Journal of Computer Sciences and Applications, 2022, Vol. 10, No. 1, 6-15
Available online at http://pubs.sciepub.com/jcsa/10/1/2
Published by Science and Education Publishing
DOI:10.12691/jcsa-10-1-2

A Review of Diabetes Datasets

Muhammad Mika'ilu Yabo[1,*], Ahamed Baita Garko[2,3], Abubakar Atiku Muslim[2], Hassan Umar Suru[2]

[1]Department of Computer Science, Shehu Shagari College of Education, Sokoto State, Nigeria
[2]Department of Computer Science, Kebbi State University of Science and Technology, Aliero, Nigeria
[3]Department Computer Science Federal University Dutse, Jigawa State, Nigeria
*Corresponding author: MYMIKAILU@GMAIL.COM

Received September 05, 2022; Revised October 07, 2022; Accepted October 16, 2022

Abstract Many intelligent healthcare systems have been developed to diagnose human diseases such as breast cancer, hepatitis, diabetes and heart diseases. Diabetes is a lifelong chronic disease that occurs when the pancreas does not produce enough insulin (Type I diabetes mellitus), or when the body's produced insulin is unable to be utilised properly (Type II diabetes mellitus), Researches that are carried out on diabetes using data mining techniques were done to predict type II diabetes mellitus using different diabetes datasets by different researchers; Pima Indians Diabetes Dataset (PIDD) is used by the majority of the researchers. The dataset (PIDD) has eight (8) attributes which limits more exploration in the field of Machine Learning (ML) for diabetes prediction. Diabetes prediction is limited because of the few attributes available in the diabetes datasets used, and these attributes play important roles in predicting diabetes mellitus types, classes and risk factors whenever a diabetes patient is diagnosed. This paper provides a systematic review of diabetes mellitus datasets, identifying the strength and weakness of the 8 attributes described in the PIDD, which is used by the most of the researchers. Furthermore, this paper has identified the need of the potential researchers in the research community to address the gap by enhancing the existing diabetes dataset attributes with additional attributes, identify the attributes required for the prediction of glucose level, diabetes Types, diabetes classes, diabetes risk factors and to develop a Model that can be used for the prediction.

Keywords: *data mining technique, healthcare systems, diabetes mellitus datasets, diabetes dataset attributes*

Cite This Article: Muhammad Mika'ilu Yabo, Ahamed Baita Garko, Abubakar Atiku Muslim, and Hassan Umar Suru, "A Review of Diabetes Datasets." *Journal of Computer Sciences and Applications*, vol. 10, no. 1 (2022): 6-15. doi: 10.12691/jcsa-10-1-2.

1. Introduction

Health informatics is a rapidly expanding field concerned with the application of computer science and information technology to medical and health data. With the ageing population in developing and developed countries, as well as the rising cost of healthcare, governments and large health organisations are becoming increasingly interested in the potentials of Health Informatics to save time, money, and human lives; however, as a relatively new field, Health Informatics does not yet have a universally accepted definition [1]. The applications of Healthcare Informatics in clinical care decision-making, is the design, development and deployment of machine using data mining techniques that can help healthcare professionals to make effective and efficient clinical decisions" [1].

Data Mining technique is among the most versatile techniques that have received a warm response in private organisations**,** government**,** enterprises and healthcare, it is mainly used in hospital for big data analytics and interpreting to smoothening the workflow of hospital management by helping health personal to serve patients better, and Surgeons to gain an insights to carry out the operation accurately with a great precision [2]. Data Mining Technology is this technology combines multiple disciplines such as statistics, probability, machine learning, and artificial intelligence to look for patterns and trends in massive amounts of data by utilising sophisticated mathematical algorithms to segment the data and assess the likelihood of future events [3,4]. Typically, the Data Mining process can detect the patient's disease with great precision by uncovering the hidden information contained in the medical data gathered; this process entails a number of processes of working via interactive and iterative data sequences for determining the primary symptoms of communicable disease/ non-communicable diseases and then treats the patient well [2]. Non-Communicable Diseases (NCDs) which include stroke, heart disease, cancer, chronic lung cancer and diabetes are responsible for almost 70% of the deaths worldwide in which diabetes is the most common disease among them and the number of patients suffering from diabetes has quadrupled since 1980 [5].

