Abstract

Introduction: Diabetes mellitus is a prevalent disease and its late diagnosis leads to dangerous complications and even death. One of the serious complications of this disease is diabetic retinopathy, the leading cause of blindness in the developed countries. Because of slowly progressive nature and lack of symptoms in the early stages of the disease, it is essential to predict the probability of developing diabetic retinopathy promptly to implement the appropriate therapy.

Methods: Our dataset contains 29 extracted features from 310 patients with type 2 diabetic disease, 155 patients of whom suffered from diabetic retinopathy. The patients were selected randomly from Motahari clinic in Shiraz, Iran between 2013 and 2014. First, the genetic algorithm, (GA) as a feature selection process, was implemented to select the most informative features (high-risk factors) for prediction of diabetic retinopathy. Then, three well-known classifiers including k-nearest neighbors (kNN), support vector machine (SVM), and decision tree (DT) were applied to the optimized dataset for classification of the two mentioned groups.

Results: Our finding showed that GA selected 13 factors for better prediction of diabetic retinopathy; these factors were the duration of the disease, history of stroke, family history, cardiac diseases, diabetic neuropathy, LDL, HDL, blood pressure, urine albumin, 2HPPG, HbA1c, FBS, and age. Given the selected risk factors, the classification accuracy was obtained 69.35%, 81.29% and 96.13% by SVM, DT, and kNN, respectively. Our results showed that kNN had the highest accuracy in the prediction of diabetic retinopathy compared to SVM and DT, and the difference between kNN and the other algorithms was statistically significant.

Conclusion: The proposed approach was compared and contrasted with recently reported methods, and it was shown that a considerably enhanced performance was achieved. This research may aid healthcare professionals to determine and individualize the required eye screening interval for a given patient.

Keywords: Diabetic Retinopathy, Feature selection, Genetic algorithm, classification

Introduction

Diabetic retinopathy is the main cause of blindness in the world. Diabetes mellitus affects the small vessels and causes extensive tissue damage in many organs of the body such as the eyes. The retina is the nerve layer behind the eye that converts the light into signals that can be interpreted by the brain as images. Diabetic retinopathy is a complex pathophysiologic process that leads to slowly progressive destruction of the blood vessels and nerve fibers of the innermost layer of the eyes. It intervenes with the proper function of the retina and finally causes blurred vision. If it is not detected and treated in the early stages, the patient will become blind finally. Since diabetic retinopathy may be asymptomatic, its timely diagnosis is the best way to prevent the loss of vision. In many cases, there is no need for treatment, but the patient should be examined regularly. In other cases, treatment is to stop the damage caused by diabetic retinopathy and improve vision, if possible.

Today, with the development of technology, huge medical data are being produced in hospitals and medical centers. Data mining is a way to analyze the data automatically, identify the hidden patterns that cannot be handled manually (1, 2), make the health department move from traditional techniques to evidence-based techniques. Several research studies used data mining techniques, feature ranking, and classification model construction in diabetic
databases (3). For example, Kandhasamy et al. (4) predicted diabetes mellitus using several classifiers such as J48 decision tree (DT), k nearest neighbors (kNN), random forest, and support vector machines (SVM). They measured the performance of these algorithms in both a database with noisy data and a database without noisy data. Their results showed that J48 classifier achieved a higher accuracy of 73.82% than the others when they used noisy data. In another case, the performance of both kNN and random forest (with an accuracy of 100%) was much better than the other algorithms. Saxena et al. (5) tried to diagnose diabetes mellitus by the kNN algorithm which proved to be cost-effective and time-saving for diabetic patients and doctors. Their results showed that as the value of k increased, the accuracy rate also increased. Vijayan et al. (6) used several algorithms to predict diabetes mellitus, where amalgam kNN and adaptive neuro-fuzzy inference system (ANFIS) provided higher classification accuracy than the others (80%). Patel et al. (7) investigated different data mining algorithm for predicting diabetes mellitus on the Pima Indians diabetes database (PIID) with the best accuracy of 70.86±0.34.

