EAI Endorsed Transactions
on Pervasive Health and Technology

Research Article **EAI.EU**

Predicting Diabetes Mellitus and Analysing Risk-Factors Correlation

Md. Faisal Faruque[1,*], Asaduzzaman[1], Syed Md. Minhaz Hossain[1,2], Md. Hasan Furhad[3] and Iqbal H. Sarker[1]

[1]Department of Computer Science and Engineering, Chittagong University of Engineering and Technology, Bangladesh
[2]Department of Computer Science and Engineering, Premier University, Chittagong, Bangladesh
[3]Canberra Institute of Technology, Reid, ACT, Australia

Abstract

INTRODUCTION: Diabetes mellitus is a common disease of the human body caused by a group of metabolic disorders where the sugar levels exceed a prolonged period, and that is very high than the usual time. It not only affects different organs of the human body but also harms a large number of the body system, in particular the blood veins and nerves.
OBJECTIVES: Early predictions of this phenomenon can help us to control the disease and also to save human life. For achieving the goal, this research work mainly explores various risk factors such as kidney complications, blood pressure, hearing loss, and skin complications related to this disease using machine learning techniques and make a decision.
METHODS: Machine learning techniques provide an efficient result to extract knowledge by constructing predicting models from diagnostic medical datasets collected from 200 diabetic patients from the Medical Centre Chittagong, Bangladesh using 16 attributes. Obtaining knowledge from such data can be useful to predict diabetes. In this work, we perform four popular machine learning algorithms, such as Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbour (KNN) and C4.5 Decision Tree (DT), on adult population dataset to predict Diabetes Mellitus.
RESULTS: C4.5 Decision Tree performs better than other algorithms for predicting diabetes with 73.5% accuracy, 72% F-measure, and 0.69 of AUC (area under ROC curve). Besides, we determine the correlation between different risk factors of Diabetes Mellitus. The highest correlation is 0.81 for blood pressure (Hypertension) complications with diabetes.
CONCLUSION: In this study, both positive and negative correlation has been established between the various risk factors and diabetes. There is a positive correlation for predicting kidney complications (Nephropathy) and blood pressure (Hypertension) complications and a negative correlation at predicting hearing loss and skin complications (diabetes dermopathy) from diabetic patients. It helps a patient to be aware of the risk factors related to diabetes.

Keywords: health informatics, machine learning, diabetes, classification, e-health services.

Received on 09 August 2019, accepted on 03 November 2019, published on 13 November 2019

doi: 10.4108/eai.13-7-2018.164173

1. Introduction

Diabetes mellitus, generally known to the people as diabetes, is a disease that immensely affects the hormone insulin and increases levels of sugar in the blood and also causes abnormality metabolism of carbohydrates. This high blood sugar has an impact on various organs of the human body and sometimes creates complications in many bodywork functionalities, in particular the blood veins and nerves. The causes of diabetes are not yet wholly discovered, and many researchers believed that both genetic elements and environmental factors are involved there. Most of the cases, diabetes occurs mainly in grown-ups, and that's why it called ‘adult-onset’ diabetes. Notably, diabetes mellitus is primarily involved with the aging process.

As indicated by the Canadian Diabetes Association (CDA), somewhere in the year from 2010 to 2020, the number of people figure out to have diabetes in Canada is

[*]Corresponding Email. faisal_uits@yahoo.com, iqbal.sarker.cse@gmail.com

from 2.5 million to around 3.7 million [1]. The current situation of the world is not different from this situation. As indicated by the International Diabetes Federation in 2013, the number of people having diabetes mellitus is 382 million [2], which are 6.6% of the total grown-up population in the world. According to the statistics of world healthcare medical data, it expects that patient of diabetic disease increases from 376 billion to 490 billion within the year 2030 [3]. Moreover, the diabetic is a conceivably independent contributing risk factor to micro-vascular entanglements.

Diabetic patients are probably more victimized against a hoisted risk of micro-vascular damage. Long-term complication effects of cardiovascular disease are the leading cause of death. This micro-vascular harm and hasty cardiovascular disease eventually prompt to retinopathy, nephropathy, and neuropathy [4].

One reason for chronic kidney disease is Diabetes Mellitus, which is evaluated by high blood glucose levels of sugar. These levels of sugar in the blood vessels mischief a large number of tiny filtering units in the kidney. This fact, in the long run, prompts to kidney failure. Around 20 to 30 percent of people with diabetes create kidney disease (diabetic nephropathy), although not all of these advances to kidney failure. A person with diabetes has more chance to attack by nephropathy, whether they use insulin or not.

There is no remedy for diabetic nephropathy, and therefore, the treatment is life-long. People with a disease like diabetes are also at risk of various kidney issues, including narrowing of the arteries to the kidneys. It is called renal artery stenosis or renovascular disease.

Many people with diabetes also have hypertension or high blood pressure. In a 2013 review, the American Diabetes Association (ADA) found that the combination of hypertension and diabetes mellitus is deadly and can substantially raise the risk of having a heart attack or stroke. A person with diabetes must make assured that his blood pressure is well controlled.

Diabetes increases the risk of developing a sudden hearing loss. It creates the impression that the patient having severe diabetes is more susceptible to hearing loss. A few studies found that high blood sugar levels can damage the tiny blood vessels in the inner ear, which affects sound reception and make it harder to hear. It restricts people of a specific age group or those living in noisy environments.

Diabetes can affect most of the parts of the body, including the skin. Fortunately, most skin conditions can be prevented or effectively treated whenever they got early. A portion of these problems are skin conditions anyone can have, but people with diabetes get more easily. These incorporate bacterial infections, fungal infections, and itching. Other skin issues happen mostly, or only the people with diabetes—these incorporate diabetes dermopathy.

Many researchers worked on the early prediction of diabetes by taking into account various risk factors related to this disease. For our analysis, we collect diagnostic datasets having 16 attributes diabetic of 200 patients such as age, diet, hypertension, the problem in vision, genetic, and so on. In the later part, we discuss these attributes with their corresponding values. Based on these attributes, we build a prediction model using various machine learning techniques to predict diabetes mellitus.

According to International Diabetes Federation (IDF) published by Atlas in 2017 [54], there are around 424.9 million diabetes patients in the world having age from 20-79 years, of whom 95% suffer from Type 2 Diabetes Mellitus (T2DM) which shown in Table 1.

Table 1. Top 5 countries for the number of people with diabetes (18-99 years), 2017

1. India	74,047, 266
2. Bangladesh	7,349,526
3. Sri Lanka	1,248,310
4. Nepal	679,207
5. Mauritius	236,795

It predicts that the number increases to 628.6 million by 2045 [55]. According to the World Health Organization (WHO) report published in 2016, the statistics of people in Bangladesh dying from this disease shown in Table 2.

Table 2. Number of people dying for diabetes disease, 2016 (WHO)

Total population: 161 million Income Group: Lower Middle

Mortality		
Number of diabetes deaths		
	Males	Females
ages 30-69	20,000	10,600
ages 70+	19,800	14,200

Over the years, there are several Machine Learning based models that deal with health-based data and context-aware based personal data [56, 58]. However, most of these models only predict the probability of a person having diabetes. Diabetes Mellitus can induce other complications like Nephropathy, Cardiovascular disease, Retinopathy, and Diabetic Foot disease. In 2017, 4 million people died all around the world due to diabetes-related risk factors, mostly because they were not monitored closely and warned. There is a scope to introduce a complete system that can correctly predict the onset of risk factors caused by T2DM using machine learning techniques, which can save thousands, if not millions, of lives around the world.

Machine learning techniques provide an efficient result to extract knowledge by constructing predicting models from diagnostic medical datasets collected from diabetic patients. Extracting knowledge from such historical data can be useful to predict diabetic patients and make correct decisions in various situations. Various machine learning techniques can predict diabetes mellitus. However, it is complicated to choose the best technique to predict based on such attributes. Thus for the study, we employ four popular machine learning algorithms such as Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbour (KNN), and C4.5 decision tree, on adult population data to predict Diabetic Mellitus.

A correlation measures the relationship between two or more variables. To evaluate the correlation between different risk factors of Diabetes Mellitus, we use statistical correlation [5] on the collected dataset attributes from chronic kidney and blood pressure disease over the diabetic disease. We have also used the CORREL function or the Analysis Toolpak add-in in Excel [6] to find the correlation.

The contributions in this study are as below:

- We collect real diagnostic datasets having various risk factors of 200 diabetes patients from the medical center.
- We make a performance comparison of different machine learning techniques and evaluate the prediction results.
- We make a correlation between different risk factors of diabetes mellitus.

In section II, we mention the related works. Our methodology is described in Section III. In Section IV, we analyze our experimental results. Finally, we conclude in Section V.

2. Related Work

Numerous works have been conducted in the area of diabetes by using machine learning techniques to extract knowledge from available medical data.

The author Sparacino et al. [7] examined the possibility of predicting future glucose levels using different machine learning techniques. To this purpose, data from 28 Type 1 diabetic volunteers were collected in 48 hours time duration and analyzed. They collected through a continuous glucose checking device with sampling intervals of 3 minutes.

Baghdadi and Nasrabadi [8] have utilized a Radial Basis Function (RBF) neural network and considered 4 intervals during the day (morning, afternoon, evening, and night) and for everyone, they made a separated neural network to predict the glucose level. In this technique, they accomplished a Root Mean Square Error (RMSE) of 0.5202 mg/dl, 0.6804 mg/dl, 0.4392 mg/dl, and 0.1134 mg/dl for the morning, afternoon, evening and night neural network, respectively.

The authors Marling et al. [9] used Support Vector Regression (SVR) algorithm by creating a model that is capable of predicting the blood glucose level of Type 1 diabetic patients. They selected a pivot point concerning one month into the patient's examination to divide data for training and testing purposes.

The author E. Georga in [10] used the SVR algorithm to come up with a model capable of accurately predicting the body covering glucose level. Information concerning the level of plasma insulin, the rate of appearance of plasma glucose level after a meal, current and past just as certain factors associated with exercise, used in such works.

