Investigating Non-Laboratory Variables to Predict Diabetic and Prediabetic Patients from Electronic Medical Records Using Machine Learning

Hamid Mukhtar[1] and Sana Al Azwari[2],
h.mukhtar@tu.edu.sa alazwari.s@tu.edu.sa
College of Computers and Information Technology, Taif University,
P. O. Box 11099, Taif, 21944, Kingdom of Saudi Arabia.

Summary
Diabetes Mellitus (DM) is one of common chronic diseases leading to severe health complications that may cause death. The disease influences individuals, community, and the government due to the continuous monitoring, lifelong commitment, and the cost of treatment. The World Health Organization (WHO) considers Saudi Arabia as one of the top 10 countries in diabetes prevalence across the world. Since most of the medical services are provided by the government, the cost of the treatment in terms of hospitals and clinical visits and lab tests represents a real burden due to the large scale of the disease. The ability to predict the diabetic status of a patient without the laboratory tests by performing screening based on some personal features can lessen the health and economic burden caused by diabetes alone. The goal of this paper is to investigate the prediction of diabetic and prediabetic patients by considering factors other than the laboratory tests, as required by physicians in general. With the data obtained from local hospitals, medical records were processed to obtain a dataset that classified patients into three classes: diabetic, prediabetic, and non-diabetic. After applying three machine learning algorithms, we established good performance for accuracy, precision, and recall of the models on the dataset. Further analysis was performed on the data to identify important non-laboratory variables related to the patients for diabetes classification. The importance of five variables (gender, physical activity level, hypertension, BMI, and age) from the person's basic health data were investigated to find their contribution to the state of a patient being diabetic, prediabetic or normal. Our analysis presented great agreement with the risk factors of diabetes and prediabetes stated by the American Diabetes Association (ADA) and other health institutions worldwide. We conclude that by performing class-specific analysis of the disease, important factors specific to Saudi population can be identified, whose management can result in controlling the disease. We also provide some recommendations learnt from this research.

Key words:
prediabetics; prediction; feature importance; feature contribution; PIMA diabetes dataset.

1. Introduction

The correct diagnosis of a disease is one of the serious challenging tasks in developed and developing countries. On the one hand, a disease or illness cannot be confirmed in the absence of appropriate medical tests. On the other hand, performing several tests to identify an illness is not only expensive and time-taking, but it also results in a huge burden on the overall medical system and the involved parties of the society. Due to limited availability of expensive resources like medical experts, laboratory tests and the required equipment, as well as due to the lack of general awareness about many common diseases, many people are undiagnosed [1] or misdiagnosed with alternative illnesses resulting in a huge gap between the government's efforts and their outcomes.

Diabetes mellitus (DM) is one of the common chronic diseases worldwide. In 2019, International diabetes Federation (IDF) announced that the number of adults having diabetes is approximately 463 million of world population [2]. Also, IDF considers the Middle East as one of the highest regions in diabetes prevalence, and the World Health Organization (WHO) places Saudi Arabia as the highest among the Middle East countries [3] and the fifth in the top ten countries known for high diabetes incidence rate in the world. It is expected that Saudi Arabia is heading to a higher position by 2035 [4]. Cost of medical treatment is also affected by the rapid growth of the number of individuals with diabetes, representing a large burden on government health expenses. According to recent estimates, the cost of diabetes incurred by the Saudi government is at 17 billion Riyals and if those with glucose intolerance (pre-diabetes) progressed at the current observed rate, the total cost would be 43 billion Riyals [5] in the coming years. Besides, Saudi Arabia is known for rapid growth in population, and it encountered a soaring economic development in the recent four decades, leading to lifestyle changes due to urbanization. These changes have led to an increasing rate of chronic diseases. However, of the many studies conducted to address the rapid growth of diabetes mellitus, most had either the objective to quantify the status of diabetics in the country [4] [6], identifying the most frequently performed self-care behaviors [7], identifying factors related to diabetes control [8] or application of mathematical [9] or machine learning models for diabetes prediction [10], etc. All these efforts are related to the

Manuscript received September 5, 2021
Manuscript revised September 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.9.3

increasing demand to enhance healthcare quality and control the elevated growth rate of diabetes in the kingdom.

We believe that the existing efforts may have their own benefits and usefulness in tackling the diabetes issues in Saudi Arabia; however, there is a need to devise mechanisms for efficient, cost-effective, and easily-available solutions for diabetes identification in the general population. Given the constant rise in the diabetic population in the country, it is imperative to not only be able to identify diabetic from non-diabetic persons, but at the same time, the growing numbers of prediabetes should also be determined.

Considering the above-mentioned context and the need of time, we are motivated to develop a solution for predicting diabetic and prediabetic patients from EHR obtained from local Saudi hospitals. Therefore, the goal of this research paper is to develop predictive classifiers and models to investigate the real diabetic patients' data gathered from different Saudi hospitals and regions, by means of different metrics. Although the obtained records include various lab test results such as cholesterol tests (HDL and LDL), and the diabetes-specific tests, (FPG and HbA1c), our objective is to use the patients' data for classification of patients into the diabetic, non-diabetic, and prediabetic groups without any medical test results, only from patient's basic health data. Previous work in this direction has used a much larger number of variables in a different context. Thus, the current work presents a novel perspective on diabetes prediction. The insights obtained from this work in prediction of Diabetes Mellitus (DM) and its associated risk factors can be useful at different levels: to support and strengthen the existing findings of DM medical research, particularly, in the context of Saudi Arabia, to assist the community in understanding the causes and prevention of the diabetes and help the government to allocate efforts in the right direction to minimize the effect of growing number of diabetes patients.

The remainder of this paper is structured as follows. In Section 2, we briefly explain the diabetes mellitus disease followed by the related work done in diabetes classification. In Section 3, we explain our research methodology from data collection and preprocessing to the development of machine learning models for diabetes prediction and classification. The results are then discussed in Section 4. Section 5 discusses the outcome and benefit of our research and Section 6 concludes this article.