Acharya et al. (8) classified normal, mild diabetic retinopathy (DR), moderate DR, severe DR, and PDR stages using the invariant bispectral features of higher-order spectra techniques. They reported an average accuracy of 82% with SVM classifier. In another work (9), they studied automatic identification of normal, mild DR, moderate DR, severe DR, and PDR stages using the blood vessels, exudates, microaneurysms, and hemorrhage features. Their proposed system demonstrated a classification accuracy of 85%. Lee et al. (10) evaluated normal, NPDR and PDR based on the three types of lesions, namely hemorrhages and microaneurysms, hard exudates, and cotton-wool spots. They yielded an accuracy of 82.6% that are comparable with those provided by human experts.

Nayak et al. (11) investigated features like blood vessels, exudates and texture to classify the normal, NPDR and PDR fundus images. They demonstrated an average accuracy of 93%, using a neural network. Larsen et al. (12) evaluated the performance of the lesion detection algorithm over subjects with and without DR. Their automated lesion detection correctly identified 90.1% retinopathy patients and 81.3% subjects without retinopathy. Yun et al. (13) have used morphological image processing techniques to extract the blood vessel area and perimeter from normal, moderate NPDR, severe NPDR, and PDR. Further, these extracted features were subjected to neural network classification and obtained an average accuracy of 84%. Tsao et al. (14) tried to build a prediction model for DR in type 2 diabetes mellitus using SVM, DT, artificial neural networks (ANN), and logistic regressions (LR). Their results demonstrated that SVM achieved better prediction performance in comparison with the other algorithms with an accuracy of 79.5% and area under the receiver operating characteristic curve (AUC) of 0.839. Their results illustrated the use of insulin and duration of diabetes as novel interpretable features in identifying the high-risk populations for diabetic retinopathy.

Voets et al. (15) attempted to develop a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs over two different benchmark datasets including EyePACS and Messidor-2. Their results showed an AUC of 0.951 on the EyePACS and 0.853 on Messidor-2. They suggested that more replication and reproduction studies were needed to validate the deep learning methods, especially for medical image analysis. Bellemo et al. (16) combined two convolutional neural networks for classifying the retinal color fundus images. A total of 4504 retinal fundus images from 3093 eyes of 1574 people from Zambia with diabetes were prospectively recruited. The AUC of their system for referable diabetic retinopathy was 0.973, with a sensitivity of 92.25% and specificity of 89.04%. They showed that the longer duration of diabetes, higher level of glycated haemoglobin, and increased systolic blood pressure were the risk factors associated with referable diabetic retinopathy.

Foshati et al. (17) analyzed different classifiers in Weka software for the prediction of diabetic retinopathy. Their results indicated that the Decorate algorithm was the most vigorous algorithm with an accuracy rate of 0.86%. They concluded that the risk factors related to this disease were age, duration of the disease, BMI, HDL level, HbA1c, FBS, 2HPPG, blood pressure, and treatment method. Rajkomar et al. (18) concluded that diagnostic error would occur in the care of every patient in his or her lifetime, and this problem was not limited to rare diseases. They emphasize that machine learning is not a new tool, like a new drug or medical device. Rather, it is the fundamental technology to process the data that exceed the capacity of the human brain.

Creation of vast amounts of healthcare data will change the nature of medical care, where the patient-doctor relationship will be enriched by additional insights from machine learning. These algorithms can assist the healthcare professionals to make better decisions. In this study, we aimed to apply efficient
pattern detections by machine learning to help the physicians identify conditions that they do not diagnose routinely. Although in recent years, different machine learning algorithms have been used for predicting diabetic retinopathy disease, it seems that studies on different risk factors, patient groups and different methods are still necessary. The purpose of this study was to predict diabetic retinopathy with well-known classification algorithms and discover the factors that may aid in predicting the possibility of developing diabetic retinopathy in a given patient affected by type 2 Diabetes Mellitus. Genetic algorithm (GA) will be used to reduce the feature dimension and select the most discriminate ones by considering the classification error as the fitness function.

Material and Method

Materials

In this study, the diabetic database was collected randomly from the Motahari clinic in Shiraz, Iran between 2013 and 2014. The dataset contained 310 diabetic patient’s information (type 2 diabetes mellitus), where only 155 patients suffered from diabetic retinopathy. The group of diabetic retinopathy patients had a mean age of 57.85 with standard deviation (SD) of 8.87, and the other group had a mean age of 55.57 with a SD of 9.78.