ALjumah et al. [11] developed a predictive analysis model using the support vector machine algorithm. In [12], Kavakiotis et al. used 10 fold cross-validation and three different algorithms, including Logistic Regression, Naive Bayes, and SVM. From all, SVM provided better performance and accuracy of 84 % than other algorithms. In [13], Zheng et al. applied Random Forest, KNN, SVM, Naive Bayes, Decision Tree, and Logistic Regression to predict diabetes mellitus at an early stage, where filtering criteria can be improved. Swarupa et al. [14] applied KNN, J48, ANN, ZeroR, and NB on various diabetes dataset where Naive Bayes outperform better accuracy of 77.01% than others. Pradeep et al. [15] applied KNN, J48, SVM, and Random Forest, where the J48 machine learning algorithm provides comparatively better performance and accuracy than others before the pre-processing technique. The classification algorithms did not evaluate using the cross-validation method. Huang et al. [16] discussed three data mining techniques, including IB1, Naive Bayes, and C4.5, to predict and control diabetes mellitus. It worked on a dataset gathered from Ulster Community and Hospitals Trust (UCHT) in the year from2000 to 2004. By applying the feature selection technique, the performance of IB1 and Naive Bayes provided a better result. In [17], Xue-Hui Meng et al. used three different data mining techniques ANN, Logistic regression, and J48 to predict diabetes using real-world data sets by collecting information by the distributed questioner. Finally, it concluded that J48 machine learning techniques provide efficient and better accuracy than others. Thirumal et al. [18] discussed four popular data mining techniques, including SVM, KNN, Naive Bayes, and C4.5. In this study, C4.5 provided better results than other techniques with an accuracy of 78.2552%. Al Jarullah A. [19] utilized the C4.5 data mining algorithm on the Pima Indians Diabetes data set [20]. He achieved an overall accuracy of 78%. Various algorithms were explained by Komi et al. [21] using different parameters such as Glucose, Blood Pressure (BP), Skin Thickness (ST), insulin, Body max index (BMI), Diabetes Pedigree function(DPF) and age where rather than choosing all parameters, only small sample datasets are used. ANN, EM, GMM, Logistic Regression, and SVM were applied to the diabetes dataset. ANN (Artificial Neural Network) shows better accuracy and performance than other algorithms. WeifengXu et al. [22] applied to different machine learning algorithm in the prediction of diabetes. From those algorithms, Random Forest (RF) provided better results than other data mining techniques. Yunsheng et al. [23] used KNN and DISKR in the prediction of diabetes, where storage space was reduced, and an instance had fewer factors eliminated.

By evacuating the exception, both performance and accuracy are increased. Sajida et al. [24] proposed the Adaboost model, which provided better performance and accuracy. Dr. Naveen & Pradeep et al. [25] applied KNN, SVM, J48, and Random Forest over 215 instances with 7 different attributes, where the J48 machine learning algorithm provided better performance and accuracy than others before preprocessing technique. Santhanam and Padmavathi [26] Genetic Algorithm, K-means, and SVM were applied and increase the accuracy value. Ramiro et al. [27] applied a fuzzy rule mechanism to decrease the possibility of wrong treatment. It helps doctors as a recommender system to give the patient the correct treatment. Different data mining algorithm applied by Saba et al. [28] and it observed from these algorithms, the Meta classifier provided higher accuracy than a single classifier. To avoid the chronic complications of diabetes, patients should control a blood glucose level as the HbA1c (3 months accumulative blood glucose level) should be less than 7% [29]. Kavi at el. [30] developed a new predicted model by using machine learning techniques. In the proposed model, the main objective is to classify diabetic patients into two classes, such as under control (HbA1c < 7%) and out of control (HbA1c > 7%).

Moreover, Diabetes mellitus is a chronic, debilitating disease, which is associated with a range of severe complications [31]. There is a strong correlation between different risk factors of diabetes mellitus [32] over kidney disease and blood pressure disease. Early detection and meticulous management to prevent complications is the major challenge of diabetic care [31].

Renal failure or kidney disease is the common risk factor correlated with chronic microvascular complications of diabetes mellitus [33]. Chronic complications vary markedly in individuals but generally increase with the duration of diabetes.

• *Chronic Kidney Disease:*

Diabetes nephropathy is the leading cause of end-stage kidney failure in the world [34]. A diabetic person with persistent albumin loss in urine and progressive renal insufficiency with or without hypertension is said to have diabetes nephropathy [34]. But it ideally depends on documentation of diabetes-specific changes in kidney biopsy material. Based on Table 3 investigation range, around 30% of people with type 1 diabetes eventually create diabetes nephropathy. Nephropathy is less common in Type 2 diabetes (15-20%) than Type 1, but due to the more significant number of Type 2 diabetes, the majority of patients with kidney failure are Type 2 diabetes [35]. Early changes may be asymptotic but later may lead to kidney failure.

Table 3. Investigation ranges of chronic kidney disease.

Gender	Serum Creatinine (mg/dL)
Man	0.7-1.3 mg/dL
Woman	0.6-1.1 mg/dL

Hypertension or blood pressure is other common risk factors that correlated with diabetes mellitus for developing heart disease, stroke, blindness, and kidney failure [36].

• *Blood Pressure Disease*:

High blood pressure causes damage to large and small blood vessels in the body. At the point when high blood pressure exists with high blood cholesterol levels or diabetes, the risk of heart attack or stroke increases many times [37]. Hypertension or blood pressure contributes to the progressing state of chronic complications of diabetes. Patients with Type 1 diabetes have hypertension is often an appearance of diabetes nephropathy. In patients with Type 2 diabetes, hypertension is often a part of metabolic syndrome. Most prospective studies with hypertensive diabetic persons have documented that reduction of blood pressure is the single most significant factor that reduces both renal disease progression and cardiovascular events [37]. Table 4 shows the investigation ranges of blood pressure disease.

Table 4. Investigation ranges of blood pressure disease.

Normal	Diastolic	Systolic
	60-89 mmHg	90-139 mmHg
Low	Diastolic	Systolic
	Less than 60 mmHg	Less than 90 mmHg
High	Diastolic	Systolic
	90-100 mmHg or Higher	10-160 mmHg or Higher

All these risk factors should be taken into account during the prevention and treatment of chronic complications of diabetes mellitus [29].

• *Hearing Loss*:

Hearing loss is a common medical health problem that influences work efficiency, functional status, social communications, personal satisfaction, and quality of life [38]. Hearing loss is one of the concerns among the growing age person infected with diabetes mellitus and the person who is working in a noisy environment in the job. Diabetes mellitus and exposure to loud noise are well-known risk factors for hearing loss. Research suggests that patients with diabetes may experience more significant hearing loss than those without the disease [39]. Below the hearing threshold range limit up to 20 decibels, is considered to be normal hearing. Several hearing loss problems can be described according to severity, as follows in Table 5.

Table 5. Investigation ranges of hearing loss.

Problem	Range in dB
Mild	20 to 40 dB
Moderate	41 to 60 dB
Severe	61 to 80 dB
Profound	Greater than 80 dB

- *Skin Problem*:

The skin issue differs significantly in side effects and severity. They can be temporary or permanent and might be painless or painful. Some have situational causes, while others might be hereditary. Sometimes, skin conditions are minor, avoidable, and others can be dangerous [40]. Diabetes can affect the small veins of the body that supply the skin with blood. The skin problem can be the first visible sign that a person has diabetes. Diabetes dermopathy causes due to the changes to the blood vessels because of diabetes. Dermopathy appears as scaly patches that are light brown or red, often occurring on the skins. The patches are sometimes called skin spots. A higher rate of this condition is found in people having retinopathy, neuropathy, or kidney disease [41]. The diagnostic range of diabetes skin problem patients is categorized in Table 6.

Table 6. Investigation ranges of skin problems.

Problem	Range in IU/L
Tested Negative	<40 IU/L
Tested Positive	>100 IU/L

The limitation of existing related work is, different dataset attributes provide different prediction results based on different machine learning classification techniques. As a result, most of the time, it is difficult to identify which algorithm is more effective and provide better results based on the dataset attribute of diabetes mellitus. In previous, it is not to be under consideration of the risk factors correlated with Diabetes Mellitus.

In this work, we make a comparison based on the performance of different machine learning classifiers and draw the prediction results based on the relevant risk factors of Diabetes Mellitus. We also build a correlation between different risk factors of diabetes mellitus.

3. Methodology

For the study, methodology comprises of few stages, which are an accumulation of diabetes dataset with the relevant attributes of the patients, such as pre-processing the attributes, to perform various machine learning classification and corresponding performance matrices utilizing such data. Finally, we analyze the risk factors correlated with diabetic dataset. In the following, we are going to discuss these phases briefly.

A. Dataset and Attributes

The dataset obtained from the diagnostic section of Medical Centre Chittagong (MCC), Bangladesh. For the study, we collect the diagnostic dataset consists of 16 attributes or risk factors of diabetes mellitus of 200 patients. We have organized the attributes and corresponding values, shown in Table 7.

The training data is categorized diversely for diabetic and non-diabetic patients. We have considered only the last result of medical tests of diabetic patients before their diagnostic of diabetes. For non-diabetic patients, we have considered all their test results throughout their previous medical history. For repetitive diagnostic tests, we consider only the results of the first test.

Table 7.Dataset Description.

SI No.	**Attributes**	**Type**	**Values**
1	Age (Years)	Numeric	{1 to 100}
2	Sex	Nominal	{Male, Female}
3	Weight (Kgs)	Numeric	{5 to 120}
4	Diet	Nominal	{Vegetarian, Non-Vegetarian}
5	Polyuria	Nominal	{Yes, No}
6	Water Consumption	Nominal	{Yes, No}
7	Excessive Thirst	Nominal	{Yes, No}
8	Blood Pressure (mmHg)	Numeric	{50 to 200}
9	Hyper Tension	Nominal	{Yes, No}
10	Tiredness	Nominal	{Yes, No}
11	Problem in Vision	Nominal	{Yes, No}
12	Kidney Problem	Nominal	{Yes, No}
13	Hearing Loss	Nominal	{Yes, No}
14	Itchy Skin	Nominal	{Yes, No}
15	Genetic	Nominal	{Yes, No}
16	Diabetic	Nominal	{Yes, No}

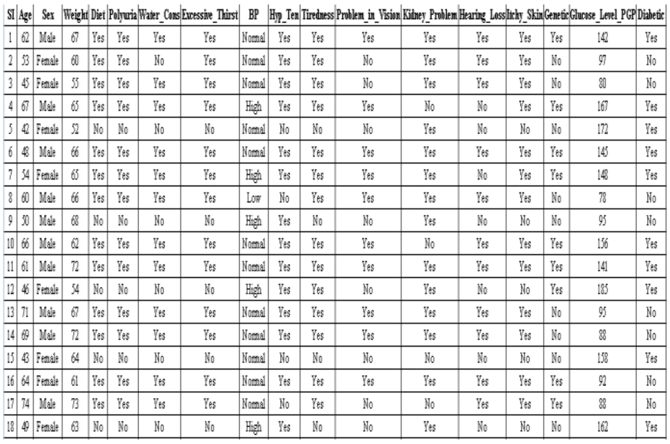
The sample clinical dataset of Type 2 diabetes patients is shown in Table 8.

