 summarize the operation of RFE method.

**Algorithm 3: Finding Best Features using Chi-square Test**

Input: Dataset , K= number of required-features

Output: Selected best-features

```
Begin
Step1: N=number of features in dataset,
Set counter i=1
Step2: Begin
    Step2.1: C[i] = Chi-Squared (feature in dataset[i])
    Step2.2: increase counter i by 1
    Step2.3: if i<=N then goto step2.1
Step3:End
Step4: Sort the C-array from low value to high
Step5: Select the k features based on the first k elements in C-array
End
```

#### 3.2.3 Combination Approach of Feature Selection

A combination method is used for lowering the features number of dataset to a least possible point to get the most important features by evaluate the results of **χ2-test** and RFE methods. Algorithm5 shows the combination approach of feature selection.

### 3.3. Classification Part

Classification part is based on use of KNN, NB, and LR classification methods to accurately diagnosis of diabetes mellitus if a patient has a diabetic disease or not. Algorithm6 explains the common steps before beginning of classification part. The most important parameter of KNN algorithm is K with value=11 by experiments. Algorithm7 presents the KNN classifier[25]. Algorithm8 introduces the NB-classifier[26]. Algorithm9 explains the Logistic Regression as classifier construction, training, and testing[27]. All these algorithms take the dataset as input, classify it, and apply the classifier. Then it checks the obtained accuracy

**Algorithm 4: Finding Best Features using FRE**

Input: Dataset , K= number of required-features

Output: Selected best-features

```
Begin
Step1:Set EPF, LWF,NF
Step2: While EPF(Primary features) not empty do
   Begin
    Step 2.1:Training using PF
    Step2.2: calculate  weight for PF
    Step2.3: Select feature with lowest weight
    Step2.4: LWF=  all feature with lowest weight
    Step2.5: EPF=EPF-LWF
    Step2.6: compute the number of elements(NF)in EPF
    End
  End
```

**Algorithm 5: Combination method to get Best Feature**

Input: Selected features from algorithm3, Selected features from algorithm4.

Output: Selected best features

```
Begin
   Step1: assign set1= Selected features from algorithm3
   Step2: assign set2= Selected features from algorithm4
   Step3: set3= set1 ∩ set2
   Step4: Selected best features= Set3
End
```

**Algorithm 6: Common Steps**

**Input:** Dataset

**Output:** clean Dataset with best feature selection

```
Begin
Step1:call algorithm1   /* filling the missing data
Step2: call algorithm2   /* no-duplicate data
Step3: call algorithm3    /* Z-score Normalization
Step4:call algorithm4    /*Chi square test as FS
Step5:call algorithm5   /* RFE as FS
Step6: break the original-dataset into 75% training-set  and 25% as testing-set
 End
```

### .4. Results's Discussion of the Proposed Method

First of all, part1 includes some tasks are performed on diabetes-disease dataset1 and dataset2 to be ready for the part2 that includes Chi-square Test and Recursive feature elimination as are filter methods to find the optimal features as shown in table4. The

Journal of Theoretical and Applied Information Technology
15th July 2022. Vol.100. No 13

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Algorithm 7: KNN classifier

**Input:** N_training, t_features, m_classes

**Output:** class label, Classification Result, Accuracy-rate

**Begin**

**Step1:**Call algorithm6 /* Common Steps of the proposed work

**Step2:** N training-set has t- features and it may be part of one of m-classes. /* **K** value is 11

**Step3.** Let O be the testing-object that needs classification.

**Step4.** Calculate distance between testing-object O and every training-set.

**Step5.** Let $d_1, d_2, \ldots, d_N$ be the resultant distances.

**Step6.** Sort distances from low to high order and identifying 1st k objects that correspond to 1st k lowest distances for the purpose of getting the set $K_C$.