One of the important points is to select the right sample size. This is extremely important since a small sample size will be problematic to generalize, while larger sample size increases the costs (17). The necessary sample size in this study is calculated as:

\[ n = \frac{N \times \left( \frac{Z_{1-\alpha}}{d} \right)^2 \times \sigma^2}{(N-1) \times d^2 + \left( \frac{Z_{1-\alpha}}{d} \right)^2 \times \sigma^2} \] (1)

where \( \sigma \) shows the variance, \( d \) is the permissible margin of error, which in turn, helps to determine the probability level and \( Z_{1-\alpha} = 1.96 \).

29 features were extracted by a specialist, which included four separate parts. The first part included the patient’s personal information such as occupation, gender, age, marital status, education; the second part was about the patient’s medical records information such as, duration of diabetes, body mass index (as a measure of body fat based on height and weight), history of stroke, cardiac diseases, diabetic neuropathy, diabetic nephropathy, family history of type 2 diabetes mellitus, and pack-year smoking history as a means of estimating a person’s lifetime exposure to tobacco. The third part included the patient’s tests information, such as urine albumin level, creatinin, eGFR (estimated glomerular filtration rate), TG, total cholesterol, low-density lipoprotein, high-density lipoprotein, HbA1c, FBS, 2HPGP, blood pressure, and the fourth part was about the treatment method of disease, based on diet, exercise, and medicine (tablet and insulin).

Urine protein levels (Urine Albumin), blood creatinine level (cr), and eGFR are the estimators of kidney function and degree of renal problem in a given patient. Triglyceride (TG), total cholesterol, low-density lipoprotein, and high-density lipoprotein tests are measures of different blood lipids. The average glucose in the previous two or three months (HbA1C), fasting blood glucose levels (FBS), blood glucose 2 hours after consuming food (2HPPG) tests are the measures of blood glucose level in a given patient regarding short and long term glycemic profile.

Methodology

In the first stage, the dataset was preprocessed for removing inconsistency and missing data. Then, for increasing the accuracy of classification, GA was used to remove the redundant features. GA was searched for an optimal subset of features. Three fitness functions based on kNN, SVM and DT algorithms were implemented. These fitness functions only used training data for the training phase and considered the fitness value as the accuracy of classification on the test data. In all stages, 10-fold cross-validation was used, which means that the total number of data was divided into ten equal parts (31 samples out of 310 samples were considered as test data and 279 samples were considered as train data). One part was a test data and the remaining part was training data, and this operation was repeated 10 times. Therefore, each instance can be used both as test and training data, and the mean and standard deviation of accuracy were calculated over 10 runs.

After discovering the most important factors in predicting diabetic retinopathy, classifiers such as SVM, DT, and kNN were applied to the optimized dataset. To determine the robustness of the features and classifiers against noise, white noise, as a disturbance with different amplitudes (10%, 20%, and 30%) was added to the test vectors. In this study, the noise amplitude was added to the maximum amplitude in each dimension. Finally, the student’s t-test was applied to determine whether the difference between the efficiency of the mentioned algorithms was significant or not. The significance level of \( P<0.05 \) was considered statistically significant. All codes in this study were designed and implemented by Matlab.
software. Figure 1 shows the simple flowchart of this study, and the techniques used are described in detail in the following sections. As Figure 1 shows, we tried to learn from observations; then, our proposed approach determined how to perform the mapping from features to labels in order to create a model that generalizes the information for never-seen-before inputs. Since feature selection by hand might be feasible with some variables, but not with large numbers of variables, GA was added for selecting important factors.

**Preprocessing**

At the data collection stage, data cleaning techniques were primarily used to detect and remove the errors and inconsistencies in the data. Most typical data errors were due to the misuse of abbreviations, data entry mistakes, duplicate records, missing values, spelling errors, outdated codes, etc. The first step was to replace the missing data where missing data refer to the data that were not available to the researcher for whatever reason, which can complicate the data analysis. An appropriate decision had to be made concerning such data. In some cases, missing data were replaced with the mean values in that category; in other cases, the missing data were replaced by a fixed value, 10000. Another step which has a logical and desirable conclusion was data normalization. Normalization refers to limiting the values of input to rectify or reduce the model error. In this study, the data were normalized, too.