The attributes of the dataset are obtained from clinical tests. The chosen clinical test result is found relevant to the diagnosis and the onset of Type 2 diabetes disease, which is the part of diabetes prevention trial studies.

B. Data Pre-processing

Data pre-processing is one of the most significant phases in the data mining process. It prepares and transforms the initial dataset. Raw data is generally incomplete and inconsistent and can produce misleading results. Thus, some pre-processing data methods can be applied to raw data before running an analysis. For example, the exact numeric value of the attributes does not have meaning to predict diabetes. This specific dataset had both nominal and real-valued attributes. We transform the numeric attribute values into nominal for finding a meaningful way to use such data.

Table 8. Sample dataset of diabetic patients.



Sl	Age	Sex	Weight	Diet	Polyuria	Water_Cons	Excessive_Thirst	BP	Hyp_Ten	Tiredness	Problem_in_Vision	Kidney_Problem	Hearing_Loss	Itchy_Skin	Genetic	Glucose_Level_PGP	Diabetic
1	62	Male	67	Yes	Yes	Yes	Yes	Normal	Yes	Yes	Yes	Yes	Yes	Yes	Yes	142	Yes
2	53	Female	60	Yes	Yes	No	Yes	Normal	Yes	Yes	No	Yes	Yes	Yes	No	97	No
3	45	Female	55	Yes	Yes	Yes	Yes	Normal	Yes	Yes	No	Yes	Yes	Yes	No	80	No
4	67	Male	65	Yes	Yes	Yes	Yes	High	Yes	Yes	Yes	No	No	Yes	Yes	167	Yes
5	42	Female	52	No	No	No	No	Normal	No	No	No	Yes	No	No	No	172	Yes
6	48	Male	66	Yes	Yes	Yes	Yes	Normal	Yes	Yes	Yes	Yes	Yes	Yes	Yes	145	Yes
7	54	Female	65	Yes	Yes	Yes	Yes	High	Yes	Yes	Yes	Yes	No	Yes	Yes	148	Yes
8	60	Male	66	Yes	Yes	Yes	Yes	Low	No	Yes	Yes	Yes	Yes	Yes	No	78	No
9	50	Male	68	No	No	No	No	High	Yes	No	No	Yes	No	No	No	95	No
10	66	Male	62	Yes	Yes	Yes	Yes	Normal	Yes	Yes	Yes	No	Yes	Yes	Yes	156	Yes
11	61	Male	72	Yes	Yes	Yes	Yes	Normal	Yes	Yes	Yes	Yes	Yes	Yes	Yes	141	Yes
12	46	Female	54	No	No	No	No	High	Yes	Yes	No	Yes	No	No	Yes	185	Yes
13	71	Male	67	Yes	Yes	Yes	Yes	Normal	Yes	Yes	Yes	Yes	Yes	Yes	No	95	No
14	69	Male	72	Yes	Yes	Yes	Yes	Normal	Yes	Yes	Yes	Yes	Yes	Yes	No	88	No
15	43	Female	64	No	No	No	No	Normal	No	No	No	No	No	No	No	158	Yes
16	64	Female	61	Yes	Yes	Yes	Yes	Normal	Yes	Yes	Yes	Yes	Yes	Yes	Yes	92	No
17	74	Male	73	Yes	Yes	Yes	Yes	Normal	No	Yes	No	No	Yes	Yes	Yes	88	No
18	49	Female	63	No	No	No	No	High	Yes	No	No	Yes	No	No	No	162	Yes

The patient's age is categorized into three, such as Young (10-25 years), Adult (26- 50 years), and Old (above 50 years).

Also, the patient's weight is categorized into three groups; they are- Underweight (less than equal 40 Kgs), Normal (41-60 Kgs), and Overweight (above 60 Kgs). At last, Blood Pressure is classified into three categories, such as Normal (120/80 mmHg), Low (less than 80 mmHg), and High (greater than 120 mmHg).

C. Apply Machine Learning Techniques

Machine learning is part and parcel of modern computer science, comprises algorithms that can learn from data; it gives a set of methods that can recognize patterns from the data and also use the patterns to generate future predictions. Machine Learning (ML) provides various techniques, methods, and tools that can help to solve diagnostic and prognostic problems in a variety of medical data. ML can be used for evaluating how important clinical parameters are and how their combinations are used for prognosis, e.g., prediction of disease progression, extraction of medical knowledge for outcome research, therapy planning, and support and above all for the patient and clinical management. ML is also used in the Intensive Care Unit and intelligent alarming, resulting in effective and efficient monitoring for the detection of regularities in the data, such as dealing with incomplete data and interpretation of continuous data.

Machine learning has incredible potential in improving the effectiveness and accuracy of decisions drawn by intelligent computer programs. Machine learning incorporates mainly concept learning and classification learning. Classification learning is the most widely used machine learning technique that includes separating the data into different segments, which are non-overlapping. Hence classification is the way toward finding a set of models that describe and recognize the class label of the data object.

Machine learning techniques also give efficient outcomes to extract knowledge by constructing predicting models from diagnostic medical datasets collected from diabetic patients. Furthermore, predicting the disease earlier leads to treating the patients before it becomes critical. Therefore, it has a significant role in diabetes research, presently like never before. Then we employ four popular machine learning classification techniques, namely Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbour (KNN), and C4.5 Decision Tree (DT), on adult population data to predict Diabetic Mellitus. We depict some explanations about these machine learning algorithms as below.

1. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a more powerful classification technique proposed by J. Platt et al. [42]. A Support Vector Machine (SVM) is an excluded classifier, formally characterized the data by separating a hyper-plane. SVM isolates entities in specified classes. It can also identify and classify an instance that is not supported by data. SVM is not caring for the distribution of acquiring data for each class. The two extensions of this algorithm are used; they are regression analysis to produce a linear function, and another one is learning to rank elements to produce a classification for individual elements.

SVM is one of the supervised learning techniques used in medical diagnosis for classification and regression [43, 45]. SVM, at the same time, minimizes the empirical classification error and maximize the geometric margin. So, SVM is called Maximum Margin Classifiers. SVM is a general algorithm based on guaranteed risk bounds of statistical learning theory, which is also called the structural risk minimization principle. SVMs can efficiently perform non-linear classification using the kernel trick. The kernel trick maps inputs into highdimensional feature spaces without explicitly realizing the feature spaces.

An SVM model represents points in space, mapped so that the different categories are divided by a clear gap [43, 44]. For instance, given a set of points belonging to either one of the two classes, an SVM finds a hyper-plane having the most significant possible fraction of points of a similar class on a similar plane. This separating hyperplane is known as the Optimal Separating Hyper-plane (OSH) that maximizes the distance between the two parallel hyperplanes and can minimize the risk of misclassifying instances of the test dataset.

For considering the overlapping points, an SVM finds a hyper-plane having the appropriate points of the same class on the same plane. This separating hyperplane is called the optimal separating hyper-plane (OSH) that maximizes the distance between the two parallel hyperplanes and can minimize the risk of misclassifying instances of the test dataset. Figure 1 shows an SVM model by using a set of training data from the sample dataset of diabetic patients.

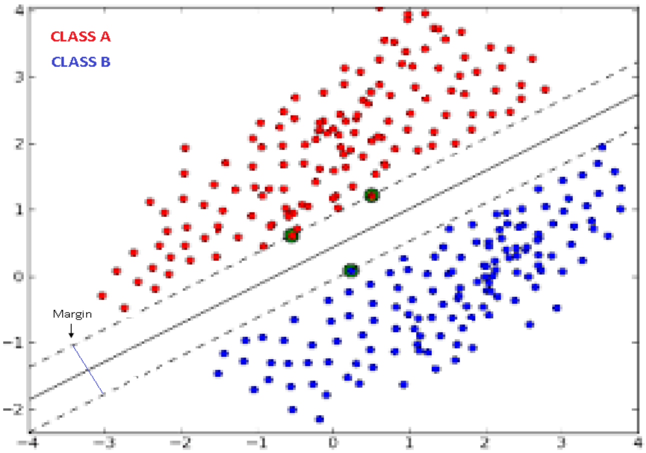


Figure 1. Maximum margin hyper-plane for SVM with samples from two classes

2. Naive Bayes Algorithm

Naive Bayes is the popular probabilistic classification technique proposed by John et al. [46]. Naive Bayes, also called the Bayesian theorem, is a simple, effective, and commonly used machine learning classifier using probabilistic results by counting the frequency and combines the value given in the data set. By using the Bayesian theorem, it assumes that all attributes are independent and based on variable values of classes. In a real-world application, the conditional independence assumption rarely holds and gives more sophisticated classifier results. The formula (1) for Bayes' Theorem is [46]:

$$P(H \mid E) = \frac{P(E \mid H) * P(H)}{P(E)} \ldots\ldots\ldots\ldots\ldots(1)$$

Here, P(H|E) is the posterior probability, the probability that a hypothesis (H) is true given some evidence (E). P(H) is the prior probability, i.e., the probability of the hypothesis being true. P(E) is the probability of the predictor, irrespective of the hypothesis. P(E|H) is the probability of the evidence when the hypothesis is true. In Naive Bayes classifier, it is acceptable that the input variables (features) are independent of each other, and all features individually contribute to the probability of the target variable. So, the existence of one feature variable does not affect the other feature variables. That is why it is called naïve. However, in real datasets, the feature variables are usually dependent on each other. So this is one of the drawbacks of the Naive Bayes classifier. Naive Bayes classifier, though, works very well for large data sets and sometimes performs better than other complicated classifiers. There are few distinct types of Naive Bayes classifiers; among them, the Gaussian Naive Bayes classifier was used in this model. The Gaussian Naive Bayes classifier accepts that the feature values are continuous, and the values of belonging to each class are normally distributed [47].

