Automatic Recognition of Descriptors helping to Cause Diabetes in Algeria

Mohammed El Amine LAZOUNI
Biomedical Engineering Laboratory
University of Tlemcen
Algeria
aminelazouni@gmail.com

Mostafa EL HABIB DAHO
Biomedical Engineering Laboratory
University of Tlemcen
Algeria
mostafa.elhabibdaho@gmail.com

Mahammed MESSAIDI
Biomedical Engineering Laboratory
University of Tlemcen
Algeria
m_messadi@yahoo.fr

Amel FEROUI
Biomedical Engineering Laboratory
University of Tlemcen
Algeria
ebm_amel@yahoo.fr

***Abstract*—The computer aided medical diagnosis systems can use a great number of very important medical data in order to help doctors in detecting different pathologies. We assume that the grater data we have, the more we facilitate and ameliorate the quality of classification. However, the classification quality does not directly depend on the size of the available database but it rather depends on its pertinence. For this, the purpose of this paper is to two different problems. The first one is the selection of the pertinent descriptors that help causing diabetes using a Random Forest feature selection approach. The second is the combination of several different machines learning algorithms (Support Vector Machine (SVM), K-Nearest Neighbor (KNN), the Multilayer Perceptron (MLP) and two Decision tree based classifiers (Classification And Regression Tree (CART), and Random Forests) in order to classify type 2 diabetic patients. We used also a majority voting method between the proposed five classifiers. In our paper, we selected an experimental database composed of 625 patients, each of whom being represented by 31 descriptors. These patients were selected in various private clinics and hospitals in western Algeria.**

Keywords—Diabetes Type2; Feature Selection Method; Database, Support Vector Machine; Random Forest; Classification And Regression Tree; K-Nearest Neighbor; Multilayer Perceptron; Majority Voting System.

I. INTRODUCTION

Diabetes is a chronic disease that cannot be completely cured, but can only be controlled. It represents a group of metabolic disease in which the patients has high blood glucose, either because insulin production is inadequate, or because the body's cells do not respond properly to insulin or both.

It is spreading very fast nowadays. In 2013 it was estimated that over 382 million people through the word had diabetes according to the International Diabetes Federation (IDF). [1]

The insulin is produce by the pancreas. It allows glucose to penetrate into the cells of the body so that it is used as a source of energy. For a non-diabetic person, the insulin plays it role and the cells have the beneficiate from the energy that they need for functioning. When the body lacks insulin or when this latter cannot play its role, as it is the case in diabetes the glucose cannot act as a fuel for the cells, and so it accumulates in the blood and causes hyperglycemia.

Diabetes is classified among the most emergent illnesses and is spreading at very high speed. Its severeness lies in its bat effects non many organs of the body: the eyes, the kidneys, the blood vessels, the brain vessels…etc.

Diabetes is classified into four broad categories: type1, type2, gestational diabetes and other specific types as the pre-diabetes for example.

In this work, our objective is to detect the factors causing type 2 diabetes in Algeria.

In literature, there are many risk factors for type2 diabetes among which the following: [2, 3]

- Persons aged 45 years and more.
- Obesity.
- Heredity.
- Physical inactivity.
- Heart diseases.
- High level of blood Cholesterol and Triglycerides.
- Pregnancy.
- Gestational diabetes or giving birth to a baby weighing more than 4 kg.
- Polycystic ovary syndrome for women.
- Certain ethnic groups (mainly Afro-Asians, Africans and Latin Americans).

Diabetes affects more than 5% of the word population. 90% of the word diabetes are subject to a pre-diabetes [1].

The diagnosis of diabetes type2 consists in classifying the patient according to two situations "diabetic or healthy patient" by analyzing a certain number of parameters which characterizes it. And sight the big number of individuals and the complexity of interpretation of the parameters; it is possible to be in front of a classification problem.

We could save that the more we increase the number of descriptors the more we facilitate and ameliorate the classification. However, the classification quality does not directly depend on the size of the available database but it rather depends on its pertinence.

