Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 6702 – 6712
Received 05 March 2021; Accepted 01 April 2021.

Data Mining Classifier for Predicting Diabetics

Dr. Manpreet Singh[1], Dr. Pankaj Bhambri[2], Dr. Inderjit Singh[3], Dr. Amit Jain[4], Er. Kirandeep Kaur[5]
[1,2] *Department of Information Technology, Guru Nanak Dev Engineering College, Ludhiana, Punjab, 141006, India*
[3,4] *Department of Computer Science & Engineering, Guru Nanak Dev Engineering College, Ludhiana, Punjab, 141006, India*
[5]*Tekki Web Solutions Pvt. Ltd, Ludhiana, Punjab, 141116, India*
[1]*mpreet@gndec.ac.in*, [2]*pkbhambri@gndec.ac.in*, [3]*inderjit26@gndec.ac.in*, [4]*amitjain_17@live.com*, [5]*kirandeep.gne@gmail.com*

Abstract— **Diabetes mellitus is very serious health problem today. If the disease is not identified and treated it can cause very dangerous health hazards. It is very important to predict the diabetic condition to control it through proper medical treatment. There are various hardware and software methods used to predict and classify the data. Various classification algorithms are used to classify the status of a person that whether a person is diabetic or not using various physical and biochemical characteristics. Different methods have different techniques and varying accuracy. In the present work a method for classifying the diabetic data is developed. It is the KNN classifier with six neighbor instance to classify any data record. KNN is lazy learning algorithm which classifies the data based on the majority of neighbors. For the better and accurate classification purpose normalization technique is used to normalize the data. Normalization is a pre processing technique that normalizes the domain value of any variable between 0 and 1. PIMA Indian dataset is used in this study, which include 768 records and 9 parameters. After classifying the data set using KNN algorithm, Cross validation method is used to compute the accuracy of the algorithm. The present work is compared with various existing method and it shows a significant enhancement in result by providing 100% accuracy which calculated through cross validation method.**

***Index Terms*—Biodiversity, Biological cells, Biological information theory, Genetic Programming, Genomics**

I. DIABETES

Diabetes is the malfunction of human body when process of converting glucose into energy is not carried out properly. Glucose is used by body cell for producing energy. The insulin is responsible for transporting glucose from blood into cell a special kind of hormone called insulin is required. This is produced by beta cells in pancreas. During diabetes the level of insulin is not efficiently produced in proper amount or the amount of insulin is not utilized properly [5]. The treatment of diabetes is must and the ignorance of this can result in various health hazards like kidney failure, heart problem, high blood pressure and blindness [6].diabetes can be classified in two categories:

- Type 1 diabetes mellitus: This type of diabetes can occur in any age but the probability and frequency is more in childhood and young. The reason behind the fact is that pancreas production of beta cell is disturbed and this leads to the deficiency of insulin production. The common symptoms are polyuria, polydipsia etc [6].

- Type 2 diabetes mellitus: This the most commonly found in the people over 40 years. This occurs because of rise in glucose level and our body is not capable of efficiently utilizing the insulin produced. This is also called as insulin resistance and hyperglycemia [7]. In human beings, the level of blood glucose is maintained with the help of pancreas which secrets two types of hormones which are named as insulin and glucagon. There are two types of cells present in pancreas which are known as alpha cells and beta cells. Beta cells are responsible for the recreation of Insulin is secreted by pancreatic beta cells whereas alpha cells are responsible for the secretions of glucagon. The concentration of the blood glucose is reduced by the beta cells and the increase in concentration of blood glucose is due to alpha cells. In type 1 diabetes, beta cells of the pancreas are damaged which lowers the production or secretions of insulin. 5–15% out of 366 million patients are diagnosed with type 1 diabetes, and also the type 1 diabetes holds the growth rate of 3.9% annually, which is making type 1 diabetes as a chronic disease.

Insulin-replacement therapy is done by injecting more than one injections everyday or using a infusion pump which continuously infuses insulin (under the skin). The closed and precise check of glucose concentration is must for a person which is suffering from type 1 diabetes. If the concentration of blood glucose is disturbed or increased then this can lead to the various health hazards which can create various long term health problems such as heart disease, blindness, kidney failure, and nerve damage. Level of glucose is generally control is using the glycated hemoglobin level (HbA1c), which is value and gives an index related to the mean blood glucose concentration during last three months. If the level of HbA1c is high then it provides that mean blood glucose concentration is also high which can represent the risk of long-term complications. The level of HbA1c should be kept below 7% for the person

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 6702 – 6712
Received 05 March 2021; Accepted 01 April 2021.

suffering from type 1 diabetes. The insulin is infused into the body in two ways through the infusion pump. One is called basal delivery while other is called bolus delivery. Basal delivery is done by the prior programming of a infusion pump, which supply the insulin continuously throughout the day according to need at a variable rate. Basal delivery supply insulin to the body require in between the daily diets schedules and also during the night and the amount of insulin supplied can be different from one patient to another depending on the condition and state of the illness. Bolus delivery supply insulin in bulk at the time of diet which allows the patient body manipulates the glucose which is taken during the diet. In this boluses are calculated prior taking the food with the help of amount of carbohydrate taken in the meal and the patient's insulin-to-carbohydrate ratio (a ratio that specifies how many grams of carbohydrate are covered by each unit of insulin). Different insulin-to-carbohydrate ratios are usually used for breakfast, lunch, and dinner [9]. Diabetes mellitus results in a notifiable difference in voice attributes. Diabetes mellitus is the disturbance or state which results in the high level of blood sugar in the human body for a long interval of time. If glucose wants to get into a cell it needs the help of insulin which is a special type of hormone produced by the pancreas. The disease of diabetes occurs when the human pancreas is unable to produce the appropriate amount of insulin or even stops producing the insulin hormone and it can also happen that the amount of insulin produced is not efficiently utilize by the insulin. These all type of conditions are referred as insulin resistance. Generally diabetes is categorized into three types. First one is known as Type 1 diabetes which generally stars in the childhood age and the human pancreas in this state is damaged which fails to produce the appropriate amount of insulin. The reason for this can be the disorder of the beta cells in the human pancreas which are responsible for producing the insulin. The Second Type of diabetes is known as type 2 diabetes. In this state then there is the lack of insulin and thus the metabolic disorder is located by the hyperglycemia in context of insulin blockage or the lower production of insulin. Overall 90% of the total diabetic case is of this type. It is a life threatening disease which can create number of hazards and complications such as cardiovascular diseases, blindness, kidney failure etc. Third type of diabetes is the state in which women who does not diagnosed with diabetes in the past has high level of blood glucose levels during the period of pregnancy. This can occur due to the presence of human placental lactogen that interact with susceptible insulin receptors and causes the rise of blood sugar levels in the body of women who is having pregnancy [10].