**Feature Selection**

Feature selection is the process of selecting the smallest subset of features that is necessary for accurate prediction. It is achieved by removal of redundant and irrelevant features. Redundant features are those which provide no further information than the currently selected features, and irrelevant features provide no useful information in any context. The benefits of feature selection are improving the performance of the classifier and providing a better understanding of the underlying process that generated the data. Using a small number of features will also save the computation time and builds a model that is better for the unseen data (19). In this study, GA is used to discover the important features with the goal of best classification of the two mentioned groups.

In GA, persons are known as genes, population

---

**Figure 1:** (a) The simple diagram of proposed approach, (b) Architecture of the GA-based feature selection where the next population is obtained by crossover and mutation operators and fitness value is considered as the classification accuracy of two mentioned groups.
is the array of persons, and diversity is the average
distance between persons (20). GA starts with the
primitive population, generated randomly. To create
the next generation, GA selects a few people as parents
based on their fitness from the current population
and uses them to create the next generation; this
new generation is called children. To get the next
generation, GA performs the following stages until
the number of generations reaches the desired level:
- Evaluating each member of the population by
  measuring its fitness.
- Selecting parents based on their fitness.
- Producing children from selected parents by
crossover and mutation operators.
- Replacing the new generation with the old
generation.

The crossover operator is analogous to
reproduction and biological crossover; the new child
(off-springs) is produced using the genetic material of
the parents. The mutation parameter of this algorithm
determines the small random change in each person.
The genetic mutation allows an algorithm to search
for wider space. In this study, the population size
was considered 100, the crossover rate 0.8, the
mutation rate 0.1, and the maximum number of
generation 100. The length of each chromosome was
29, corresponding to the number of input, and the
fitness function of this algorithm was the accuracy of
predictive algorithms.

Classification

Three well-known algorithms were applied for
classification task: kNN, SVM and DT (21, 22). In
k-NN classification, each sample was classified by
a plurality vote of its neighbors, the sample being
assigned to the class most common among its k
nearest neighbors (k is a positive integer, typically
small). In this study, k was set to 5.

In SVM, the original finite-dimensional space
was mapped into a much higher-dimensional space,
presumably making the separation easier in that
space. Then, the best hyperplane must be selected, the
one that represents the largest separation (margin)
between the two classes. The hyperplane was selected
based on the maximum distance to the nearest data
point on each side.

DT algorithm works by choosing the finest feature
to divide the data and expand the leaf nodes of
the tree until the ending condition is met. In these tree
structures, leaves represent class labels and branches
represent conjunctions of features that lead to those
class labels. After the construction of the decision tree,
a tree-pruning phase can be accomplished to reduce
the size of the decision tree. Pruning is supported by
trimming the branches of the initial tree in a way
that it increases the interpretation capability of the
decision tree.

Results

As described in the materials section, the dataset
contained 29 features; the mean and SD of some
features for each group are shown in Table 1. In
the first stage, all features were considered for the
classification task, and the classification accuracy
was estimated using 10-fold cross-validation. Table 2
(right column) shows the classification accuracy of
65.60%, 80.64% and 80.65% by SVM, DT, and kNN
classifiers, respectively.

In the second stage, GA was applied to select the
most discriminative features. The selected features by
GA included 13 important features: the duration of the
disease, history of stroke, family history, presence of
cardiac diseases, diabetic neuropathy, blood pressure,
urine albumin, 2HPPG, HbA1c, FBS, HDL, LDL, and
age. To better evaluate the performance of the GA, we
drew the variation of best and mean fitness values of
the population for three algorithms. Figure 2 shows the

Figure 2: The best mean fitness for (a) SVM, (b) kNN, (c) DT
classifiers; it can be seen that the GA can improve the mean
fitness of the population (or the quality of individuals) during
generations, especially for kNN classifier.
best and mean fitness of population for SVM, DT and kNN algorithms during 100 generations. Our results showed that the lowest and mean error for the SVM algorithm were 0.214 and 0.246, respectively. For the kNN algorithm, the lowest error was 0 and the mean error was 0.00714. Finally, the lowest error for DT algorithm was 0.0714 and the mean error was 0.079.