In the literature, there have been different studies in the field of artificial intelligence where many classification techniques are used in order to assist doctors in different specialties. [4]

In order to discover key descriptors and latent knowledge, data mining techniques were applied. A large number of studies performed on diabetes classification have applied reducing parameters. To improve computational efficiency, the feature selection technique, the combination of five machines learning and a majority voting algorithm are used in this research to rank the important descriptors affecting diabetic control and to classify the diabetics patients.

Although many have tried to approach diabetic's health problems through data mining techniques, there are many constraints and difficulties in this [5]. The main difficulty arises in data collection. Because hospitals have not used electronic medical record systems for long, it is very hard to collect a dataset large enough for such research. Moreover, in the university hospital setting, physicians and medical practitioners have occasionally transferred to other medical institutions, which might cause different patient care protocols, including physical examinations and interviews formats. Furthermore, because of over confidence in their health condition or for other personal reasons, some outpatients visit the hospital and private clinics irregularly.

These circumstances may lead to very irregular, incomplete, or missing data in the clinical setting, and thus make it difficult to extract meaningful information from the data.

In this paper, we focus on the objective of a variable feature selection, while maintaining the medial interpretation and a good classification.

The remaining parts of this paper are organized as follows: in section 2 we present the different work done for the recognition of type 2 diabetes in the literature. Section 3, the description of the proposed algorithm used in this study. In section 4, we describe the collected database and we discuss its different parameters. The results obtained in applications are given in Section 5. This section also includes the discussion of these results in specific and general manner.

Finally in Section 6, we conclude the paper with summarization of results by emphasizing the importance of this study.

II. State of the Art

Diabetes is a disease that can lead to the development of serious complications and early death. [1]

For this reason, many researchers are interested to study this disease and to detect the descriptors that can cause this latter.

In this section, we will first present some work of researchers whose goal was to detect the disease of diabetes in a general way using different machines algorithm. After we will present the different works carried out on diabetes using variable selection techniques.

In Ref. [6], Michael Klompas et al. have created a new surveillance algorithm able to detect the type 1 diabetic patients versus the type 2 using structured Electronic Health Record (EHR) data. The collected database contains the data of 700,000 patients. Although the database used in this work is gigantic, the disadvantage of this work is that it does not use selection and classification techniques.

In Ref. [7], a feature selection algorithm is run for the selection process. Secondly, the deep learning model has a deep neural network which employs a Restricted Boltzmann Machine (RBM) on Pima database. This algorithm will help to predict diabetes with much more precision.

Maham et al [8] have proposed an application of Automatic Multilayer Perceptron which is combined with an outlier detection method Enhanced Class Outlier Detection using distance based algorithm to create a prediction framework named as Enhanced Class Outlier with Automatic Multi layer Perceptron (ECO-AMLP). The used database is PIMA dataset and the obtained accuracy rate are 88.7%

A.P.Shingade et al proposed in [9] A Review on Implementation of Algorithms for Detection of Diabetic Retinopathy (DR). The goal of this work was to provide assistance to doctors in order to detect the diabetic retinopathy using three machines algorithm which are the Support Vector Machine, the Multi-Layer Perceptron and the K-Nearest Neighbor. The obtained results by all machines learning approaches are very satisfactory. [9]

In the same field, Application of Random Forests Methods to Diabetic Retinopathy Classification Analyses has been proposed by Ramon Casanova et al. in Ref [10]. We have introduced Random Forest algorithm. Their results suggest that RF methods could be a valuable tool to diagnose DR diagnosis and evaluate its progression.

In the same context, Subramani Mani et al. proposed in [11] Type 2 Diabetes Risk Forecasting from EMR Data using Machine Learning. The use of different selection and classification techniques of data in order to detect type 2 diabetes, which are: the RF, the SVM, the K-NN, the CART, the Gaussian Naïve Bayes and the Logistic Regression. In this study, authors have collected a dataset of 2280 patients each of them is represented by 17 descriptors. The results obtained by the RF, and the CARTS approaches are the best.

In [12], the authors have applied the Random Forest algorithm on 10 different databases. Among them, PIMA diabetes database. An error rate of 0.2387 was obtained for this latter.