Diabetes mellitus can be diagnosed invasively or non- invasively. But these two methods have some drawbacks like patients preparation, piercing of skin which can result in various infections and demands the presence of skilled technicians etc. So due to this reason the second method is mostly adopted to diagnose the diabetes in the diabetic person.. Diabetes mellitus can be diagnosed through a non-invasively technique with the analysis of human facial block color features with a sparse representation classifier. The diabetes can also be detected form the hair element as the hair element level in the diabetic patient is significantly varying as compared to the healthy person.

II. CLASSIFICATION, NORMALIZATION AND KNN

As reviewed from the different aspects and different point of view the classification is having a significant importance in the analysis and other research fields. The data collected is generally of heterogeneous and have different format and different representation and also the set of operation that can be performed is different. Data classification is the effort made to classify or to assign the classes by clustering the collected data into a homogeneous group which is having the common characteristics. The classification is done depending on the fact that what kind of analysis is to be performed on the classified data. There is a classifier which is used to represent the classification criteria, rules, and mathematical expression used for the purpose of classification. The classification process is carried out in the context to study the nature and scope of the data. As the classified sets of the data is related to some specific task which is to be performed on that particular data and hence in this way the nature and behavior of the data can be studied easily and efficiently and a meaningful prediction can be made. The classification task also provides the assist for data cleansing and to enhance the data quality by removing the various anomalies in the homogeneous group after the classification. The various kinds of data cleansing operation like removal of redundant, irrelevant content can be performed to obtain a precise and concise data [6].

Normalization is one of the most important parts of the presented work. Generally normalization refers to the transformation of the given data into a desirable range of values. When the range of values of a given parameters of dataset differs significantly, then it is very difficult to proceed with the correct analysis and hence it is impossible to incur the correct and some meaningful facts. In terms of machine learning it is very difficult for the algorithm to learn when the values in the domain of any variable or parameter differ with a significant gap. Whereas the normalized data helps to carry out the perfect and accurate machine learning. There are various normalization techniques available to normalize the data but technique used in this work is generally known as minimum and maximum normalization which is also recognized as feature scaling the advantage of normalization is that the values of the data set is transformed between 0 and 1. Each and every value of the dataset is normalized and hence the corresponding normalized value is calculated. The newly normalized value is denoted with Z, which is given by:

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 6702 – 6712
Received 05 March 2021; Accepted 01 April 2021.

$$Z = \frac{x - min(X)}{max(X) - min(X)} \quad (1)$$

So in the given equation Z is the newly normalized value which lies between 0 and 1, X is the actual value and min(X) is the minimum value in the domain range and max(X) is the maximum value in the domain range.

The KNN is a classifying algorithm that is used to classify the given data set into different classes based on the previously supplied data to the algorithm. Whenever the algorithm is used to classify any record into any class, the voting is done and the neighbor of the records are included in the voting and the class which has majority of neighbor is assigned to the record. The variable K in this algorithm specifies the number of neighbor that should be taken into the account in order to classify any instance of data set. The neighbor in the algorithm is generally taken into account on the basis of distance and that distances is generally called Euclidean distance. The accuracy and the performance of the algorithm generally depend on the value of K, which defines the number of neighbor which will decide the assign class to a given data set. The optional value of K generally lies in the range of 3-10. The Euclidean distance is given by:

$$E.D = \sqrt{\sum_{i=}^{k} (x_i - y_i)^2} \quad (2)$$

III. R SOFTWARE

R is statistical analyzing software languages which also carry out reporting and graphical representation of the processed or manipulated content and comes with variety of packages which helps to carry out the analyzing task directed towards a specific region or a specific field of interest [1]. It is a open source tool which is adapted by variety of platform like windows, Linux, Macintosh and provides an extension to users to extend and enhance this package software up to the level which can be accepted by the global community of users [2]. All the fact and discussion is grooming around data. This software gives a versatile scope of data analyzing which is basically application oriented .This is a flexible tool which is now a day's used by the data scientists to reaches a new level of research and to stroke out the hidden and unallocated facts about that particular research field. R provides the handful dataset or we can say data formats which can be used and manipulated as desired by the tool operator and these data formats are vector, list, data frame, matrix along with the basic formats like integers, logical, complex and characters. The work done in the R software is comparative to session based environment, so the list of commands which we use or execute can be recalled at any time but when the workspace is closed then the session also terminates unless we save it. The command and the comment line in this software package have their specific syntax following which we can successfully execute any command. The important part is that the command line of this software is case sensitive, hence the upper and lowercase has their own meaning and task [2].

