Computational Approach to Clinical Diagnosis of Diabetes Disease: A Comparative Study

Deepak Gupta[1], Ambika Choudhury[1], Umesh Gupta[1], Priyanka Singh[2], Mukesh Prasad[2]

[1] Computer Science & Engineering, National Institute of Technology Arunachal Pradesh, Yupia, India
[2] School of Computer Science, FEIT, University of Technology Sydney, Sydney, Australia

Abstract: Diabetes is one of the most prevalent non-communicable diseases and is the 6th leading cause of death worldwide. It's a chronic metabolic disorder which has no cure, however, it is a highly treatable condition, if diagnosed and managed on time to avoid its complications. This paper explores and compares various machine learning (ML) approaches that can help in determining the risk of diabetes at an early stage and aid in improving the medical diagnosis of diabetes. The paper considers two real-world datasets one is a diabetic clinical dataset (DCA) collected from a medical practitioner in the state of Assam, India during the year 2017-2018 and other is public PIMA Indian diabetic dataset. To analyze the various machine learning techniques on DCA and PIMA Indian diabetic datasets for the classification of diabetic and non-diabetic patients, different classifiers like perceptron, Gaussian process, linear discriminant analysis, quadratic discriminant analysis, statistical gradient descent, ridge regression classifier, support vector machines, k-nearest neighbors, decision tree, naïve Bayes, logistic regression, random forest and ELM for multiquadric, RBF, sigmoid activation functions are used. The results of numerical experiments suggested that logistic regression yields better performance in comparison to the other techniques.

Keywords: Machine learning; diabetes; support vector machines; logistic regression; k-nearest neighbors; ELM.

1. Introduction

Diabetes or Diabetes Mellitus is a non-communicable disease which belongs to a group of metabolic disorders. In 2019, around 9.3 percent of the global adult population suffered from diabetes and it is expected to rise almost 11 percent by the year 2045 [26]. Diabetes is a chronic condition that occurs when the pancreas does not produce sufficient amount of insulin, or specific cells do not respond to insulin [72] resulting in elevated sugar levels in the blood. Insulin is a hormone that regulates sugar levels in the blood. Persistent high blood sugar levels can cause harm to various organs of the body like heart, blood vessels, eyes, kidneys and nerves. Diabetes is mainly of three types: type 1 diabetes, type 2 diabetes and gestational diabetes [94].

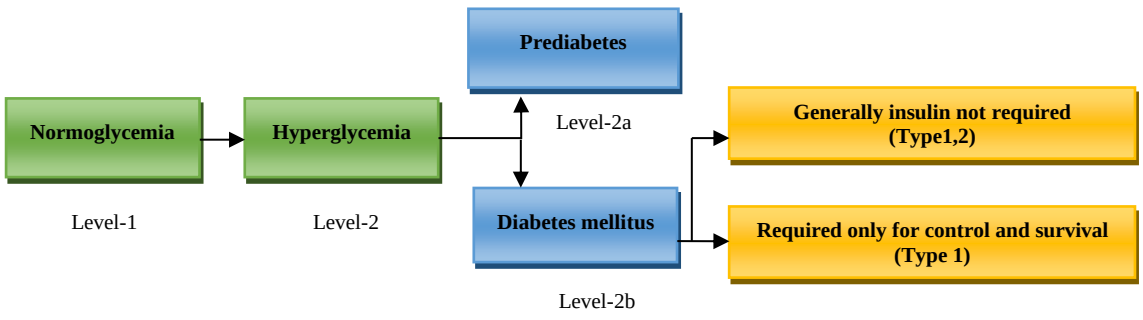

Figure 1: Graphical representation of Type -1 and Type-2 diabetes stage wise.

Type 1 diabetes is also called juvenile diabetes which results due to insufficient insulin production and requires daily insulin administration. Type 2 diabetes occurs in adults due to ineffective use of insulin by the body. This type 2 is the commonest type of diabetes found worldwide. Gestational diabetes occurs mainly during pregnancy and it gets resolved after the baby is delivered [108]. However, it is estimated that one in four live births is affected by hyperglycemia in pregnancy [54]. The graphical representation is plotted in Figure 1.

According to the World Health Organization [107], 422 million people worldwide suffered from diabetes in 2014 and estimated deaths attributed to it were 1.6 million making it the 6th leading cause of death in 2016. According to estimates from the International Diabetes Federation 2019 [46], 79% of adults with diabetes were living in low and middle-income countries and cases of diabetes will rise to 700 million by 2045 worldwide. From an economic perspective, diabetes caused at least USD 760 billion in health expenditure in 2019. IDF further adds that 1 in 2 (232 million) people with diabetes were undiagnosed. Underdiagnosis of diabetes increases the risk for delay in treatment and this, in turn, increases the risks of developing kidney disease, blindness, nerve damage, heart diseases, stroke and blood vessel damage. Although diabetes is incurable, it can be treated and its consequences can be avoided or delayed with diet, physical activity, medication and regular screening and treatment for complications. Hence, its utmost importance to diagnose diabetes early and accurately, so that the patients can be treated on time resulting in improved health outcomes and reduced economic burden of disease on the health systems. Therefore, it is of interest to undertake a study that helps in improving the accuracy of clinical diagnosis of diabetes. Various machine learning methods have been commonly adopted to diagnose the most prevalent illness, such as diabetes, hepatitis, cancer, tumour and many more in recent times. Several supervised, semi-supervised and unsupervised learning are involved in early detection of risk to diagnose diabetes either type 1, 2 or gestational diabetes. Polat&Güneş [77] have suggested a two-phase system, which is the combination of principal component analysis (PCA) and adaptive neuro-fuzzy inference system (ANFIS) to improve the diagnosis of diabetic disease where dimensionality reduction is done by PCA and diagnosing diabetes through automatic ANFIS system, However, PCA is not well capable to discriminate the features of the diabetes datasets efficiently. In 2008, Polat et al. [78] improved their PCA-ANFIS model by considering a hybrid approach of generalized discriminant analysis (GDA) and least square support vector machine (LSSVM) to detect diabetes effectively [78] in which GDA can work properly to discriminate the features. However, GDA-LSSVM suffers from lower generalization performance. In the field of diabetic retinopathy identification, many researchers have applied artificial neural network (ANN) and supervised learning approaches such as support vector machine (SVM), Gaussian Mixture Model (GMM), k-nearest neighbors (k-NN), and AdaBoost [1, 3, 19, 80, 87, 93, 101] and used digital fundus images [2, 30, 33, 110] to prominently handle the retinopathy diabetes. A multilayer neural network structure using the Levenberg–Marquardt (LM) algorithm and probabilistic neural network structure is also involved in the prediction of diabetes disease [99]. Temurtas et al. [99] have proposed two models as MNN-LM and PNN which is tested over PIMA Indian datasets which have shown better generalization performance in compared to other conventional neural networks as well as PCA based models but it doesn't consider the advantages of different validation based approaches.

Support vector machine is one of the most prominent machine learning algorithms which is applied directly or combined with other algorithms like the modified cuckoo search for diagnosing the diabetes disease. SVM is an approach which is based on the principle of structural risk minimization principle in contrast to neural network principles. There are many works in the literature related to SVM, some of them are SVMs with single nucleotide polymorphisms (SNPs) to handle type 2 diabetes which is generating better prediction performance [12]. For more study, one can follow:[12, 13, 35, 45, 50, 56, 60, 98, 112]. Chikh et al. [22] improved the accuracy of diagnosing diabetes disease using k-nearest neighbors based on fuzzy memberships with an artificial immune recognition system2 (AIRS2) to give the importance of each data samples. Mani et al. [67] have tested EMR data instead of PIMA Indian diabetic dataset for diagnosing the diabetes disease using two linear classifiers, one sample-based classifier, two decision tree-based classifiers, and one kernel-based classifier which gives a comparable performance. Lee et al. [61] applied naïve Bayes and logistic regression algorithms for the

prediction of fasting plasma glucose (FPG) location, which will be helpful to diagnose the type 2 diabetes [7], but overburden with the computational cost. Marini et al. [70] proposed an approach for type-1 diabetes using a dynamic Bayesian network as per DCCT/EDIC study however it is not a generalized model for diabetes. Marini et al. [71] developed a continuous-time Bayesian network (CTBN) for T2D cohort, which is an important network fitting medical literature. Joshi and Alehegn suggest another approach in the year 2017, where some known predictive algorithms are applied to diabetes data to cluster and predict symptoms [45], which predicts quickly the diabetes. Alic et al. [8] used an artificial neural network (ANN) and Bayesian networks for the categorization of diabetes and cardiovascular diseases to attain high accuracy. Teliti et al. [98] discussed the presence of risk factors for the development of microvascular complications of type 2 diabetes. Image processing techniques have some of the reliable options for type-2 diabetes diagnosis that already tested on iris infrared images [88] but it will not well performed if noise is present in the datasets. Wu et al. [109] has proposed a model that takes a series of pre-processing methods and incorporates two such machine learning algorithms namely the improved k-means and the logistic regression algorithm to make better correctness of the predictive ideal, and to assemble the ideal robust for many datasets [109]. Yadav et al. [111] have proposed a neural network-based health tracking framework for diabetes disease prediction using ANN but not sufficient prediction performance. Alade et al. [4] also used ANN, Bayesian regulation algorithm and backpropagation method to develop and train a paradigm for the prediction of diabetes in women who are pregnant [4], still facing generalization problem. Dagliati et al. [25] have suggested how various data mining and machine learning techniques [38] impacts in clinical medicine for deriving diabetic's complications. Ijaz et al. [47] have developed a mixed prediction model to deal with type 2 diabetes as well as hypertension. In 2019, some useful research work towards binary classification has been done by Tiwari &Melucci [96]; Tiwari & Melucci [97]; Jaiswal et al. [48] in the direction of medical data with machine learning approaches. In 2020, there is also a great interest in the field of research such as early detection the diabetic's [5, 21, 33, 37] by several sustainable machine learning models for type-1, type 2 [55, 58, 62, 63, 74]and retinopathy diabetes as follows [82, 95, 106]. Many researchers have presented a comparative performance analysis over diabetic datasets by using several machine learning approaches such as [51, 53, 57, 59, 64, 68, 81, 91,92].