In this paper, we target three distinct objectives: the database construction, Recognition the best descriptors helping to cause diabetes in Algeria with Random Forest Approach, and finally data classification using five machines learning algorithms.

For the data used in our experiments, we propose to use a large clinical database composed of 625 patients, which have been collected manually form several hospitals and private clinics.

III. The Proposed Algorithm

A general schema for predictive model building able to select the best descriptors and classify in the same time the patients who have diabetes diseases is presented in Figure 1 below.

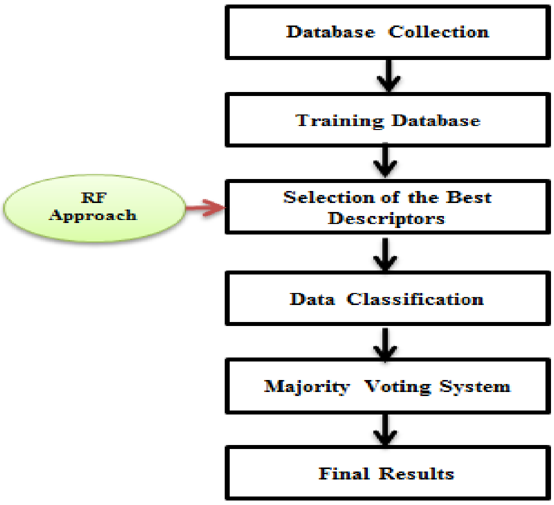

Figure 1: A General Schema for the proposed approaches

After the realization of the learning step in our database with 10 cross validation technique. The steps of the selection of the pertinent descriptors that help causing diabetes are realized using a Random Forest feature selection approach. In the state of the art, the most used Feature Selection method by different researchers is the Random Forests algorithm. The description of this latter is illustrated in the next section. After that, a combination of classifiers is made for the purpose of detecting whether the patient is diabetic or not. Finally, a majority voting method is used between the proposed classifiers (SVM, MLP, K-NN, CART, and RF) in order to get closer to the maximum of the medical advice.

A. *The Random Forest Approach:*

A Random Forest is an ensemble method introduced by Breiman in 2001 [13]. The idea of ensemble methods is to combine several classifiers to build the best models.

Breiman proposed to use the Bagging, but for each data set generated, the growth of the tree is processed with a random selection among the variables at each node. He uses the approaches developed by L. Breiman, [14] and Amit and Geman [15] to generate a set of trees doubly disrupted using a randomization operating both on the training sample and at internal partitions. Each tree is thus generated at first from a sub sample (a bootstrap sample) of the complete training set, similar to the techniques of bagging.

Then, the tree is constructed using the Classification And Regression Tree (CART) methodology with the difference that at each node the selection of the best split based on the Gini index is performed not on the complete set of attributes M but on a randomly selected subset of it.

The size F of this subset is established prior to the execution of the procedure ($1\leq F\leq M$) [16].

The tree is then developed to its maximum size without pruning. During the prediction phase, the individual to be classified is propagated in every tree of the forest and labeled according to the CART rules. The whole forest prediction is provided by a simple majority voting approach of the class assignments of individual trees.

The algorithm of the Random Forest method (for classification and prediction) is as follows:

1. Draw ntree bootstrap samples from the original data.
2. For each of the bootstrap samples, grow an unpruned classification or regression tree, with the following modification: at each node, rather than choosing the best split among all predictors, randomly sample mtry of the predictors and choose the best split from among those variables. (Bagging can be thought of as the special case of random forests obtained when mtry = p, the number of predictors.)
3. Predict new data by aggregating the predictions of the ntree trees (i.e., majority votes for classification, average for regression). [17]

An estimate of the error rate can be obtained, based on the training data, by the following:

1. At each bootstrap iteration, predict the data not in the bootstrap sample (what Breiman calls “out-of-bag”, or OOB, data) using the tree grown with the bootstrap sample.
2. Aggregate the OOB predictions. (On the average, each data point would be out-of-bag around 36% of the times, so aggregate these predictions.) Calcuate the error rate, and call it the OOB estimate of error rate. [17]

After selected the dexcriptor’s importance with an expected degree of importance by The Random Forest approach, we performed the step of classification using five machines learning algorithm and a majority voting system.