It provides various types of statistical computing like time series analysis , linear and non-linear modeling, and graphical representation and manipulation and also provide the platform to the data mining and data scientists to carry out the research task and infer new facts from the existing content[1]. It provide an interface to the online database by importing that database or part or small chunk of that database and to carry out the analysis work and hence enables to manipulate the data in desired way and to represent the output in desired form and to and formulate the new results [2].

It provides the set of packages which can be installed by setting crane mirror and hence the user can use the routines of that package. It also provides the extension platform to the users to build up their own packages and also to debug the existing one. These packages allow the users to use the routines and functions of in order to carry out those particular computations. It provides the graphical assist to manipulate or to analyze the set of data and to represent the processed data set in graphical form like graphs and scatter plots. It also provides the set of tools for analyzing the data for a specific purpose. This software also provides a window to interface with other programming languages like java, c, python etc and hence also enables to integrate the packages written in any other language in to R [2].

The R software provides the biological data computations and also provide the interface to the online data by importing the biological data into R data frames and other datasets and enables to find out the solutions for the various biological problems and hence inferring some significant results make the scope of this Software package as an important tool in this region of research [8]. In Finance, this software provides various tools for analyzing the financial data like time series and stock exchange and we can access, visualize, and optimize the data related to financial aspects. The R software almost covers up the every field of operation which includes marketing, big data analysis, drug enhancement, production and manufacturing, sports and health science, graphics, data mining and many other field which can be focused for analyzing through R software for different purpose.

IV. LITERATURE REVIEW

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 6702 – 6712
Received 05 March 2021; Accepted 01 April 2021.

Reference [11] tends to optimize the diagnosis of any kind of disease using novel processing technique and k nearest neighbor classifier by using machine learning. The process of machine learning is decomposed into three major parts named as data collection, pre processing diagnosis. In the given work various experiments were done using UCI machine learning repository diabetes database which was collect from PIMA Indian heritage which consist of 768 samples. Data pre-processing technique are very helpful to optimize the output which is directly proportional to the quantity and quality of the dataset which is going to serve as input to the model. K nearest neighbor is very efficient and easy machine learning algorithm which classifies the data based on previously stored data. The basic logic behind the working is to assign a class to any new case based on the majority of neighbor. The observation part in this work signifies that proposed work shows 100% accuracy which was evaluated through K fold cross validation. Hence the model tends to be a effective assistant to any medical expert which could use it to optimally diagnose the chronic disease and can refer the patient for the corresponding medical treatment.

Reference [12] proposed the system which ascertains the measurement of various body parts such as heart, kidney, and eye maladies and so on. Firstly, the author collected dataset, then Eclat theorem algorithm is connected which compute control and uncontrolled condition of diabetes. After applying algorithm Eclat frequent pattern are carried thereby the pattern range comparison is done with standard dataset forming prediction of severity results on organ. In which data is collected from PIMA Indian diabetes data set that includes different attributes like age, sex, BMI or test results of diabetes like patient is diabetic or not. The total accurate measurement and attributes are carried out from recent test papers. Next the ranges were recognized by standard range for diabetes ECLAT algorithm. Eclat is an equivalence class clustering and bottoms up cross section traversal algorithm. It needs lesser space than apriority item set are little in number so, it is appropriate for little dataset and also requires less for frequent pattern era. In the end we can say that if the patient suffers from diabetes from previous year then Eclat algorithm easily calculates control and uncontrolled condition of diabetes as well as the measurement of severity on an organ is calculated.

Reference [13] implemented the random forest algorithm. It includes a lot of freely decision tree and also satisfied with their estimates by averaging under great completion in setting that requires number of variables is much larger in size comparatively than observation. It utilizes Bootstrap due to statistical theory there by freely re sampling method to extract multiple versions of the sample set from original training data set that makes a decision tree model for each ones. Mainly there are some steps are required for prediction using Bootstrap re sampling technique. it establish k samples which are cover 2/3 part of the real data set and rest of the data is known as OOB that is out of bag which can be used as test data for feature space. Further, k sample create k decision tree with lots of nodes of each ones and also mathematically steps must be followed until the minimum correction for the OOB data set to be obtained. So, in last we can say to predict the test samples using some techniques with majority rule voting mechanism. However, these samples easily indicates whether or not patient have diabetes. So it can greatly helpful for medical industry for diagnosis.

Reference [14] shows that BPNN is advanced and accurate architecture for detection and prediction of diabetes as well as it is user friendly software tool which is built in MATLAB under GUI which acts as a media between patient and doctor who tackle this software for better classification. Secondly, in the absence of doctor it can easily tells the assistant also whether someone is diabetic or not within just few seconds. So, it is beneficial for further treatment to anyone who are indulge with diabetic. This network consist of three layers in which first one having an input with 8 parameters and one layer is hidden with 10 neurons and one output is generated to produce good outcome. Then the csv file contains different parameters for classified into groups having same features which used as input to diagnosis. Theses 8 parameters are loaded by GUI trained with the help of BPNN and clearly output is displayed. Basically, there is additional option is provided for single and multiple patients. So, here is text box on GUI where the doctor can get results of any patient just by entering patient serial number.

Reference [15] shows an efficient and simple detection of classification for diabetic patients have been presented. The proposed applied weighted classifier forecast a recently developed applied Manhattan distance formula for partition which identifies different parameters on the behalf of its priority level such as low, high and medium risk. It is clearly noted that 403 record. In which 120 low risk patient, 114 medium risk patient and 44 at medium risk patient. The performance evolution of the classifier depends on these parameters up to great extent.