For simplifying prior and posterior probability calculation, among 200 real diagnostic datasets, we have considered 14 training datasets with four different attributes of the diabetic patients. The attributes are blood pressure, kidney disease, skin disease, and hearing loss. If any patient suffers from blood pressure and kidney disease and at the same time do not infect from skin disease and hearing loss problem, by using Naive Bayes classifier, we have found out the probability value of 'Yes' (0.48) is higher than the probability value of 'No' (0.11). It means that the patient has diabetes.

Naive Bayes algorithm is used for binary and multiclass classification and can also be trained on a small dataset, which is a huge advantage. It is also swift and scalable. Moreover, it mitigates the problem arising from the curse of dimensionality to some degree. However, as mentioned before, it makes the unrealistic presumption that the input variables are free of one another. It is not the case in real-life datasets, where there can be many complex relationships between the feature variables.

3. K-Nearest Neighbour Algorithm (KNN)

K-nearest neighbour is a simple classification and regression algorithm that used the non-parametric method proposed by Aha et al. [48]. The algorithm incorporates all valid attributes and classifies new attributes based on their resemblance measure. To determine the distance from the point of interest to points in the training data set, it uses a tree-like data structure. The value of k is always a positive integer of the nearest neighbour. The nearest neighbours are chosen from a set of class or object property values.

KNN is a standout amongst the most basic and straight forward data mining techniques.

It is called Memory-Based Classification, as the training examples should be in the memory at run-time [49]. When dealing with continuous attributes, the distinction between the attributes is calculated using the Euclidean distance. The equation (2) gives the Euclidean distance between two x and y points:

$$\text{Euclidean} = \sqrt{\sum_{i=1}^{K}(xi-yi)^2}\text{.............}(2)$$

KNN generally manages with continuous attributes; however, it can also deal with discrete attributes. The study additionally shows that K-Nearest-Neighbour is a standout amongst the most broadly utilized data mining techniques in classification problems. Its simplicity and generally high convergence speed make it a popular choice. However, the main disadvantage of KNN classifiers is the enormous memory necessity expected to store the entire sample. When the sample is large, the response time on a sequential computer is also significant. KNN classifier separates the training data into smaller subsets, and building a model for each subset, then applying voting to classify testing data, can enhance the classifier's performance. Figure 2 shows a simple KNN classifier algorithm (with, k=1) that applied to 200 diabetic patient datasets obtained from data pre-processing.

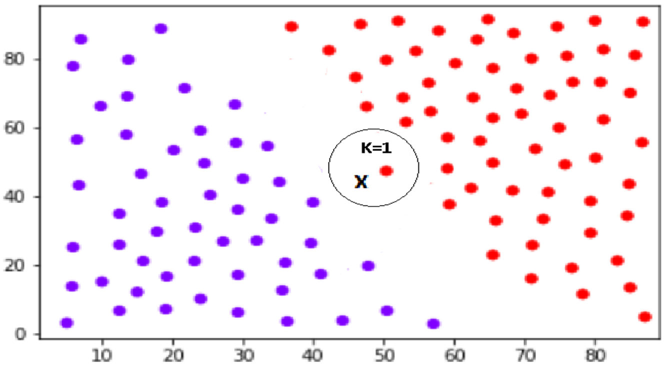


Figure 2. KNN classifier with diabetes data

4. Decision Tree

Decision Tree is the best classification procedure which can make a comparison among the most popular classification technique and easy to understand for knowledge and information systems [50]. It is a learning system, which uses a 'Divide and Conquers' procedure to classify instances. Most of the cases real data contain noisy information, and the divide and conquer approach to deal with building trees will, in general, incorporate such inconsistencies with the classifier, which prompts to lower prediction accuracy on unseen test data. For overcoming such over-fitting, Decision Trees are suitable [59]. Pruning of the decision tree is a helping way to understand easily as well as the execution of test data.

A Decision Tree is a robust classification technique to predict diabetes. Most of the information represented in limited discrete areas called the 'classification.' Each of the discrete areas and features of the specific domain is called a class. In Decision Tree, an input feature of the class attribute is labeled with the internal node, and the leaf node of the tree is labeled by attribute, and each attribute associated with a target value. The highest among information gain for all the attributes are evaluated in each node of the tree.

Numerous popular decision tree algorithms are available to classify diabetes data, including ID3, J48, C4.5, C5, CHAID, and CART. In project work, the C4.5 Decision Tree algorithm has been chosen to measure performance analysis of the diabetes data. C4.5 provides extended features of the ID3 Decision Tree algorithm proposed by Ross Quinlan et al. [51]. This learning method can be used to diagnose medical data for predicting the value of the decision attribute based on information gain. C4.5 evaluates and selects the attribute value of the data that separates the tested data into subset data, which enriched the class in each branch of the decision tree. The normalized information gain decides to pick the highest value attribute in constructing the tree. As a result, the significant risk factors of the diabetes mellitus attributes are arranged from root nodes to child nodes downwards with comparatively week attributes. In this process, the tree structures are constructing.

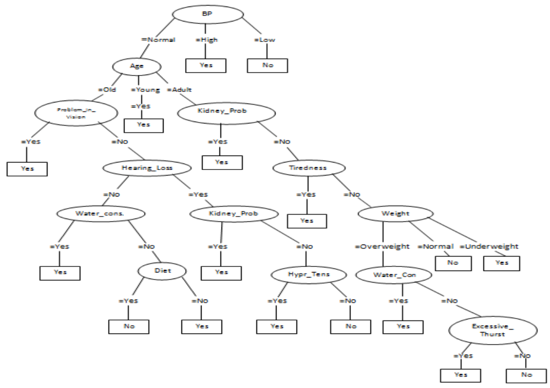


Figure 3. C4.5 decision tree

Figure 3 shows the Decision Trees are built-in C4.5 by using a set of training datasets of diabetic patients. By using the datasets, the information gain is calculated for all the attributes that we have considered in our work to measure the performance analysis of machine learning techniques.

According to Figure 4, after defining the problem, we have to collect the relevant data from the Diagnostic Data Storage. We then pre-process the data to build the prediction model. After that, various machine learning techniques are applied to the training dataset. Finally, the test dataset is used to measure the performance of the techniques for choosing the best classifier to predict diabetes mellitus.

D. Risk Factors Correlation

Progression of diabetes mellitus is strongly correlated to several complications, the leading causes of chronic kidney and blood pressure disease. It is well known that

DM covers a wide area of different pathophysiological conditions. The most widely common complications are divided into micro and macrovascular disorders, including diabetic nephropathy, retinopathy, neuropathy, and cardiovascular disease. Because of high DM increase mortality and morbidity. Its related complications need to be prevented. That's why it is essential to eliminate several risk factors related to long term diabetes complications; as a result, longevity can be increase. A correlation measures the relationship between two or more variables indicating the risk factors of DM. To evaluate the correlation between different risk factors of DM, statistical correlation [5] can be applied to the data set attributes that we have considered of chronic kidney disease, blood pressure, hearing loss, and skin problem over the diabetes disease.

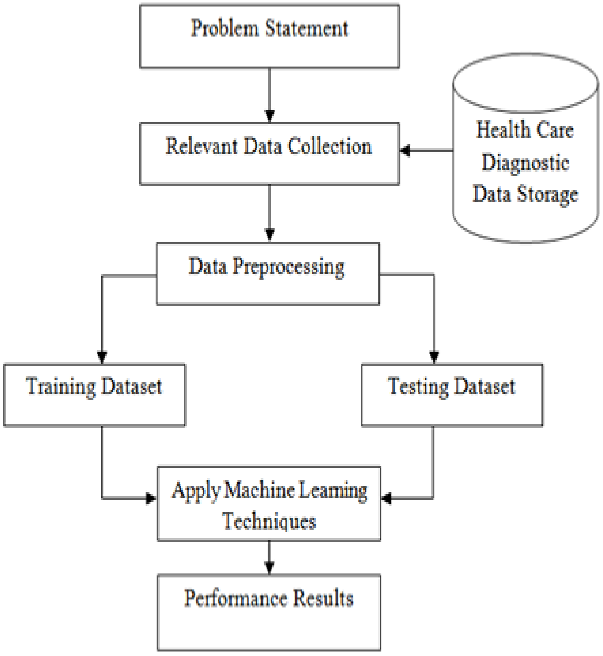


Figure 4. The overall process of our work.

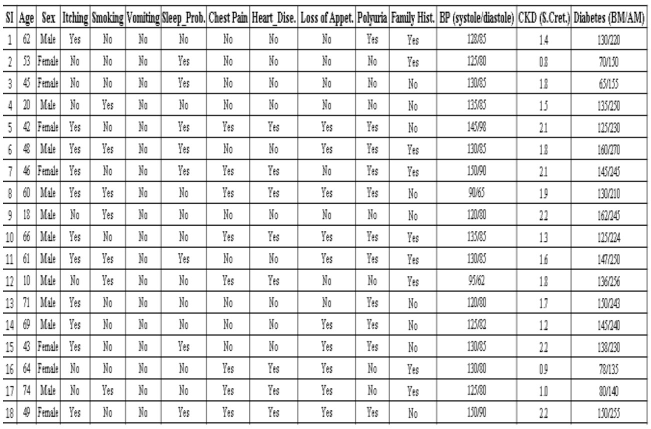
- *Correlation between diabetes and Kidney disease:*

One reason for chronic kidney disease is diabetes mellitus, which is characterized by high blood glucose (sugar) levels. Over time, the high amounts of sugar in the blood vessels that harm millions of small filtering units in the kidney. It is, in the long run, prompts to kidney failure.

Approximately 20 to 30 percent of people with diabetic enlarge to kidney disease (diabetic nephropathy), although not all of these people progress to kidney failure. A person with diabetes is at risk of nephropathy, whether they depend on insulin or not. The risk is correlated to the length of time the person has diabetes. There is no heal for diabetic nephropathy, and therefore the treatment is life-long. Diabetic peoples are also at risk of other kidney problems, including narrowing of the arteries to the kidneys called renal artery renovascular disease.

To build a correlation between chronic kidney disease and diabetic patients, the attributes that we have considered are- age, sex, blood pressure, itching, vomiting, trouble sleeping, chest pain, smoking, heart disease, loss of appetite, too much urine, breath problem, and family history. The sample clinical dataset of diabetic nephropathy patients is shown in Table 9.

Table 9. Sample dataset of diabetic nephropathy patients.