B. *The Machine Learning Algorithm:*

In this study, we used five different machines learning algorithms and a majority voting system. These algorithms are: Two sample-based classifier (K-nearest neighbor, and Multi Layer Perceptron), one kernel based classifier (Support Vector Machine) and two decision tree based classifiers (CART and Random Forests).

We have chosen to use five different machine learning approaches because in this work we have used a new database.

The choice of these five machine learning approaches is justified by:

The Neural networks, is known for its power of learning and their remarkable ability to derive meaning from complicated or imprecise data, can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. Also, it presents better generalization ability, especially on noisy data. But, the disadvantage of this approach is that it is considered as a black box, i.e. it does not give interpretability for doctors.

The K-Nearest Neighbor (KNN) is a very simple classifier that works well on basic recognition problems. It is very easy to implement its algorithm.

The CART algorithm is useful tools in decision tree methodology. It is well-known by providing powerful classification, in addition to a set of interpretable rules because it gives a pictorial view of the results obtained.
There are several advantages of CART. One of them is that trees are more efficient at dealing with high dimension data than parametric regression techniques. Additionally, it is able to flexibly deal with missing data. [18]

The Support Vector Machine (SVM), is known for its good management in multi-class cases, High accuracy, its minimization of the empirical error, treats the big data like the medical field, and with an appropriate kernel they can work well even if you're data isn't linearly separable in the base feature space.
Despite its many advantages, the SVM algorithm remains a black box that does not offer interpretability for doctors.

Among the advantages of the Random Forest approach is: a good classification rate, reduction of the computing time: very important in large dimension (big data) and runs efficiently in this latter, estimate of what descriptors are pertinent in the classification, provides effective methods for estimating missing data, generated forests can be saved for future use on other data [19], Obtaining a greater variety of models. [12]

In this work, we suggest to use a Majority Voting method (MV) among the proposed classifiers in order to optimize the classification results.
The majority voting systems are widely used in computer aided diagnosis in order to minimize errors. The aim of majority voting algorithms is to determine in any given sequence of votes whether there is a class with more votes than all the others, and if so, to determine this class as the most favorable one.

IV. THE COLLECTED DATABASE

The patients' in this database was collected from different private clinics and hospitals from the western of Algeria from year 2017 to 2018 as part of data clinical management. .
The existing database contains 625 patients (227 are healthy and 398 Type 2 diabetic patients'); 408 males and 217 females. The patients' age ranged from six month to 87 years, with only 93 patients younger than 21. The mean value of the patients' age is 58.5. Therefore, the work concentrates on the older group. This is consistent with the fact that diabetes mainly affects the older age groups (>60) within the population.
Representation by the geographical histogram of the Algeria city's is illustrated in the following figure2. It shows that patients collected in our database are from different regions (west, central, south and north) which prove the robustness of our data base.

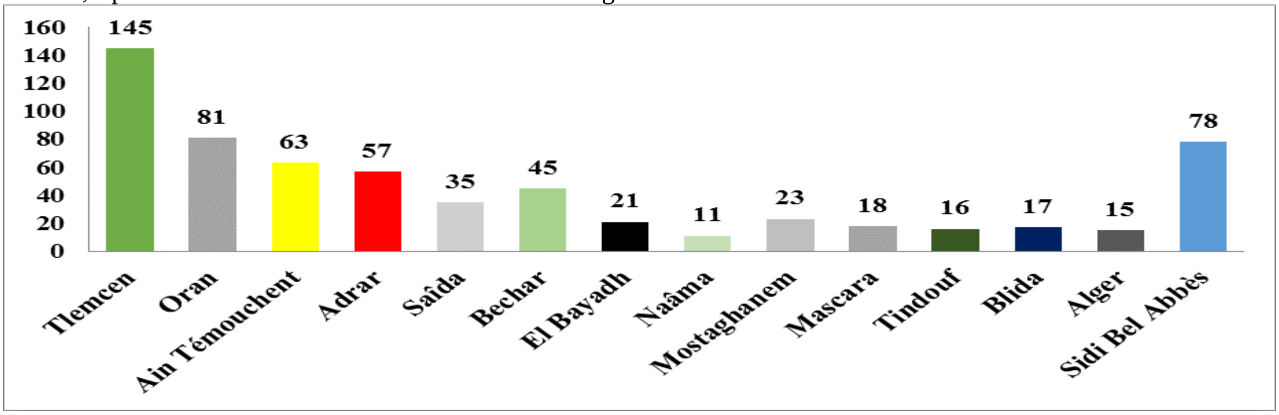

Figure2. The collected patients as Algerian cities

A set of 33 descriptors including patient characteristics, treatment, complication care, physical and laboratory findings were present in the database.