2. Background and Related Work

2.1 Diabetes Mellitus (DM)

Diabetes mellitus (DM) is a set of endocrine disorders resulting in high levels of blood glucose in the human body due to deficiency in insulin excretion or insulin action and sometimes both. It causes direct and indirect complications responsible for significant illness and death [6]. There are different types of diabetes, but the most common ones are type 1 diabetes (T1D) and type 2 diabetes (T2D). Type 2 (T2D) is the most popular form of diabetes; around 90% of diabetic patients are (T2D). The remaining 10% is classified as type 1 (T1D) or gestational diabetes that may occur during pregnancies. Main causes of T2D are low physical activity, lifestyle, and nutritional habits. T1D is caused due to destruction of Langerhans islets holding β-cells of pancreas [11]. Since diabetes is a chronic disease, patients are required to monitor glucose level and physical lifestyle lifelong.

Some selective features from diabetes data can positively affect the performance of prediction models. As the disease is related to the production and control of insulin in the body, it is used as a treating agent for diabetes [12]. The blood test for the measurement of Hemoglobin A1c (HbA1c) level is clinically significant in prediabetes and diabetes diagnosis [13]. Similarly, the glucose in plasma of fasting subjects is accepted as a diagnostic criterion for diabetes [14]. Moreover, according to American Diabetes Association (ADA) there is more harmony between blood tests such as FPG and HbA1c when compared to two types of blood tests in separation of HbA1c [15]. Thus, as we will see later, HbA1c and FPG laboratory tests are sufficient to determine if a person is diabetic, prediabetic or non-diabetic. The availability of both tests, or anyone of them, especially HbA1c is enough to know the diabetic condition of a person. If a physician has results of any of these, further investigation of diabetes evaluation may not be needed in most cases.

It has only become challenging for a physician to know the diabetic status of a patient in the absence of these tests. And the utility of machine learning for classifying a person into diabetic, prediabetic, or non-diabetic is there when the prediction can be made in the absence of such tests. Unfortunately, most of the existing work achieve good results of diabetes prediction only when they include these tests in their input to the machine learning model [16] [17] [18] [19] [20] [21]. Such predictions cannot be of use in real-world situations, where inclusion of these features in the models imply taking medical blood tests, and in absence of these tests, making it impossible to identify the diabetic status of a patient.

Finally, while classification of a person into diabetic or non-diabetic is relatively simple, many existing models do not consider the prediabetes stage to find out the associations that are relevant to developing Diabetes Mellitus. From a medical point of view, it is possible to avoid DM disease at this early stage or at least control its complications [22]. For example, Individuals with a certain range of FPG and HbA1c, i.e., $100 \leq FPG \leq 125$mg/dl and $5.4\% \leq HbA1c \leq 6.4\%$, are considered as prediabetic patients [13]. Their

early diagnosis can help in preventing their transition to the diabetic or in their recovery into the non-diabetic stage.

2.2 Related Work

Othmane et al. [23] applied and evaluated four machine learning algorithms (decision tree, K-nearest neighbors, artificial neural network, and deep neural network) to predict patients with or without type 2 diabetes mellitus. These techniques were trained and tested on two diabetes databases: one obtained from Frankfurt hospital (Germany), and the other one, the openly available, well-known Pima Indian dataset[1]. These datasets contained the same features composed of risk factors and some clinical data such as number of pregnancies, glucose levels, blood pressure, skin thickness, insulin, BMI, age, and diabetes pedigree function. The results compared using different similarity metrics gave classification accuracy of more than 90% and up to 100% in some cases. The limiting factor of their approach was the inclusion of glucose (FPG) and insulin in the training data of the model. Since both these factors can determine diabetes with almost certainty, as discussed above, their inclusion in the input data discloses the vital information to the machine learning model, rendering its prediction useless.

Lai et al. [24] built a predictive model to better identify Canadian patients at risk of having Diabetes Mellitus based on patient demographic data and the laboratory results. Their data included the patient features age, sex, fasting blood glucose, body mass index, high-density lipoprotein, triglycerides, blood pressure, and low-density lipoprotein. They built predictive models using Logistic Regression and Gradient Boosting Machine (GBM) techniques achieving good results. But inclusion of so many laboratory tests that readily indicate the patient as diabetic or non-diabetic is not useful in the machine learning predictions. By looking at these results, a physician can immediately classify the patient as diabetic or non-diabetic without any aid from a computer model.

In a similar fashion, many other approaches train their model on the data that contain a feature in the input data, which leak the classification criteria and, thus, result in very good performance. For example, in [16] [17] [18] [19] [20] [21]) the authors used the Pima Indian diabetes dataset by modifying the preprocessing steps, applying different algorithms, and adjusting their hyperparameters to generate good results. But in all these approaches, the machine learning model contained an important attribute (e.g., FPG) during the training process, which was decisive in the classification of a patient as diabetic or not. Thus, the model was exposed prematurely to an attribute that can alone predict the result. Despite this, many research works have compared the performance of several machine learning approaches for diabetes classification and concluded that a certain type of algorithms can give better results for prediction without considering the issue of data pollution due to input [25] [26].

Several other studies have used the National Health and Nutrition Examination Survey (NHANES)[2] from the US center for disease control (CDC) for prediction of diabetes or some other disease. The NHANES data was initiated in 1999 and is growing every year in the number of records as well as the variables it considers in its surveys. These studies, while utilizing the main NHANES dataset, use some subset of variables for disease prediction or classification tasks. For example, Yu et al. [27] identified fourteen important variables – age, weight, height, BMI, gender, race and ethnicity, family history, waist circumference, hypertension, physical activity, smoking, alcohol use, education, and household income – for training their machine learning models. Using two different classification schemes, they achieved 83.5% and 73.2% results for the area under the receiver operating characteristic (ROC) curve. Semerdjian and Frank [28] added two more variables – cholesterol and leg length – in their analysis. . By applying an ensemble model using the output of five classification algorithms they were able to predict the onset of diabetes with an AUC of 83.4%. In both these studies, the number of variables (14 and 16) were significantly higher than would normally be available in most EHRs. Also, hospitals supporting the record of these variables may not have the values for all these variables for maximum patients. This limits the generic or wide applicability of the approaches.