**Step7.** Let $x_r$ be the number of the objects in a set $K_C$ belong to a class r (r= 1, 2,…, m). The object is assigned to class λ if, $\pi_\lambda = \max \{ x_1, x_2, \ldots, x_m \}$

Step8: If obtained Accuracy is **NOT** acceptable, then go to Step2

END

independent feature has high-value of **χ2**-test means this feature is ignored but low-value of RFE means that feature is independent and ignored because the important of feature is based on the rank of this feature. Part3 as classification includes KNN, NB,

Algorithm 8: NB classifier

**Input:** D:training dataset of n features with attribute with n value A1 00 An, X:testing-set, C:classes

**Output:** class label of testing-set, Classification result, Accuracy

**Begin**

**Step1:**Call algorithm6 /* Common Steps of the proposed work

**Step2.** Calculate class prior probabilities using the following

$P(C) = \frac{|C_i,D|}{|D|}$ (5)

Where $|C_i, D|$ is the number of training features of the category $C_i$ in D.

**Step3.** To decrease calculation in computing $P(X|C_i)$, the conditional independence of the class hypothesis is performed. i.e. all features are independent and there aren't any relationships between features as the following $P(X|C_i) = \prod_{k=1}^{m} P(x_k, C_i)$ (6)

**Step4.** Calculate mean ($\mu_{ci}$) and standard deviation ($\sigma_{ci}$), of trainingfeatures values of class Ci. Employ those two values:

$g(x, \mu, \boldsymbol{\sigma}) = \frac{1}{\sqrt{2\pi}} e^{(x-\mu)^2} 2\sigma^2$ and $P(X_K|C_i) = g(X_K, \mu_{ci}, \sigma_{ci})$ (7)

**Step5.** For predicting the class label of the testing set X, for every class $C_i$estimates $P(X_K|C_i)P(C_i)$. The predication is done using $P(X_K|C_i)\,P(C_i) > P(X_K|C_j)\,P(C_j)$, for $(i \le j \le m, j \ne i)$ (8)

Step6: If obtained Accuracy is **NOT** acceptable, then go to Step2

END

and LR algorithms which the results of this part depend strongly on part2 as shown in table5 and fig3. Table6 illustrates result of classifier algorithms after the **χ2**-test as a feature selection process.

In all cases the selected attributes are indicated by 1, and the removed attributes indicated by 0. The results of Dataset1 using Chi-square test gave KNN of 98%,97% and 96% accuracy[28] with 5,6 and 7-feature selection respectively. LR and NB are achieved 98% accuracy with 5 and 6- features

Algorithm 9: LR Algorithm

**Input:** training-set, testing-set, p

**Output:** Class label ,Classification Result, Accuracy

**Begin**

**Step1:** Call algorithm6 /* Common Steps of the proposed work

**Step2**: for each row in training-set *logit= 0*

**Step2.1**: for each I of weight factor in a row

*logit=logit+weight[I]*x[I]*

end for

**Step2.2**: Calculate Sigmoid of *logit* using

$P(logit) = \frac{1}{1+e^{(-logit)}}$ ……(9)

**Step2.3:** Update weight for each j weight

*Weight[j]+=rate_of_learning*(Y-p(logit))*x[j]*

end for

**Step2.4**: Update likelihood function using eq.10

$like \mathrel{+}= Y * log\left(\frac{1}{1+e^{(-log\ )}}\right) + (1-Y)\ log\left(1 - \frac{1}{1+e^{(-logit)}}\right)$ ….(10)

**Step3:** for each sample in testing_set

For each weight factor i

*z = z+weight[i]*x[i]*

end for

**Step3.1:** calculate predication for testing set using

$P(z) = \frac{1}{1+e^{(-z)}}$ …… (11)

**Step3.2**: If ( $p \ge 0.5$ ) : successful ( coded as 1 )

Else ( $p < 0.5$ ) : failed ( coded as 0)

/* $0 < p < 1$ represent threshold variable that is typically equal to 0.5.

**Step4:** If obtained Accuracy is **not** acceptable, then go to Step2

End

selection. So the results of PIMA-dataset2 state that KNN of 84% accuracy with 7-features. LR achieved 80% accuracy with 5-features, while 77% accuracy of 6 and 7-feature selection. NB achieved 89% accuracy with 6-features while had 86% accuracy with 5 and 7-features selection.