the descriptors used in our database in order to detect if the patient is diabetic or not are: sex, age, medical backgrounds: hypertension, respiratory failure, heart failure, ecg parameters: heart rate (bpm), steadiness of heart rate, pacemaker, atrioventricular block, left ventricular hypertrophy, myocardial infarction, level of blood cholesterol, level of blood triglycerides, prothrombin ratio, glycated hemoglobin, blood sugar (g/l), number of birth, polycystic ovary syndrome, family history, history of gestational diabetes, skin color, body mass index, number of hours of physical activity, number of cigarettes per day, alcohol consumption, type of drug to take, cerebrovascular accident, asa (american society of anesthesiologists) scores , oxygen saturation (%), blood pressure: systole and diastole (mmHG) and resulting classes.

There was some difficulty in data collection: the patients have not always taken all the tests at the visit, and their visits were very irregular. It was inevitable that there will be incomplete, noisy and inconsistent values.

Therefore some process was necessary; the principal transformation deals with discretization which allows the application of data mining methods for discrete attribute values.

V. RESULTS AND DISCUSSIONS

In practice, it is very useful to have information from the used the data variables. We should choose the needed variables to explain the output results.

This extracted information can be of great help in interpreting the data. They can also be used to build better classifiers: a classifier built using only the useful variables can be more powerful than a classifier built with additional noisy variables.

In this paper we focus our work on random bits for variables selection task.

It’s very simple to implement, when selecting a subset of explanatory variables (of importance) from a large number of them.

The implementation of Random Forest uses two main parameters: The most important parameter is the number m of variables randomly selected at each node of the tree. We called it mtry and it can vary from 1 to “p” (observations) here we fixed it equal to $\sqrt{(p-1)}$ depending on the setting of Breiman [20].

We can also adjust the number of trees of the forest. This parameter is named Ntrees and its final value is taken by building trees until the error research a small fixed value.

The results as can been see in Figure 3, show that at 250 trees the error rate converge to 0.11 which give an 89 % classification rate.

And beyond this value the error remains nearly constant.

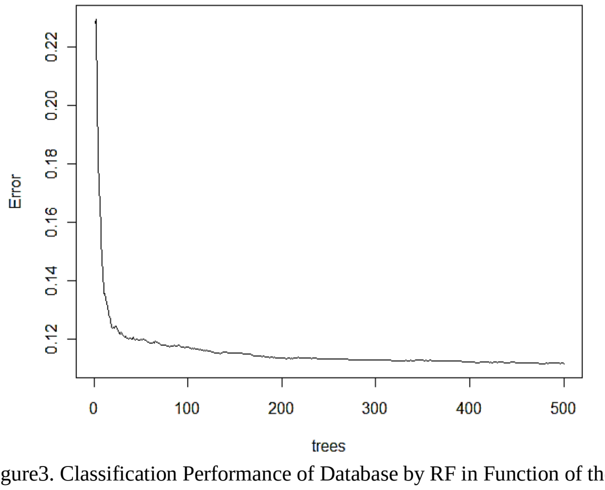

Figure3. Classification Performance of Database by RF in Function of the Number of Trees

After calculating the variables importance for our database the results obtained are presented in the following table with a descending order as follow:

TABLE 2. THE DEGREE OF IMPORTANCE FOR THE DIFFERENT DESCRIPTORS OBTAINED BY THE RF METHOD

Descriptors	*Degree of importance*	*Descriptors*	*Degree of importance*
Glycated hemoglobin	0.86	Myocardial Infarction (MI)	0.41
Blood sugar	0.8	Skin color	0.41
Family history	0.79	Left Ventricular Hypertrophy	0.37
Body Mass Index (BMI)	0.74	Atrioventricular Block	0.33
Age	0.72	Oxygen saturation (%)	0.31
Level of blood Triglycerides	0.71	Steadiness of Heart Rate	0.31
Level of blood Cholesterol	0.69	Respiratory failure	0.30
Hypertension	0.65	Number of birth	0.21
Heart failure (HF)	0.61	Prothrombin Ratio	0.19
Systole (mmHg)	0.61	Pace maker	0.17
Diastole (mmHg)	0.6	Cerebrovascular Accident (CVA)	0.13
Physical Activity	0.55	Sex	0.05
History of gestational diabetes	0.49	Type of drug to take	0.05
ASA Scores	0.47	Alcohol consumption	0.02
Heart rate value	0.46	Polycystic ovary syndrome	0
Number of cigarettes per day	0.43		

The obtained results concerning the selection of variables according to their importance using the random forests, confirm the medical approach.

They show clearly the importance of Glycated Hemoglobin and the blood sugar value in the identification of diabetes type 2.

From the medical point of view, these two values are the most used by doctors in order to determine either the patient is diabetic or not.

Most of the time, for a diabetic patient with an overweight, the value of the level of blood Cholesterol and triglycerides are not provides balanced.

Also the descriptors Age shows the fact that the majority of the Algerian population are elderly people. And it confirms with the values given by the International Diabetes Federation (IDF), “The global prevalence of diabetes among people aged 60 to 79 years is 18.6%”. [21]

The family history descriptor's is a very important parameter because diabetes is a hereditary disease that can be transmitted from one generation to another.

Most of the time, a diabetic patient may suffer from heart disease. So, the descriptors history of hypertension, blood pressure (systolic and diastolic) and heart failure are highly correlated.

The aim of this research concentrates on two principal points:

1. Identifying significant factors influencing diabetes type2 control;

2. Classifying patients after selecting the best descriptors influencing on diabetes.

.

The following table (table 3) summarizes the obtained results in the five experiments. In each experiment we set a threshold level of Degree of importance. In the first experiment, we propose a threshold ≥ 0.01, our algorithm used 31 descriptors. In the second, we propose a threshold ≥ 0.3, our algorithm used 23 descriptors. In the third experiment, we propose a threshold ≥ 0.6, our algorithm used 11 descriptors. In the fourth experiment, we propose a threshold ≥ 0.7, our algorithm used 06 descriptors. Finally in the last experiment, we propose a threshold ≥ 0.8, our algorithm used just 02 descriptors

TABLE 3. THE OBTAINED RESULTS USING THE MACHINES LEARNING ALGORITHM AND THE MAJORITY VOTING SYSTEM

Experiment 1	**The Obtained Results Using a threshold ≥ 0.01**					
(31 descriptors)	***SVM***	***MLP***	***K-NN***	***CART***	***RF***	***MV***
Classification Rate	80.77	79.83	77.72	83.07	83.94	85.33
AUC	0,661	0,544	0,508	0,699	0.693	
Experiment 2	**The Obtained Results Using a threshold ≥ 0.3**					
(23 descriptors)	***SVM***	***MLP***	***K-NN***	***CART***	***RF***	***MV***
Classification Rate	87.01	81.27	80.59	83.05	88.12	92.94
AUC	0,679	0,688	0,594	0,707	0.799	
Experiment 3	**The Obtained Results Using a threshold ≥ 0.6**					
(11 descriptors)	***SVM***	***MLP***	***K-NN***	***CART***	***RF***	***MV***
Classification Rate	88.65	81.89	82.06	90.38	89.91	94.24
AUC	0.759	0.804	0.663	0,812	0.877	
Experiment 4	**The Obtained Results Using a threshold ≥ 0.7**					
(6 descriptors)	***SVM***	***MLP***	***K-NN***	***CART***	***RF***	***MV***
Classification Rate	87.99	81.54	81.12	86.81	89.69	93.18
AUC	0.711	0,707	0,613	0.732	0.865	
Experiment 5	**The Obtained Results Using a threshold ≥ 0.8**					
(2 descriptors)	***SVM***	***MLP***	***K-NN***	***CART***	***RF***	***MV***
Classification Rate	86.61	80.11	77.87	84.02	85.32	88.55
AUC	0,694	0,627	0,588	0,703	0.749	

The performances of the five experiments were evaluated on the basis of Classification rate criterion and the Area Under the Curve (AUC).