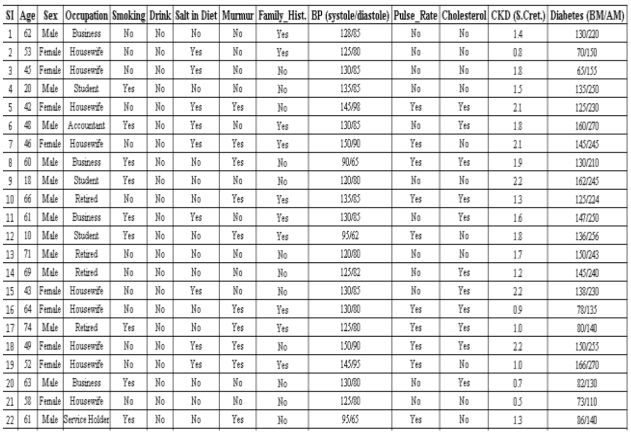
Sl	Age	Sex	Itching	Smoking	Vomiting	Sleep_Prob.	Chest Pain	Heart_Dise.	Loss of Appet.	Polyuria	Family Hist.	BP (systole/diastole)	CKD (S.Cret.)	Diabetes (BM/AM)
1	62	Male	Yes	No	No	No	No	No	No	Yes	Yes	128/85	1.4	130/220
2	53	Female	No	No	No	Yes	No	No	No	No	Yes	125/80	0.8	70/150
3	45	Female	No	No	No	Yes	No	No	No	No	No	130/85	1.8	65/155
4	20	Male	No	Yes	No	No	No	No	No	No	No	135/85	1.5	135/250
5	42	Female	Yes	No	No	Yes	Yes	Yes	Yes	Yes	No	145/98	2.1	125/230
6	48	Male	Yes	Yes	No	Yes	No	No	Yes	Yes	Yes	130/85	1.8	160/270
7	46	Female	Yes	No	No	Yes	Yes	Yes	No	Yes	Yes	150/90	2.1	145/245
8	60	Male	Yes	Yes	No	No	Yes	Yes	Yes	Yes	No	90/65	1.9	130/210
9	18	Male	No	Yes	No	No	No	No	No	No	No	120/80	2.2	162/245
10	66	Male	Yes	No	No	No	Yes	Yes	Yes	Yes	Yes	135/85	1.3	125/224
11	61	Male	Yes	Yes	No	Yes	No	No	Yes	Yes	Yes	130/85	1.6	147/250
12	10	Male	No	Yes	No	No	Yes	Yes	No	No	Yes	95/62	1.8	136/256
13	71	Male	Yes	No	No	No	No	No	No	Yes	No	120/80	1.7	150/243
14	69	Male	Yes	No	No	No	No	No	Yes	Yes	No	125/82	1.2	145/240
15	43	Female	Yes	No	No	Yes	No	No	Yes	Yes	No	130/85	2.2	138/230
16	64	Female	No	No	No	No	Yes	Yes	Yes	No	Yes	130/80	0.9	78/135
17	74	Male	No	Yes	No	No	Yes	Yes	Yes	No	Yes	125/80	1.0	80/140
18	49	Female	Yes	No	No	Yes	Yes	Yes	Yes	Yes	No	150/90	2.2	150/255

- *Correlation between diabetes and blood pressure disease:*

Many people with diabetes mellitus also have hypertension or blood pressure disease. Blood pressure disease is known as a "silent killer" since it often has no cleared symptoms, and many people are uninformed they have it. According to the 2013 review [36], the American Diabetes Association (ADA) found that a combination of hypertension and diabetes mellitus is particularly deadly and can significantly raise the risk of having a heart attack or stroke. A person with diabetes must control blood pressure.

In our comparative study to build a correlation between blood pressure disease and diabetic patients the attributes that we have considered are- age, sex, occupation, smoking, blood pressure (both systolic and diastolic), pulse rate, drink, family member, salt in diet, murmur, and cholesterol. The sample clinical dataset of diabetes blood pressure patients is shown in Table 10.

Table 10. Sample dataset of diabetic blood pressure patients.



SI	Age	Sex	Occupation	Smoking	Drink	Salt in Diet	Murmur	Family_Hist.	BP (systole/diastole)	Pulse_Rate	Cholesterol	CKD (S.Cret.)	Diabetes (BM/AM)
1	62	Male	Business	No	No	No	No	Yes	128/85	No	No	1.4	130/220
2	53	Female	Housewife	No	No	Yes	No	Yes	125/80	No	No	0.8	70/150
3	45	Female	Housewife	No	No	Yes	No	No	130/85	No	No	1.8	65/155
4	20	Male	Student	Yes	No	No	No	No	135/85	No	No	1.5	135/250
5	42	Female	Housewife	No	No	Yes	Yes	No	145/98	Yes	Yes	2.1	125/230
6	48	Male	Accountant	Yes	No	Yes	No	Yes	130/85	No	Yes	1.8	160/270
7	46	Female	Housewife	No	No	Yes	Yes	Yes	150/90	Yes	No	2.1	145/245
8	60	Male	Business	Yes	No	No	Yes	No	90/65	Yes	Yes	1.9	130/210
9	18	Male	Student	Yes	No	No	No	No	120/80	No	No	2.2	162/245
10	66	Male	Retired	No	No	No	Yes	Yes	135/85	Yes	Yes	1.3	125/224
11	61	Male	Business	Yes	No	Yes	No	Yes	130/85	No	Yes	1.6	147/250
12	10	Male	Student	Yes	No	No	Yes	Yes	95/62	Yes	No	1.8	136/256
13	71	Male	Retired	No	No	No	No	No	120/80	No	No	1.7	150/243
14	69	Male	Retired	No	No	No	No	No	125/82	No	Yes	1.2	145/240
15	43	Female	Housewife	No	No	Yes	No	No	130/85	No	Yes	2.2	138/230
16	64	Female	Housewife	No	No	No	Yes	Yes	130/80	Yes	Yes	0.9	78/135
17	74	Male	Retired	Yes	No	No	Yes	Yes	125/80	Yes	Yes	1.0	80/140
18	49	Female	Housewife	No	No	Yes	Yes	No	150/90	Yes	Yes	2.2	150/255
19	52	Female	Housewife	No	No	Yes	Yes	Yes	145/95	Yes	No	1.0	166/270
20	63	Male	Business	Yes	No	No	No	No	130/80	No	Yes	0.7	82/130
21	58	Female	Housewife	No	No	No	No	No	125/80	No	No	0.5	73/110
22	61	Male	Service Holder	Yes	No	No	Yes	No	95/65	Yes	No	1.3	86/140

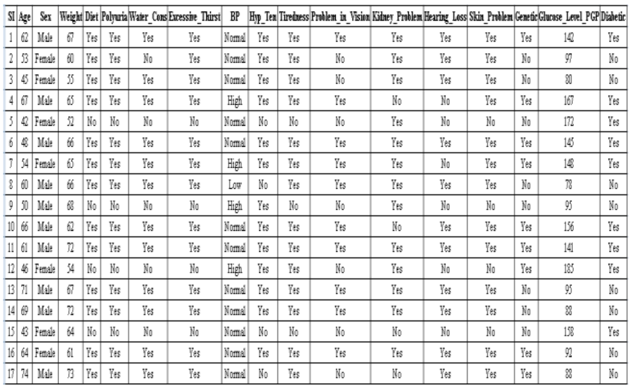
- *Correlation between diabetes and hearing loss disease:*

Diabetes is associated with a risk of hearing loss. Type 2 diabetes may be an independent risk factor for hearing loss. Because high blood sugar effects of hyperglycemia may damage the cochlea. Signs and symptoms that commonly occur in Type 2 diabetes can be related to the immediate effects of hyperglycemia or hypoglycemia (blurred vision and excessive thirst, for example). Many patients may not realize the relation between their hearing impairment and their diabetes condition. According to the National Institutes of Health [39], Hearing loss is common in people with diabetes. To build a correlation between hearing loss and diabetic patients the attributes that we have considered are - age, sex, weight, diet, polyuria, water consumption, excessive thirst, blood pressure, hypertension, tiredness, the problem in vision, kidney problem, hearing loss, skin problem, genetic and diabetic. The sample clinical dataset of diabetes hearing loss patients is shown in Table 11.

- *Correlation between diabetes and skin problem disease:*

Long term Type 2 diabetes with hyperglycemia or high blood glucose, tends to be associated with poor circulation, which decreases blood stream to the skin. It can also affect blood vessels and nerves. The capacity of the white platelets to fend off infections is also decreased in the face of elevated blood sugar. Diminished blood circulation can prompt changes in the skin's collagen. It changes the skin's texture, appearance, and ability to heal.

Table 11. Sample dataset of diabetic hearing loss patients.



SI	Age	Sex	Weight	Diet	Polyuria	Water_Cons	Excessive_Thirst	BP	Hyp_Ten	Tiredness	Problem_in_Vision	Kidney_Problem	Hearing_Loss	Skin_Problem	Genetic	Glucose_Level_PGP	Diabetic
1	62	Male	67	Yes	Yes	Yes	Yes	Normal	Yes	Yes	Yes	Yes	Yes	Yes	Yes	142	Yes
2	53	Female	60	Yes	Yes	No	Yes	Normal	Yes	Yes	No	Yes	Yes	Yes	No	97	No
3	45	Female	55	Yes	Yes	Yes	Yes	Normal	Yes	Yes	No	Yes	Yes	Yes	No	80	No
4	67	Male	65	Yes	Yes	Yes	Yes	High	Yes	Yes	Yes	No	No	Yes	Yes	167	Yes
5	42	Female	52	No	No	No	No	Normal	No	No	No	Yes	No	No	No	172	Yes
6	48	Male	66	Yes	Yes	Yes	Yes	Normal	Yes	Yes	Yes	Yes	Yes	Yes	Yes	145	Yes
7	54	Female	65	Yes	Yes	Yes	Yes	High	Yes	Yes	Yes	Yes	No	Yes	Yes	148	Yes
8	60	Male	66	Yes	Yes	Yes	Yes	Low	No	Yes	Yes	Yes	Yes	Yes	No	78	No
9	50	Male	68	No	No	No	No	High	Yes	No	No	Yes	No	No	No	95	No
10	66	Male	62	Yes	Yes	Yes	Yes	Normal	Yes	Yes	Yes	No	Yes	Yes	Yes	156	Yes
11	61	Male	72	Yes	Yes	Yes	Yes	Normal	Yes	Yes	Yes	Yes	Yes	Yes	Yes	141	Yes
12	46	Female	54	No	No	No	No	High	Yes	Yes	No	Yes	No	No	Yes	185	Yes
13	71	Male	67	Yes	Yes	Yes	Yes	Normal	Yes	Yes	Yes	Yes	Yes	Yes	No	95	No
14	69	Male	72	Yes	Yes	Yes	Yes	Normal	Yes	Yes	Yes	Yes	Yes	Yes	No	88	No
15	43	Female	64	No	No	No	No	Normal	No	No	No	No	No	No	No	158	Yes
16	64	Female	61	Yes	Yes	Yes	Yes	Normal	Yes	Yes	Yes	Yes	Yes	Yes	Yes	92	No
17	74	Male	73	Yes	Yes	Yes	Yes	Normal	No	Yes	No	No	Yes	Yes	Yes	88	No

Diabetes can hurt the body skin in two ways:

1. If blood glucose is high, at that point, the body loses liquid or fluid. With less fluid in the body, result skin can get dry. Dry skin can be itchy, causing the body to scratch and make it sore. Cracks cause infections and allow germs. It increases as the blood sugar becomes high. The skin of the legs, feet, elbows, and other places in the body can get dry [52].