One ulterior motive in this paper is to present a comparative study for the estimation of the most viable classification technique for diagnosing whether a patient is diabetic or non-diabetic. In this paper, several popular classification techniques have been considered for diabetes disease prediction, namely perceptron (P), Gaussian process (GP), linear discriminant analysis (LDA), quadratic discriminant analysis(QDA), stochastic gradient descent (SGD), ridge regression classifier (RC), support vector machines (SVM), k-nearest neighbors (k-NN), decision tree (DT), naïve Bayes (NB), logistic regression(LR), random forest (RF) and Extreme learning machine (ELM) for multiquadric, RBF, sigmoid activation functions. These algorithms are tested on PIMA Indian diabetic [76] as well as the DCA dataset. The detail description of the DCA dataset is described in the next section. A comparative analysis of all concerned algorithms is also presented, and the result is analysed based on accuracy, sensitivity (recall), specificity (true negative rate), precision, NPV, FP rate, RME, F_1-measure, G-mean and Matthews correlation coefficient (MCC). Here the applicability of different machine learning approaches is also validated through two famous graphs such as receiver operating characteristic curve (ROC) and precision-recall curve (PRC) where the area under ROC curve (AUC_ROC), as well as average precision score (AP), are calculated. McNemar statistical test is also performed on DCA dataset. One can observe that logistic regression shows better performance among other approaches for both PIMA Indians diabetes and the DCA dataset. The value of AUC_ROC and AP for both the datasets are maximum for logistic regression.

The rest of the paper is organized as follows: Section 2 describes the DCA dataset, section3 introduces the different classification algorithms, section 4 describes the proposed diabetic diagnosis system, section 5 evaluates the performance of algorithms on both datasets PIMA Indian diabetes and DCA and finally conclusion is concluded in section 6.

2. Description of Diabetic Clinical Assam dataset

In this paper, the 2017-2018 diabetic clinical dataset (DCA) is gathered from medical practitioner clinical information from Assam State, India. The dataset consists of 174 instances and 11 attributes (including 1 class attribute), where a total of 113 diabetic and 61 non-diabetic patients. These attributes are as follows:

1) Age (years)
2) Sex (1 for female and 2 for male)
3) History of family diabetes (0 for yes and 1 for no)
4) Weight(kilogram)
5) The decision on of blood pressure (mmHg) (3 for high, 2 for normal and 1 for low)
6) Height (meters).
7) Systolic blood pressure(mmHg).
8) Body mass index(kilogram/meter^2).
9) Diastolic blood pressure (mmHg).
10) Blood sugar level (mg/dl).
11) Class (1 fordiabetic and 0 for non-diabetic).

In DCA dataset, the minimum and maximum value of all these attributes are as shown in Table 1.

Table 1. Minimum and Maximum value of the attributes of DCA dataset

S. No.	Types of attributes	Minimum value	Maximum value
1	Age defining in years.	18	75
2	Sex defining either Male or Female.	1	2
3	Family history (History of diabetes in family).	0	1
4	Weight measured in kg.	36	90
5	The decision of blood pressure measured in mmHg.	1	3
6	Height measured in meters.	1.54	1.75
7	Systolic blood pressure measured in mmHg.	90	180
8	Body mass index measured in kg/meter^2.	14.1	33.5
9	Diastolic blood pressure measured in mmHg.	60	110
10	Blood sugar level measured in mg/dl.	61	425

2.1 Statistical Relationship

In the DCA dataset, the attributes are statistically correlated to diabetic or non-diabetic class by using the statistical r correlation formula as follows:

$$r = \frac{\frac{1}{N}\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{N}\sum (x_i - \bar{x})^2}\sqrt{\frac{1}{N}\sum (y_i - \bar{y})^2}}$$

where $\bar{x}$ and $\bar{y}$ are $\sum_{i=1}^{N} x_i \Big/ N$ and $\sum_{i=1}^{N} y_i \Big/ N$.

Here r lies between -1 to 1. If $0 < r < 1$ then positive correlation occurs and if $-1 < r < 0$ then negative correlation occurs, otherwise there is no correlation between two variables. Table 2 shows the statistical correlation of all attributes of DCA dataset.

Table 2. Statistical correlation of attributes for DCA dataset

S. No.	Types of attributes	Correlation parameter (r)	Status
1	Age defining in years.	0.3355	Positive correlation
2	Sex defining either Male or Female.	0.162	Positive correlation
3	Family history (History of diabetes in family).	-0.0219	Negative correlation
4	Weight measured in kg.	0.2695	Positive correlation
5	The decision of blood pressure measured in mmHg.	0.2148	Positive correlation
6	Height measured in meters.	0.1619	Positive correlation
7	Systolic blood pressure measured in mmHg.	0.1342	Positive correlation
8	Body mass index measured in kg/meter^2.	0.2563	Positive correlation
9	Diastolic blood pressure measured in mmHg.	0.1372	Positive correlation
10	Blood sugar level measured in mg/dl.	0.6487	Positive correlation

Figure 2 shows that nine attributes out of ten attributes are positively correlated and one attribute is negatively correlated. It means that along with the Blood sugar level other attributes are also important to build the predictive models for the classification of the diabetes disease.

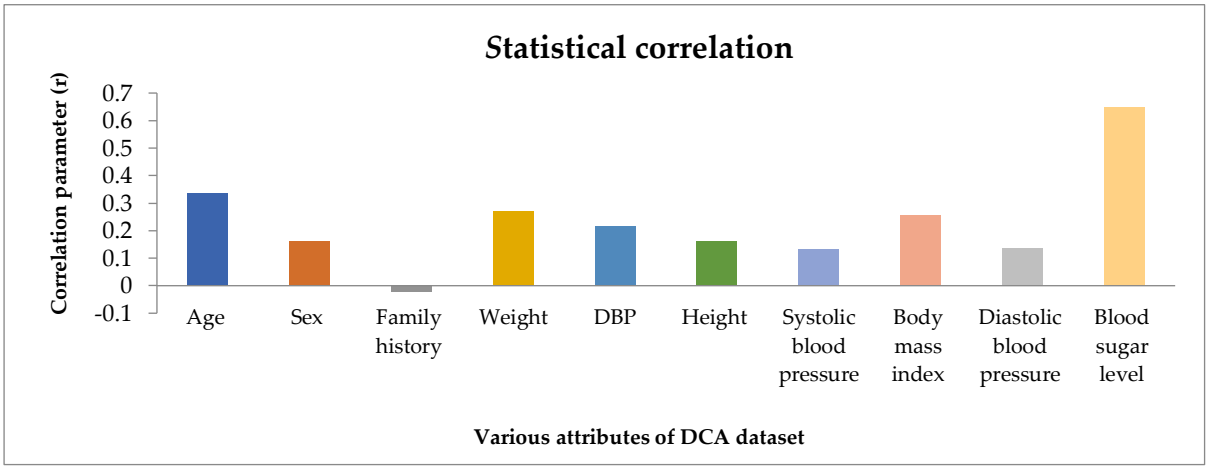

Figure 2: Graphical representation of correlation parameter(r)for the attributes of DCA dataset

3. Computational algorithms

In this paper, some well-known machine learning algorithms like perceptron, Gaussian process, linear discriminant analysis, quadratic discriminant analysis, statistical gradient descent, ridge regression classifier, support vector machines, k-nearest neighbors, decision tree, naïve Bayes, logistic regression, random forest and ELM for multiquadric, RBF, sigmoid activation function discuss with their pros and cons. Further employ on PIMA Indian diabetic dataset[76] and DCA dataset to find the class labels of the patient whether a particular patient is diabetic or non-diabetic.

3.1. Perceptron

It is the famous learning algorithm which was developed for a binary classifier [86]. It is also a threshold function where input feature value is mapped to the output value, either diabetic or non-diabetic. It is defined as

Definition of Perceptron

INPUT: $\{x_i \mid i = 1,2,3,..,N\}$ = input training data

v = test data.

Procedure:

$$H(x) = \begin{cases} 0 & otherwise \\ 1 & if \;\; w.x + b > 0 \end{cases}, \qquad (1)$$

where w is the weight vector and b is the bias vector; $H(x)$ is the either 0 and 1 which can easily classify the diabetic and non-diabetic patients.

OUTPUT: Class label v .

For more details, one can follow [59, 64, 66].

3.2. Gaussian process (GP)

The Gaussian process is one of the learning classifiers which is used to calculate the similarity between different points as a stochastic process and help in binary classification using Gaussian distribution [27]. It is defined as: $(X_{t_1},..., X_{t_b})$ is Gaussian if and only if

Definition of GP

INPUT: $\{X_{t_i} \mid i = 1,2,3,..,N\}$ = input training data

v = test data.

Procedure:

$$E\left(\exp\left(a\sum_{b=1}^{n} q_b X_{t_b}\right)\right) = \exp\left(-\frac{1}{2}\sum_{b,c}\rho_{bc} q_b q_c + a\sum_{b=1}^{n}\sigma_b q_b\right), \qquad (2)$$

where a denotes the imaginary unit; ρ_{bc} is the covariances of the variables in the process; σ_b is the mean of the variables in the process; q_b denotes all equalities.

OUTPUT: Class label v .

For a more detailed description, one can read [68, 73, 83-84].

3.3. Linear discriminant analysis (LDA)

It is one of the popular classifier named linear discriminant classifier [23] in which some prerequisite have assumed such that the conditional probability density functions(PDFs) are normally distributed over mean and covariance. By following this assumption, the Bayes optimal solution is to forecast the non-diabetic patients for binary classification if the log of the likelihood ratio is higher in compared to one threshold (Q) . It is defined as for $H(x) > Q$:

Definition of LDA

INPUT: $\{X = x_i \mid i = 1,2,3,..,N\}$ = input training data

v = test data.

Procedure:

$$H(X) = (X-\rho)^t \sum_{\rho}^{-1}(X-\rho) + \ln|\sum_{\rho}| - (X-\sigma)^t \sum_{\sigma}^{-1}(X-\sigma) - \ln|\sum_{\sigma}|, \qquad (3)$$

where the mean parameters are $(\rho, \sum_{\rho})$; the covariances parameters are $(\sigma, \sum_{\sigma})$.

OUTPUT: Class label v.

For further study, one can follow [11, 27, 105].

3.4. Quadratic discriminant analysis (QDA)

It is another variant of LDA where the prerequisite assumption was not considered. It classifies the different data points by using the quadric surface [23, 100]. They follow the likelihood ratio test, which is defined as taking consideration of $H(x) > Q$ as:

Definition of QDA

INPUT: $\{X = x_i \mid i = 1,2,3,..,N\}$ = input training data

v = test data.

Procedure:

$$H(X) = \frac{(\sqrt{|2\pi\sum_b|})^{-1} \exp(-\frac{1}{2}(X-\sigma_b)^t \sum_b^{-1}(X-\sigma_b))}{(\sqrt{|2\pi\sum_a|})^{-1} \exp(-\frac{1}{2}(X-\sigma_a)^t \sum_a^{-1}(x-\sigma_a))}, \qquad (4)$$

where a, b denotes the diabetic and non-diabetic class datapoints; $H(x)$ is the likelihood ratio.

OUTPUT: Class label v.

For further study, one can follow [34, 66, 100].

3.5. Stochastic gradient descent (SGD)

It is an iterative approach for classification which apply some smoothness function known as stochastic gradient descent (SGD). In SGD [9, 85], the standard gradient is changed by an estimate from a randomly chosen subset of patients datasets.