As shown in Table 3, the third experiment provided the better performances among all the other experiments.

While the obtained results in the first one are the lowest performances among all the other experiments. The results for the other experiments are ranked in ascending order as follow: experiment 4, experiment 2 and experiment 5.

The AUC value of the all experiments was reasonable (~0.7). The CART classifier had the best AUC in the first experiment. For all the other experiments, the Random forests classifier had the best AUC.

In case of using all database (625 patients), the results are presented in ascending order as follow: RF, CART, SVM, MLP and finally K-NN.

However, when using only the half of our database (300) in the same experiment (experiment 3), we obtain a different order of results (SVM, RF, CART, K-NN, and finally MLP). Therefore, we can conclude the results depend mainly from the size of the database.

Since the size of our database may be increased in the future, we propose to apply all the machine learning algorithms and the majority voting systems in order to select the best related result.

Also, we notice that, for all the experiments, the five machines algorithm used on our database have given a good classification rate but the majority voting approach outperforms all of them.

For the all test, the best results are those given by the Support Vector Machine, the Multilayer Perceptron, the Random Forest and the CART Decision tree classifier. The classification rate of K-Nearest Neighbor is the lowest among the four machines learning techniques.

Our database does not contain the cases of women who have Polycystic ovary syndrome. For this reason, the RF technique has found importance degree of this descriptor is null.

The degree of importance of the Alcohol consumption descriptor equal to 0.02. But in the medical literature, that descriptor affects too much on the diabetic type 2 disease. This result is justified by the number of drinkers alcohol patients collected in our database that tend towards zero.

If we compare the obtained results during the third and the fourth experiments, we note that the results found in the third were better than those obtained on the fourth experiment. Recall that the number of descriptors used in the third experiment is 11 against only 06 descriptors in the 4th. This means that the 05 descriptors which are not used are also very important. These descriptors are: Level of Blood Cholesterol, Medical backround of Hypertension and Heart failure (HF), and diastole.

The results of the fifth experiment show that the most relevant descriptors to determine whether the patients have diabetes type 2 diseases or not are the value of the Blood sugar and glycated hemoglobin. However, according to the medical literature, these two parameters are the most important.

VI. Conclusion

The development of the medical computer aided diagnosis system is becoming today a very motivating research field. Indeed, numerous researchers working in the field of artificial intelligence are trying to suggest interpretable intelligent automatic systems ready to help doctors in their routine clinical work.

The two objectives of this paper are: to detect the best descriptors helping to Cause Diabetes type2 in Algeria and to classify the type 2 diabetic patients.

For this, we have used the Random Forest feature selection approach in order to select the pertinent descriptors.

The advantage of this technique is to depend only on a very limited number of parameters to be executed, which makes their exploration easier. The two main features of the method are its ability to select important variables and the number of trees forming the whole forest prediction.

The obtained results show that, the classification quality does not directly depend on the size of the available database but it rather depends on its pertinence.

eart failure (HF), systole and diastole) and highlight the 02 more informative which are (Glycated hemoglobin, Blood sugar) and beyond totally ignore the other noisy variables.

By applying the Random Forest feature selection approach, we can conclude that four descriptors can affect the Algerian population for the diabetes type 2 diseases, with a good Glycated hemoglobin and good blood sugar value, most diabetes complications can be avoided.

Beyond these two factors: the strict control of the blood pressure (systole and diastole) especially for people over 45 years, the control of the weight and the practice of physical exercise may preventable diabetes type2 disease.

In our future works, we plan to enrich our database while augmenting the number of patients whose number is reduced in our database, for example the women with the Polycystic ovary syndrome for the purpose of obtaining better results. And make it available online for a general use by other researchers.