2. Nerve damage can decrease the amount of body sweat. Sweating may cause dry skin in the feet and legs [52].

To build a correlation of diabetic patients, the attributes that we have to consider of a skin problem patients are age, sex, weight, diet, polyuria, water consumption, excessive thirst, blood pressure, hypertension, tiredness, the problem in vision, kidney problem, hearing loss, skin problem, genetic and diabetic. The sample clinical dataset of diabetic skin problem patients is shown in Table 12.

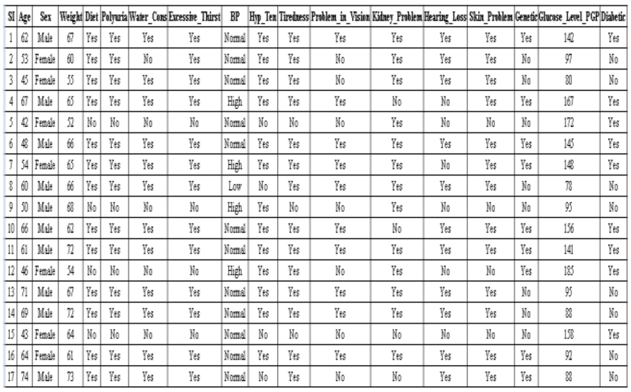
4. Experimental Results and Discussion

After the implementation of different Machine Learning models, the next step is to measure the performance of the classification techniques.

It is done by running the models on the test dataset, which was set aside earlier. The test dataset comprised of the original data for diabetic patients. N-fold (N-10) cross-validation technique was done for cardiovascular disease of the original data in the test dataset. To measure the performance N-fold (N=10) cross-validation [53] technique can be used. Cross-validation techniques can follow the following mechanism:

a) The test dataset is separated into N-folds, where each fold is used for classifying the testing and training data to predict the model.
b) Repeats N times until completing the procedure for the testing and training data.

Table 12. Sample dataset of diabetic skin problem patients.



Sl	Age	Sex	Weight	Diet	Polyuria	Water_Cons	Excessive_Thirst	BP	Hyp_Ten	Tiredness	Problem_in_Vision	Kidney_Problem	Hearing_Loss	Skin_Problem	Genetic	Glucose_Level_PGP	Diabetic
1	62	Male	67	Yes	Yes	Yes	Yes	Normal	Yes	Yes	Yes	Yes	Yes	Yes	Yes	142	Yes
2	53	Female	60	Yes	Yes	No	Yes	Normal	Yes	Yes	No	Yes	Yes	Yes	No	97	No
3	45	Female	55	Yes	Yes	Yes	Yes	Normal	Yes	Yes	No	Yes	Yes	Yes	No	80	No
4	67	Male	65	Yes	Yes	Yes	Yes	High	Yes	Yes	Yes	No	No	Yes	Yes	167	Yes
5	42	Female	52	No	No	No	No	Normal	No	No	No	Yes	No	No	No	172	Yes
6	48	Male	66	Yes	Yes	Yes	Yes	Normal	Yes	Yes	Yes	Yes	Yes	Yes	Yes	145	Yes
7	54	Female	65	Yes	Yes	Yes	Yes	High	Yes	Yes	Yes	Yes	No	Yes	Yes	148	Yes
8	60	Male	66	Yes	Yes	Yes	Yes	Low	No	Yes	Yes	Yes	Yes	Yes	No	78	No
9	50	Male	68	No	No	No	No	High	Yes	No	No	Yes	No	No	No	95	No
10	66	Male	62	Yes	Yes	Yes	Yes	Normal	Yes	Yes	Yes	No	Yes	Yes	Yes	156	Yes
11	61	Male	72	Yes	Yes	Yes	Yes	Normal	Yes	Yes	Yes	Yes	Yes	Yes	Yes	141	Yes
12	46	Female	54	No	No	No	No	High	Yes	Yes	No	Yes	No	No	Yes	185	Yes
13	71	Male	67	Yes	Yes	Yes	Yes	Normal	Yes	Yes	Yes	Yes	Yes	Yes	No	95	No
14	69	Male	72	Yes	Yes	Yes	Yes	Normal	Yes	Yes	Yes	Yes	Yes	Yes	No	88	No
15	43	Female	64	No	No	No	No	Normal	No	No	No	No	No	No	No	158	Yes
16	64	Female	61	Yes	Yes	Yes	Yes	Normal	Yes	Yes	Yes	Yes	Yes	Yes	Yes	92	No
17	74	Male	73	Yes	Yes	Yes	Yes	Normal	No	Yes	No	No	Yes	Yes	Yes	88	No

c) According to N-folds cross-validation, we partition the data into 10-folds, where each fold is nearly the same with other folds in the dataset.
d) Execution of each iteration contains nine folds as a training set to adapt the model, and the remaining 1 fold known as the testing set is used for evaluating performance.
e) In the end, learning scheme techniques performed 10 times on training data sets, and lastly, the prediction accuracy averages for 10 data sets. Various performance metrics, such as precision, recall, F-measure, and accuracy, are described as follow:

A. Evaluation Metric

If TP belongs to true-positive rate and FP belongs to false-positive rate then according to [53] the formal definition of precision is in equation (3),

$$\text{Precision} = \frac{TP}{TP + FP} \text{................}(3)$$

Furthermore, recall is defined as below where FN represents the false-negative rate [53] and represented in equation (4)

$$\text{Recall} = \frac{TP}{TP + FN} \text{..................}(4)$$

The F-measure can be evaluated using the value of precision and recall and defined as below [53] and represented in equation (5)

$$\text{F-measure} = \frac{2 * \mathrm{Re}\,call * \mathrm{Pr}\,ecision}{\mathrm{Pr}\,ecision + \mathrm{Re}\,call} \text{.............}(5)$$

On top of that, we also calculate the accuracy based on the correctly classified instances performed by machine learning techniques. Formally, Accuracy is calculated as below [53] and represented in equation (6)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \text{.....}(6)$$

B. Comparison Results

The performance of different machine learning techniques has been shown in Table 13, based on precision, recall, and F-measure. The table shows the results of various machine learning techniques such as SVM, NB, KNN, and C4.5.As information gain helps to construct the trees with attributes of the highest gain to lowest in a downward fashion, it is evident that C4.5 achieves better results than other classifiers to predict diabetes mellitus. According to Figure 6, C4.5 achieves 72% precision, 74% recall, and 72% F-measure on this dataset, which is higher than other learning techniques. This experimental result provides evidence that C4.5 Decision Tree performs well on medical datasets to predict diabetes mellitus based on various risk factors discussed in the earlier section.

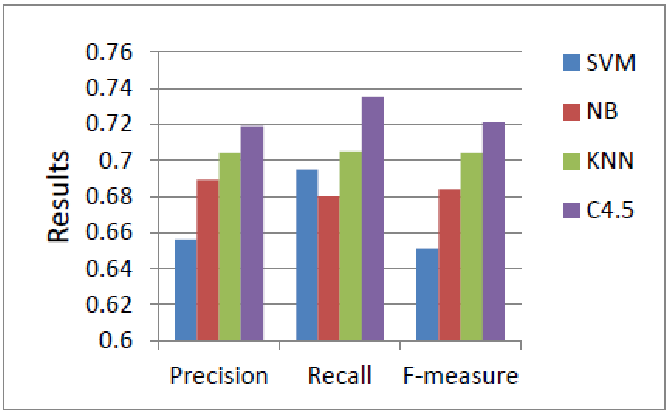


Figure 5. Predictions Results of Various Machine Learning Techniques

In addition to precision, recall, and F-measure, we also calculate the direct accuracy rate in the percentage of all these classifiers shown in Figure 7. If we observe Figure 5, we see that the C4.5 decision tree technique outperforms other techniques to predict Diabetes Mellitus.

Table 13. Comparison of prediction results of various machine learning techniques

Algorithm	Precision	Recall	F-measure
SVM	0.66	0.70	0.65
NB	0.69	0.68	0.68
KNN	0.70	0.71	0.70
C4.5	0.72	0.74	0.72

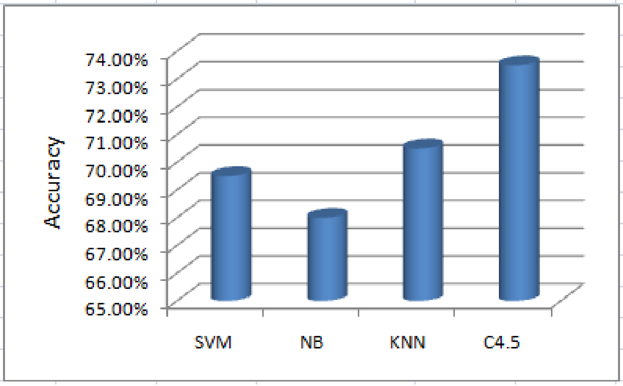


Figure 6. Accuracy Results of Various Machine Learning Techniques

Based on Figure 6, it can be observed that C4.5 Decision Tree achieves a better accuracy of 73.5% to predict diabetes mellitus utilizing a given medical dataset.