*Table4: χ2-Test and RFE Results of Dataset*

| # | Feature No | Feature name | Chi-square Test | RFE |
|---|---|---|---|---|
| Dataset1 | 1 | Gender | 0.584781434 | 0.0 |
| | 2 | Age | 0.018740373 | 0.145813806 |
| | 3 | HbA1C | 1.51E-08 | 0.643129425 |
| | 4 | Glucose | 2.1812E06 | 0.4056762672 |
| | 5 | HDL Cholesterol | 0.546116202 | 0.052177187 |
| | 6 | LDL Cholesterol | 0.332325304 | 0.027210943 |
| | 7 | Total Cholesterol | 0.369979616 | 0.064042661 |
| | 8 | Triglyceride | 0.021615962 | 0.161872419 |
| | 9 | Creatinine | 0.668813144 | 0.09803126 |
| Dataset2 | 1 | Pregnancy | 0.006680855 | 0.059964804 |
| | 2 | Glucose | 0.000297356 | 0.145154164 |
| | 3 | Blood pressure | 0.230610704 | 0.047338336 |
| | 4 | Skin thickness | 0.097142405 | 0.00738906 |
| | 5 | Insulin | 0.006499296 | 0.123023086 |
| | 6 | BMI | 0.061729437 | 0.016359698 |
| | 7 | DPF | 0.147204411 | 0.0 |
| | 8 | Age | 0.000415715 | 0.121383742 |

Notice that of table5, the NB achieved the highest accuracy of (96%) using Dataset1 free of feature selection, while KNN and LR are achieved the accuracy of (90%) and (96%) respectively while NB was the highest rate of accuracy(81) using Dataset2 free of features selection. In table6, results of KNN,

LR, and NB achieved the highest accuracy of (98%) using Dataset1with Chi-square test. Table7 illustrates the result of classifier algorithms after the RFE as a feature selection process, when using dataset1 based on RFE as feature selection gave KNN of 90% accuracy, 94% accuracy, and 98% accuracy when use five, six, and 7-features respectively.

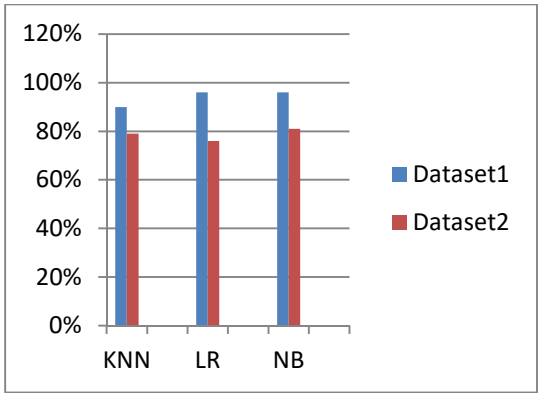


*Figure3: Classification methods Accuracy*

The accuracy of using LR was upper than 94%, and NB gave the accuracy was 95%. On other side the results of Dataset2 based RFE as feature selection gave KNN achieved 80% accuracy with 7-features. LR achieved 80% accuracy with 5 and 6-features, while achieved 79% accuracy with 7-features. NB achieved 83% accuracy when use 5-features, and 86% accuracy when use 6 and 7-features. Table8 and fig.4 introduce the final-result of classifier-part that deals with Dataset1 and Dataset2 based on combining-**χ2**test-RFE as feature selection. The results of this combination features gave better results with less number of features such that when Dataset1 was use, the classifier KNN, LR, and NB achieved 98% accuracy with four, five and six features. On other hand the results of Dataset2 gave KNN achieved 86% accuracy with 4-features, while achieved 81% accuracy with 6-features. LR achieved 90% accuracy with 4-features, while 86% accuracy with 5-features and achieved 87% accuracy with 6-features. NB achieved 83% accuracy with 4-features, 87% accuracy with 5-features and 90% accuracy with 6-features. Therefore all accuracy rates of classifier methods based the combination feature selection were better than the rates free of the feature selection. Table9 shows the results of comparative of the proposed method with some previous researches.