The study by Dinh et al. [29] used the NHANES dataset and various machine learning algorithms to predict variables that are a major cause for the development of diabetes and cardiovascular diseases. They also considered the prediction of prediabetes and undiagnosed diabetes. Logistic regression, support vector machines, random forest, and gradient boosting algorithms were used to classify the data and predict the outcome for the diseases. Authors used ensemble models by combining the performance of the weaker models to improve the accuracy. In diabetes classification, they used 123 variables, and achieved good prediction performance. A distinguishing aspect of their work was that the dataset was further categorized into laboratory dataset (containing laboratory results) versus no laboratory (survey data only) dataset. Laboratory results were any feature variables within the dataset that were obtained via blood or urine tests. The purpose of non-laboratory dataset was to enable performance analysis of

[1] http://www.kaggle.com/uciml/pima-indians-diabetes-database

[2] http://wwwn.cdc.gov/nchs/nhanes/

Table 1. The set of features selected in our dataset for classification of diabetic and prediabetic patients

No.	Feature Name	Feature Type	Feature Description
1	Date of birth	Date	Values in date format
2	Gender	Binominal	F: Female, M:Male
3	Height	Numerical	Values in Centimeter (cm)
4	Weight	Numerical	Values in Kilograms (kg)
5	Hypertension (HTN)	Binominal	Values as Yes, No
6	Fasting Plasma Glucose (FPG)	Numerical	Lab test results measured in mmol/L
7	Hemoglobin A1c (HbA1c)	Numerical	Lab test results measured in percentage (%).
8	High-density lipoprotein (HDL)	Numerical	Lab test results in mmol/L
9	High-density lipoprotein (LDL)	Numerical	Lab test results in mmol/L
10	Physical Activity Level	Categorical	Values in L: Low, M: Medium, H: High
11	Diagnosis start date	Date	Values in date format
12	Primary diagnosis	Categorical	Values available in ICD10 code format.
13	Secondary diagnosis	Categorical	Values available in ICD10 code format.
14	Primary diagnosis full name	Categorical	Values indicate diagnosis full name
15	Secondary diagnosis full name	Categorical	Values indicate diagnosis full name
16	Region	Categorical	Values indicate the region of the patient whether in central, western or eastern region.

machine learning models in cases where laboratory results were unavailable for patients, supporting the detection of at-risk patients based only on a survey questionnaire. According to their results, waist size, age, self-reported weight, leg length, and sodium intake were five major predictors for diabetes patients. The study found that machine learning models based on survey questionnaires can give automated identification mechanisms for patients at risk of diabetes. In non-laboratory data, the most important features included waist size, age, weight (self-reported and actual), leg-length, blood-pressure, BMI, household income, etc. [29]. The exact number of variables used in non-laboratory data is not reported by the authors, and, thus, it cannot be concluded if their approach can be useful in general situations.

3. Materials and Methods

In this work, several supervised learning algorithms, specifically classification algorithms, have been applied to evaluate our proposal. Its validation has been carried out using traditional metrics such as accuracy or precision, as well as classic and new metrics designed specifically to evaluate classification models induced from imbalanced data.

3.1 Data collection

The anonymized EHRs have been acquired from five different Saudi hospitals across three regions: Central region, the Western region, and the Eastern region. It contains data for around 3000 patients and collected for more than two years from 2016 until 2018 through different departments such as outpatient, inpatient, and emergency. Moreover, the obtained dataset consists of 17 features of numerical, binominal, polynomial, and date type. The initial features along with a brief description of each are listed in Table 1.

3.2 Data Preprocessing

In the data preprocessing phase, data is prepared to be suitable for cleansing and classification. The work in this phase is divided into two sub phases. First phase deals with how data is prepared for cleansing. In the second phase, data is cleansed using normalization (smaller range) and discretization (intervals of numeric values). Transformation of some columns (features) were performed; for example, birth date was used to generate the age of the patient. The Body Mass Index (BMI) was calculated for each adult patient from their height and weight. The general formula for BMI calculation divides the weight by the square of height (kg/m2).

In addition, many patients were missing important feature values like FPG and HbA1c. Since both features were used to initially classify a person as diabetic, prediabetic, or non-diabetic, to establish the ground truth, all the patients who did not have these feature values were removed. Replacing the missing values for both features would not be helpful since the number missing these features was so high. After filtering the patients, our dataset decreased to 225 eligible patients for classification.

Table 2. The total number of instances in the initial dataset according to the features: diabetic (Y) and non-diabetic (N)

Class		Gender		Region		Age		PAL		BMI		HTN	
N	86	F	54	Central	31	(18 – 30)	6	L	99	Normal	15	NO	46
		M	32	Eastern	27	(31 – 45)	28	H	3	Overweight	29	YES	78
				Western	28	(> 45)	53	M	22	Obese	80		
Y	113	F	70	Central	42	(18 – 30)	15	L	63	Normal	8	NO	24
		M	43	Eastern	67	(31 – 45)	32	H	3	Overweight	27	YES	51
				Western	4	(> 45)	65	M	9	Obese	40		

However, 43 out of 225 patients were missing HDL and LDL values. HDL is considered as "Good Cholesterol" – higher HDL means better state – while LDL is considered as "Bad Cholesterol". Therefore, lower LDL values are desirable. In the experiments, when HDL and LDL values were used, the records with missing values were dropped. Fasting Plasma Glucose (FPG) and Hemoglobin A1c (HbA1c) values were also transformed using American Diabetes Association (ADA) as reference for the different value ranges [30].

3.3 Subject Exclusion

In this study, we excluded subjects whose age was less than 19 years to focus on the prediction of T2D by reducing the chances of T1D, which usually develops in children and adolescents. Previous work [29] [28] [27] also excluded similar data as well as data indicated as gestational diabetes, which is relevant to pregnant women; however, since we lack information on pregnancy, we did not perform this step.

3.4 Features Selection

Of the 16 features mentioned in Table 1, we had to remove some to reduce the number of features to the minimum required for classification. We proceeded as follows. The date of birth was replaced by age. The height and weight were replaced by the BMI feature calculated by us. All the features containing diagnosis information (primary, secondary and their full names) were removed. The diagnosis was given in ICD10 codes. ICD10 is an abbreviation of the International Classification of Diseases where 10 represents the 10th revision[3]. The diagnosis of patients includes multiple diagnoses, most importantly is type 1 diabetes (T1D), type 2 diabetes (T2D) and prediabetes. Presence of ICD10 codes expose the same problem as by the FPG and HbA1c test results. Hence, their removal was necessary. Finally, the region and diagnosis start date features were also removed.