We plan also to use some other feature selection approach and other machine learning techniques, in order to test the robustness of our database.

VII. References

[1] The International Diabetes Federation (IDF) http://www.idf.org/about-diabetesAccessed 2013 January 10.

[2] Risérus U, Willett WC. "Dietary fats and prevention of type 2 diabetes". Progress in Lipid Research . 2009. 48 (1): 44–51.

[3] Touma, C; Pannain, S. "Does lack of sleep cause diabetes?". Cleveland Clinic journal of medicine. 2011. 78 (8): 549–558.

[4] M. A Lazouni, M. A Chikh, and S Mahmoudi. “A New Computer Aided Diagnosis System for Pre-Anesthesia Consultation”. Journal of Medical Imaging and Health Informatics. Vol. 3, pp 1–9, 2013.

[5] Vosoulipour A, Teshnehlab M and Moghadam H.A . “Classification one Diabetes Mellitus Date-set Based one Artificial Neural Networks and ANFIS” Biomed 2008, Proceedings 21, pp. 27-30.

[6] Michael Klompas, Emma Eggleston, Jason McVetta, Ross Lazarus, Lingling Li, and Richard Platt. Automated Detection and Classification

[7] Sushant Ramesh, H. Balaji, N.Ch.S.N Iyengar and Ronnie D. Caytiles,” Optimal Predictive analytics of Pima Diabetics using Deep Learning”, International Journal of Database Theory and Application Vol.10, No.9 (2017), pp.47-62

[8] Maham Jahangir, Hammad Afzal, Mehreen Ahmed, Khawar Khurshid, Raheel Nawaz, “ECO-AMLP: A Decision Support System using an Enhanced Class Outlier with Automatic Multilayer Perceptron for Diabetes Prediction”. 2017. Machine Learning.

[8] Temurtas H, Yumusak N, and F. Temurtas: A comparative study on diabetes disease diagnosis using neural networks. Expert systems with applications, volume 36 Issue. pp 8610- 8615. May 2009.

[9] A.P.Shingade, A.R.Kasetwar. A Review On Implementation Of Algorithms For Detection Of Diabetic Retinopathy. International Journal of Research in Engineering and Technology Volume: 03 Issue: 03. Mar-2014.

[10] Ramon Casanova, Santiago Saldana, Emily Y. Chew , Ronald P. Danis, Craig M. Greven, Walter T. Ambrosius. Application of Random Forests Methods to Diabetic Retinopathy Classification Analyses. PLOS ONE Journal. pp 1—8. Vol 9(6). 2014.

[11] Subramani Mani, Yukun Chen, Tom Elasy, Warren Clayton, and Joshua Denny. Type 2 Diabetes Risk Forecasting from EMR Data using Machine Learning. AMIA Annu Symp. 2012. Pp 606–615.

[12] M. EL HABIB DAHO and M A. CHIKH. “Classification and Recognition of Biomedical Data with Ensemble Methods ». Abou Bekr Belkaid Tlemcen University. Doctoral Thesis. 2015.

[13] Breiman, “Random forests," Machine Learning, vol. 45, no. 1, pp. 5—32, 2001.

[14] L. Breiman, “Bagging predictors," Machine Learning, vol. 24, no. 2, pp.123--140, 1996.

[15] Y. Amit and D. Geman, “Shape quantization and recognition with randomized trees," Neural Computation, vol. 9, no. 7, pp. 1545--1588, 1997.

[16] N. Sirikulviriya and S. Sinthupinyo, “Integration of rules from a random forest." in International Conference on Information and Electronics Engineering IPCSIT vol.6 (2011) IACSIT Press, Singapore, 2011.

[17] A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

20] Leonard Gordon. Using Classification and Regression Trees (CART) in SAS® Enterprise MinerTM For Applications in Public Health. Data Mining and Text Analytics. 2013

[18] M. Walker. “Random Forests Algorith”. September 2013. Data Science Central.

[19] http://www.therapeutique-dermatologique.org/spip.php?article1234

[20] L. Breiman, “Random forests,” Machine Learning, vol. 45, pp. 5–32, 2001.

[21] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. (references)

[22] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.