Algorithm for SGD

INPUT: $\{u_i \mid i = 1,2,3,..,N\}$ = input training data

v = test data.

Procedure:

Step 1: Select an input vector of u and correspondingly its learning rate λ.

Step 2: Repeat step 3 to 5 until an approximation minimum is achieved.

Step 3: Mixed all input examples without following any order.

Step 4: Start loop from 1 to total datapoints do

Step 5: Calculate $u := u - \lambda \nabla G_i(u)$, where $\nabla G_i(u)$ is the true gradient of loss function $G_i(u)$ at i^{th} example.

OUTPUT: Class label v.

For more description, this literature will be helpful [16-17, 113].

3.6. Ridge regression classifier (RC)

211 Ridge regression has been used as a classifier by many researchers such as Saunders et al. [89];
212 Vago et al. [103]; Jiang et al. [49]; Pratt et al. [79]; Rajkumar et al. [81]; Chakravarti et al. [20]. The
213 definition of ridge regression classifier is defined as:

Definition of RC

INPUT: $\{X = x_i \mid i = 1,2,3,..,N\}$ = input training data

v = test data.

Procedure:

$$\chi^{ridge} = \arg\min_{\chi \in R} \| Y - X\eta \|_2^2 + \delta \| \chi \|_2^2, \quad (5)$$

where X is the number of features; χ is the coefficient or beta; $\eta = \sqrt{\chi_0^2 + ... + \chi_n^2}$.

OUTPUT: Class label v .

214

215 *3.7. k-nearest neighbors*

216 k-nearest neighbors [32] is a simple non-parametric algorithm. Here k is the total number of
217 nearest neighbors. New cases are classified depending on the neighborhood of the feature space.
218

Algorithm for k-NN

INPUT: $\{v_i \mid i = 1,2,3,..,N\}$ = labelled training data

v = test data

Procedure:

for $i = 1,2,3,..,N$ $i = 1,2,3,..,N$ determine the range between v and v_i .

if $i \le k$ then comprise v_i in k -nearest neighbors

else if

remove the far of the k -nearest neighbors and comprise v_i in k -nearest neighbors

endif

end for

class label of v =label of the majority of the k -nearest neighbors class.

OUTPUT: Class label v .

219 k -NN has merits such as easy implementation, faster training, and versatility; but it also has some
220 drawbacks such as the curse of dimensionality [32].

221 *3.8. Decision tree*

222 The algorithm for decision tree C4.5 [10] is as follows:

Algorithm for Decision tree

INPUT: $\{x_i \mid i = 1,2,3,..,N\}$ = labelled training data

v = test data

Procedure:

Step 1: Create a root node

(a) Entropy $H(S) = -\sum p(x) \log\, p(x) = -(\frac{positive}{total}) \log(\frac{positive}{total}) - (\frac{negative}{total}) \log(\frac{negative}{total})$

where S is sample space.

The value of entropy will be 0 if all the members belong to the same class; on the other hand, entropy will be 1 when 50% of the members belong to one particular class, and the other 50% belongs to another class.

(b) The following step is to select the attribute A which gives us the information gain $IG(S, A)$ with the highest possibility and will be chosen as the root node.

$$IG(S, A) = H(S) - \sum_{i=0}^{n} P(x) * H(x)$$

Here, x denotes the feasible values for an attribute and $P(x)$ is denoted as the probability of the event x.

Step2:

If all the instances are found to be positive,

return leaf node as 'positive'.

Else if all the instances are negative

return leaf node as 'negative'.

endif

Remove the attribute which yields the highest information gain from the group of attributes.

Repeat the process till the last attribute or the decision tree achieves all the leaf nodes.

OUTPUT: Class label v.

Decision trees have the advantage of easy interpretations; they work with both with numerical and categorical data; they have a moderate computational burden (when compared with other methods such as support vector machines); they are easy to interpret by a human because of the intuitive flow-chart structure of their models [10]. For more study one can follow: [15, 52].

3.9. Support vector machine (SVM)

In Support vector machines[24], each data is intrigued as a point in the space of n dimension; N number of training instances is considered. Each instance is signified by tuple (p_i, q_i) $(i = 1,2,..., N)$ where $p_i = (p_{i1}, p_{i2},..., p_{in})^t$ correlate to the group of an attribute for the ith instance. $q_i \in \{-1,1\}$ denotes the label of class. Thus the margin of a decision of a linear classifier can be written as $w \cdot p + b = 0$. Here, w and b are unknown.

A simple Linear SVMs-train algorithm [24] is given below:

Algorithm for SVM

INPUT: Training samples $\{\overrightarrow{p_1}, \overrightarrow{p_2},..., \overrightarrow{p_n}\}^t$

Procedure:

1. Class label $\{q_1, q_2,..., q_n\}$,
2. Maximize over α_k : $L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{i'=1}^{N} \alpha_i \alpha_{i'} q_i q_{i'} p_i^T p_{i'}$

subject to $0 \le \alpha_i \le \frac{1}{2N}$ and $\sum_{i=1}^{N} \alpha_i q_i = 0$.

3. By using the values of α_k, one can find the values of w and b as

$$w = \sum_{i=1}^{N} \alpha_i q_i p_i \ , \alpha_i [q_i (w \cdot p_i + b) - 1] = 0 \text{ and } \alpha_i \ge 0 .$$

OUTPUT: Optimal hyperplane which classifies the data point.

SVM is a very popular classification algorithm to its ability of mapping complex, non linearly separable data into a higher dimensional space where they become linearly separable. However, SVM requires higher computational resources when compared to other algorithms such as logistic regression or naive Bayes, as well as a more complex parameter tuning [36,40].

3.10. Naïve Bayes

Naive Bayes classifier is a probabilistic model which follows the Bayes theorem with independent assumptions.

$$P(B \mid A) = \frac{P(A \mid B)}{P(A)}, \qquad (6)$$

where A represents already prior event and B is a present event or dependent variable.

After calculating the $P(B \mid A)$, Naive Bayes algorithm [75] considers and counts the total number of cases such as M where both events occur together and such as N where the prior event occurs only. In the second and final step M is divided by N.

Naive Bayes algorithms have several merits like easy implementation, a fast estimate of test data, and performed well in ambiguously explicit input variables in contrast to numerical variables, in both binary and multiclass predictions [75, 29].

3.11. Logistic regression

Logistic regression [43, 90,28] also called the Logit model.

Definition of LR

INPUT: $\{X = x_i \mid i = 1,2,3,..,N\}$ = input training data

v = test data.

Procedure:

$$\log it(P_1) = \ln\left(\frac{P_1}{1-P_1}\right) = \omega_0 + \omega_1 X_1 + ... + \omega_n X_n = \omega_0 + \omega_1 X_1 + ... + \omega_n X_n$$

$$= \omega_0 + \sum_{k=1}^{n} \omega_k X_k \qquad (7)$$

where, ω_0 represents the intercept and $\omega_1, \omega_2, ..., \omega_n$ are the coefficients related to the variable $X_1, X_2, ..., X_n$.

A bipartition variable has two values like yes (1)/no (0), diabetic/non-diabetic, alive/dead that signifies the presence or absence of some event. $X_1, X_2, ..., X_n$ are the independent variables and may be continuous, bipartition, discrete or combination [26, 89].

$$P_1(X) = \frac{1}{1 + e^{-\log it(P_1(X))}} = \frac{1}{1 + e^{-\left(\omega_0 + \sum_{k=1}^{n} \omega_k X_k\right)}}, \qquad (8)$$

OUTPUT: Class label v .

Maximum likelihood estimation model (MLE) is generally considered to estimate the coefficients which begin with some random input estimation of coefficients and finds the variation in magnitude and direction of the efficient. After finding the first function, all remaining coefficients have to be tested and update the new estimation function. As the MLE model is iterative, this process continues until convergence is reached [41, 43, 90,102].

3.12. Random forest (RF)

Random forest is also one the viable machine learning algorithm which is used for both classification and regression [14, 42, 65]. The algorithm of random forest is written as:

Algorithm for RF

INPUT: $\{v_i \mid i = 1,2,3,..,N\}$ = labelled training data

v = test data

Procedure:

Step 1: find the bootstrap samples from diabetic patients data.

Step 2: Develop an unpruned classification tree with a random sample of the predictors and select the best split from among those variables.

Step 3: Forcast the new data by considering the majority votes of classification.

OUTPUT: Class label v .

For more details, one can follow [18, 46, 48, 91].

3.13. Extreme learning machine (ELM)

Extreme learning machine (ELM) is one of prominent machine learning algorithm which is also used for classification and many predictions based problem with one or many hidden layer nodes as well as no tuning is required for parameters [44]. The output function $f(.)$ of ELM for η hidden nodes is presented as

Definition of ELM

INPUT: $\{X = x_i \mid i = 1,2,3,..,\eta\}$ = input training data

v = *test data.*

Procedure:

$$y_k = f(x_k) = \sum_{g=1}^{\eta} \omega_g U(a_g, b_g, x_k) \quad \text{for } k = 1,..., l_o, \tag{9}$$

Here, $w = (w_1, w_2, ..., w_\eta)^t$ is the output weight vector. One can write the Eq.(9) as $Hw = y$, where

$$H = \begin{bmatrix} U(a_1,b_1,x_1) & ... & U(a_\eta,b_\eta,x_1) \\ . & ... & . \\ U(a_1,b_1,x_{l_o}) & ... & U(a_\eta,b_\eta,x_{l_o}) \end{bmatrix}_{l_o \times \eta} .$$

OUTPUT: Class label v .

In ELM, many activation functions are used such as multiquadric, RBF, Sigmoid for hidden node. The definition of the activation function is as:

For ELM(Multiquadric): $U(a,b,x) = \sqrt{(\| x-a \|^2 + b^2)}.$

For ELM(RBF): $U(a,b,x) = \exp(-b \| x-a \|^2).$

For ELM(Sigmoid): $U(a,b,x) = 1/(1+\exp(-(ax+b)).$

4. The framework of Diabetic diagnosis process

In this section, a systematic procedure of diabetic diagnosis process [6, 104] is discussed as a framework of the classification model for diagnosis of diabetes in Figure 3. There are several steps which should be followed under this framework in such a way:

Step1: The process of the study starts by collecting the dataset named DCA.