After initial feature selection, the dataset obtained consisted of 9 features: age, BMI, gender, hypertension (HTN), physical activity level (PAL), lab tests of lipoprotein levels (HDL and LDL), fasting plasma glucose (FPG) and Hemoglobin A1c (HbA1c). We would like to mention that age, BMI, HDL, LDL, FPG and HbA1c were all numerical features, while gender (M or F), HTN (Yes or No), and PAL (L, M, or H) were categorical features containing text or literals. As our implementation is done in the scikit-learn library[4], whose methods require numerical data for efficient processing, we converted the categories to numerical values. Instead of replacing text by numbers (e.g., L:0, M:1, H:2), we used one-hot encoding to prevent the implicit ordering caused by the numeric values. Table 2 details two class details of the patients in the dataset.

At this stage, our data processing steps were finished. Before starting the analysis, it was imperative to identify each record as representing data for a diabetic, prediabetic or non-diabetic patient. In other words, each record was to be labelled with an appropriate class.

3.5 Label Assignment

The appropriate references to use for evaluating diabetes were the "Standards of Medical Care in Diabetes – 2018" from the American Diabetes Association (ADA) [31] and considering the algorithm proposed by the American Association of Clinical Endocrinologists (AACE) [30]. Two medical experts were also consulted who provided guidance in the diagnosis of diabetes and prediabetes including the factors related to predict the development of diabetes among people. Their suggestions agreed with the ADA and AACE specifications. Based upon these criteria, FPG and HbA1c laboratory tests were used to classify patients into three classes (categories): Diabetic, Prediabetic, and Non-Diabetic. Table 4 summarizes the ranges for FPG and HbA1c used for the three classes. An algorithm was used to automatically label all the records in the dataset with either of these classes using the criteria in the table. For example, a person was labelled diabetic if his/her FPG $\geq$ 126 OR HbA1c $\geq$ 6.5. In the similar way, prediabetic and non-diabetic were assigned labels. The data preprocessing steps, and labeling resulted in the division of our dataset by various features as shown in Table 3.

[3] http://eicd10.com/

[4] http://scikit-learn.org/

Table 3. The total number of instances in the final dataset according to the features: diabetic (D), prediabetic (P) and non-diabetic (N)

Gender		Region		Age		PAL		BMI		HTN		Label	
F	124	Central	46	(18-30)	9	L	99	Normal	15	NO	46	N	10
		Eastern	54	(31-45)	41	H	3	Overweight	29	YES	78	P	44
		Western	24	(>45)	75	M	22	Obese	80			D	70
M	75	Central	27	(18-30)	13	L	63	Normal	8	NO	24	N	11
		Eastern	40	(31-45)	19	H	3	Overweight	27	YES	51	P	21
		Western	8	(>45)	42	M	9	Obese	40			D	43

Table 4. The grouping of FPG and HbA1c ranges for classifying patients into diabetic, prediabetic, and non-diabetic classes

Class	FPG Range	HbA1c Range
Normal	<100 mg/dl (5.6 mmol/L)	<5.7% (39 mmol/mol)
Prediabetes	100 – 125 mg/dl (5.6 – 6.9 mmol/L)	5.7 – 6.4% (39 – 47 mmol/mol)
Diabetic	≥ 126 mg/dl (7.0 mmol/l)	≥ 6.5% (48 mmol/mol)

We can observe some imbalance of the data for the three classes. The diabetic instances (n=113) are more than the prediabetic (n=65) and non-diabetic (n=21). The combined non-diabetic and prediabetic data (n=86) could be used as a single class of non- diabetic as done in existing approaches, and the imbalance could have been avoided, but the analysis and prediction of pre-diabetic is an important aspect of our work, so we preferred to use resampling approach for dealing with the imbalanced datasets to obtain reliable results. We also observe that there are more instances of females (n=124) as compared to males (n=75) and the Western region has comparatively less instances than the other two regions, so we dropped the region attribute.

3.6 Model Development

Our aim was to use a minimum number of features, particularly the non-laboratory features, to classify a patient into one of the three classes. At the same time, first, it was important to ensure if the initially selected features containing the laboratory test were sufficient to perform the classification correctly. Thus, we divided our experiments into three different cases: I, II, and III.

In case I, all the 9 features – including the laboratory test results for LDL, HDL, HbA1c, and FPG – were included. This would serve as a base case. Next, in case II, the laboratory test results of HbA1c and FPG were excluded, and the remaining 7 features were used. Finally, in case III, the LDL and HDL laboratory tests were also excluded and only five features age, BMI, gender, HTN, and PAL were used for classification. In all these cases, the original labeling of data was kept intact.

Although the classification can be done by any machine learning classifier, by relying on a single classifier, we could obtain incorrect results due to algorithm configuration or optimization issues. As there are many machine learning classifiers available, we chose three of them to conduct our experiments so that their performance can be validated against one another. The selected classifiers included Support Vector Machines (SVM), Decision Trees (DT) and Random Forest (RF). The rationale for choosing these is based on their previous performance reports in similar situations [27] [26] [10]. As our objective was to understand the factors contributing to the classification, we chose not to use any neural networks-based classifier in our analysis due to their "black-box" nature of interpretation of the model [29] [32].

4. Results

To measure the performance of each classifier, we used the widely-accepted performance statistics: accuracy, precision, recall, and F1-score [33].

4.1 Performance of Machine Learning Classifiers

Table 5 describes the comparative performance of the three classifiers against each performance metric in the three cases. In the first case (case I), we can see that by utilizing the complete feature set that includes the laboratory results for FPG, HbA1c, HDL, and LDL, the classification results are comparable to existing approaches for diabetes classification [10] [26] [29] [27]. We observe that the decision tree and random forest classifiers have better performance in all metrics compared to SVM. This is because SVM does not perform well in multi-class analysis. By default, they are good in binary classification and need configuration, which we did not apply [34].

The results could have been improved further by optimizing the hyperparameters [35]. But this was not the main objective of our research. As discussed before, case I was only designed to serve as a base case to ensure that our

Table 5. Three cases in our research with or without FPG, HbAlc, LDL and HDL lab tests results. The **same model** of each classifier was used in all cases.

	Case I Data without FPG and HbA1c features				**Case II Data without FPG, HbA1c, LDL and HDL**			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
SVM	0.78	0.78	0.71	0.73	0.65	0.75	0.54	0.59
Decision Trees	0.59	0.58	0.59	0.59	0.51	0.52	0.55	0.52
Random Forest	0.76	0.71	0.69	0.7	0.59	0.55	0.57	0.56

data can be used for the classification task. So, we move on to the next cases.