2. Diabetes

Diabetes is a lifelong illness that occurs as a result of lack of insulin hormone or ineffectiveness of insulin hormone. Nowadays, diabetes is becoming one of the most serious diseases, which its frequency keep on increasing in the world and varies from one community to another based on age, gender, race, dietary habits, genetic characteristics and environmental factors [3,6]. Insulin as a hormone transports glucose to the bodycells from bloodstream to be used as energy, if this energy not efficiently consumed by the bodycells, the excess can cause major health issues [7]. There are two main types of diabetes that is Type I and Type II. According to health experts, diabetes occurs when the human body's gland called the pancreas cannot produce enough insulin (Type I diabetes), and the produced insulin cannot be used by the cell of the body (Type 2 diabetes) [6,8,9].

2.1. Type I Diabetes

Type I diabetes was previously known as Insulin-Dependent Diabetes Mellitus. This Type of diabetes can be formed at any age but needs to be diagnosed for those that are below 20 years of age. This Type of diabetes is formed when insulin-producing cells or beta cells in the pancreas get destroyed [10]. Type I diabetes is caused by a damage to pancreatic and insulin-producing beta cells at the end of an autoimmune process, or due to unfamiliar disorders [6]. In general, 5-10% of diabetes cases in the community constitute cases of Type I diabetes [8]. Patients with Type I have insulin deficiency; they must take insulin hormone from outside for life [6,8,9].

2.2. Type II Diabetes

Type II diabetes was known as Non-Insulin-Dependent Diabetes Mellitus as it was diagnosed in patients above 20 years of age [10]; or is one of the most common metabolic illnesses and is characterised by a deficiency in the generation of insulin secretion via the pancreatic islet β-cells [11]. Type II diabetes is generally associated with obesity and physical immobility. At the basis of the disease, genetically predisposed individuals have lifestyle-related insulin resistance and decreased insulin secretion over time [6]. More than 90% of diabetes cases diagnosed worldwide is type II diabetes [7].

3. Related Research Work

Breault [12], in his research work, Pima Indian Diabetic Dataset (PIDD) was used to test the developed model using data mining algorithms (naïve Bayes classifier) to accurately predict diabetic patient status of been positive or negative, where out of 392 complete cases, guessing all are non-diabetic gives an accuracy of 65.1%. The main objective of the research conducted by Parthiban et al. [13], was to predict the chances of the diabetic patient getting heart disease, the researchers applied Naïve Bayes technique, which produces an optimal prediction model, they proposed a system, which predicts attributes such as age, sex, blood pressure and blood sugar and the chances of a diabetic patient getting a heart disease, the dataset used in Parthiban et al. [13] work was a diabetic clinical dataset collected from Chennai of about 500 patients. Padmaja [14] aimed at find out the characteristics that determine the presence of diabetes and to track the maximum number of women who have diabetes. The researcher used clustering and attributes oriented induction techniques to track the characteristics of the women who have from diabetes using PIDD. Rajesh and Sangeetha [15], applied data mining techniques to classify diabetes clinical data and predict the likelihood of a patient who has diabetes or not, using the Decision Tree Algorithm (DTA) and PIDD.

Rahim [16] developed a preliminary classification and screening system using SVM algorithm for diabetic retinopathy, with the use of “Eye Fungus Images” diabetes dataset by focusing on the detection of the earliest signs of diabetic retinopathy, which are the microaneurysms. Hina et al. [5] carried out a research, using PIDD dataset and preprocessed it to a more meaningful structure; the data mining tool opted for the research was WEKA. In the research, different classifying algorithms such as Naïve Bayes, Multi-Layer Perceptron, Decision Tree, ZeroR, Random Forest, and Regression were applied to predict the chances of a diabetic patient being positive or negative. Neilesh and Gandhi [17] used a new feature selection method along with SVM in developing a model, in their work, the feature selection method was considered one of the best methods to improve the prediction accuracy of diabetes patient for having positive or negative, Pima Indian Diabetes Dataset was taken from the UCI Repository was used to test and train the developed Model. Vijayan and Anjali [18] developed a decision support system that used the different base classifiers (SVM, NB, Decision Stump, DT and AdaBoost algorithms), PIDD was used to test and train the developed system. Miss and Megha [19], conducted research where Back Propagation Neural Network (BPNN), and Graphical User Interface (GUI) were built using MATLAB, Pima Indian Diabetes Dataset was used by the researchers to test their proposed methodology. Mohebbi et al. [20] developed an approach using deep learning for detection of Type 2 diabetes patient status using Continue Glucose Monitoring (CGM) signals dataset collected from 9 patients. Machine Learning Techniques such as Logistic Regression, Multi-Layer Perceptron, and Convolutional Neural Network were applied in the research, the dataset was divided into training and testing dataset, with 1 to 6 patients CGM signals used as a training dataset, and 7 to 9 patients CGM signals were used as a test dataset.