Reference [16] proposed a novel breath analysis system. The gadget is convenient, quick and simple to work and also it uses commercial sensors to distinguish acetone in human breath. Lots of experiments were done with novel breath test for both inpatient and outpatient approved the exactness of the framework. However, the sensor cluster is mainly key point of our framework as we can say that fully experiment depends on the ability of the sensor. It seven commercial sensors were used to built the device which were selected based on their performance in various experiments then it simulates with acetone by grouping precision. Other several sensors are proved to have higher significance. At the end the outcomes demonstrate the framework was capable to recognize healthy and diabetes samples with straight forward element extraction and classification algorithm. This framework could contribute a

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 6702 – 6712
Received 05 March 2021; Accepted 01 April 2021.

stage toward a down to earth framework that we can also infer from the outcome that breath samples controls diabetic more similar to healthy people whereas samples of uncontrolled diabetic patient have clearly big difference from the test of healthy ones.

Reference [17] shows a dangerous disease which can be responsible for various chronic complications and also it can increase significantly among the individual of different age so, health care have been using data mining techniques such as decision tree induction to detect and predict risk factors related to health issues as well as to follow some rules and measures to cure chronic disease. It makes easy for the medical expert to understand the complications and to follow the require treatment procedure with the help of available data. This technique depends on input that can be fetch carefully along with pre-processing of the same for the purpose of collect and efficient further estimation process so, decision tree is much better as compare to statistical methods because it shows higher significant and accurate results. This model works in two half in the first one the model tends to make effects to predict whether a person can develop a comorbibty whereas in the second half it works on to predict that what kind of comborbibty it could be. The result part signifies that regression technique were good over decision tree. In any case sampling and cost analysis investigation methodologies show sign improvement. Various experiments were preformed on a collection of 14162 T2DM patients during the year 2009 to 2012 and the observation made was that 3459 from the total were comorbibty patients. The model shows a significant figure of accuracy which was noted 87% in the initial phase of the Meta classifier model, whereas the accuracy in the last phase tends to be 68%.

Reference [18] presented the hidden pattern from medical data. The work tends to optimize the amalgam demonstrate for characterizing PIDD (Pima Indian Diabetic Database). In the model K mean collaborates with KNN to carryout multi step pre-processing. The proposed model carryout effect to enhance the quality of data by eliminating noisy data in order the enhance the working of algorithm and to increase accuracy and efficiency of the algorithm. In order to root out the incorrectly classifier cases K mean clustering was implementing whereas, missing values were adjusted using mean and medians. After obtaining a clear and correct dataset through various pre-processing KNN algorithms was applied. The value of K for the algorithm is decided by the quality of data. The noise effect is eliminated up to great extent. In the classification process if the value of large. The objective of the model was to accurately and efficiently classify the PIMA Indian diabetic database through amalagam KNN algorithm by defining optimal value of K and using various pre- processing techniques along with different values of k. It was observed that for the greater value of K the specified model tends to approach the accuracy of 97.4% and the result was compared with simple KNN and cascade K mean and KNN for the same value of K and the classification accuracy of the specified model was significantly good as compared to other model.

Reference [19] presented the particular point of care testing gadget for analysis of diabetes mellitus. It depends optical investigation which permits high sensitivity as compare to various available electrochemical detection POCT DM devices and also it a compact device with little dimension which make it portable and easily opera table and the cost of production for the device was also low. All these features make this device a very significant tool for the monitoring of diabetes. Since they have high sensitivity low per test cost and are impacted by outer elements for example temperature and pH variation. So, the device tends to be optimal hardware for diabetic monitoring. The POCT device development enables consistent and accurate diagnose of diabetes by using spectrophotometric analysis by optical absorption. However, due to significant reduction in cost of electronic components. The device tends to the more practical for the field application. The result obtained from the device depicts efficient detection linearity. So, the device was a perfect and optimal instrument for the purpose of continues and connotative determination of glucose concentration along with the significant accuracy while determining the glucose concentration the standard deviation for each case was less than 10%. Hence, it makes sure that the developed device tends to be the most useful hardware.

Reference [20] proposed to develop a connection among diabetes and caveolae by the mean of picture analysis technique. caveolae are the specialized membrane micro domain made out of shorts of less. The caveolae picture consists of various attributes which includes high noise, low complexity and uneven brightness. A large portion of all caveolae is extremely tiny and contrasting tissues. In this way, the hybrid image processing techniques were optimizing. The pre- processing part consist of eliminating noise and increasing the contrast, eliminating islets and cleaning tiny fragments in the picture, predicting the edges the with dynamic adaptive cally algorithm, using mathematical technique to extract integrated caveolae, defining the caveolae with help of master driven technique. Highlight the features of caveolae for representation and description. So, the conclusion signifies that the model capable to address the issue of high noisy and non uniform back round problem. The implementation of the model have to depict relationship between diabetes and the number of caveolae and afterword data mining technique was used to establish co relation between diabetes and caveolin.

Reference [21] suggested a decision support system for diagnosing diabetes. The system was composed by utilizing net beans 7.1's GUI features and was created by input parameter rule bases, computer data base and primary

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 6702 – 6712
Received 05 March 2021; Accepted 01 April 2021.

indication by establishing special computational algorithm by rule base. It uses MySQL server to design and implement database. The database of the model is composed of the parameters which are the various risk factors, sign and symptoms related to the diabetes and also the physical attributes of patient. The nature of the model is user friendly which is obtained by menu driven approach. In order to predict the result and to make a decision about the illness, level if its risk, the model ask some question to the user who intract with the model. The user is needed to answer the question either in yes or no based on which the decision is found. The observation made in this model conclude that special computational algorithm by rule base (SCRAB) is used to predict diabetes with some define set of input rule base and asking some question to user regarding physical attribute and of some basic sign and symptoms. The model tends to be a cost effective development as eliminates the expenditure of various medical test, reports and also it was a user friendly interactive system.