In case II, when FPG and HbA1c laboratory results were removed, the same classifiers were applied again without any adjustment for data loss or compensation. The performance dropped suddenly for all metrics from as minimum as 20% to as much as 40%. This was expected because both these tests were used as criteria to assign class labels and excluding them removed all the primary information related to the classification task. In general, a change in the data, especially reduction of attributes, usually requires updating the classifier parameters; but we used the same parameters. Here, we see that SVM and random forest have similar performance but decision tree performance is much worse. This can be interpreted by the fact that random forest is an ensemble learning technique [36] that uses several decision trees simultaneously for performance evaluation and different combinations of decision trees lead to better results in comparison with that produced by a single tree.

Finally, in case III, we excluded all the four laboratory results, and used the five non-laboratory features (age, gender, BMI, PAL, and HTN). We see that the performance of the classifier dropped further, in some cases more than 15% of drop was noted as compared to case II. This means that the cholesterol level test (HDL and LDL) has some relationship with diabetes. This is also in agreement with the existing findings [37]. The performance of the decision tree was just better than chance (50%) and its reason has been stated in the evaluation of case II: we did not make any changes in the model of classifiers used for the prediction tasks in case I.

4.2 Interpretation of Diabetes Classification without Laboratory Results

Although the performance of the machine learning models was not good in case III, instead of their prediction performance in the classification task, we are interested in analyzing the predictive power of each variable among the five features that we selected in case III: age, gender, BMI, HTN and PAL. For this, we carry out the feature importance analysis as follows.

Fig. 1 depicts the role played by each feature in the overall classification of diabetes/prediabetes/non-diabetes. The values are obtained from the coefficients assigned to these features by the random forest algorithm. We can see that BMI and age of a person play an important role in the classification, while the physical activity level (PAL) plays the least important role in the overall classification. To understand how these features play an important role, we consider Fig. 2.

Fig. 2 specifies the impact of each of the features on the three individual classes when the overall impact is considered simultaneously. The variable age plays an important role in the classification process, but it has a major impact in predicting prediabetes and the least in predicting non-diabetic people, when considering this variable alone. Compared to age, BMI has slightly less impact on prediabetes, which indicates that increased weight plays a role in diabetes. A previous study in Saudi Arabia [38] has found a small correlation between BMI and DM (r=.149) while in our case the correlation is even smaller (r=.015). The same study has found a slightly higher correlation between DM and HTN (r=.366), while in our case, it is still smaller (r=.168). These differences may be due to the difference in the size of data as well as the time periods the data was collected.

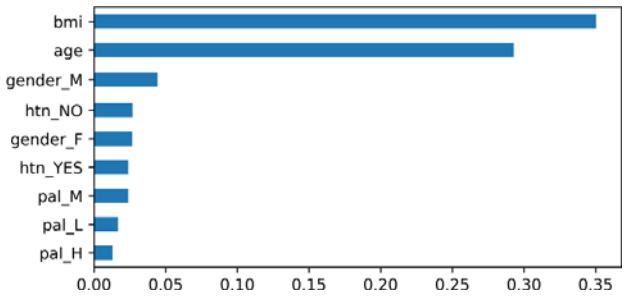

Figure 1. Feature importance of five variables according to their role in diabetes classification

It can be seen in Fig. 2 that HTN (hypertension=Yes or No) has an almost similar role in all the classes. That is why it is not among top variables with respect to feature importance in Fig. 1. Like HTN, the Gender of a person (M or F) has no impact on diabetes classification. However, this is because the model for feature importance considers the overall contribution with respect to all variables in all classes. To find the individual role played by the variables in each class independently, we now consider the feature importance of each variable in each class.

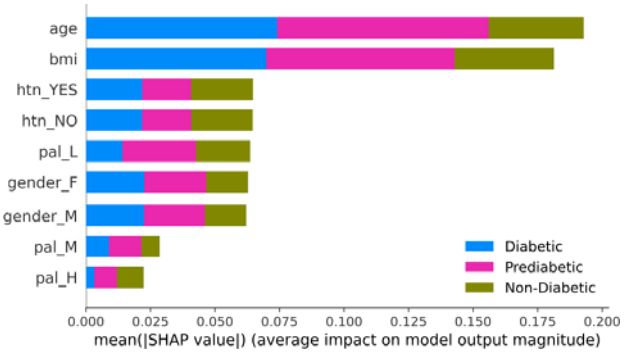

Figure 2. The contribution played by five variables according to their role in classifying diabetic, prediabetic, or non-diabetic persons

Fig. 3–5 depict the importance played by each of the five variables in each class. PAL is one variable that has the highest differences in the classification. For diabetes (Fig. 3), positive values of pal_L and pal_M suggest that lower and medium values of PAL play a role in predicting a person as diabetic, while negative value of pal_H imply that lack or absence of higher PAL contributes more to the other classes. The comparison is fully understood when we consider the feature importance for prediabetic and non-diabetic classes (Fig. 4 and Fig. 5, respectively). For prediabetics, lower values of PAL (i.e., higher feature importance for pal_L are an important factor in its prediction, while in non-diabetics, higher values of PAL (greater feature importance of pal_H or higher level of physical activity) contribute to their prediction. From the combined analysis of PAL, we conclude that when PAL is high (pal_H), it contributes more to non-diabetic; when PAL is medium (pal_M), it contributes to both prediabetic and non-diabetic with more impact on the latter; finally, for low levels of PAL (pal_L), we see its higher impact on prediabetes. The overall analysis concludes that lower levels of PAL develop into prediabetic, and together with age and BMI they give rise to diabetes. Higher levels of PAL, on the other hand, are an important factor in non-diabetic people.