Francesco et al., [21] used Pima Indian Diabetic Dataset and WEKA tool to predict diabetic patients status of being positive or negative using different machine learning techniques such as Decision Tree, JRip, Multilayer Perceptron, Random Forest, HoeffdingTree, and BayesNet. Maham et al., [22], in their research, Pima Indian Diabetes Dataset was used to test their developed Model for the classification of diabetic patients as positive or negative using Multilayer Perceptron technique. Wenqian et al., [23] used Pima Indian Diabetes Dataset, K-means for data reduction and Decision Tree as a classifier to predict the status of diabetic patient. The graph-based approach was developed by Mangrulkar [24] where retinal image are classify by

dividing retinal vessels in to two types as arteries and veins. In his research, the use of retinal vessels extracted for vascular changes detection is the most important phase. The decision tree algorithm having a 10-fold cross-validation method and K-means algorithm that was developed using WEKA is used by the researcher. The patient's retinal image was used to find artery vein ratio. Sidong et al., [25] in their research, five different techniques of machine learning were used for diabetes diagnosis and preprocessing of data, and those techniques include DNN, Logistic Regression, Decision Tree, SVM, and Naïve Bayes, Pima Indian Diabetic Dataset was used to calculate the accuracy of cross-validation. A diabetes prediction model with dropout was developed by Ashiquzzaman [26], for diabetes prediction using deep learning neural networks that had a fully connected layer plus dropout layers. Pima Indian Diabetic Dataset was used to train and test the proposed Model. Deepti and Dilip [27] developed a model having three different machine learning algorithms, those machine learning algorithms include Decision Tree, SVM, and Naïve Bayes, for diabetes status prediction of the target class of 1 as positive and 0 as negative, Pima Indian Diabetic Dataset was used to train and test the proposed Model. A data mining techniques for the prediction of Type 2 diabetes mellitus was developed by Han et al., [28] using Pima Indian Diabetes Dataset to test their proposed Model for predicting status of diabetic patients by reducing dataset complexity and analysing the medical implication of every attribute and their correlation with diabetes mellitus. Safial and Islam [29] developed a model using deep neural network with five-fold cross-validation and ten-fold cross-validation to diagnose diabetes using Pima Indian Diabetes Dataset. Ayon and Islam [30] developed a strategy for diagnosis of diabetes using a deep neural network by training its attributes in a five-fold and ten-fold cross-validation fashion, Pima Indian Diabetes (PID) dataset used in the research for the prediction of diabetes patient status. Naz and Ahuja [31] presents a methodology for diabetes prediction using PIMA dataset. The researchers use Decision Tree (DT), Naive Bayes (NB), Artificial Neural Network (ANN), and Deep Learning (DL) techniques for the prediction of diabetes patient status. Alshammari et al., [11] created a machine learning model that is capable of predicting diabetes with high performance, the researchers used the BigML platform to train four machine learning algorithms, namely, Deepnet, Decision Tree, Ensemble and Logistic Regression using collected dataset from Ministry of National Guard Hospital Affairs (MNGHA) at Saudi Arabia from 2013 to 2015, and the dataset attribute is for tested adult patients for Hemoglobin A1c (HgbA1c). In their research, HgbA1c result is used to determine the patient's diabetic status where patient is classified as diabetic if HgbA1c value is grater or equal to seven, and patient is classified as non-diabetic if HgbA1c value is less than seven.