Reference [22] introduced a new strategy to indentify diabetes which relays on the method of estimation of urine smell by utilizing an electronic nose. Glucose was mixed into a pure urine sample for the purpose of making artificial urine in order to simulate the condition of diabetic patient. E-nose consists of eight commercial gas sensors which act as sensing elements for e-nose. In order to analysis the data accurately principle component analysis PCA and cluster analysis CA were used. Principle component analysis PCA helps to recognize the pattern which defines the level of glucose in urine. The various samples were tested under varying temperature ranging from 35 to 45 degree Celsius. The temperature change affects the evaporation rate of the solution which effects sensor response. Further, to analysis the distinction of PCA points. In each experiment cluster analysis (CA) was perform. The cluster analysis conforms that sample measured with increased temperature increases the ability of electronic nodes to make prediction about the disease. The result of these analyses defines the concentration of glucose in the urine. The contracted e-nose is capable of measuring urine order by applying the temperature than >45 degree Celsius. The conclusion of the proposed work was that urine generally contains very less amount of glucose but the people suffering from high blood sugar have amount of glucose greater than the required limit in their urine. This is because of the diabetes. The order of urine with high glucose is also different as compare to the normal person urine. So, the developed e-nose can be a useful and important tool to diagnose diabetic nature of any patient.

Reference [23] attempted to locate the parameters causing development of diabetes up to five years after gestational diabetes mellitus (GDM) and building a prediction model. To do so (OGTT) was performed five years after GDM, which is a75-g oral glucose tolerance test which was performed on 362 women, but the women already found affected with diabetes were not considered. All but 21 women had results from follow-up at 1–2 years, while 84 women were lost from that point. Predictive parameters were located with the help of logistic regression analysis and it was found that Five years after GDM, 28/362 women (8 %) were affected with diabetes and 187/362 (52 %) had normal glucose tolerance (NGT). Of the latter, 139/187 (74 %) also had NGT at 1- to 2-year follow-up. In simple regression analysis, using NGT at 1–2 years and at 5 years as the reference, diabetes at 1- to 2-year follow-up or later was clearly associated with easily assessable clinical variables, such as BMI at 1- to 2-year follow-up, 2-h OGTT glucose concentration during pregnancy, and non-European origin ($P < 0.0001$). A prediction model based on these variables resulting in 86 % correct classifications, with an area under the receiver-operating characteristic curve of 0.91 (95 % CI 0.86–0.95), was applied in a function-sheet line diagram illustrating the individual effect of weight on diabetes risk. Thus the results highlight the importance of BMI as a potentially modifiable risk factor for diabetes after GDM. Our proposed prediction model performed well, and should encourage validation in other populations in future studies.

Reference [7] summarized that Environmental and genetic parameters are responsible for the causing any illness or body malfunctioned disease. A genetic disorder in a disease is caused by abnormalities in an individual's genetic material (genome). It is of great importance to know how these genetic factor cause diseases. Differential gene expression analysis is one of the most significant method that helps for the study of genetic factors causing diseases. A method for recognizing differentially expressed genes which are responsible for Type-2 diabetes mellitus using micro array data for diabetes with parental history and healthy is presented. The objective of this method is to find out multivariate and univariate outliers using a mathematical terms such as Mahalanobis Distance, Minimum Co-variance Determinant (MCD) and other useful statistical approaches. The given approach is implemented on micro array data collected from two different samples. First one is diabetes with parental history and the second from healthy. By this method 1579 genes which can be expressed differentially was located. Before doing analysis the collected data from the samples used to build up a micro array was normalized by a normalization technique called Loess Normalization. After the systematic implementation of this procedure 1579 differentially expressed genes were recognized from 39400 genes which were analyzed between healthy and diabetic with parental history. The advance classification of these identified genes can be done, which is the part of future research in order to identify the candidate genes which are responsible for causing type-2 diabetes.

Reference [8] discussed that Now a day diabetes is one of the most critical and threatening health issue which is faced by majority of people and hence if appropriate and efficient effort for controlling and curing this problem is not

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 6702 – 6712
Received 05 March 2021; Accepted 01 April 2021.

taken then it can cause various health complexion in the social being society. Various methods to detect or to diagnose this problem are adopted but breath analysis is one of the efficient methods that provide a simple, straight, accurate and valuable approach which provides a descriptive clinical care for the disease. This journal analyzes the concentration of acetone levels in breath for monitoring blood glucose levels and in this way prediction for diabetes can be made. This approach relay upon support vector mechanism in order to identify and classify normal and diabetic samples. In order to accomplish the task ten samples of acetone levels are collected which are classified into healthy, type 1 diabetic and type 2 diabetic. This paper represents the effort made to enhance the previous classification approach (healthy and diabetic) to a more descriptive level for the efficient prediction (healthy, type 1 and type 2 diabetic). In future efforts can be made to design and develop the sensory array for analyzing concentrations of acetone in breath samples.

Reference [24] analyzed that Blood glucose dynamics is one of the most core and necessary part in the control and management of diabetes. Continuous monitoring of the blood glucose enables a automated analyze of blood glucose dynamics. This journal present the consequences of the combined time-frequency analysis for short-time periods applied to time series of the blood glucose for diabetes which depends on insulin. These approaches help to extract the structural component, the pattern of blood glucose dynamics and help to identify the nature of glucose levels. Blood glucose is a important factor having a valuable clinical temporal pattern. Wigner-Ville distribution is an appropriate method to analyze temporal change of blood glucose. These outcomes are very helpful to diagnose diabetes mellitus, in blood glucose and help to manage the problem with a efficient and optimal medical treatment.

Reference [25] mentioned that Data mining is widely used in variety of application and in bioinformatics it has a great impact in analyzing biomedical data. Rapid-I's Rapid Miner is tool which provides the methods to analyze a Pima Indians Diabetes Data Set which gathers facts and detail of individual with and without developing diabetes. Various types of processing of data prior to the implementation including attribute identification and selection, outlier removal, data normalization and numerical discretization, visual data analysis, hidden relationships discovery, and a diabetes prediction model construction is done which is the core objective.