The presence of hypertension (HTN=Yes) is an important feature for classifying a person as diabetic and its negative values for prediabetic and non-diabetic imply lack of

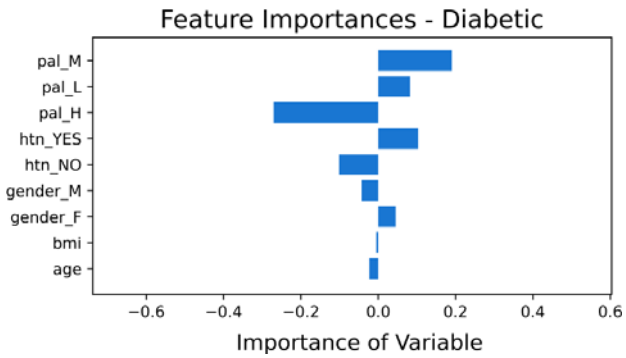

Figure 3. Importance of each feature in predicting a person as diabetic

hypertension contributes to their prediction. The female gender (gender_F) has more important role in diabetes class, either because some females may have gestational diabetic, and we do not have data on that or because 60% diabetic patients are females (51/85) as compared to 40% males (31/85) in the diabetic class obtained after labeling. More data and additional variables may be needed to explore this effect. We also see that in the individual contributions of variables, age and BMI play the least important role while they are the most important ones when considering in the overall classification (Fig. 1). This is due to the complex interplay of variables when used together. As such, the importance of selected features does not imply that the remaining features do not play a role in data analysis. The non-significant features might not add up much individually but removing multiple such features is likely to decrease the performance evaluation.

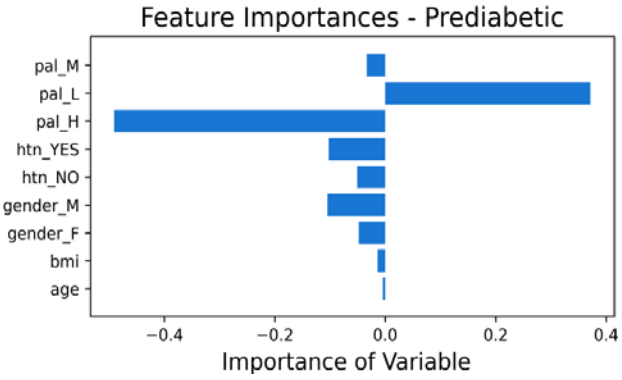

Figure 4. Importance of each feature in predicting a person as prediabetic

5. Discussion

In the previous section, we established that some physiological features could play an important role in the overall classification of a person into diabetic, prediabetic, or non-diabetic classes. The features also contribute to the

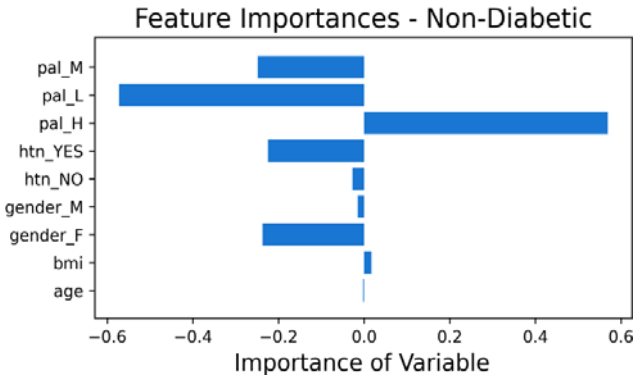

Figure 5. Importance of each feature in predicting a person as non-diabetic

prediction of individual classes such as the different role played by the PAL feature in each class.

When compared to existing approaches, we can identify some distinguishing features of our approach. First, from the literature review, many existing approaches build a learning model incorrectly by including such features in the model training that can help in the classification task alone and including them in the training process undermines the predictive capability of the remaining features. Such variables include the various lab tests usually prescribed by the medical experts to confirm the glucose or sugar level in the body and the availability of these tests do not require application of machine learning to classify patients.

On the other hand, Dinh et al. [29] went in the right direction of using non-laboratory data for diabetes classification. However, their data had hundreds of features and even after removing the various laboratory tests, they were left with a much higher number of features (the exact number is not known). Finding these many features in real-world data is rarely possible. So, we proposed a mechanism whereby only with 5 available physiological features, we can infer the role played by them in the classification of a person into any of the three classes: diabetic, prediabetic, or non-diabetic.

The identification of contribution of each feature through the feature importance is significant in the current analysis. Mostly, a correlation analysis is performed to identify such hidden patterns from data. However, as can be seen in Fig. 6, visualizing the correlation, or seeing the correlation values for different features does not reveal the same information as we have inferred from our results. We can only see that FPG and HbA1c have high correlation (r=.75), which is evident from the fact that they both have the same purpose; and that BMI and PAL are negatively correlated (r=-.61), a well-known fact but which does not contribute to diabetes classification in general. Thus, feature importance depict better contribution in diabetes classification.

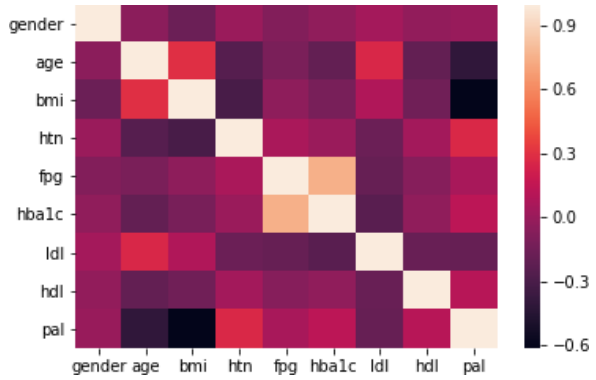

Fig 6 Correlation between the 9 features used in our analysis

The performance accuracy of a classification task mainly depends on the availability of large amounts of data [39]. Unfortunately, our final dataset had only 199 records and after removing the missing values found for LDL and HDL features, we had only 162 records with complete feature values. With such small-scale data, there are limited options to test the available classifiers as well as configuration of their various hyperparameters. That is why, we could not optimize our classifiers for the given small dataset. The current study used BMI as a feature that was calculated from height and weight of the patient. Most of the existing work uses the values for height and weight [29] [39] [28]. This was done to reduce the number of features in view of a small dataset, but in the future, we may evaluate weight and height individually for achieving better classification. In the future, we can work on the predictive performance of our classifiers on relatively bigger datasets.

After establishing the feature importance of various features, we can even utilize black-box approaches like machine learning or deep learning and achieve state-of-the-art performance evaluation results [16] [17] [18].