Bhoia et al., [32] developed a diabetes prediction model of females in Pima Indians heritage, using a binary classification problem with Pima Indians Diabetic Dataset, supervised learning techniques such as Support Vector Machine (SVM), classification tree (CT), Naïve Bayes (NB), k-Nearest Neighbour (k-NN), AdaBoost (AB), Random Forest (RF), Logistic Regression (LR), and Neural Network (NN) have been used in the research. The researchers use k-fold cross-validation to carry out the process of training and testing. Islam et al., [33] in their research, they use a type II diabetes dataset taken from Bangladesh Demographic and Health Survey, 2011 containing 1569 respondents where 127 respondents are diabetes. For diabetes risk factors prediction, Six ML-based supervised classifiers as support vector machine, random forest, linear discriminant analysis, logistic regression, k-nearest neighbourhood, bagged classification and regression tree (Bagged CART) have been adopted in the research. Manimaran &Vanitha [10] in their research work, Maldives Bureau of Statistics (MBS) dataset was used, the dataset was collected from various districts to predict diabetes Disease using Data Mining Classification Techniques such as Multilayer Perceptron, Bayesian networks, Decision tree, and Fuzzy Lattice Reasoning (FLR), the dataset contains 1024 complete instances with 26 Parameters, and the data was gathered from answers to Questionnaires given during the research work. The main objective of the questionnaire was to converse on a set of parameters for the diagnosis of diabetes risk factors in patients. Alpan & İlgi [34]; Chaves & Marques [3], in their work, propose a comparative analysis of data mining techniques for early diabetes diagnosis, that is to predict only diabetes risk factors based on the attributes available in the dataset, both researchers use a publicly accessible dataset of Sylhet Diabetes Hospital, Portugal containing 520 instances, 17 attributes. Alpan & İlgi [34] used WEKA tool, Bayesian Network, Naïve Bayes, Random Tree, Random Forest, k-NN, SVM techniques. The Results Obtained by the researchers indicated that k-NN performed the highest accuracy with 98.07% and this algorithm is the best method to identify and classify diabetes diseases on the studies dataset., while Chaves & Marques [3] uses six classification algorithms, namely k-nearest neighbours (kNN), Naive Bayes, Random Forest, Neural Network, AdaBoost, and Support Vector Machine (SVM), the Results Obtained by the researchers indicated that Neural Networks is good for diabetes prediction as the Model presents a Specificity of 97.5%., an Area Under the Curve (AUC) of 98.3%, F1-Score, Sensitivity and Precision of 98.4% and accuracy of 98.1%.

Table 1. Summary of the related literature to diabetes prediction showing strengths and weaknesses

S_n	Authors	Techniques	Dataset	Tools / Languages	Strengths	Weaknesses
1	[14]	Clustering and Attribute Oriented Induction Techniques	Pima Indian Diabetes Dataset	-	This research aimed at finding out the characteristics that determine the presence of diabetes and tracking the maximum number of women who have diabetes.	The research is restricted to only predicting a positive or negative of type II diabetes without predicting Type I, classes of type II diabetes or predicting risk factors that lead to diabetes.