**5. Conclusions:**

Feature selection Part is enhanced the performance of the diagnosis model and that enhancement depend on the type of the feature selection method. Part3 as classifier process based on NB achieved the maximum accuracy=98% on Dataset1and use IG as feature selection, while KNN, LR achieved maximum accuracy=96%. KNN, LR, and NB given accuracy=98% on Dataset1 use combination **χ2**test-RFE as feature selection. So KNN achieved the accuracy of=81% on Dataset2 without feature selection, while LR achieved the accuracy of=76%, NB achieved the accuracy =85%. KNN achieve a maximum accuracy=85% on Dataset2-Pima using combination feature selection, while LR, NB could achieve a maximum accuracy=90% .Diabetes diagnosis requires feature selection stage because without this stage, the diagnostic accuracy rate decreases to 76%.The process of selecting the properties has a large impression on the proposed work. The outcomes have proven that the combination method for selecting the characteristics proposed by the researcher is better than Chi-Square method and Recursive Feature Elimination method each alone.
The limitation that faced the research is the process of collecting data on the DM and purifying the dataset

*Table5:Classifire Results free of Feature Selection*

| # | Method | PR-Rate | RE-Rate | FScore | ACC-Rate | CM-Values | |
|---|---|---|---|---|---|---|---|
| Dataset1 | KNN | 91% | 87% | 89% | 90% | 20 | 2 |
| | | | | | | 3 | 25 |
| | LR | 95% | 95% | 95% | 96% | 20 | 1 |
| | | | | | | 1 | 28 |
| | NB | 96% | 96% | 96% | 96% | 22 | 1 |
| | | | | | | 1 | 26 |
| Dataset2 | KNN | 85% | 89% | 86% | 79% | 44 | 8 |
| | | | | | | 6 | 10 |
| | LR | 93% | 77% | 85% | 76% | 42 | 3 |
| | | | | | | 12 | 13 |
| | NB | 87% | 88% | 88% | 81% | 47 | 7 |
| | | | | | | 6 | 10 |

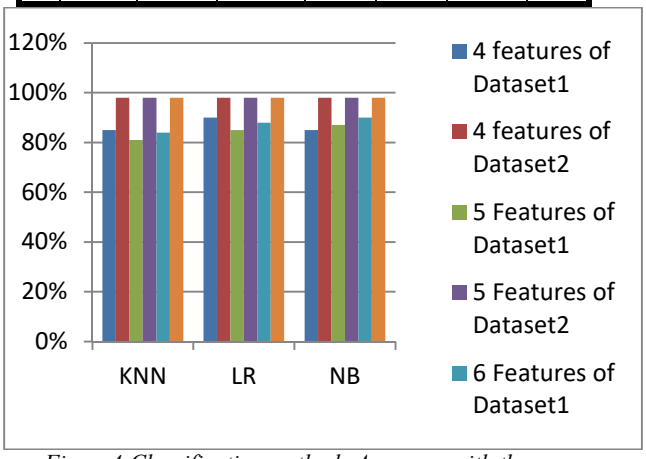


*Figure4:Classification methods Accuracy with the Combination Feature Selection*

## REFERENCES

[1] Z. Punthakee , R. Goldenberg, P. Katz, "Definition, Classification and Diagnosis of Diabetes, Prediabetes and Metabolic Syndrome" , Diabetes Canada Clinical Practice Guidelines Expert Committee, Canadian Journal of Diabetes, s 42 (2018) S10–S15.

[2] A. Azrar, M. Awais, Y. Ali, and K.Zaheer ,"Data mining models comparison for diabetes prediction", Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 8, pp. 320–323, 2018, doi: 10.14569/ijacsa.2018.090841.

[3] R.Taylor," Types of Diabetes Mellitus", WebMD, December08,2021,https://www.webmd.com/diabetes/guide/types-of-diabetes-mellitus

[4] B. S. Lal, "DIABETES : CAUSES , SYMPTOMS AND TREATMENTS DIABETES : CAUSES , SYMPTOMS AND TREATMENTS", no. December, 2016.

[5] Sh. H.Shaker , A.Saeid ,R.Ogla," IID3 Classifier To Diagnosis Of High Blood Glucose Levels During Pregnancy", Webology,ISSN: 1735-188X, Volume 19, Number 2, 2022.