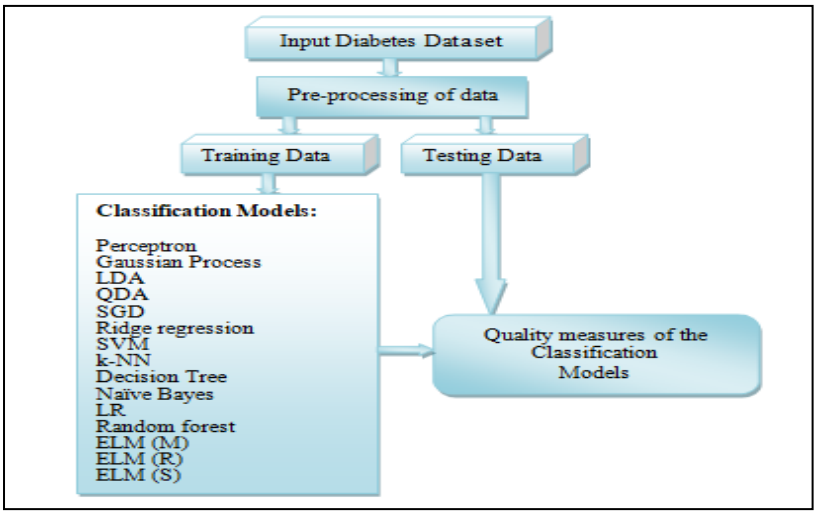

Figure3: Framework of classification model for diagnosis of diabetes

Step2: After collecting the initial dataset, pre-processing is done, such as calculating the body mass index (BMI), which is computed as weight (kg) dividing by height (meter) squared. Further, some attributes like sex, family history, the decision of blood pressure are encoded.

Step3: The whole PIMA Indian diabetic or DCA data set is divided into train data and test data.

Step4: After selecting the size of train and test data, train data is passed as the input to any one of the machine learning algorithms such as perceptron, Gaussian process, linear discriminant analysis, quadratic discriminant analysis, statistical gradient descent, ridge regression classifier, support vector machines, k-nearest neighbors, decision tree, naïve Bayes, logistic regression, random forest and ELM for multiquadric, RBF, sigmoid activation function.

Step5: After the completion of training, we will have a resultant classifier and various quality measures such as accuracy, sensitivity (recall), specificity (true negative rate), precision, NPV, FP rate, RME, F1-measure, receiver operating characteristic curve, precision-recall curve, G-mean and Matthews correlation coefficient (MCC). To find the class label of any patient, we will satisfy the test sample to

the resultant classifier and get the output either diabetic or non-diabetic. One can also validate the data through various quality measures.

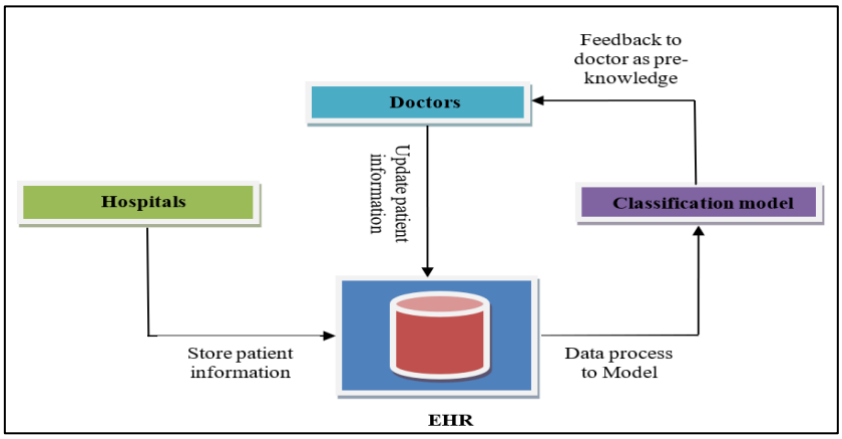

Figure 4: Framework of the diabetic diagnosis process

4.1. The flow of data from electronic health record (EHR) to model:

In this subsection, we have described the framework of diabetic diagnosis process in which patient information will store in the electronic health record (EHR) from the different hospitals. Further, the stored information is applied to the classification model to get the decision classifiers. This decision classifier will respond to doctors regarding the patient's output class, either diabetic or non-diabetic. This output will be given to the doctor as the input in the form of pre-knowledge, which will help the doctors for deciding on the diagnosis of the disease. In the next step, the doctor will take action according to output and also update the information in the EHR for prospects. The framework of a diabetic diagnosis process is shown in Figure 4.

5. Results and discussion

In this paper, 15 classification techniques have been applied for the categorisation of diabetic class or non-diabetic class on the DCA dataset as well as PIMA Indian diabetic dataset [76]. To evaluate the performance of the models, we have computed various quality measures like accuracy, sensitivity (recall), specificity (true negative rate), precision, NPV, FP rate, RME, F_1-measure, receiver operating characteristic curve, precision-recall curve, G-mean, Matthews correlation coefficient (MCC), the area under the curve (AUC_ROC) and average precision score (AP) for both datasets. In DCA dataset, training and testing are split on the ratio of 60:40. We have used ten-fold cross-validation in SVM. The analysis is being investigated by the help of these quality measures where the presence of diabetes are considered to be of positive class, and the absence of diabetes is considered to be of negative class [31]. Here, true positive (TP) depicts the number of instances where the non-existence of diabetes is concluded as the non-existence of diabetes. False-positive (FP) depicts the number of instances where the existence of diabetes is concluded as the non-existence of diabetes. True negative (TN) depicts the number of instances where the existence of diabetes is concluded as the existence of diabetes. False-negative (FN) depicts the number of instances where the non-existence of diabetes is concluded as the existence of diabetes.

Quality measures are described underneath [39, 69]:

Quality measures	Definition
Accuracy:	$(TP+TN)/(TP+TN+FP+FN)$,

Recall (Sensitivity):	$TP/(TP+FN)$,
Specificity:	$TN/(TN+FP)$,
Precision(P):	$TP/(TP+FP)$,
Negative predictive value(NPV):	$TN/(TN+FN)$,
False-positive rate(FPR):	$FP/(FP+TN)$,
Rate of misclassification (RME):	$(FP+FN)/(TP+TN+FP+FN)$
F_1-Measure:	$2*(P*R)/(P+R)$
G-mean:	$sqrt(P*R)$
Matthews's correlation coefficient (MCC):	$\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$
Areas under ROC curve (AUC_ROC):	$\frac{1+R-FPR}{2}$
Average precision score (AP)	$\sum_j (R_j - R_{j-1}) * P_j$

315

316 *5.1. Performance evaluation of diabetes clinical Assam data*

317 The confusion matrix of the DCA dataset for the 15 classification techniques is reported in Table 3
318 and the graphical representation of these values are depicted in Figure 5.

319 **Table 3.**Values of TP, FP, TN and FN for perceptron, GP, LDA, QDA, SGD, RC, SVMs, k-NN, DT, NB, LR, RF,
320 ELM(Multiquadric), ELM(RBF) and ELM(Sigmoid) for the DCA dataset

Models	**TP**	**FP**	**TN**	**FN**
Perceptron	2	0	25	47
Gaussian process	25	6	2	41
LDA	23	11	4	36
QDA	27	16	0	31
SGD	21	7	6	40
Ridge classifier	27	10	0	37
SVM	30	2	39	3
k-NN	27	8	0	39
Decision tree	26	4	1	43
Naïve bayes	26	6	1	41
Logistic regression	27	4	0	43
Random forest	24	3	3	44
ELM (Multiquadric)	35	10	23	6
ELM (RBF)	35	9	23	7
ELM (Sigmoid)	36	7	26	5

321

From Figure 5, one can see that the ELM (Sigmoid) is having the maximum TP value, whereas the support vector machine (SVM) has the maximum TN values among all algorithms. From Table 4, it can be concluded that Logistic regression has shown better performance among all classification techniques with better accuracy, FP_rate, RME, AUC_ROC and AP using the DCA dataset. Logistic regression seems, therefore, an excellent choice to predict the class label of any patient as diabetes or non-diabetes. The ROC curve has been plotted in Figure 6 for the DCA dataset, where one can see that logistic regression has shown better performance among all reported approaches. Here, our collected DCA dataset is imbalanced, so, we have plotted the precision-recall curve (PRC), which is shown in Figure 7. PRC is a curve which maintains the balance between true positive rate and false-positive rate. One can conclude that the Logistic regression has shown better results in comparison to other algorithms on this set.

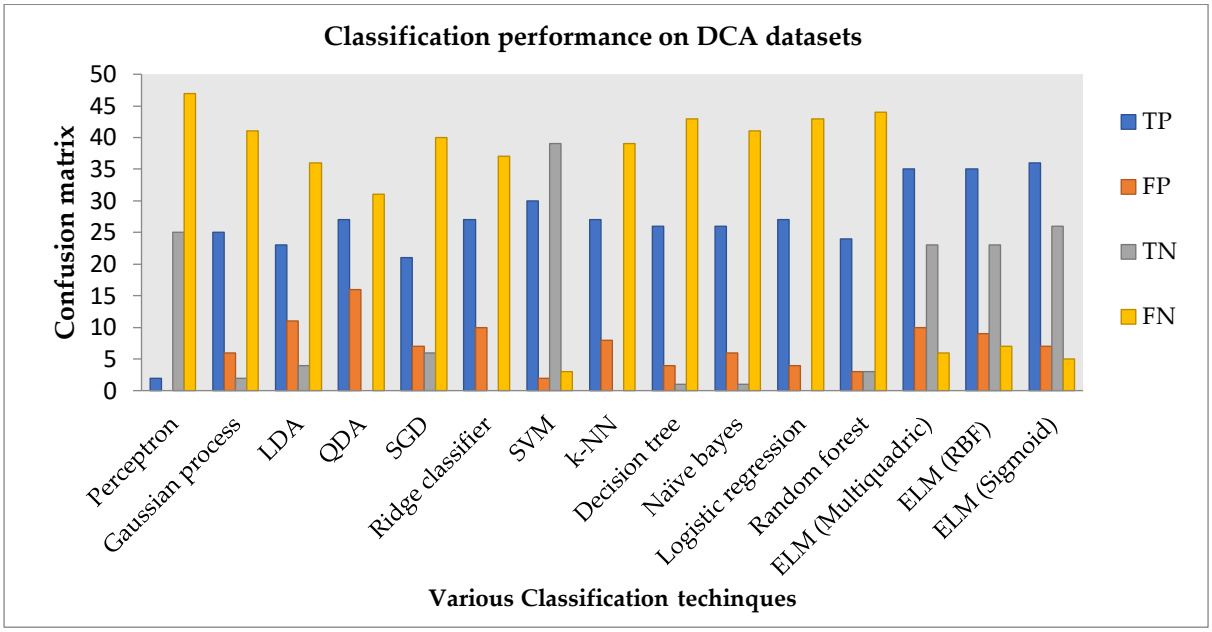

Figure 5. Classification performance of the perceptron, GP, LDA, QDA, SGD, RC, SVMs, k-NN, DT, NB, LR, RF, ELM(Multiquadric), ELM(RBF) and ELM(Sigmoid) in terms of TP, FP, TN and FN on DCA dataset.

5.1.1 McNemar statistical test

For the statistical significance of the results of 15 reported approaches, we have performed the McNemar test [35]. In this test, we obtain p-value which is tabulated in Table 4. From this McNemar test, one can see that the p-value for all the reported approaches is significant except naïve Bayes (NB) at the significance level $(\alpha < 0.1)$. we can say our calculated results over DCA datasets are significant.