Reference [26] proposed a new method for the purpose of predicting the risk of type-2 diabetes, which tends to be a cost sensitive learning. The data set used in this model was collected from Pima Indian data set and another one was of Tabriz from Iran. After implementing the proposed method, its comparison was done with various other cost sensitive methods. In this method the major objective was to optimize the feature cost and misclassification cost. In the proposed work the cost for instance which was truly predicted, was taken as zero, whereas, false negative and false positive tends to have different ratios. For the purpose of comparing this method with various other methods, the evaluation criteria used was 5-fold cross validation. Unit variation and zero mean were used in order to normalize the feature. On the other hand any missing value was adjusted by filling it through the average of 5 close values. Missing neighbor values was adjusted by utilizing the K nearest neighbor. In order to compare the proposed method with various other methods, the criteria of comparison selected was cost per person. When compared using described criteria, the method tends to be the most effective and efficient and was better in major set of conditions from other methods, when different types of cost matrix was considered.

Reference [27] performed the effort to predict the likelihood state of diabetes by analyzing the various physical parameters and other lifestyle activities of a person. The various lifestyle activities that were included in the work are such as eating habits, sleeping habits, body mass index and waist circumference. The mathematical calculations which were carried out to establish a relation between diabetes and these lifestyle activities are chi square test, CART (classification and regression tree) and machine learning was also performed on the proposed data and also the method was evaluated using cross validation. The data set used in the work was collected by making survey and asking various questions through questionnaires, which were prepared by taking the help of doctors. Two types of questionnaires were prepared for diabetic and non diabetic nature respectively. The observation part signifies that when CART model was implemented for the purpose of making valuable prediction, the model approaches the accuracy of 75% and indicates that the various lifestyle activities like blood pressure, junk food, late sleeping and other genetic factor are responsible for diabetes.

Reference [28] proposed the method in which PIMA Indian data set was used with newly proposed two models named as KNN and Fuzzy KNN along with Fuzzy c-mean clustering was introduced, which clusterise the data and afterwards KNN classification approach was applied , whereas in the second approach, Fuzzy c-mean clustering was clubbed with Fuzzy KNN classification model. By analyzing both the models, it was found that the second model was more precise as compared to the first one. The number of neighbors decided in both the models was 3. When the model was tested on PIMA Indian dataset, it showed a tremendous figure of accuracy, which was the evaluation criterion for the models. The observations made were that KNN classification model along with Fuzzy c-mean clustering shows 97.02%accuracy, whereas the second model, which was Fuzzy KNN classification model along with Fuzzy c-mean clustering, shows 99.25% accuracy.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 6702 – 6712
Received 05 March 2021; Accepted 01 April 2021.

V. PROBLEM FORMULATION

ADiabetes is the malfunction of human body when process of converting glucose into energy is not carried out properly. Glucose is used by body cell for producing energy. The insulin is responsible for transporting glucose from blood into cell a special kind of hormone called insulin is required. This is produced by beta cells in pancreas. During diabetes the level of insulin is not efficiently produced in proper amount or the amount of insulin is not utilized properly [5]. The treatment of diabetes is must and the ignorance of this can result in various health hazards like kidney failure, heart problem, high blood pressure and blindness. In human beings, the level of blood glucose is maintained with the help of pancreas which secrets two types of hormones which are named as insulin and glucagon. There are two types of cells present in pancreas which are known as alpha cells and beta cells. Beta cells are responsible for the secretion of Insulin is secreted by pancreatic beta cells whereas alpha cells are responsible for the secretion of glucagon. The concentration of the blood glucose is reduced by the beta cells and the increase in concentration of blood glucose is due to alpha cells. In type-1 diabetes, beta cells of the pancreas are damaged which lowers the production or secretion of insulin. 5–15% out of 366 million patients is diagnosed with type-1 diabetes, and also the type-1 diabetes holds the growth rate of 3.9% annually, which is making type-1 diabetes as a chronic disease. The analysis is carried out in the context to study the nature and scope of the data. As the classified sets of the data is related to some specific task which is to be performed on that particular data and hence in this way the nature and behavior of the data can be studied easily and efficiently and a meaningful prediction can be made.

VI. METHODOLOGY

The methodology is the presented work is to development and algorithm which can classify the diabetic data and should have a significant accuracy. For this purpose a strong bioinformatics and data analysis tool “R” is used. R provides a handful set of commands and methods to carry out the analysis and to perform desirable transformation on the given data. R is window interfere software which is used in wide range of domains and is very flexible and user-friendly. Now a day’s R is widely used in many common and complex tasks like data analysis, statics, bioinformatics, and mathematics and in other graphical interpretation application. R comes with role of packages which includes various functions and methods which can perform different tasks. The desired package can be downloaded and installed in order to use the functions of that package. R is a open source software which also enables the user to develop his own package and publish it. The flexibility of the tool can be defined as it provides wide range of data type such as vector, data frame, fasta format files and CSV files.