[6] K. Thangadurai and N. Nandhini, "Comparison of data mining algorithms for prediction and diagnosis of diabetes mellitus", vol. 7, no. 5, pp. 221–224, 2016.

[7] K. Saravanapriya and J. Bagyamani, "Performance Analysis of Classification Algorithms on Diabetes Dataset", *Int. J. Comput. Sci. Eng.*, vol. 5, no. 9, pp. 15–20, 2017, doi: 10.26438/ijcse/v5i9.1520.

[8] K. Akyol and B. Sen, "Diabetes Mellitus Data Classification by Cascading of Feature Selection Methods and Ensemble Learning Algorithms", *Int. J. Mod. Educ. Comput. Sci.*, vol. 10, no. 6, pp. 10–16, 2018, doi: 10.5815/ijmecs.2018.06.02.

[9] M. F. Faruque, Asaduzzaman, and I. H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus", *2nd Int. Conf. Electr. Comput. Commun. Eng. ECCE 2019*, pp. 1–4, 2019, doi: 10.1109/ECACE.2019.8679365.

[10] R. Kumarmangal, and et.al. ," An Enhanced Machine Learning Framework for Type 2 Diabetes Classification Using Imbalanced Data with Missing Values", Wiley Hindawi, Volume 2021, Article ID 9953314, 21 pages https://doi.org/10.1155/2021/9953314

[11] Y. Meliz, Y. Belma, and T. Özgür, " Classification Comparison of Machine Learning Algorithms Using Two Independent CAD Datasets", Mathematics journal,10,311. https://doi.org/10.3390/math10030311

[12] Sun, Y.L., Zhang, D.L., " Machine learning techniques for screening and diagnosis of diabetes: a survey", . Tehnic˘ki vjesnik. 26 (3), 872–880,2019.

[13] M. Saar and F. Provost, "Handling missing values when applying classification models," Journal of Machine Learning Research, vol. 8, pp. 1625–1657, 2007.

[14] Y. Mitchell, and et.al. "Unsupervised feature selection using swarm intelligence and consensus clustering for automatic fault detection & diagnosis in heating ventilation and air conditioning systems." Applied Soft Computing 34 (2015): 402-425.

[15]D.umais, and et.al, "Inductive learning algorithms and representations for text categorization." In Proceedingsof the seventh international conference on Information and knowledge management, pp. 148-155. ACM, 1998.

[16] T. Sharma, A. Sharma, V. Mansotra," Performance Analysis of Data Mining Classification Techniques on Public Health Care Data", 2016. Available online: https://www.researchgate.net/publication/313571291P (accessed on 22 August 2021).

[17]D.Abdlqader, A.Abdlazeez,D.Zeebaree,." ML Supervised Algorithms of Gene Selection :Review".Technol.Rep.Kansai Univ. 2020, 62.

[18] A. Dwivedi," Performance evaluation of different machine learning techniques for prediction of heart disease",Neural Comput. Appl. 2016, 29, 685–693.

[19] F. Ahmed, and et. Al, " A hybrid machine learning framework to predict mortality in paralytic ileus patients using electronic health records (EHRs). J. Ambient. Intell. Humaniz. Comput. 2021, 12, 3283–3293.

[20] S. Ashraf, T. Ahmed, Z.Aslam, and D. Muhammad," Conversion of adverse data corpus to shrewd output using sampling metrics" . Vis. Comput. Ind. Biomed. Art 2020, 3, 19.

[21] T. Wuest,D. Weimer, C. Irgens, K. Thoben, " Machine learning in manufacturing: Advantages, challenges, and applications. Prod. Manuf. Res. 2016, 4, 23–45.

[22] C. Guo ,"Enhancing Face Identification Using Local Binary Patterns and K-Nearest Neighbors". J. Imaging 2017, 3, 3.

[23] "Data. World. datasets. open data: Pima Indian Diabetes Database" [Online] . Aialable: data. world./ data- society/ Pima - indians - diabetes - database.