5.1 Comparative evaluation

The current work is a continuation of previous research efforts. Ahmad et al. [40] used hierarchical clustering for feature selection. Using comparison of five machine learning classifiers, they showed that FPG-labelled dataset had better classification accuracy than the HBA1c-labelled dataset. The SVM classifier had the highest accuracy compared to logistic regression, decision tree, random forest, and ensemble of classifiers. Overall, it is evident that by using feature selection before applying machine learning to the dataset, we can improve the classification accuracy.

5.2 Recommendations

With the insights from the current work, we can present some recommendations. First, we can see that with limited physiological data, patients can be prescreened for diabetes and in case of their classification into prediabetic, they can be advised to carry out laboratory tests and make appropriate changes to their lifestyle. Second, accurate recording of physiological data is important and should be enforced by hospitals and local clinics for any visiting patients for better opportunities to diagnose patients-at-risk. Third, EHRs should contain some additional features of the patients as well. In our current work, we only had access to the hypertension feature of a patient as being Yes or No. But in practice, the patient's blood pressure is recorded as diastolic and systolic values. Similarly, temperature, vision, waste size, etc. are some other features that can be recorded with commonly available instruments in every clinic. So, these and other features should be recorded for each patient to improve the diagnostic process. Finally, the government could enforce prescreening of diabetes and people should be aware of diabetes risk-factors and its prevention in case of prediabetes without the need for going through the procedures of carrying out lab tests.

5.3 Limitations of the work

We can also identify a few limitations of our work. First, as data availability is an important issue in health science research, although our data concerned 3000 patients, the final size of data was very small; with large data, we may have better insights. Second, the data was obtained in the context of Saudi Arabia. It would be interesting to test our approach on similar datasets from other countries/regions of the world. Third, due to lack of the data, we could not focus on improving the classification performance of our approach. With more data, better classifiers can be trained, evaluated, and optimized.

6. Conclusion

The prevalence of diabetes is not only a burden for the governments in terms of the associated expenditures, but it is also a lifelong strain on diabetic patients. Due to lack of data in the EHR of a patient, diabetes classification is a challenging task without the required laboratory tests. In this work, we identified some physiological features from patient's basic health data that can be used as a prescreening test for classifying a patient as diabetic, prediabetic, or non-diabetic without the immediate availability of relevant laboratory tests. With data from other countries, our approach could be generalized, which may have important implications in the healthcare community. The prescreening of diabetes could be rapid, people could be more aware and educated about their lifestyles and the government expenditures could be reduced alongside the decrease in the significant burden on hospitals due to the prevalence of diabetes. With the ability to predict the onset of prediabetes, necessary steps can be taken to avoid the diabetic stage of millions of people who get undiagnosed due to limited resources and lack of awareness. This can not only improve the person's quality of life but also result in a positive impact on the healthcare system. Several recommendations have been proposed in this article in this regard.

References

[1] M. J. Alomar, K. R. Al-Ansari, and N. A. Hassan, "Comparison of awareness of diabetes mellitus type ii with treatment's outcome in term of direct cost in a hospital in Saudi Arabia," *World journal of diabetes,* vol. 10, no. 8, p. 463, 2019.

[2] P. Saeedi, I. Petersohn, P. Salpea, B. Malanda, S. Karuranga *et al.,* "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the international diabetes federation diabetes atlas," *Diabetes research and clinical practice,* vol. 157, p. 107843, 2019.

[3] N. Cho, J. Shaw, S. Karuranga, Y. Huang, J. da Rocha Fernandes *et al.,* "Idf diabetes atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045," *Diabetes research and clinical practice,* vol. 138, pp. 271–281, 2018.

[4] K. Al-Rubeaan, H. A. Al-Manaa, T. Khoja, N. Ahmad, A. Alsharqawi et al., "The saudi abnormal glucose metabolism and diabetes impact study (saudi-dm)," Annals of Saudi Medicine, vol. 34, pp. 465 – 475, 2014.

[5] M. AlMazroa, "Cost of diabetes in saudi arabia," Iproceedings, vol. 4, no. 1, p. e10566, 2018.

[6] A. Alotaibi, L. Perry, L. Gholizadeh and A. Al-Ganmi, "Incidence and prevalence rates of diabetes mellitus in saudi arabia: An overview," Journal of Epidemiology and Global Health, vol. 7, pp. 211 – 218, 2017.

[7] A. M. Saad, Z. M. Younes, H. Ahmed, J. A. Brown, R. M. Al Owesie et al., "Self-efficacy, self-care and glycemic control in saudi arabian patients with type 2 diabetes mellitus: A cross-sectional survey," Diabetes research and clinical practice, vol. 137, pp. 28–36, 2018.

[8] M. A. Alsuliman, S. A. Alotaibi, Q. Zhang and P. K. Durgampudi, "A systematic review of factors associated with uncontrolled diabetes and meta-analysis of its prevalence in saudi arabia since 2006," Diabetes/Metabolism Research and Reviews, p. e3395, 2020.

[9] E. Almutairi, M. Abbod and T. Itagaki, "Mathematical modelling of diabetes mellitus and associated risk factors in saudi arabia." International Journal of Simulation–Systems, Science & Technology, vol. 21, no. 2, 2020.

[10] A. H. Syed and T. Khan, "Machine learning-based application for predicting risk of type 2 diabetes mellitus (T2DM) in saudi arabia: A retrospective cross-sectional study," IEEE Access, 2020.

[11] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas et al., "Machine learning and data mining methods in diabetes research," Computational and Structural Biotechnology Journal, vol. 15, pp. 104 – 116, 2017.