S_n	Authors	Techniques	Dataset	Tools / Languages	Strengths	Weaknesses
2	[12]	Naïve Bayes classifier	Pima Indian Diabetes Dataset	ROSETTA software	The researchers tested data mining algorithms to predict the accuracy in predicting diabetic status from the 8 variables given. out of 392 complete cases, with the accuracy of predicting the diabetic status of 82.6% on the initial random sample.	Limited to only predicting a positive or negative of type II diabetes without predicting Type I, and classes of type II diabetes or predicting risk factors that lead to diabetes.
3	[13]	Naïve Bayes	Chennai Records of Diabetic Patients.	WEKA	The main objective of their research is to predict the chances of the diabetic patient getting heart disease, that is Class H of Type II Diabetes Mellitus	It predicts only Class H of Type II Diabetes Mellitus, The prediction of Type I, Other Classes of Type II and Risk factors of having Type I or Type II diabetes were not addressed.
4	[13]	SVM, NB, and DT algorithms	Pima Indian Diabetes Dataset	-	They have applied data mining techniques to classify Diabetes Clinical data and predict the likelihood of a patient suffering with diabetes or not	The research is restricted to only predicting a positive or negative of type II diabetes without predicting Type I, classes of type II diabetes or predicting risk factors that lead to diabetes.
5	[17]	Feature Selection, SVM	Pima Indian Diabetes Dataset	-	The researchers improved the prediction efficiency, and reduced complexity for feature selection using K-means clustering and F-score	Limited to only predicting a positive or negative of type II diabetes without predicting Type I, and classes of type II diabetes or predicting risk factors that lead to diabetes.
6	[18]	SVM, NB, Decision Stump, DT and AdaBoost algorithms	Pima Indian Diabetes Dataset	MATLAB and WEKA	They developed a decision support system that used the different base classifiers to predict diabetes. Missing values in the dataset were fulfilled by replacing them with the mean value.	Uses PIDD dataset with just 8 attributes therefore, the research is restricted to only predicting a positive or negative of type II diabetes without predicting Type I, and the classes of type II diabetes or predicting risk factors that lead to diabetes.
7	[16]	SVM	dataset of eye Fungus images	-	Developed classification and screening system for Predicting diabetic retinopathy (Class R of Type II) by detection of microaneurysms at an early stage	It predicts only Class R of Type II Diabetes Mellitus, The prediction of Type I, Other Classes of Type II and Risk factors of having Type I or Types II diabetes were not addressed.
8	[19]	Back Propagation Neural Network (BPNN)	Pima Indian Diabetes Dataset	MATLAB	The researchers developed a methodology that can be used for predicting diabetic patients' status of positive or negative.	Limited to only predicting a positive or negative of type II diabetes without predicting Type I, and the classes of type II diabetes or predicting risk factors that lead to diabetes.
9	[5]	Naïve Bayes, MLP, Decision Tree, ZeroR, Random Forest, and Regression	Pima Indian Diabetes Dataset	WEKA	They developed an intelligent model built with comprehensive results by altering the dataset to a more meaningful structure, applied Machine Learning Techniques to depict the chances of a diabetic patient to be positive or negative.	Limited to only predicting a positive or negative of type II diabetes without predicting Type I, and the classes of type II diabetes or predicting risk factors that lead to diabetes.
10	[20]	Logistic Regression, Multi-Layer Perceptron, and Convolutional Neural Network algorithms	CGM Signals	MATLAB	The researchers developed a deep learning approach for the detection of Type 2 diabetes, using CGM signals collected from 9 patients as a dataset.	Limited to only predicting a positive or negative of type II diabetes without predicting Type I, and the classes of type II diabetes or predicting risk factors that lead to diabetes.

S_n	Authors	Techniques	Dataset	Tools / Languages	Strengths	Weaknesses
11	[21]	Decision Tree, Multilayer Perceptron, Random Forest, HoeffdingTree, and BayesNet.	Pima Indian Diabetes Dataset	WEKA	The researchers developed a model to predict the status of diabetic patients and diabetes diagnosis using different machine learning techniques	Uses PIDD dataset with just 8 attributes therefore, the research is restricted to only predict a positive or negative of type II diabetes without predicting Type I, and the classes of type II diabetes or predicting risk factors that lead to diabetes.
12	[22]	Multi-Layer Perceptron (MLP)	Pima Indian Diabetes Dataset	WEKA	They developed a model for predicting diabetes, and the preprocessing of the diabetes dataset was done by detecting outliers with the help of enhanced class outlier-based methods	Limited to only predicting a positive or negative of type II diabetes without predicting Type I, and the classes of type II diabetes or predicting risk factors that lead to diabetes.
13	[23]	Decision Tree (DT) classifier.	Pima Indian Diabetes Dataset	WEKA	They developed a model for predicting diabetes, using a Decision tree algorithm having a 10-fold cross-validation	Uses PIDD dataset with just 8 attributes therefore, the research is restricted to only predict a positive or negative of type II diabetes without predicting Type I, and the classes of type II diabetes or predicting risk factors that lead to diabetes.
14	[24]	The graph-based approach	Scanned Retinal Images	MATLAB	The researcher developed a model that extracts retinal vessels to detect vascular changes, that was used to calculate the artery vein ratio. Artery-to-vein ratio was used to determine Diabetes status	It predicts only Class R of Type II Diabetes Mellitus, The prediction of Type I, Other Classes of Type II and Risk factors of having Type I or Types II diabetes were not addressed.
15	[25]	DNN, Logistic Regression, Decision Tree, SVM, and Naïve Bayes	Pima Indian Diabetic Dataset	-	In their research, five different techniques of machine learning were used for diabetes diagnosis, prediction and preprocessing of data, by calculating the accuracy of cross-validation.	Limited to only predicting a positive or negative of type II diabetes without predicting Type I, and the classes of type II diabetes or predicting risk factors that lead to diabetes.
16	[26]	Deep Learning Neural Network	Pima Indian Diabetes Dataset	Python	They developed a prediction system with dropout for diabetes prediction by training and testing the proposed Model.	Uses PIDD dataset with just 8 attributes therefore, the research is restricted to only predict a positive or negative of type II diabetes without predicting Type I, and the classes of type II diabetes or predicting risk factors that lead to diabetes.
17	[27]	Decision Tree, SVM, and Naïve Bayes	Pima Indian Diabetes Dataset.	WEKA	They Developed a model having three different machine learning algorithms for diabetes prediction.	Limited to only predicting a positive or negative of type II diabetes without predicting Type I, and the classes of type II diabetes or predicting risk factors that lead to diabetes.
18	[28]	Improved K-means and Decision Tree as a classifier.	Pima Indian Diabetes Dataset	WEKA	They developed a model that was used to reduce dataset complexity and analysed the medical implication of every attribute along with the correlation of having Positive or Negative diabetes mellitus.	Uses PIDD dataset with just 8 attributes therefore, the research is restricted to only predict a positive or negative of type II diabetes without predicting Type I, and the classes of type II diabetes or predicting risk factors that lead to diabetes.