- Data Pre-Processing: The data set used in the presented work is collected from UCI repository which consists of 768 records which are clinically tested and have different type’s parameters which are related to the physical and biological factors of human. During the pre-processing of this data set the data set is analyzed from different point of view and the attempts are made to establish some meaningful relation between the different parameters of the data set. In order to ensure the correct analysis, the data set is checked for any null values and it is also made sure that all the parameters are in supported data type and in the correct domain. Pre-processing also includes the study of the structure of the data which provides the minimum and maximum value of any parameters and also provides total number of instances of the data set. And after that data set is summarized which provides the detailed information above the data set.
- Normalization: Normalization is one of the most important parts of the presented work. Generally normalization refers to the transformation of the given data into a desirable range of values. When the range of values of a given parameters of dataset differs significantly, then it is very difficult to proceed with the correct analysis and hence it is impossible to incur the correct and some meaningful facts. In terms of machine learning it is very difficult for the algorithm to learn when the values in the domain of any variable or parameter differ with a significant gap. Whereas the normalized data helps to carry out the perfect and accurate machine learning. There are various normalization techniques available to normalize the data but technique used in this work is generally known as minimum and maximum normalization which is also recognized as feature scaling the advantage of normalization is that the values of the data set is transformed between 0 and 1. Each and every value of the dataset is normalized and hence the corresponding normalized value is calculated. The newly normalized value is denoted with Z, which is given by:

$$Z = \frac{x - min(X)}{max(X) - min(X)} \quad (3)$$

Table I. PARAMETER SPECIFICATION TABLE

Colounm name	Data Type	Description
Pregnant	INTEGER	This field describes that how many times a woman conceive pregnancy
Plasma glucose	INTEGER	This field describes the plasma glucose level of the person
Diastolic BP	INTEGER	This field describes the diastolic blood pressure of a person
Triceps fold	INTEGER	This field describes the triceps skin fold of a person
Insulin	INTEGER	This field describes the insulin level of person

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 6702 – 6712
Received 05 March 2021; Accepted 01 April 2021.

Bmass Index	NUMBER	This field describes the body mass index of a person
Pedigere fxn	NUMBER	This field describes the pedigree analysis of a person
Age	INTEGER	This field describes the age of a person
Class	INTEGER	This field describe that whether a person is diabetic or not

- Data Partitioning: The process of machine learning is initializing by training the algorithm and then applies the algorithmic computations in order to test the algorithm for classifying the data. After obtaining the normalized data set, the next phase is to separate the data or to make partition of the data into training set and testing set. The data consist of total 768 instances. In order to generate training and testing set, 65% of the records are reserved for the training purpose, which makes 500 instances of the total 768 records whereas 35% records are kept in testing set, which makes 269 record of the total. So, 768 records are splitter into two parts containing 500 records in the testing part and 269 records in the training part.
- Applying KNN Algorithm: The KNN is a classifying algorithm that is used to classify the given data set into different classes based on the previously supplied data to the algorithm. Whenever the algorithm is used to classify any record into any class, the voting is done and the neighbor of the records are included in the voting and the class which has majority of neighbor is assigned to the record. The variable K in this algorithm specifies the number of neighbor that should be taken into the account in order to classify any instance of data set. The neighbor in the algorithm is generally taken into account on the basis of distance and that distances is generally called Euclidean distance. The accuracy and the performance of the algorithm generally depend on the value of K, which defines the number of neighbor which will decide the assign class to a given data set. The optional value of K generally lies in the range of 3-10.

$$E.D = \sqrt{\sum_{i=}^{k}(x_i - y_i)^2} \quad (4)$$

- Cross Validation: Cross validation is the method to find out the accuracy of the applied algorithm. In the present work the accuracy of the algorithm is calculated by the cross table method which is a built in function in R software. In this method the input given to the function is the classified data by the algorithm and the actual data. This function calculates the accuracy of the algorithm on the basis of four variables. These variables are known as true negative, true positive, false negative and false positive. This method calculate all the true positives, all the true negatives, all the false positives and all the false negatives and based on these parameters provide the accuracy of the algorithm. The equation to calculate the accuracy is given by:

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \quad (5)$$

Table II. CROSS VALIDATION TABLE

TEST SET	PREDICT SET	
0	True Negative 0	False Negative 1
1	True Negative 1	False Positive 0

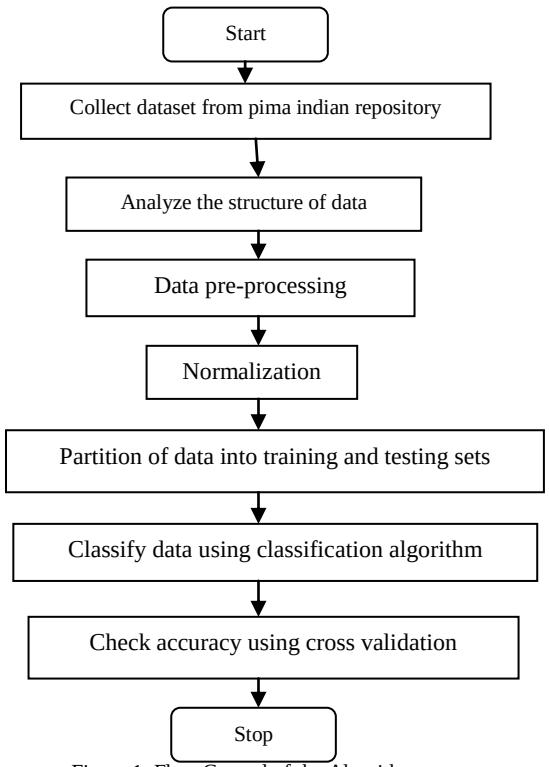

Figure 1. Flow Control of the Algorithm

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 6702 – 6712
Received 05 March 2021; Accepted 01 April 2021.

- Calculating memory: After performing the alignment process the next task is to calculate the amount of memory taken by the algorithm to perform the alignment process. For this purpose a built in function in R software is used to compute the amount of memory taken. Name of the function is object.size(), this function takes the alignment function as an argument and returns the memory used by any process to complete its task. Generally the unit of memory return by the function is byte by default and it also enables to manipulate the output unit into desired format.
- Calculating time: As specified in the objectives time bound is also an important factor in determining the efficiency of the algorithm. Hence after analyzing the memory requirements the next goal is to calculate and analyze time required by the algorithm to accomplish the task of sequence alignment. For the purpose of calculating time taken by the algorithm to align the sequences in built function of R software is used. The name of the function is system.time(). This function takes the object as argument and calculates the time taken by the object to complete its task. Actually this function further call a function named as proc.time() which calculate the time taken by any process to complete its task and return the CPU time to the user by which the execution time of any function in R software can be calculated.