[24] S. Ahmed, A.Tariq,"COMPARISON OF DATA MINING ALGORITHMS FOR DIAGNOSIS OF DIABETES MELLITUS", International Journal of Computer Science and Engineering (IJCSE) ISSN (P): 2278–9960; ISSN (E): 2278–9979 Vol. 10, Issue 2, Jul–Dec 2021; 1–8

[25] S. Selvkumr, K. Kannan, & S. Gothaichiyar, "Prediction of Diabetes Diagnosis Using Classification Based Data Mining Techniques", *Int. J. Stat. Syst.*, vol. 12, no. 2, pp. 183–188, 2017

[26] Nurhayati and A.Rahman, "Implementation of Naive Bayes and K-Nearest Neighbor Algorithm for Diagnosis of Diabetes Mellitus", *Proc. 13th Int. Conf. Appl. Comput. Appl. Comput. Sci. (ACACOS '14)*, pp. 117–120, 2014.

[27] M. Soukla, and et al.," Prediction Model Using Logistic Regression", *Int. J. Eng. Res. Technol.*, vol. 7, no. 04, pp.12–19, 2019.

[28] C. Oprea, "Performance Evaluation of the Data Mining Classification Methods", Analele Univ. Constantin Brâncuşi din Târgu Jiu Ser. Econ., vol. 1, no. Special number-Information society and sustainable development, pp. 249–253, 2014.

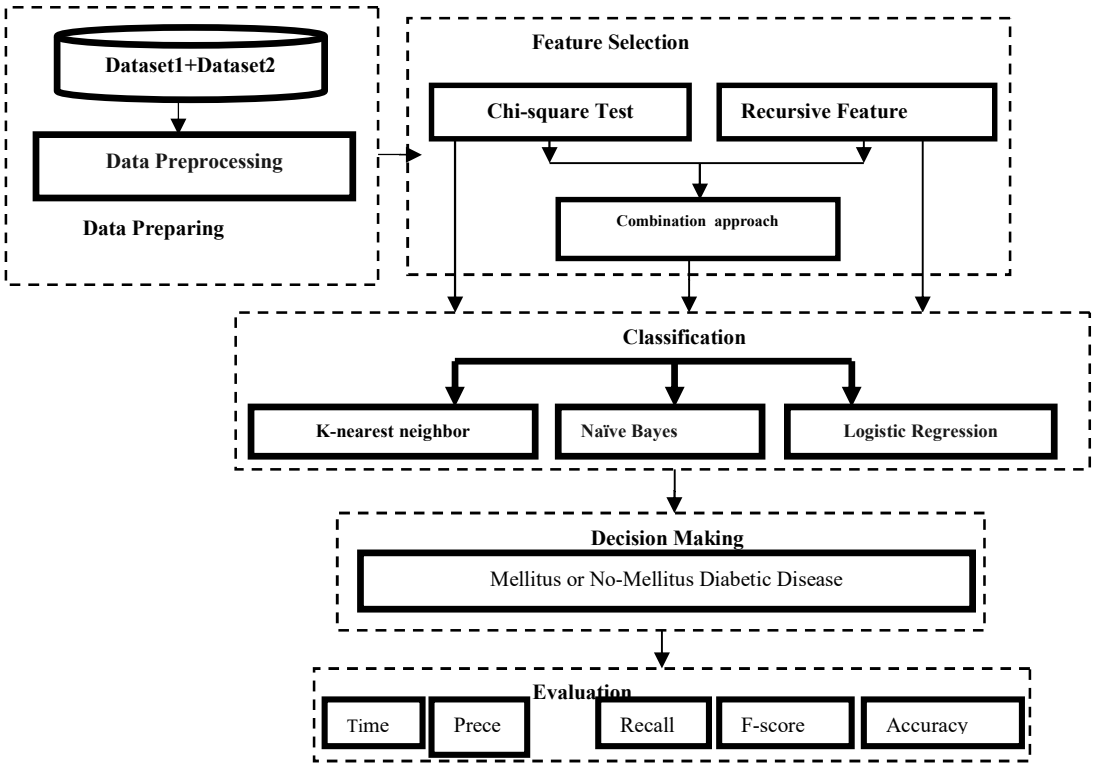


*Figure2: Main steps of proposed Diagnosis Diabetes Mellitus*

*Table9: Comparative of accuracy rate*

| Year-Reference# | | Accuracy |
|---|---|---|
| 2016[6] | | 78.1% |
| 2017[7] | | 80% |
| 2018[8] | | 73.88% |
| 2019 [9] | | 73% |
| 2021[10] | | 98 % |
| Proposed method (combination FS) | Dataset1 | 98% |
| | Dataset2 | 90% |