5.2. Performance evaluation on PIMA Indian diabetic dataset

In this section, one more dataset named PIMA Indian diabetic [76] is also used with 15 machine learning techniques for the classification of diabetes or non-diabetes patient. The dataset consists of 768 instances and 9 attributes.

Table 5. Values of TP, FP, TN and FN for perceptron, GP, LDA, QDA, SGD, RC, SVMs, k-NN, DT, NB, LR, RF, ELM(Multiquadric), ELM(RBF) and ELM(Sigmoid)for PIMA Indians diabetic dataset.

Models	TP	FP	TN	FN
Perceptron	94	40	57	40

Gaussian process	112	37	39	43
LDA	137	35	14	45
QDA	104	47	47	33
SGD	143	68	8	12
Ridge classifier	138	35	13	45
SVM	135	17	42	37
k-NN	120	33	31	47
Decision tree	120	28	31	52
Naïve bayes	126	39	25	41
Logistic regression	137	34	14	46
Random forest	124	31	27	49
ELM (Multiquadric)	48	18	133	32
ELM (RBF)	45	16	136	34
ELM (Sigmoid)	45	15	136	35

The whole dataset is divided into training and testing set as 70: 30, respectively. The values of TP, FP, TN, FN corresponding to all concerned algorithms on PIMA Indian diabetic dataset are tabulated in Table 5 respectively and depicted the values as a bar graph in Figure 8.

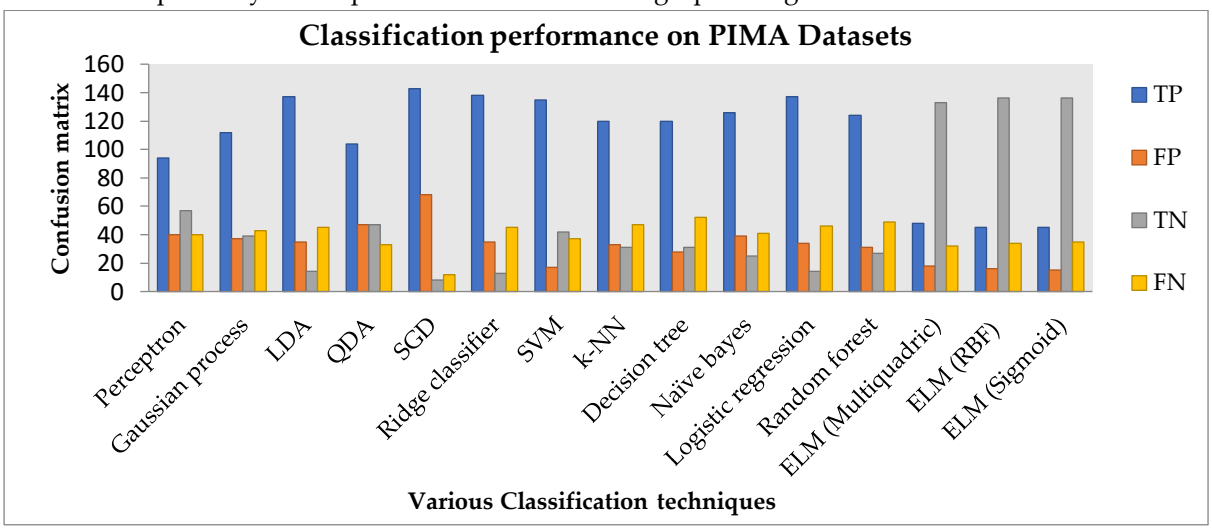

Figure 8. Classification performance of the perceptron, GP, LDA, QDA, SGD, RC, SVMs, k-NN, DT, NB, LR, RF, ELM(Multiquadric), ELM(RBF) and ELM(Sigmoid) in terms of TP, FP, TN and FN for PIMA Indian dataset.

The ROC curve has been drawn for the standard PIMA Indian dataset in Figure 9. From Figure 9, one can say that LDA is performing well in comparison to others. In Figure 10, one can see the precision-recall curve of PIMA Indian dataset. All the quality measures are calculated and tabulated in Table 6 based on the predicted results by using perceptron, Gaussian process, linear discriminant analysis, quadratic discriminant analysis, statistical gradient descent, ridge regression classifier, support vector machines, k-nearest neighbors, decision tree, naïve Bayes, logistic regression, and random forest, Extreme learning machine (multiquadric), Extreme learning machine (radial basis function), and Extreme learning machine (sigmoid). From these results, one can say that similar to the previous case; logistic regression is performed better among other comparative approaches using PIMA Indian

diabetic dataset [76]. The accuracy of logistic regression is 79.22%, which is high in comparison to others. So, logistic regression may be the better choice in comparison to other models.

6. Conclusions

A systematic effort has been made in this paper to predict real-world problems like diabetes disease by identifying machine learning approaches. The DCA dataset has been used for comparative analysis of the machine learning techniques based on accuracy, recall, F_1-measure, specificity, rate of misclassification, precision, negative predicted value, false-positive rate, G-mean, ROC curve, PRC curve and MCC. Among the 15 algorithms, logistic regression yields higher outcomes in terms of accuracy and MCC for the classification of diabetic and non-diabetic samples. Similarly, we have applied these algorithms on PIMA Indian diabetes dataset and again found logistic regression to be the top performer in terms of accuracy and various parameters. One can propose to improve these performances by a hybrid approach and expand the scope of the present study by including more samples (patients) in the data sets.

Acknowledgements

The real-world dataset evaluated in this study is collected from clinical data within duration 2017-2018 of Dhubri District, Assam, India. We are greatly thankful to *Dr Ramendra Nath Choudhury*, Retd. Joint Director of Health Services, Assam, India for providing the real-world dataset of both diabetic and non-diabetic patients.

References

1. Acharya, R., Chua, C.K., Ng, E.Y.K., Yu, W. and Chee, C., 2008. Application of higher order spectra for the identification of diabetes retinopathy stages. *Journal of Medical Systems*, 32(6): 481-488.
2. Acharya, U.R., Lim, C.M., Ng, E.Y.K., Chee, C. and Tamura, T., 2009. Computer-based detection of diabetes retinopathy stages using digital fundus images. *Proceedings of the institution of mechanical engineers, part H: journal of engineering in medicine*, 223(5): 545-553.
3. Akram, M.U., Khalid, S. and Khan, S.A., 2013. Identification and classification of microaneurysms for early detection of diabetic retinopathy. *Pattern Recognition*, 46(1): 107-116.
4. Alade, O.M., Sowunmi, O.Y., Misra, S., Maskeliūnas, R. and Damaševičius, R., 2017. A Neural Network Based Expert System for the Diagnosis of Diabetes Mellitus. *In International Conference on Information Technology Science*. Springer, Cham,14-22.
5. Albahli, S., 2020. Type 2 Machine Learning: An Effective Hybrid Prediction Model for Early Type 2 Diabetes Detection. *Journal of Medical Imaging and Health Informatics*, 10(5), pp.1069-1075.
6. Alfian, G., Syafrudin, M., Ijaz, M., Syaekhoni, M., Fitriyani, N. and Rhee, J., 2018. A personalized healthcare monitoring system for diabetic patients by utilizing BLE-based sensors and real-time data processing. *Sensors*, 18(7), p.2183.
7. Ali, R., Hussain, J., Siddiqi, M., Hussain, M. and Lee, S., 2015. H2RM: A hybrid rough set reasoning model for prediction and management of diabetes mellitus. *Sensors*, 15(7), pp.15921-15951.
8. Alić, B., Gurbeta, L., and Badnjević, A., 2017. Machine learning techniques for classification of diabetes and cardiovascular diseases. *6th Mediterranean Conference on Embedded Computing* (MECO), Bar, 1-4.
9. Amari, S.I., 1993. Backpropagation and stochastic gradient descent method. *Neurocomputing*, 5(4-5), pp.185-196.
10. Argentiero, P., Chin, R., and Beaudet, P., 1982. An automated approach to the design of decision tree classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1: 51-57.
11. Balakrishnama, S. and Ganapathiraju, A., 1998. Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing*, 18, pp.1-8.

12. Ban, H.J., Heo, J.Y., Oh, K.S. and Park, K.J., 2010. Identification of type 2 diabetes-associated combination of SNPs using a support vector machine. *BMC Genetics*, 11(1): 26.
13. Barakat, N., Bradley, A.P. and Barakat, M.N.H., 2010. Intelligible support vector machines for diagnosis of diabetes mellitus. *IEEE transactions on information technology in biomedicine*, 14(4):1114-1120.
14. Barandiaran, I., 1998. The random subspace method for constructing decision forests. IEEE Trans. *Pattern Anal. Mach. Intell*, 20(8), pp.1-22.
15. Bashir, S., Qamar, U., Khan, F.H. and Javed, M.Y., 2014, December. An efficient rule-based classification of Diabetes using ID3, C4. 5, & CART ensembles. *In 2014 12th International Conference on Frontiers of Information Technology* (pp. 226-231). IEEE.
16. Bottou, L., 1998. Online learning and stochastic approximations. *On-line learning in neural networks*, 17(9), p.142.
17. Bottou, L., 2010. Large-scale machine learning with stochastic gradient descent. *In Proceedings of COMPSTAT'2010* (pp. 177-186). Physica-Verlag HD.
18. Butwall, M. and Kumar, S., 2015. A data mining approach for the diagnosis of diabetes mellitus using random forest classifier. *International Journal of Computer Applications*, 120(8).
19. Carrera, E. V., González, A. and Carrera, R., 2017. Automated detection of diabetic retinopathy using SVM. *IEEE XXIV International Conference on Electronics, Electrical Engineering and Computing* (INTERCON), Cusco, 1-4.
20. Chakravorti, T., Nayak, N.R., Bisoi, R., Dash, P.K. and Tripathy, L., 2019. A new robust kernel ridge regression classifier for islanding and power quality disturbances in a multi distributed generation based microgrid. *Renewable Energy Focus*, 28, pp.78-99.
21. Chaki, J., Ganesh, S.T., Cidham, S.K. and Theertan, S.A., 2020. Machine Learning and Artificial Intelligence based Diabetes Mellitus Detection and Self-Management: A Systematic Review. *Journal of King Saud University-Computer and Information Sciences.*
22. Chikh, M.A., Saidi, M. and Settouti, N., 2012. Diagnosis of diabetes diseases using an artificial immune recognition system2 (AIRS2) with a fuzzy k-nearest neighbors. *Journal of medical systems*, 36(5): 2721-2729.
23. Cohen, J., Cohen, P., West, S.G. and Aiken, L.S., 1983. Applied multiple regression. *Correlation Analysis for the Behavioral Sciences*, 2.
24. Cortes, C. and Vapnik, V., 1995. Support-vector networks. *Machine learning* 20(3): 273-297.
25. Dagliati, A., Marini, S., Sacchi, L., Cogni, G., Teliti, M., Tibollo, V., De Cata, P., Chiovato, L. and Bellazzi, R., 2018. Machine learning methods to predict diabetes complications. *Journal of diabetes science and technology*, 12(2): 295-302.
26. Diabetes, 2019. Available at:https://www.statista.com/statistics/271464/percentage-of-diabetics-worldwide/ (accessed April 10, 2020).
27. Dogantekin, E., Dogantekin, A., Avci, D. and Avci, L., 2010. An intelligent diagnosis system for diabetes on linear discriminant analysis and adaptive network based fuzzy inference system: LDA-ANFIS. *Digital Signal Processing*, 20(4), pp.1248-1255.
28. Dreiseitl, S. and Ohno-Machado, L., 2002. Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, 35(5-6): 352-359.
29. Dudley, R.M., 1989. Real Analysis and Probability (Wadsworth & Brooks/Cole: Pacific Grove, CA). CA. *Mathematical Reviews* (MathSciNet): MR91g, 60001.
30. Faust, O., Acharya, R., Ng, E.Y.K., Ng, K.H. and Suri, J.S., 2012. Algorithms for the automated detection of diabetic retinopathy using digital fundus images: a review. *Journal of medical systems*, 36(1): 145-157.
31. Fielding, A.H., and Bell, J.F., 1997. Review of methods for the assessment of prediction errors in conservation presence/absence models: *Environmental Conservation* 24(1):38-49.
32. Friedman, J. H., Baskett F., and Shustek, L. J., 1975. An algorithm for finding nearest neighbors. *IEEE Transactions on Computers*, 10: 1000-1006.