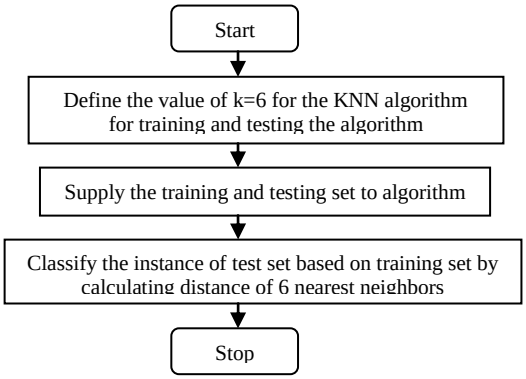

Figure 2. Working of KNN Algorithm

VII. RESULT AND DISCUSSION

The part of the work elaborates the result of the whole work or project. R software is used to build up the working model and the main agenda was to build up a model that can classify the diabetic data. The data set was collected from Puma Indian Repository, which provides the clinical record of the patient along with their various physical and biological characteristics which are having some kind of relation with the diabetes and hence we have used this data set as input to our model. This data set is a CSV file which is given as input. The KNN models used to predict the diabetic nature of any instances. For this purpose we take into account 6 neighbors and further more to implement the proposed model we partition the given data set into training and the testing part and after training the model we test the model on the test set and it classify the data on the basis of training provided to it and then we check the accuracy of the algorithm through cross validation and it noted that all the instance that were to classified correctly and gives us the 100% accuracy of our algorithm. Afterwards the proposed model was compared with other state of arts and various comparison factors were taken into account and the proposed model was found to be significantly better. Normalization of the data was done in order to supply an accurate and appropriate data for algorithm to make efficient learning. The feature scaling normalization was used in this model, which re-transform the variable into the range the variable into the range of 0 and 1 and this helps the machine learning to efficiently learn and then to predict the new instance correctly based on the prior learning.

TABLE III. COMPARISON WITH VARIOUS OTHER TECHNIQUES

Method	Accuracy	Reference
Proposed Normalization +KNN=6	100%	Current study
K-mean + ANN	99.20%	
ANN + K-mean, k=2	97.45	
K-mean + DT	93.33%	Karegowda *et al* (2012b)
K-mean + KNN, k=5	96.68%	Karegowda *et al* (2012a)
ARTMAP-IC	81.0%	Carpenter and Markuzon (1998)
GRNN	80.21%	Kayaer and Yildirim (2003)
MLNN-GDA	77.60%	Kayaer and Yildirim (2003)
MLNN-GDA-NN	76.56%	Kayaer and Yildirim (2003)
MLNN-GDA+ADLR	77.60%	Kayaer and Yildirim (2003)
MLNN-LM	77.08%	Kayaer and Yildirim (2003)

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 6702 – 6712
Received 05 March 2021; Accepted 01 April 2021.

MLNN-LM	82.37%	Temurtas, Yumusak and Termutas (2009)
PNN	78.13%	Temurtas *et al*, (2009)
CBNN-GDA	81.0%	Aibinu, Salami and Shafie, (2010)
CBNN-CAR	81.28%	Aibinu, Salami and Shafie, (2011)
PCA-ANFIS	89.47%	Polat and Gunes 2007
LS-SNM	78.1%	Polat, Gunes and Aslan (2008)
GDA-LS-SNM	79.16%	Polat, Gunes and Aslan (2008)

VIII. CONCLUSION AND FUTURE SCOPE

In this work, the model for predicting the diabetic and non-diabetic nature of person is designed. Diabetes is a chronic disease which can cause very critical health hazards and because of this, diagnose of this disease is very important. KNN is a machine learning model which can classify the data based on training provided to it. So, in this model KNN is used to predict or to classify the diabetic nature of any person. To build up the proposed model, R software, which is very strong statistical analysis tool and also provides a number of bioinformatics and other application is used to R software provides a versatile applications and very flexible working environment. The software offers a numbers of packages which can be used for dealing with different kind of data and to achieve different objectives. So in this work KNN provides the appropriate functionally which is required to classify the given dataset. The data set consist of 768 instances and then after the partition of the whole data set was done and 500 instances were taken as training set and 268 instances were taken as testing set. Firstly the algorithm was trained with the help of training set and then it was tested on the on the testing set and it was noted that the algorithm is having a efficient performance and provides a good accuracy by classifying all the instances of the testing set correctly and hence provides the 100% accuracy which signifies that the algorithm goes through a proper training session by considering 6 neighbor into account and on the basis of that all the test set instances were classified correctly.

As the present work is used to make prediction about diabetic and non-diabetic data with a good and significant accuracy, the future scope of this work is to extend this model to a higher platform and to integrate the prescription model which can also advise the valuable control measure in order to control the disease and can also help the patient to get rid of it. So the next level of this proposed work is to build a complete working model which can diagnose and can also provide valuable control measure for the diabetes control without involving any medical expert.

ACKNOWLEDGMENT

This work is executed for the completion of Dissertation Work of Masters of Technology in Computer Science and Engineering. Facilities available at Guru Nanak Dev Engineering College were used for the submission of Dissertation to Inder Kumar Gujral Punjab Technical University, Kapurthala, Punjab, India. Dr. Manpreet Singh and Dr. Pankaj Bhambri are the supervisors for this research work.