In the results, Area under the Receiver Operating Characteristic (ROC) curve of the SVM, NB, KNN, and C4.5 Decision Tree algorithms are 0.56, 0.65, 0.67 and 0.69 respectively, which is shown in Table 14. From Figure 7, the confidence band of the curve is clearly shown for C4.5 Decision Tree rather than other techniques. With the features of the information gain criterion, C4.5 Decision Tree achieves better accuracy for ROC. The experimental results prove that for the diabetic dataset, the area under ROC for the C4.5 algorithm performs best in four learning techniques is 0.69.

Table 14. Comparison of AUC of the four models

Algorithm	The area under the ROC Curve(AUC)
SVM	0.56
NB	0.65
KNN	0.67
C4.5	0.69

To determine the correlation between different risk factors of diabetes mellitus, we collect the dataset consists of various attributes or risk factors of kidney disease and diabetes mellitus of 200 diabetic patients. In the diagnostic dataset results, blood glucose levels (after the meal) are significantly increased for diabetic patients. Serum creatinine levels observed significantly low in non-diabetic patients and high in diabetic patients. A positive correlation (0.72) is found between serum creatinine and blood glucose level for diabetic patients, which shown in Figure 8.

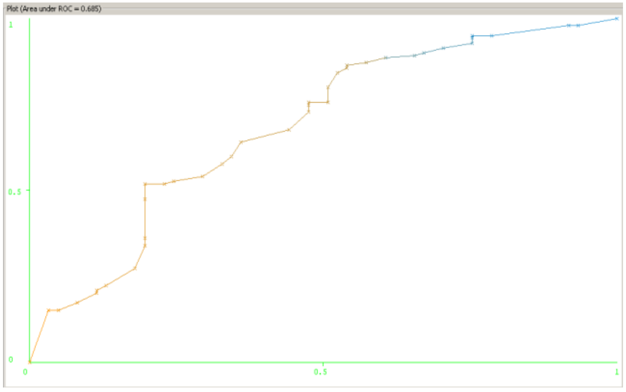


Figure 7. ROC Curve for C4.5 decision tree

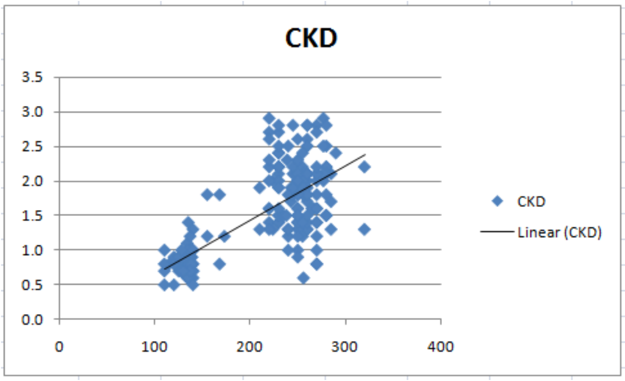


Figure 8. Correlation between Kidney disease and diabetic patients

Diabetes mellitus and blood pressure frequently coexist. Formation o blood glucose (after and before the meal) and blood pressure (systole/diastole) have recently reported as disease markers for diabetes and hypertension, respectively. This study is aimed to find the correlation between diabetes mellitus and blood pressure. An equal number of people of different ages and sex are selected to test. Their blood glucose (after and before the meal), serum creatinine, pulse rate, and cholesterol levels are measured by spectrophotometer. The dataset attributes are correlated by statistical methods. Notably, we see that blood glucose (after and before the meal) levels, as well as blood pressure (systole/diastole) levels, are significantly high for diabetic hypertensive patients. A significant positive correlation (0.81) is found between blood glucose levels and BP levels in diabetic hypertensive patients, shown in Figure 9. These findings suggest that the combination of hypertension and diabetes can be deadly, and together they can enhance the risk of a heart attack or stroke.

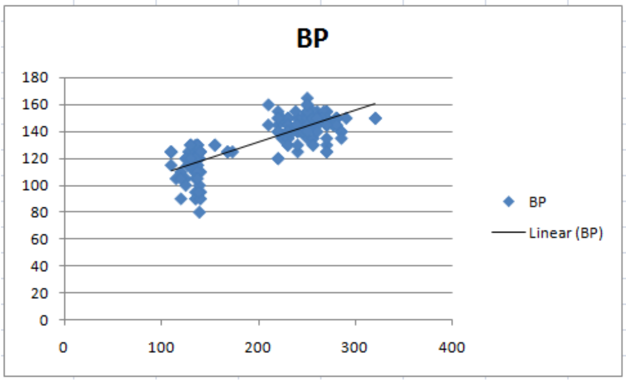


Figure 9. Correlation between blood pressure and diabetic patients

We have analyzed the correlation between diabetes mellitus and hearing loss patients. A negative correlation (-0.72) is found between blood glucose levels and diabetic hearing loss patients, shown in Figure 10. These results suggest that hearing loss and diabetes mellitus are comparatively weak correlated.

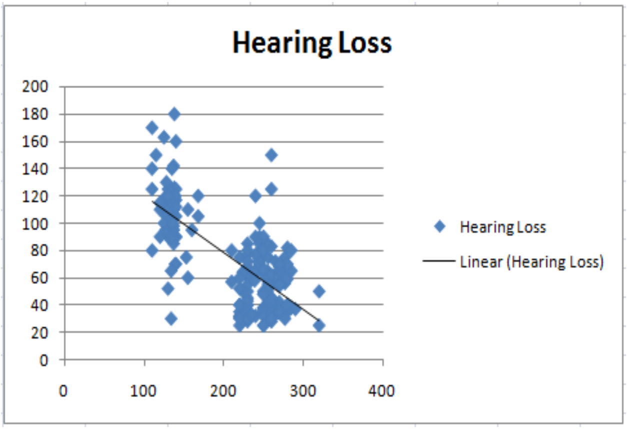


Figure 10. Correlation between hearing loss and diabetic patients

We also evaluate the correlation between diabetes mellitus and skin problems. A negative correlation (-0.76) is found between blood glucose levels and diabetic skin problem patients, shown in Figure 11. These results suggest that skin problem and diabetes mellitus is nearly weak correlated.

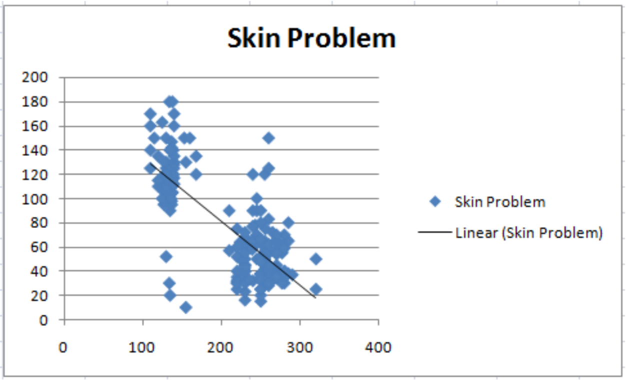


Figure 11. Correlation between skin problem and diabetic patients

In the future, we can collect more data and make decisions based on their correlation with other diseases respective by males and females by considering the concept of recent pattern analysis [57] for building more effective models.

5. Conclusion

In this work, we have explained how Machine Learning can be adopted in clinical diagnostics to predict the probability of diabetes-induced complications. It is done using different Machine Learning algorithms under various circumstances. Knowledge extraction from real health care dataset can be useful to predict diabetes mellitus. To predict diabetes mellitus effectively, we have performed our experiments using four popular machine learning algorithms, such as Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbour (KNN) and C4.5 Decision Tree on the adult population. From the experimental results, we can make the decision that C4.5 Decision Tree is significantly superior to other machine learning techniques on diabetes data. We also find a positive correlation at predicting kidney complications (Nephropathy) and blood pressure (Hypertension) complications and find a negative correlation at predicting hearing loss and skin complications (diabetes dermopathy) from the diabetic patients. For the study, we have collected a diagnostic dataset having 16 attributes diabetic of 200 patients. The experimental results assist the health care centre to make better clinical decisions on diabetes. It is also helpful for individuals to control diabetes.

Acknowledgments

It is our gratitude to give thanks and acknowledge to Dr. Shahed Parvez, Director, Medical Centre Chittagong, and Dr. Mohammad Maruf Faruqi, RMO, ICU, Medical Centre Chittagong, Bangladesh for providing the real diagnostic data and for their cooperation throughout our work.

References

[1] Morteza, M., Franklyn, P., Bharat, S., Linying, D., Karim, K., and Aziz G. 2015. Evaluating the Performance of the Framingham Diabetes Risk Scoring Model in Canadian Electronic Medical Records. Canadian Journal of diabetes 39, 30(April. 2015), 152-156.

[2] V., A. K., and R., C., 2013. Classification of Diabetes Disease Using Support Vector Machine. International Journal of Engineering Research and Applications. 3, (April. 2013), 1797-1801.

[3] Carlo, B G., Valeria, M., and Jesús, D. C., 2011. The impact of diabetes mellitus on healthcare costs in Italy. Expert review of pharmacoeconomics & outcomes research. 11, (Dec. 2011),709-19.

[4] Nahla B., Andrew, P. B., and M., N. B., 2010. Intelligible support vector machines for diagnosis of diabetes mellitus. Information Technology in Biomedicine, IEEE Transactions. 14, (July. 2010), 1114-20.

[5] V. Vapnik, "The Nature of Statistical Learning Theory." NY: Springer- Verlag. 1995.

[6] Akinnola N. AKINTUNDE "Path Analysis Step by Step Using Excel."

[7] G. Sparacino, F. Zanderigo, S. Corazza, A Maran, A Facchinetti, and C. Cobelli. "Glucose concentration can be predicted in time from continuous glucose monitoring sensor time-series. Biomedical Engineering", IEEE Transactions on, 54(5): 931{937, May 2007. ISSN 0018-9294. DOI: 10.1109/TBME.2006.889774.

[8] G. Baghdadi and AM. Nasrabadi. "Controlling blood glucose levels in diabetes by neural network predictor. In Engineering in Medicine and Biology Society", 29th Annual International Conference of the IEEE, pages 3216{3219, Aug 2007. DOI: 10.1109/IEMBS.2007.4353014.

[9] C. Marling, M. Wiley, R. Bunescu, J. Shubrook, and F. Schwartz. "Emerging applications for intelligent diabetes management." Artificial Intelligence Magazine, 33(2):67, 2012.