33. Gadekallu, T.R., Khare, N., Bhattacharya, S., Singh, S., Reddy Maddikunta, P.K., Ra, I.H. and Alazab, M., 2020. Early detection of diabetic retinopathy using PCA-firefly based deep learning model. *Electronics*, 9(2), p.274.
34. Giancardo, L., Meriaudeau, F., Karnowski, T.P., Li, Y., Garg, S., Tobin Jr, K.W. and Chaum, E., 2012. Exudate-based diabetic macular edema detection in fundus images using publicly available datasets. *Medical image analysis*, 16(1): 216-226.
35. Giraudo, A., Grassi, S., Savorani, F., Gavoci, G., Casiraghi, E. and Geobaldo, F., 2019. Determination of the geographical origin of green coffee beans using NIR spectroscopy and multivariate data analysis. *Food control*, 99, pp.137-145.
36. Giveki, D., Salimi, H., Bahmanyar, G. and Khademian, Y., 2012. Automatic detection of diabetes diagnosis using feature weighted support vector machines based on mutual information and modified cuckoo search. *arXiv preprint arXiv*:1201.2173.
37. Gómez-Peralta, F., Abreu, C., Cos, X. and Gómez-Huelgas, R., 2020. When does diabetes start? Early detection and intervention in type 2 diabetes mellitus. *Revista Clínica Española* (English Edition).
38. Gregori, D., Petrinco, M., Bo, S., Rosato, R., Pagano, E., Berchialla, P. and Merletti, F., 2011. Using data mining techniques in monitoring diabetes care. The simpler the better?. *Journal of medical systems*, 35(2), pp.277-281.
39. Gupta, U. and Meher, P., 2020. Statistical Analysis of Target Tracking Algorithms in Thermal Imagery. *In Cognitive Informatics and Soft Computing* (pp. 635-646). Springer, Singapore.
40. Gupta, U. and Gupta, D., 2019. Lagrangian Twin-Bounded Support Vector Machine Based on L2-Norm. *In Recent Developments in Machine Learning and Data Analytics* (pp. 431-444). Springer, Singapore.
41. Hajmeer, M. and Basheer, I., 2003. Comparison of logistic regression and neural network-based classifiers for bacterial growth. *Food Microbiology*, 20(1): 43-55.
42. Ho, T.K., 1995, August. Random decision forests. *In Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278-282). IEEE.
43. Hosmer Jr, D.W., Lemeshow, S. and Sturdivant, R.X., 2013. *Applied logistic regression*. John Wiley & Sons, 398.
44. Huang, Guang-Bin, Qin-Yu Zhu, and Chee-Kheong Siew, 2006. "Extreme learning machine: theory and applications." *Neurocomputing* 70.1-3: 489-501
45. Huang, Y. and Nashrullah, M., 2016. SVM-based Decision Tree for medical knowledge representation. *International Conference on Fuzzy Theory and Its Applications* (iFuzzy), Taichung, 1-6.
46. IDF, 2019. Available at:https://www.idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html - IDF Diabetes Atlas 9th Edition 2019 (accessed April 10, 2020).
47. Ijaz, M., Alfian, G., Syafrudin, M. and Rhee, J., 2018. Hybrid Prediction Model for Type 2 Diabetes and Hypertension Using DBSCAN-Based Outlier Detection, Synthetic Minority Over Sampling Technique (SMOTE), and Random Forest. *Applied Sciences*, 8(8), p.1325.
48. Jaiswal, A.K., Tiwari, P., Kumar, S., Gupta, D., Khanna, A. and Rodrigues, J.J., 2019. Identifying Pneumonia in Chest X-Rays: A Deep Learning Approach. *Measurement.*
49. Jiang, Y., Zhou, S., Jiang, Y., Gong, J., Xiong, G. and Chen, H., 2011, June. Traffic sign recognition using ridge regression and Otsu method. *In 2011 IEEE Intelligent Vehicles Symposium* (IV) (pp. 613-618). IEEE.
50. Joshi, R. and Alehegn, M., 2017. Analysis and prediction of diabetes diseases using a machine learning algorithm: Ensemble approach. *International Research Journal of Engineering and Technology* (IRJET) 4(10): 426-435.
51. Kandhasamy, J.P. and Balamurali, S., 2015. Performance analysis of classifier models to predict diabetes mellitus. *Procedia Computer Science*, 47, pp.45-51.
52. Karegowda, A.G., Punya, V., Jayaram, M.A. and Manjunath, A.S., 2012. Rule based classification for diabetic patients using cascaded k-means and decision tree C4. 5. *International Journal of Computer Applications*, 45(12), pp.45-50.
53. Karun, S., Raj, A. and Attigeri, G., 2019. Comparative Analysis of Prediction Algorithms for Diabetes. *In Advances in Computer Communication and Computational Sciences* (pp. 177-187). Springer, Singapore.

54. Karuranga, S., Fernandes, J. D. R., Huang, Y., and Malanda, B. 2018. *IDF DIABETES ATLAS: https://diabetesasia.org/content/diabetes_guidelines* (accessed September 5, 2018).
55. Kaur, H. and Kumari, V., 2020. Predictive modelling and analytics for diabetes using a machine learning approach. *Applied computing and informatics*.
56. Khalil R. M. and Al-Jumaily A. 2017. Machine learning based prediction of depression among type 2 diabetic patients. *12th International Conference on Intelligent Systems and Knowledge Engineering* (ISKE), Nanjing, 1-5.
57. Kim, K.S., Choi, H.H., Moon, C.S. and Mun, C.W., 2011. Comparison of k-nearest neighbor, quadratic discriminant and linear discriminant analysis in classification of electromyogram signals based on the wrist-motion directions. *Current applied physics*, 11(3), pp.740-745.
58. Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A. and Stiglic, G., 2020. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Scientific reports*, 10(1), pp.1-12.
59. Kumar Dewangan, A. and Agrawal, P., 2015. Classification of diabetes mellitus using machine learning techniques. *International Journal of Engineering and Applied Sciences*, 2(5).
60. Kumari, V.A. and Chitra, R., 2013. Classification of diabetes disease using support vector machine. *International Journal of Engineering Research and Applications*, 3(2):1797-1801.
61. Lee, B.J., Ku, B., Nam, J., Pham, D.D., and Kim, J.Y., 2014. Prediction of fasting plasma glucose status using anthropometric measures for diagnosing type 2 diabetes. *IEEE Journal of biomedical and health informatics* 18(2): 555-561.
62. Leon, K., Primoz, K., Leona, C., Sheikh, A. and Gregor, S., 2020. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Scientific Reports (*Nature Publisher Group*), 10(1).
63. Levi, O.U., Webb, F. and Simmons, D., 2020. Diabetes detection and communication among patients admitted through the Emergency Department of a Public Hospital. *International Journal of Environmental Research and Public Health*, 17(3), p.980.
64. Li, C.P., Zhi, X.Y., Jun, M.A., Zhuang, C.U.I., Zhu, Z.L., Zhang, C. and HU, L.P., 2012. Performance comparison between Logistic regression, decision trees, and multilayer perceptron in predicting peripheral neuropathy in type 2 diabetes mellitus. *Chinese medical journal*, 125(5), pp.851-857.
65. Liaw, A. and Wiener, M., 2002. Classification and regression by randomForest. *R news*, 2(3), pp.18-22.
66. Liou, D.R., Liou, J.W. and Liou, C.Y., 2013. Learning Behaviors of Perceptron. *iConcept Press*.
67. Mani, S., Chen, Y., Elasy, T., Clayton, W. and Denny, J., 2012. Type 2 diabetes risk forecasting from EMR data using machine learning. *In American Medical Informatics Association (AMIA) annual symposium proceedings*. 606.
68. Maniruzzaman, M., Kumar, N., Abedin, M.M., Islam, M.S., Suri, H.S., El-Baz, A.S. and Suri, J.S., 2017. Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. *Computer methods and programs in biomedicine*, 152, pp.23-34.
69. Maniruzzaman, M., Rahman, M.J., Ahammed, B. and Abedin, M.M., 2020. Classification and prediction of diabetes disease using machine learning paradigm. *Health Information Science and Systems*, 8(1), p.7.
70. Marini, S., Trifoglio, E., Barbarini, N., Sambo, F., Di Camillo, B., Malovini, A., Manfrini, M., Cobelli, C. and Bellazzi, R., 2015. A Dynamic Bayesian Network model for long-term simulation of clinical complications in type 1 diabetes. *Journal of biomedical informatics*, 57: 369-376.
71. Marini, S., Dagliati, A., Sacchi, L., and Bellazzi, R., 2016. "Learning T2D evolving complexity from EMR and administrative data using Continuous time Bayesian networks." *In 9th International Conference on Health Informatics, HEALTHINF 2016-Part of 9th International Joint Conference on Biomedical Engineering Systems and Technologies*, BIOSTEC. SciTePress.
72. Mohan V., Sandeep S., Deepa R., and Shah B., Varghese C, 2012. Epidemiology of type 2 diabetes: Indian scenario. *Indian journal of medical research* 136(4): 705-718.
73. Murray I., 2008. Introduction to Gaussian Processes. University of Toronto. *Available on: https://www.cs.toronto.edu/~hinton/csc2515/notes/gp_slides_fall08.pdf*.