S_n	Authors	Techniques	Dataset	Tools / Languages	Strengths	Weaknesses
19	[29],	Deep Learning Neural Network with five-fold cross-validation and ten-fold cross-validation	Pima Indian Diabetes Dataset	Python	A Pima Indian Diabetes Dataset was used to predict diabetes.	Limited to only predicting a positive or negative of type II diabetes without predicting Type I, and the classes of type II diabetes or predicting risk factors that lead to diabetes.
20	[30]	Deep Learning Neural Network with five-fold cross-validation and ten-fold cross-validation	Pima Indian Diabetes (PID) dataset	-	They developed a strategy for diagnosis of diabetes and prediction of diabetes with prediction accuracy.	Uses PIDD dataset with just 8 attributes therefore, the research is restricted to only predict a positive or negative of type II diabetes without predicting Type I, and the classes of type II diabetes or predicting risk factors that lead to diabetes.
21	[31]	Naive Bayes (NB), Decision Tree (DT) (Artificial Neural Network (ANN), and Deep Learning (DL)	Pima Indians diabetic dataset (PIDD)	-	They present a methodology for diabetes prediction using a diverse machine learning algorithm that resulted to the range of 90–98% of accuracy).	Limited to only predicting a positive or negative of type II diabetes without predicting Type I, and the classes of type II diabetes or predicting risk factors that lead to diabetes.
22	[11]	Deepnet, Decision Tree, Ensemble and Logistic Regression	Collected from the Ministry of National Guard Hospital Affairs (MNGHA) in Saudi Arabia	-	Recall, Precision, Accuracy, F-measure and PhiCoefficient) were examined in the research; the dataset attribute is for adult patients who had tested for Hemoglobin A1c (HgbA1c) and labelling patients as diabetic relied on the results of the HgbA1c.	Limited to only predicting a positive or negative of type II diabetes without predicting Type I, and the classes of type II diabetes or predicting risk factors that lead to diabetes.
23	[9]	DT, KNN, RF, NB, AB, LR, SVM	Pima Indian Diabetes (PID) dataset	WEKA	They developed a methodology that can be used to predict diabetes, for feature selection, they dropped three features and used five input features (Glucose, BMI, Insulin, Pregnancy, and Age) and one output feature (outcome).	Uses PIDD dataset with just 8 attributes therefore, the research is restricted to only predict a positive or negative of type II diabetes without predicting Type I, and the classes of type II diabetes or predicting risk factors that lead to diabetes.
24	[32]	Decision tree (DT), Support Vector Machine (SVM), k-Nearest Neighbour (k-NN), Naïve Bayes (NB), Random Forest (RF), Neural Network (NN), AdaBoost (AB) and Logistic Regression (LR).	Pima Indian Diabetes (PID) dataset	-	The k-fold cross-validation was used by the researchers for training and testing, to determine the results of area under the curve (AUC), F1, precision, classification accuracy (CA), and recall of used supervised learning algorithms and determine the suitable algorithm for diabetes prediction	Limited to only predicting a positive or negative of type II diabetes without predicting Type I, and the classes of type II diabetes or predicting risk factors that lead to diabetes.
25	[33]	Support Vector Machine, Random Forest, Linear Discriminant Analysis, Logistic Regression, K-Nearest Neighborhood, bagged classification and regression tree.	Bangladesh demographic and health survey diabetes dataset	-	The researchers uses two statistical tests as independent, t for continuous and chi-square for categorical variables to determine patients diabetes status.	Limited to only predicting a diabetes patient status without predicting Type I, and the classes of Type II diabetes or predicting risk factors that lead to diabetes.
26	[10]	Multilayer Perceptron, Bayesian networks, Decision tree, and Fuzzy Lattice Reasoning (FLR)	MVS dataset	-	The researchers propose a methodology that can be used to predict risk factors of diabetes disease.	Limited to only predicting risk factors that lead to diabetes, but predicting a positive or negative of Type I and II diabetes, classes of Type II diabetes is not addressed.