After determining the most important factors, the mentioned classifiers were applied to the new dataset. Therefore, 13 important features were considered and the other features were removed from the dataset. Table 2 (left column) showed classification rates of 69.35%, 81.29% and 96.13% by SVM, DT, and kNN, respectively. As Table 2 shows, kNN obtained the highest accuracy, with more than 14% difference in comparison with SVM and DT classifiers. The difference between kNN and two other classifiers was evaluated by student’s t-test. Our results revealed that there was a significant difference between the accuracy of kNN and other algorithms, P=0.034 for SVM and P=0.00008 for DT classifiers.

The ROC and AUC (area under the ROC graph) are the criteria for better evaluation of classification algorithms. The ROC curve was created by plotting the true positive rate against the false-positive rate at various threshold settings. Figure 3 shows the ROC diagram for SVM, kNN and DT algorithms, respectively. For more comparison, the AUC value was calculated: 81.20 for kNN, 51.92 for SVM, and 80.46 for DT. Our results showed that kNN had a higher AUC value compared to DT and SVM algorithms where the higher the value of AUC, the

<table>
<thead>
<tr>
<th>Characteristics name</th>
<th>Patients without retinopathy</th>
<th>Patients with retinopathy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occupation</td>
<td>2.12±1.28</td>
<td>2.09±1.13</td>
</tr>
<tr>
<td>Gender</td>
<td>1.64±0.48</td>
<td>1.67±0.47</td>
</tr>
<tr>
<td>Age</td>
<td>55.57±9.78</td>
<td>57.85±8.87</td>
</tr>
<tr>
<td>Marital status</td>
<td>1.21±0.61</td>
<td>1.08±0.27</td>
</tr>
<tr>
<td>Education</td>
<td>1.88±1.11</td>
<td>1.99±1.25</td>
</tr>
<tr>
<td>Duration of the disease (year)</td>
<td>7.13±7.48</td>
<td>11.65±7.94</td>
</tr>
<tr>
<td>BMI (Kg/m²)</td>
<td>27.75±4.16</td>
<td>27.09±4.35</td>
</tr>
<tr>
<td>History of stroke</td>
<td>0.04±0.19</td>
<td>0.07±0.26</td>
</tr>
<tr>
<td>Presence of cardiac diseases</td>
<td>0.15±0.36</td>
<td>0.28±0.45</td>
</tr>
<tr>
<td>Diabetic neuropathy</td>
<td>0.03±0.17</td>
<td>0.11±0.31</td>
</tr>
<tr>
<td>Diabetic nephropathy</td>
<td>0.06±0.23</td>
<td>0.15±0.36</td>
</tr>
<tr>
<td>Previous wound leg</td>
<td>0.64±0.25</td>
<td>0.12±0.33</td>
</tr>
<tr>
<td>Family history</td>
<td>0.50±0.57</td>
<td>0.63±0.48</td>
</tr>
<tr>
<td>Smoking</td>
<td>0.25±0.43</td>
<td>0.27±0.45</td>
</tr>
<tr>
<td>Cigarette pack year</td>
<td>3.24±0.43</td>
<td>2.55±0.45</td>
</tr>
<tr>
<td>Urine Albumin (mg/gram)</td>
<td>35.46±0.30</td>
<td>71.79±1.23</td>
</tr>
<tr>
<td>Creatinine (mg/dl)</td>
<td>0.99±0.29</td>
<td>1.11±0.60</td>
</tr>
<tr>
<td>eGFR</td>
<td>75.93±0.45</td>
<td>71.95±0.65</td>
</tr>
<tr>
<td>TG</td>
<td>173.28±93.28</td>
<td>175.91±80.40</td>
</tr>
<tr>
<td>Total cholesterol</td>
<td>193.18±57.89</td>
<td>185.55±52.15</td>
</tr>
<tr>
<td>Low-density lipoprotein</td>
<td>108.5±43.40</td>
<td>109.28±37.11</td>
</tr>
<tr>
<td>High-density lipoprotein</td>
<td>50.03±27.47</td>
<td>47.52±23.48</td>
</tr>
<tr>
<td>HbA1c (%)</td>
<td>8.20±2.12</td>
<td>8.92±2.54</td>
</tr>
<tr>
<td>FBS (mg/dl)</td>
<td>162.30±63.80</td>
<td>195.63±95.26</td>
</tr>
<tr>
<td>2HPPG (cm/dl)</td>
<td>245.87±104.21</td>
<td>274.2±101.75</td>
</tr>
<tr>
<td>Treatment method</td>
<td>1.47±0.94</td>
<td>1.41±0.75</td>
</tr>
<tr>
<td>Bp (Systolic) (mm)</td>
<td>137.5±21.19</td>
<td>144.6±24.64</td>
</tr>
<tr>
<td>Bp (Diastolic) (hg)</td>
<td>79.8±11.58</td>
<td>82.45±14.27</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Without feature selection</th>
<th>With feature selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>69.35±4.87</td>
</tr>
<tr>
<td>DT</td>
<td>81.29±1.36</td>
</tr>
<tr>
<td>kNN</td>
<td>96.13±5.65</td>
</tr>
</tbody>
</table>
better the performance of the classifier.