74. Naz, H. and Ahuja, S., 2020. Deep learning approach for diabetes prediction using PIMA Indian dataset. *Journal of Diabetes & Metabolic Disorders*, 19(1), pp.391-403.
75. Parthiban, G., Rajesh, A. and Srivatsa, S.K., 2011. Diagnosis of heart disease for diabetic patients using naive Bayes method. *International Journal of Computer Applications*, 24(3): 7-11.
76. PIMA, 2019. University of California, Irvine Learning Repository, *Available on https://archive.ics.uci.edu/ml/machine-learning-databases/pima-indians-diabetes/.*
77. Polat, K. and Güneş, S., 2007. An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to the diagnosis of diabetes disease. *Digital Signal Processing*, 17(4):702-710.
78. Polat, K., Güneş, S. and Arslan, A., 2008. A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine. *Expert systems with applications*, 34(1):482-487.
79. Pratt, H., Coenen, F., Broadbent, D.M., Harding, S.P. and Zheng, Y., 2016. Convolutional neural networks for diabetic retinopathy. *Procedia Computer Science*, 90, pp.200-205.
80. Priya, R. and Aruna, P., 2013. Diagnosis of diabetic retinopathy using machine learning techniques. *ICTACT Journal on Soft Computing*, 3(4): 563-575.
81. Rajkumar, M., Charulatha, P., Bindu, P.H. and Kiruthika, A.V., 2019. Diagnosis of Diabetic Retinopathy using Machine Learning algorithms.
82. Rakhonde, A.N., Kshirsagar, P.R. and Marve, S.M., 2020. Diabetes Retinopathy Disease Detection using Convolution Neural Network. *Test Engineering and Management*, pp.4431-4434.
83. Rasmussen, C.E., 2003, February. Gaussian processes in machine learning. *In Summer School on Machine Learning* (pp. 63-71). Springer, Berlin, Heidelberg.
84. Rasmussen, C.E., 2006. *CKI Williams Gaussian processes for machine learning*.
85. Robbins, H. and Monro, S., 1951. A stochastic approximation method. *The annals of mathematical statistics*, pp.400-407.
86. Rosenblatt, F., 1957. The perceptron, a perceiving and recognizing automaton Project Para. *Cornell Aeronautical Laboratory*.
87. Roychowdhury, S., Koozekanani, D.D. and Parhi, K.K., 2014. DREAM: diabetic retinopathy analysis using machine learning. *IEEE Journal of biomedical and health informatics*, 18(5): 1717-1728.
88. Samant, P. and Agarwal., R., 2018. Machine learning techniques for medical diagnosis of diabetes using iris images. *Computer Methods and Programs in Biomedicine* 157: 121-128.
89. Saunders, C., Gammerman, A. and Vovk, V., 1998. *Ridge regression learning algorithm in dual variables*.
90. Schumacher, M., Roßner, R. and Vach, W., 1996. Neural networks and logistic regression: Part I. *Computational Statistics & Data Analysis*, 21(6), pp.661-682.
91. Singh, A.K., 2019. A Comparative Study on Disease Classification using Machine Learning Algorithms. Available at SSRN 3350251.
92. Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., and Johannes, R.S., 1988. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *In Proceedings of the Symposium on Computer Applications and Medical Care*, IEEE Computer Society Press, 261-265.
93. Sopharak, A., Dailey, M.N., Uyyanonvara, B., Barman, S., Williamson, T., Nwe, K.T. and Moe, Y.A., 2010. Machine learning approach to automatic exudate detection in retinal images from diabetic patients. *Journal of Modern Optics*, 57(2):124-135.
94. Sumangali, K., Geetika B. S. R., and Ambarkar H., 2016. A classifier based approach for early detection of diabetes mellitus. *International Conference on Control, Instrumentation, Communication and Computational Technologies* (ICCICCT), Kumaracoil, 389-392.
95. Thyde, D.N., Mohebbi, A., Bengtsson, H., Jensen, M.L. and Mørup, M., 2020. Machine Learning-Based Adherence Detection of Type 2 Diabetes Patients on Once-Daily Basal Insulin Injections. *Journal of Diabetes Science and Technology*, p.1932296820912411.

96. Tiwari, P. and Melucci, M., 2018, October. Towards a quantum-inspired framework for binary classification. *In Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (pp. 1815-1818). ACM.
97. Tiwari, P. and Melucci, M., 2019. Towards a Quantum-Inspired Binary Classifier. *IEEE Access*, 7, pp.42354-42372.
98. Teliti, M., Cogni, G., Sacchi, L., Dagliati, A., Marini, S., Tibollo, V., De Cata, P., Bellazzi, R. and Chiovato, L., 2018. Risk factors for the development of microvascular complications of type 2 diabetes in a single-centre cohort of patients. *Diabetes and Vascular Disease Research*, 1479164118780808.
99. Temurtas, H., Yumusak, N. and Temurtas, F., 2009. A comparative study on diabetes disease diagnosis using neural networks. *Expert Systems with Applications*, 36(4): 8610-8615.
100. Tharwat, A., 2016. Linear vs. quadratic discriminant analysis classifier: a tutorial. *International Journal of Applied Pattern Recognition*, 3(2), pp.145-180.
101. Usher, D., Dumskyj, M., Himaga, M., Williamson, T.H., Nussey, S. and Boyce, J., 2004. Automated detection of diabetic retinopathy in digital retinal images: a tool for diabetic retinopathy screening. *Diabetic Medicine*, 21(1):84-90.
102. Vach, W., Roßner, R. and Schumacher, M., 1996. *Neural networks and logistic regression*.
103. Vago, E. and Kemeny, S., 2006. Logistic ridge regression for clinical data analysis (a case study). *Applied ecology and environmental research*, 4(2), pp.171-179.
104. Vashist, S., 2013. Continuous glucose monitoring systems: a review. *Diagnostics*, 3(4), pp.385-412.
105. Venables ,W. N.,Ripley, B.D., 2002. *Modern applied statistics with S*.Springer.
106. Woldaregay, A.Z., Launonen, I.K., Albers, D., Igual, J., Årsand, E. and Hartvigsen, G., 2020. A Novel Approach for Continuous Health Status Monitoring and Automatic Detection of Infection Incidences in People With Type 1 Diabetes Using Machine Learning Algorithms (Part 2): A Personalized Digital Infectious Disease Detection Mechanism. *Journal of Medical Internet Research*, 22(8), p.e18912.
107. W.H.O., 2013 "Diagnostic criteria and classification of hyperglycaemia first detected in pregnancy – WHO Publications", *WHO/NMH/MND*/13.2.
108. WHO., 2019. Available at: https://www.who.int/health-topics/diabetes#tab=tab_1/(accessed April 10, 2020).
109. Wu, H., Yang S., Huang Z., He J. and Wang X., 2018. Type 2 diabetes mellitus prediction model based on data mining. *Informatics in Medicine Unlocked* 10: 100-107.
110. Xu, K., Feng, D. and Mi, H., 2017. Deep convolutional neural network-based early automated detection of diabetic retinopathy using fundus image. *Molecules*, 22(12), p.2054.
111. Yadav, B., Sharma, S., and Kalra, A., 2018. Supervised Learning Technique for Prediction of Diseases. *Intelligent Communication, Control and Devices, Advances in Intelligent Systems and Computing*, 624: 357-369.
112. Yu, W., Liu, T., Valdez, R., Gwinn, M. and Khoury, M.J., 2010. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC medical informatics and decision making*, 10(1): 16.
113. Zhang, T., 2004. Solving large scale linear prediction problems using stochastic gradient descent algorithms, ICML, 2004. *Complexity of Tetris-Packing Problem Gilbert Young* (California State Polytechnic University).

614 **Table 4.**Classification performance measure indices of the perceptron, GP, LDA, QDA, SGD, RC, SVMs, k-NN, DT, NB, LR, RF, ELM(Multiquadric),ELM(RBF) and ELM(Sigmoid) for DCA dataset

Models	**Accuracy**	**Recall**	**TN rate**	**Precision**	**NPV**	**FP_rate**	**RME**	**F1-Measure**	**Gmean**	**MCC**	**AUC_ROC**	**AP**	**p- value**
Perceptron	0.6622	0.0408	1	1	0.3472	0	0.6351	0.0784	0.202	0.119	0.92	0.96	0.710347
Gaussian process	0.8919	0.3788	0.25	0.8065	0.0465	0.75	0.6351	0.5155	0.5527	0.2336	0.54	0.66	0.617075
LDA	0.7973	0.3898	0.2667	0.6765	0.1	0.7333	0.6351	0.4946	0.5135	0.2771	0.93	0.96	0.546494
QDA	0.7838	0.4655	0	0.6279	0	1	0.6351	0.5346	0.5406	0.446	0.99	0.99	0.77283
SGD	0.8243	0.3443	0.4615	0.75	0.1304	0.5385	0.6351	0.4719	0.5082	0.1524	0.89	0.94	0.267257
Ridge classifier	0.8649	0.4219	1	1	0.2128	0	0.5	0.5934	0.6495	0.2996	0.98	0.99	0.288844
SVM	0.9324	0.9091	0.9512	0.9375	0.9286	0.0488	0.0676	0.9231	0.9232	0.8632	0.91	0.94	0.0036094
k-NN	0.8919	0.4091	0	0.7714	0	1	0.6351	0.5347	0.5618	0.3675	0.96	0.97	0.683091
Decision tree	0.9324	0.3768	0.2	0.8667	0.0227	0.8	0.6351	0.5252	0.5715	0.2163	0.94	0.95	0.371093
Naïve bayes	0.9054	0.3881	0.1429	0.8125	0.0238	0.8571	0.6351	0.5253	0.5615	0.2771	0.95	0.98	1
Logistic regression	0.9459	0.3857	0	0.871	0	1	0.6351	0.5346	0.5796	0.2815	1	1	0
Random forest	0.9189	0.3529	0.5	0.8889	0.0638	0.5	0.6351	0.5052	0.5601	0.0834	0.97	0.99	0.220671
ELM (Multiquadric)	0.7757	0.8537	0.697	0.7778	0.7931	0.303	0.2162	0.814	0.8149	0.5607	0.67	0.66	6.151E-05
ELM (RBF)	0.7892	0.8333	0.7188	0.7955	0.7667	0.2813	0.2162	0.814	0.8142	0.5571	0.75	0.73	0.0036094
ELM (Sigmoid)	0.8378	0.878	0.7879	0.8372	0.8387	0.2121	0.1622	0.8571	0.8574	0.6709	0.89	0.88	0.504985