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

*Table 6: Results of classifier Algorithms with χ2-test Feature Selection*

| # | Selected features | Method | PR-Rate | RE-Rate | FScore | ACC-Rate | CM-Values | |
|---|---|---|---|---|---|---|---|---|
| Dataset1 | 110001110(5) | KNN | 95% | 83% | 89% | 90% | 20 | 1 |
| | | | | | | | 4 | 25 |
| | | LR | 95% | 90% | 93% | 94% | 19 | 1 |
| | | | | | | | 2 | 28 |
| | | NB | 98% | 98% | 98% | 98% | 26 | 0 |
| | | | | | | | 1 | 23 |
| | 111001110(6) | KNN | 96% | 92% | 94% | 94% | 22 | 1 |
| | | | | | | | 2 | 25 |
| | | LR | 95% | 95% | 95% | 96% | 19 | 1 |
| | | | | | | | 1 | 29 |
| | | NB | 98% | 96% | 98% | 98% | 25 | 0 |
| | | | | | | | 1 | 24 |
| | 111011110(7) | KNN | 90% | 86% | 88% | 90% | 19 | 2 |
| | | | | | | | 3 | 26 |
| | | LR | 95% | 95% | 95% | 96% | 19 | 1 |
| | | | | | | | 1 | 29 |
| | | NB | 98% | 96% | 98% | 98% | 26 | 0 |
| | | | | | | | 1 | 23 |
| Dataset2 | 10010111(5) | KNN | 79% | 89% | 84% | 77% | 42 | 11 |
| | | | | | | | 5 | 12 |
| | | LR | 91% | 81% | 86% | 80% | 44 | 4 |
| | | | | | | | 10 | 12 |
| | | NB | 88% | 88% | 88% | 83% | 42 | 6 |
| | | | | | | | 6 | 16 |
| | 10110111(6) | KNN | 83% | 88% | 85% | 78% | 43 | 9 |
| | | | | | | | 6 | 11 |
| | | LR | 91% | 81% | 86% | 80% | 43 | 4 |
| | | | | | | | 10 | 13 |
| | | NB | 94% | 86% | 90% | 86% | 44 | 3 |
| | | | | | | | 7 | 16 |
| | 10111111(7) | KNN | 83% | 90% | 86% | 80% | 43 | 9 |
| | | | | | | | 5 | 12 |
| | | LR | 91% | 80% | 85% | 79% | 43 | 4 |
| | | | | | | | 11 | 12 |
| | | NB | 93% | 86% | 90% | 86% | 43 | 3 |
| | | | | | | | 7 | 17 |