To determine the robustness of the features and classifiers against noise, we added white noise as a disturbance with different amplitudes to the test vectors. The classification accuracy of each classifier against different amplitudes of the noise is shown in Figure 4 and Table 3. Our results showed that the accuracy of all the three algorithms decreased significantly.

### Discussion

Our results showed that the proposed approach is well suited for the meaningful prediction of diabetic retinopathy disease. In recent years, numerous treatment plans have been proposed for early diagnosis of retinopathy disease and these methods have been successful literally, but still, its late diagnosis causes serious problems, often when patients become aware that their illness has progressed. Therefore, the

---

**Figure 3:** ROC Chart for kNN, SVM and DT algorithms; Roc is a graphical plot that illustrates the diagnostic ability of the classifier as its discrimination threshold is varied

**Figure 4:** The accuracy of kNN, SVM and DT algorithms with different noise levels, it can be seen additive noise decreases the classification accuracy of all algorithms

**Table 3:** The Mean±SD of classification accuracy of different classifiers with additive noise

<table>
<thead>
<tr>
<th>Noise levels</th>
<th>SVM</th>
<th>Decision Tree</th>
<th>kNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>65.38±6.39</td>
<td>66.67±4.74</td>
<td>67.31±5.06</td>
</tr>
<tr>
<td>20%</td>
<td>61.72±7.18</td>
<td>57.31±4.21</td>
<td>61.07±5.93</td>
</tr>
<tr>
<td>30%</td>
<td>59.24±7.36</td>
<td>56.77±6.26</td>
<td>60.65±6.07</td>
</tr>
</tbody>
</table>
The proposed approach is suggested as a complementary tool along with traditional methods to solve this problem.

In this study, first, different clinical information (29 features) are considered for the prediction of diabetic retinopathy. Our best-reported accuracy (80.65±2.24) was obtained by kNN with 10-fold cross-validation. In the next stage, GA was applied to improve the efficiency of classification algorithms and discover the important factors regarding the probability of the presence of diabetic retinopathy in a given patient. Combination of GA and kNN classifier increased the accuracy to 96.13±±5.65. The GA highlighted 13 features (or risk factors) including duration of the disease, history of stroke, family history, presence of cardiac diseases, diabetic neuropathy, blood pressure, presence of albuminuria (either micro or microalbuminuria), 2HPPG, HbA1c, FBS, HDL, LDL, and age. Our results showed that these features had more discriminative information for predicting the diabetic retinopathy; this is in line with the findings of epidemiologic and clinical studies that have investigated the risk factors associated with diabetic retinopathy (23-28).