615

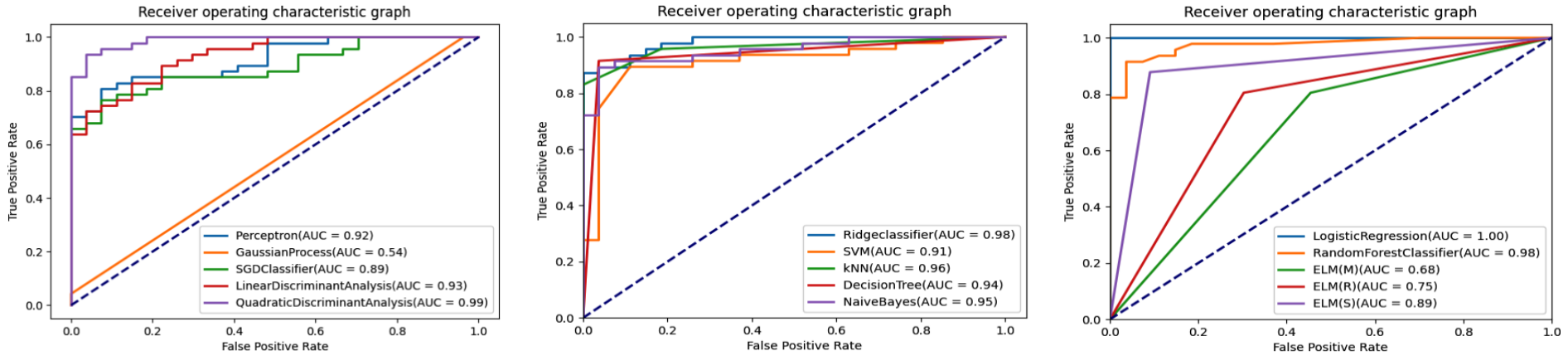

Figure6: ROC curve of the perceptron, GP, LDA, QDA, SGD, RC, SVMs, k-NN, DT, NB, LR, RF, ELM(Multiquadric), ELM(RBF) and ELM(Sigmoid) classifiers for DCA.

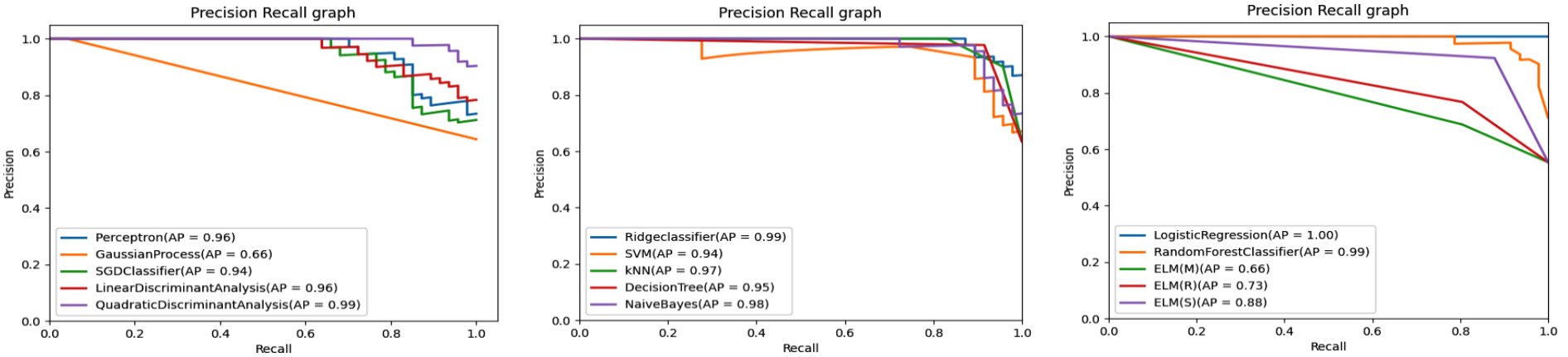

Figure7: The precision-recall curve for DCA dataset using classification result of the perceptron, GP, LDA, QDA, SGD, RC, SVMs, k-NN, DT, NB, LR, RF, ELM(Multiquadric), ELM(RBF) and ELM(Sigmoid) in terms of precision and recall

Table 6. Classification performance measure indices of the perceptron, GP, LDA, QDA, SGD, RC, SVMs, k-NN, DT, NB, LR, RF, ELM(Multiquadric), ELM(RBF) and ELM(Sigmoid) for PIMA Indian diabetic dataset.

Measures	**Accuracy**	**Recall**	**TN rate**	**Precision**	**NPV**	**FP_rate**	**RME**	**F_1-Measure**	**Gmean**	**MCC**	**AUC_ROC**	**AP**
Perceptron	0.5801	0.7015	0.5876	0.7015	0.5876	0.4124	0.3463	0.7015	0.7015	0.2891	0.57	0.55
Gaussian process	0.671	0.7226	0.5132	0.7517	0.4756	0.4868	0.3463	0.7369	0.737	0.2315	0.58	0.4
LDA	0.7879	0.7527	0.2857	0.7965	0.2373	0.7143	0.3463	0.774	0.7743	0.0361	0.84	0.73
QDA	0.5931	0.7591	0.5	0.6887	0.5875	0.5	0.3463	0.7222	0.723	0.2675	0.62	0.45
SGD	0.671	0.9226	0.1053	0.6777	0.4	0.8947	0.3463	0.7814	0.7907	0.0465	0.69	0.54
Ridge classifier	0.7922	0.7541	0.2708	0.7977	0.2241	0.7292	0.3463	0.7753	0.7756	0.0233	0.84	0.73
SVM	0.7435	0.7849	0.7119	0.8882	0.5316	0.2881	0.2338	0.8334	0.835	0.4567	0.64	0.47
k-NN	0.7229	0.7186	0.4844	0.7843	0.3974	0.5156	0.3463	0.75	0.7507	0.1921	0.73	0.55
Decision tree	0.7446	0.6977	0.5254	0.8108	0.3735	0.4746	0.3463	0.75	0.7521	0.2028	0.72	0.53
Naïve bayes	0.7229	0.7545	0.3906	0.7636	0.3788	0.6094	0.3463	0.759	0.759	0.1438	0.75	0.61
Logistic regression	0.7922	0.7486	0.2917	0.8012	0.2333	0.7083	0.3463	0.774	0.7745	0.0373	0.84	0.73
Random forest	0.7489	0.7168	0.4655	0.8	0.3553	0.5345	0.3463	0.7561	0.7573	0.1682	0.82	0.67
ELM (Multiquadric)	0.783	0.6	0.8808	0.7273	0.8061	0.1192	0.2165	0.6575	0.6606	0.5064	0.74	0.57
ELM (RBF)	0.7839	0.5696	0.8947	0.7377	0.8	0.1053	0.2165	0.6428	0.6482	0.4997	0.72	0.57
ELM (Sigmoid)	0.7865	0.5625	0.9007	0.75	0.7953	0.0993	0.2165	0.6429	0.6495	0.5026	0.74	0.58

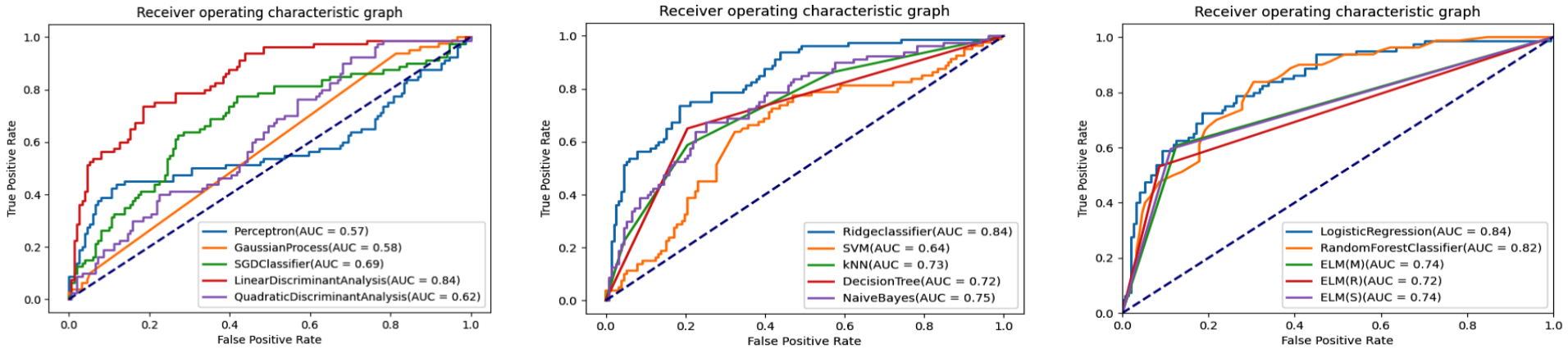

Figure9: ROC curve of the perceptron, GP, LDA, QDA, SGD, RC, SVMs, k-NN, DT, NB, LR, RF, ELM (Multiquadric), ELM (RBF) and ELM (Sigmoid) classifiers for PIMA dataset.

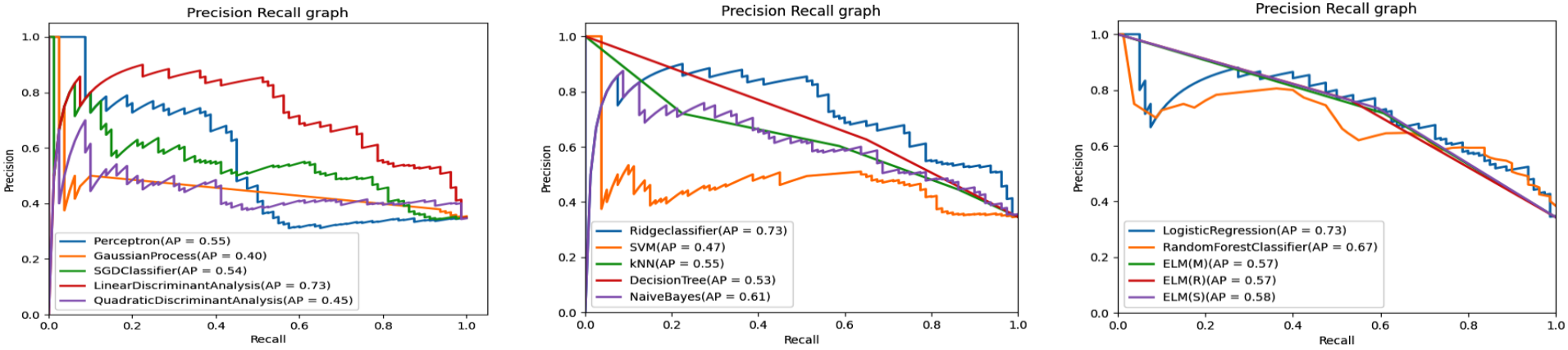

Figure10: Precision-Recall curve (PRC) of the perceptron, GP, LDA, QDA, SGD, RC, SVMs, k-NN, DT, NB, LR, RF, ELM(Multiquadric), ELM(RBF) and ELM(Sigmoid)classifiers for PIMA dataset.