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

*Table7: Results of Classifier Algorithms with RFE Feature selection*

| # | Selected features | Method | PR-Rate | RE-Rate | FScore | ACC-Rate | CM-Values | |
|---|---|---|---|---|---|---|---|---|
| Dataset1 | 010101110(5) | KNN | 97% | 96% | 98% | 98% | 22 | 0 |
| | | | | | | | 1 | 27 |
| | | LR | 97% | 97% | 98% | 98% | 19 | 1 |
| | | | | | | | 0 | 30 |
| | | NB | 97% | 98% | 98% | 98% | 21 | 0 |
| | | | | | | | 1 | 28 |
| | 011101110(6) | KNN | 96% | 96% | 96% | 97% | 20 | 1 |
| | | | | | | | 1 | 28 |
| | | LR | 97% | 97% | 98% | 98% | 19 | 1 |
| | | | | | | | 0 | 30 |
| | | NB | 97% | 98% | 98% | 98% | 21 | 0 |
| | | | | | | | 1 | 28 |
| | 011111110(7) | KNN | 96% | 97% | 96% | 96% | 21 | 0 |
| | | | | | | | 2 | 27 |
| | | LR | 97% | 96% | 96% | 96% | 19 | 1 |
| | | | | | | | 1 | 29 |
| | | NB | 98% | 98% | 98% | 97% | 21 | 0 |
| | | | | | | | 1 | 28 |
| Dataset2 | 10110011(5) | KNN | 81% | 90% | 85% | 79% | 45 | 8 |
| | | | | | | | 6 | 11 |
| | | LR | 91% | 81% | 86% | 80% | 43 | 4 |
| | | | | | | | 10 | 13 |
| | | NB | 96% | 85% | 89% | 86% | 44 | 2 |
| | | | | | | | 8 | 16 |
| | 1011011(6) | KNN | 91% | 83% | 87% | 81% | 43 | 4 |
| | | | | | | | 9 | 14 |
| | | LR | 89% | 79% | 84% | 77% | 42 | 5 |
| | | | | | | | 11 | 12 |
| | | NB | 94% | 90% | 92% | 89% | 44 | 3 |
| | | | | | | | 5 | 18 |
| | 11111011(7) | KNN | 84% | 93% | 87% | 84% | 43 | 8 |
| | | | | | | | 3 | 14 |
| | | LR | 89% | 76% | 82% | 74% | 41 | 5 |
| | | | | | | | 13 | 11 |
| | | NB | 96% | 85% | 90% | 86% | 45 | 2 |
| | | | | | | | 8 | 15 |

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

*Table8: Results of part3 based combining Feature Selection*

| # | Selected features | Method | PR-Rate | RE-Rate | FScore | ACC-Rate | CM-Values | |
|---|---|---|---|---|---|---|---|---|
| Dataset1 | 010001110(4) | KNN | 98% | 97% | 98% | 98% | 25 | 0 |
| | | | | | | | 1 | 24 |
| | | LR | 95% | 95% | 95% | 96% | 19 | 1 |
| | | | | | | | 1 | 29 |
| | | NB | 98% | 96% | 98% | 98% | 21 | 0 |
| | | | | | | | 1 | 28 |
| | 011001110(5) | KNN | 98% | 96% | 98% | 98% | 25 | 0 |
| | | | | | | | 1 | 24 |
| | | LR | 96% | 98% | 98% | 98% | 19 | 1 |
| | | | | | | | 0 | 30 |
| | | NB | 98% | 96% | 98% | 98% | 21 | 0 |
| | | | | | | | 1 | 28 |
| | 011011110(6) | KNN | 98% | 95% | 98% | 98% | 20 | 0 |
| | | | | | | | 1 | 29 |
| | | LR | 96% | 98% | 98% | 98% | 19 | 1 |
| | | | | | | | 0 | 30 |
| | | NB | 98% | 96% | 98% | 98% | 21 | 0 |
| | | | | | | | 1 | 28 |
| Dataset2 | 10010011(4) | KNN | 92% | 91% | 91% | 86% | 48 | 5 |
| | | | | | | | 5 | 12 |
| | | LR | 87% | 85% | 86% | 90% | 49 | 3 |
| | | | | | | | 5 | 13 |
| | | NB | 91% | 84% | 88% | 83% | 42 | 4 |
| | | | | | | | 8 | 16 |
| | 10110011(5) | KNN | 85% | 86% | 85% | 79% | 44 | 8 |
| | | | | | | | 7 | 11 |
| | | LR | 92% | 89% | 90% | 86% | 48 | 4 |
| | | | | | | | 6 | 12 |
| | | NB | 96% | 86% | 90% | 87% | 44 | 2 |
| | | | | | | | 7 | 16 |
| | 10111011(6) | KNN | 88% | 87% | 87% | 81% | 45 | 6 |
| | | | | | | | 7 | 12 |
| | | LR | 90% | 92% | 91% | 87% | 47 | 5 |
| | | | | | | | 4 | 14 |
| | | NB | 96% | 96% | 93% | 90% | 45 | 2 |
| | | | | | | | 5 | 18 |