[10] E. Georga, V. Protopappas, D. Polyzos, and D. Fotiadis. "Predictive modeling of glucose metabolism using free-living data of type 1 diabetic patients. In Engineering in Medicine and Biology Society (EMBC)", 2010 Annual International Conference of the IEEE, pages 589{592, Aug 2010. DOI: 10.1109/IEMBS.2010.5626374.

[11] Abdullah A. Aljumah, "Application of data mining: Diabetes health care in young and old patients, Journal of King Saud University" - Computer and Information Sciences, Volume 25, Issue 2, July 2013, Pages 127-136.

[12] Kavakiotis, Ioannis, Olga Tsave, AthanasiosSalifoglou, NicosMaglaveras, IoannisVlahavas, and IoannaChouvarda. "Machine learning and data mining methods in diabetes research." Computational and structural biotechnology journal (2017).

[13] Zheng, Tao et al. "A machine learning-based framework to identify type 2 diabetes through electronic health records." International journal of medical informatics 97 (2017): 120- 127.

[14] Rani, A. Swarupa, and S. Jyothi. "Performance analysis of classification algorithms under different datasets." In Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on, pp. 1584-1589. IEEE, 2016.

[15] Kandhasamy, J. Pradeep, and S. Balamurali. "Performance analysis of classifier models to predict diabetes mellitus." Procedia Computer Science 47 (2015): 45-51.

[16] Y. Huang, P. McCullagh, N. Black, R. Harper, "Feature selection and classification model construction on type 2 diabetic patients 'data", Artificial Intelligence in Medicine 41 (3) (2015) 251–262.

[17] Meng, X. H., Huang, Y. X., Rao, D. P., Zhang, Q., & Liu, Q. (2013). "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors." The Kaohsiung journal of medical sciences, 29(2), 93-99.

[18] Thirumal, P. C., & Nagarajan, N. "Utilization of data mining techniques for the diagnosis of diabetes mellitus-a case study." ARPN Journal of Engineering and Applied Science, 10(2015).

[19] A. Al Jarullah, "Decision tree discovery for the diagnosis of type II diabetes, in Innovations in Information Technology (IIT)," 2011 International Conference on, 2011, pp. 303–307.

[20] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, R. S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," Johns Hopkins APL Technical Digest 10 (1988) 262–266.

[21] Komi, Messan, Jun Li, YongxinZhai, and Xianguo Zhang. "Application of data mining methods in diabetes prediction." In Image, Vision, and Computing (ICIVC), 2017 2nd International Conference on, pp. 1006-1010. IEEE, 2017.

[22] Xu, Weifeng, Jianxin Zhang, Qiang Zhang, and Xiaopeng Wei. "Risk prediction of type II diabetes based on the random forest model."In Advances in Electrical, Electronics, Information, Communication, and Bio-Informatics (AEEICB), 2017 Third International Conference on, pp. 382-386. IEEE, 2017.

[23] Song, Yunsheng, Jiye Liang, Jing Lu, and Xingwang Zhao. "An efficient instance selection algorithm for k nearest neighbor regression." Neurocomputing 251 (2017): 26-34.

[24] Perveen, Sajida, Muhammad Shahbaz, Aziz Guergachi, and Karim Keshavjee. "Performance analysis of data mining classification techniques to predict diabetes." Procedia Computer Science 82 (2016): 115-121.

[25] Pradeep, K. R., and N. C. Naveen. "Predictive analysis of diabetes using the J48 algorithm of classification techniques." In Contemporary Computing and Informatics (IC3I), 2016 2nd International Conference on, pp. 347-352. IEEE, 2016.

[26] Santhanam, T., and M. S. Padmavathi. "Application of K-means and genetic algorithms for dimension reduction by integrating SVM for a diabetes diagnosis." Procedia Computer Science 47 (2015): 76-83.

[27] Meza-Palacios, Ramiro, Alberto A. Aguilar-Lasserre, Enrique L. Ureña-Bogarín, Carlos F. Vázquez-Rodríguez, Rubén Posada-Gómez, and Armín Trujillo-Mata. "Development of a fuzzy expert system for the nephropathy control assessment in patients with type 2 diabetes mellitus." Expert Systems with Applications 72 (2017): 335-343.

[28] Bashir, Saba, Usman Qamar, Farhan Hassan Khan, and M. YounusJaved. "An Efficient Rule-Based Classification of

Diabetes Using ID3, C4. 5, & CART Ensembles." In Frontiers of Information Technology (FIT), 2014 12th International Conference on, pp. 226-231. IEEE, 2014.

[29] K. Meena, N. Vijayalakshmi, "An Analysis of Risk Factor for Diabetes using Data Mining Approach," Indian Journal of Public Health Research and Development, Vol. 6, Issue No. 2, pp 112-117, April-June 2015.

[30] Varsha Kavi and Divyesh Joshi, "A Survey on Enhancing Data Processing of Positive and Negative Association Rule Mining," International Journal of Computer Sciences and Engineering, Volume-02, Issue-03, Page No (139-143), Mar -2014.

[31] J. Lindstrom and J. Tuomilehto, "The Diabetes Risk Score: A practical tool to predict type 2 diabetes risk," Diabetes Care, 26:3 (2003), 725-731.

[32] Ajay Meshram, Karuna Kachhawa, Vijay Gujar, Pradeep Bokariya-"Correlation Of Dyslipidemia And Type 2 Diabetes Mellitus Amongst The People Of Vidarbha Region Of India". IOSR Journal Of Pharmacy, (e)-ISSN: 2250-3013, (p)-ISSN: 2319-4219 Volume 6, Issue 1 (January 2016), PP. 45-50.

[33] Mustafa Z. Mahmoud, Omer A. Mahmoud, Maram A. Fagiri - "Chronic renal failure secondary to diabetes mellitus." Int. J Case Rep Images 2017;8 (2):124–128.

[34] Dabla PK.- "Renal function in diabetic nephropathy," World Journal Diabetes 2010 May 15;1(2):48–56.

[35] Remuzzi G, Schieppati A, Ruggenenti P. Clinical practice, "Nephropathy in patients with type 2 diabetes", International Journal of Data Mining 2012 Apr 11;346(15):1145–51.

[36] P. Pawelczak, R. Venkatesha –"American Heart Association, Heart diseases, and Stroke Statistics," IEEE International Conference on Machine Learning, Cape Town, South Africa, pp. 1-5, 2010 2009.

[37] Stokes, J., W.B. Kannel, P.A. Wolf, R.B. D'Agostino, and L.A. Cupples, 1989. "Blood pressure as a risk factor for cardiovascular disease: The Framingham Study 30 years of follow-up, Hypertension", IEEE International Conference on hypertension patient 31: I13-I18.

[38] Ersin Elbasi, "Determination of Diabetic Patient's Hearing Sensitivity using Data Mining Techniques," IEEE Transaction on Evolutionary Computation, Special Issue on Artificial Immune System, Volume 6, Issue 3, pp. 239-251, 2012.

[39] Ebrahim Darvishi, "Prediction of diabetes hearing loss patients using machine learning approach," JNCI Journal of National Cancer Inst., Volume 99, Issue 4, pp.268-289, 2017.

[40] Xie, J., & Wang, C., "Classification of Skin Disease using Ensemble Data Mining Techniques," International Research Journal of Engineering and Technology, 2(8), 1544-1547., 2017.

[41] Michelle Duff, "Cutaneous Manifestations of Diabetes Mellitus," In 2017 IEEE Imaging Systems and Techniques (IST) (pp. 1-5).

[42] Platt, John C. "12 fast training of support vector machines using minimal sequential optimization." Advances in kernel methods (1999): 185-208.

[43] Cortes, C., Vapnik, V., "Support-vector networks," Machine Learning, 20(2), pp. 273-297, 1995.

[44] Christopher J.C. Burges. "A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery," Springer, 2(2), pp.121-167, 1998.

[45] V. Vapnik, "The Nature of Statistical Learning Theory." NY: Springer- Verlag. 1995.

John, George H., and Pat Langley. "Estimating continuous distributions in Bayesian classifiers." Proceedings of the Eleventh Conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 1995.

[46] Zhang, H. (2004). "The optimality of Naive Bayes." International conference artificial intelligence, 1(2), 3.

[47] Aha, David W., Dennis Kibler, and Marc K. Albert. "Instance-based learning algorithms." Machine learning 6.1 (1991): 37-66.

[48] Alpaydin, E. (1997), "Voting over Multiple Condensed Nearest Neighbors," Artificial Intelligence Review, p. 115–132.

[49] Jiang L, Li C, Cai Z. "Learning the decision tree for ranking. Knowledge and Information Systems" in IEEE INFOCOM, pp. 400-406, 2002

[50] Ross Quinlan (1993). "C4.5: Programs for Machine Learning". Morgan Kaufmann Publishers, San Mateo, CA.

[51] Michelle Duff, "Cutaneous Manifestations of Diabetes Mellitus," In 2017 IEEE Imaging Systems and Techniques (IST) (pp. 1-5).

[52] Witten, I. H. et al. (1999). "Weka: Practical machine learning tools and techniques with Java implementations."

[53] K. Ogurtsovaa, J.D. da Rocha Fernandesa, Y. Huanga, U. Linnenkampa, L. Guariguataa, N.H. Choa,b, D.Cavana, J.E. Shawc, L.E. Makaroff "IDF Diabetes Atlas: Global estimates for the prevalence of diabetes for 2015 and 2040". http://dx.doi.org/10.1016/j.diabres.2017.03.024 0168-8227/2017 Elsevier B.V.

[55] World Health Organization, "Definition, diagnosis, and classification of diabetes mellitus and its complications. Report of a WHO Consultation. Geneva: WHO," World Health Organization, pp. 1-66, 2016.

[56] Sarker, I.H. Context-aware rule learning from smartphone data: survey, challenges, and future directions. *J Big Data* **6,** 95 (2019). https://doi.org/10.1186/s40537-019-0258-4.

[57] Sarker, I.H., Colman, A. & Han, J. RecencyMiner: mining recency-based personalized behavior from contextual smartphone data. *J Big Data* **6,** 49 (2019). https://doi.org/10.1186/s40537-019-0211-6

[58] Sarker, Iqbal H. et al. “Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage.” *Journal of Big Data* 6 (2019): 1-28.

[59] I. H. Sarker, A. Colman, J. Han, A. I. Khan, Y. B. Abushark, and K. Salah, “BehavDT: A Behavioral Decision Tree Learning to Build User-Centric Context-Aware Predictive Model,” Mobile Networks and Applications, 2019.