For more comparison, the summary of several previous studies (Tsao et al.(14), Zheng (29), Yan Liu et al. (30), Yun et al. (31), Thapa et al. (32), Bellemo et al.(16), Foshati et al.(17), Ting et al.(33), Magliah et al. (34)) are illustrated in Table 4. As Table 4 shows, some factors are repeated in different studies. These factors are the duration of illness, age, blood pressure, HbA1c, lipids. It shows the importance of these factors for predicting this disease. However, there are differences between the risk factors in Table 4; these differences may refer to an expert’s opinion, available dataset, etc. In these studies, different methods have been used to find the risk factors of diabetic retinopathy. These methods are divided into machine learning-based and cross-sectional based methods. Comparison of the reported performance measure in Table 4 shows that we could achieve better results compared to the other studies although it would not be the best way for comparing different studies unless we run all methods on the same data sets.

Many machine learning algorithms have been used to predict this disease, but it is difficult to choose an efficient algorithm for clinic centers; the right approach can be selected based on the efficiency of the algorithms and selecting important factors risk. Our findings demonstrated that GA could be an effective way of selecting important features and improving the accuracy of the kNN algorithm, and our suggested predicting approach is low-cost and easy.

It must be considered that for a serious illness, like diabetic retinopathy, high accuracy in prediction is of utmost importance and it is less costly than other methods, such as retinal fundus images. Probably, in some communities that have poor access to

<table>
<thead>
<tr>
<th>Authors</th>
<th>Features</th>
<th>Method</th>
<th>Performance measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>This study</td>
<td>Duration of the disease, history of stroke, family history, Cardiac diseases, diabetic neuropathy, blood pressure, urine albumin, 2HPPG, HbA1c, FBS, HDL, LDL, and age.</td>
<td>kNN with a genetic algorithm</td>
<td>96.13%</td>
</tr>
<tr>
<td>Tsao et al.</td>
<td>Insulin, duration of diabetes</td>
<td>Support vector machine</td>
<td>79.50%</td>
</tr>
<tr>
<td>Zheng</td>
<td>Obesity, glycemic, C peptide, lipids, time, renal, blood pressure and metabolic factors</td>
<td>Logistic- Ordinal regression</td>
<td>——</td>
</tr>
<tr>
<td>Liu et al.</td>
<td>Younger age, longer diabetes duration, higher SB, oral medicine use and insulin use, higher postprandial blood glucose (PBG), HbA1c, triglyceride, and low-density lipoprotein</td>
<td>A cross-sectional</td>
<td>——</td>
</tr>
<tr>
<td>Yun et al.</td>
<td>Glycemic control, diabetes duration, age, and albuminuria</td>
<td>Multivariate Cox Proportional Hazards models</td>
<td>——</td>
</tr>
<tr>
<td>Thapa et al.</td>
<td>Duration of diabetes, hypertension, and alcohol consumption</td>
<td>population-based, cross-sectional</td>
<td>——</td>
</tr>
<tr>
<td>Bellemo et al.</td>
<td>Duration of diabetes, a higher level of glycated hemoglobin, and increased systolic blood pressure</td>
<td>neural networks</td>
<td>92.25%</td>
</tr>
<tr>
<td>Foshati et al.</td>
<td>Age, duration of the disease, BMI, HDL, HbA1c, FBS, 2HPPG, blood pressure, treatment method</td>
<td>Decorate</td>
<td>86.72%</td>
</tr>
<tr>
<td>Ting et al.</td>
<td>Age, diabetes duration, HbA1c, and systolic blood pressure</td>
<td>Deep learning</td>
<td>——</td>
</tr>
<tr>
<td>Magliah et al.</td>
<td>Sex, type of diabetes mellitus (DM), obesity, and smoking, the duration of DM, hemoglobin A1c level, uncontrolled diabetes, hypertension, dyslipidemia, nephropathy, insulin treatment, and age</td>
<td>A cross-sectional</td>
<td>——</td>
</tr>
</tbody>
</table>
ophthalmologists for regular follow up, it can help the primary physician for selecting those patients that need to be visited by ophthalmologist urgently.

**Conclusion**

Our results were compared and contrasted with recently reported methods and it was revealed that a considerably enhanced performance was achieved. Thus, the proposed approach can be applied as a complementary tool for specialists to predict and diagnose diabetic retinopathy as soon as possible. It may aid the healthcare professionals to determine and individualize the required eye screening interval for a given patient.

**Conflict of Interest:** None declared.

**References**