[12] H. Yki-Järvinen, "Combination therapies with insulin in type 2 diabetes," Diabetes care, vol. 24, no. 4, pp. 758–767, 2001.
[13] D. Sacks, "A1c versus glucose testing: A comparison," Diabetes Care, vol. 34, pp. 518 – 523, 2011.
[14] W. H. Organization, "World health organization: definition and diagnosis of diabetes mellitus and intermediate hyperglycemia: report of a WHO/IDF consultation," 2006.
[15] A. D. Association, "2. classification and diagnosis of diabetes: standards of medical care in diabetes—2019," Diabetes care, vol. 42, no. Supplement 1, pp. S13–S28, 2019.
[16] Q. Wang, W. Cao, J. Guo, J. Ren, Y. Cheng et al., "DMP_MI: An effective diabetes mellitus classification algorithm on imbalanced data with missing values," IEEE Access, vol. 7, pp. 102 232–102 238, 2019.
[17] P. Kaur and R. Kaur, "Comparative analysis of classification techniques for diagnosis of diabetes," in Advances in Bioinformatics, Multimedia, and Electronics Circuits and Signals. Springer, Singapore, 2020, pp. 215–221.
[18] R. H. Devi, A. Bai and N. Nagarajan, "A novel hybrid approach for diagnosing diabetes mellitus using farthest first and support vector machine algorithms," Obesity Medicine, vol. 17, p. 100152, 2020.
[19] H. Abbas, L. Alic, M. Erraguntla, J. Ji, M. Abdul-Ghani et al., "Predicting long-term type 2 diabetes with support vector machine using oral glucose tolerance test," bioRxiv, 2019.
[20] M. S. Kadhm, I. W. Ghindawi and D. E. Mhawi, "An accurate diabetes prediction system based on k-means clustering and proposed classification approach," International Journal of Applied Engineering Research, vol. 13, no. 6, pp. 4038-4041, 2018.
[21] S. Afzali and O. Yildiz, "An effective sample preparation method for diabetes prediction," Int. Arab J. Inf. Technol., vol. 15, pp. 968–973, 2018.
[22] P. Tuso, "Prediabetes and lifestyle modification: time to prevent a preventable disease." The Permanente journal, vol. 18, no. 3, pp. 88–93, 2014.
[23] O. Daanouni, B. Cherradi and A. Tmiri, "Type 2 diabetes mellitus prediction model based on machine learning approach," in The Proceedings of the Third International Conference on Smart City Applications. Springer, Cham, 2019, pp. 454–469.
[24] H. Lai, H. Huang, K. Keshavjee, A. Guergachi and X. Gao, "Predictive models for diabetes mellitus using machine learning techniques," BMC endocrine disorders, vol. 19, no. 1, pp. 1–9, 2019.
[25] B. Alic´, L. Gurbeta and A. Badnjevic, "Machine learning techniques for classification of diabetes and cardiovascular diseases," in 2017 6th Mediterranean Conference on Embedded Computing (MECO), Bar, Montenegro, 2017, pp. 1–4.
[26] S. Uddin, A. Khan, M. E. Hossain and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," BMC Medical Informatics and Decision Making, vol. 19, no. 1, pp. 1–16, 2019.
[27] W. Yu, T. Liu, R. Valdez, M. Gwinn and M. J. Khoury, "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes," BMC medical informatics and decision making, vol. 10, no. 1, p. 16, 2010.
[28] J. Semerdjian and S. Frank, "An ensemble classifier for predicting the onset of type ii diabetes," arXiv preprint arXiv:1708.07480, 2017.
[29] A. Dinh, S. Miertschin, A. Young and S. Mohanty, "A data-driven approach to predicting diabetes and cardiovascular disease with machine learning," BMC Medical Informatics and Decision Making, vol. 19, 2019.
[30] A. D. Association, "Standards of medical care in diabetes—2018 abridged for primary care providers," Clinical diabetes: a publication of the American Diabetes Association, vol. 36, no. 1, p. 14, 2018.
[31] H. Rodbard, P. Jellinger, J. Davidson, D. Einhorn, A. Garber et al., "Statement by an American association of clinical endocrinologists/american college of endocrinology consensus panel on type 2 diabetes mellitus: an algorithm for glycemic control," Endocrine practice, vol. 15, no. 6, pp. 540–559, 2009.
[32] J. V. Tu, "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes," Journal of clinical epidemiology, vol. 49, no. 11, pp. 1225–1231, 1996.
[33] R. Caruana and A. Niculescu-Mizil, "Data mining in metric space: an empirical analysis of supervised learning performance criteria," in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, Seattle, USA, 2004, pp. 69–78.
[34] A. Mathur and G. M. Foody, "Multiclass and binary svm classification: Implications for training and classification users," IEEE Geoscience and remote sensing letters, vol. 5, no. 2, pp. 241–245, 2008.
[35] M. Feurer and F. Hutter, "Hyperparameter optimization," in Automated Machine Learning. Springer, Cham, 2019, pp. 3–33.
[36] C. Zhang and Y. Ma, "Random Forests" in Ensemble machine learning: methods and applications. London, UK: Springer Science+Business Media, pp. 157-176, 2012. [Online]. Available: https://link.springer.com/content/pdf/10.1007/978-1-4419-9326-7.pdf
[37] H. Nasri and M. Yazdani, "The relationship between serum LDL-cholesterol, HDL-cholesterol and systolic blood pressure in patients with type 2 diabetes." Kardiologia polska, vol. 64, no. 12, pp. 1364–8, 2006.
[38] J. Gutierrez, A. Alloubani, M. Mari and M. Alzaatreh, "Cardiovascular disease risk factors: Hypertension, diabetes mellitus and obesity among Tabuk citizens in Saudi Arabia," The open cardiovascular medicine journal, vol. 12, p. 41, 2018.
[39] L. Zhou, S. Pan, J. Wang and A. V. Vasilakos, "Machine learning on big data: Opportunities and challenges," Neurocomputing, vol. 237, pp. 350–361, 2017.
[40] H. F. Ahmad, H. Mukhtar, H. Alaqail, M. Seliaman, A. Abdulaziz. Investigating Health-Related Features and Their Impact on the Prediction of Diabetes Using Machine Learning, Applied Sciences, Vol. 11, No. 3. 2021. DOI:10.3390/app11031173.

Hamid Mukhtar is an Associate Professor at Taif University. He has a Ph.D. in computer science from Telecom SudParis, France. His research interests include healthcare management, machine learning and deep learning.

Sana Al Azwari received the BS in computer science from Taif University, Taif, Saudi Arabia, in 2004, and the MS and PhD in information sciences from Strathclyde University, Glasgow, UK, in 2010 and 2016, respectively. She joined the Information Technology Department, Taif University, Taif, Saudi Arabia, as an assistant professor in 2017. At present, she is the Vice Dean of the college of Computer and Information Technology, Taif University. Her current research interests include data science, big data, machine learning, data mining, ontologies, and the semantic Web. Dr. Al Azwari is an ambassador of Women in Data Science committee and is an international science ambassador for Strathclyde University, Glasgow, UK. She awarded the King Abdulaziz and his Companions Foundation for the Gifted Award, Saudi Arabia, in 2006.