S_n	Authors	Techniques	Dataset	Tools / Languages	Strengths	Weaknesses
27	[34]	Bayesian Network, Naïve Bayes, Random Tree, Random Forest, k-NN, SVM	Sylhet Diabetes Hospital Dataset, Portugal	WEKA	The researchers developed a model to predict risk factors of diabetes disease.	The research only identifies people with a higher risk of having diabetes, without predicting Type I, Type II and classes of Type II diabetes mellitus.
28	[3]	Naive Bayes, Neural Network, AdaBoost, k-nearest neighbours (kNN), Random Forest, and Support Vector Machine (SVM)	Sylhet Diabetes Hospital Dataset, Portugal	WEKA	The researchers propose a methodology to predict risk factors of diabetes disease.	Limited to only predicting risk factors that lead to diabetes, but predicting a positive or negative of Type I and II diabetes, classes of type II diabetes is not addressed.

This paper has reviewed 28 papers and identified that many researches were done using different diabetes datasets; and the majority of the researchers give more impasses on predicting type II diabetes mellitus, this happened because of the limitation of attributes available in the datasets used as described in Table 2 below.

Table 2. Summary of Some Diabetes Related Datasets used by the authors

S_N	Dataset	Number of Authors used the Dataset
4	Bangladesh Demographic and Health Survey, Dataset	1
	Chennai Records of Diabetic Patients	1
3	Continue Glucose Monitoring (CGM) signals dataset	1
	Eye Fungus images dataset	1
2	Maldives Bureau of Statistics (MVS) dataset	1
5	MNGHA Dataset	1
6	Pima Indian Diabetes Dataset (PIDD)	19
6	Scanned Retinal Images Dataset	1
7	Sylhet Diabetes Hospital Dataset, Portugal	2

The Table 2 above confirmed that the majority of the reviewed papers used the Pima Indian Diabetes Dataset (PIDD). Therefore, this paper will consider PIDD attributes and suggest how the attributes limitations can be improved so that diabetes prediction can be enhanced to predict not only diabetes status, but also to predict patient Type of diabetes, class of diabetes and associated risk factors. This PIDD is available online from the URL https://data.world/data-society/pima-indians-diabetes-database, where all patients are females and of at least 21 years old, with 768 instances, 8 attributes and 1 outcome [31,32,35].

Table 3. PIDD attributes with their descriptions [31,32,35]

S_N	PIDD Attributes	Attributes Description
1	Age	Age of participants
2	BMI	Body Mass Index (weight in kg/(height in m)^2)
3	Diabetes pedigree Function	An appealing attributed used in diabetes prognosis
4	Diastolic Blood pressure	It is when blood applies into arteries between heart)(mm Hg)
5	Glucose	Is a concentration of Plasma glucose which is 2hrs in an oral glucose tolerance test
6	Pregnancy	Number of times a participant is pregnant
7	Skin Thickness	Triceps skinfold thickness (mm). It concluded by the collagen content
8	Serum Insulin	2-Hour serum insulin (mu U/ml)
9	Outcome	Diabetes class variable, (Yes or 1) represent the patient is diabetic and (No or 0) represent patient is not diabetic

Table 3 above described the 8 attributes that are available in PIDD with their descriptions, as stated below:-

1) **Age:** This attribute shows patient's number of years, in numeric, of each instance, and the range of it value is from 21 to 81, where the average age value is 33 in PIDD.
2) **Pregnancies:** This attribute shows number of times a patient gets pregnant, in numeric, of each instance, and the range of it value is from 0 to 17, where the average pregnant value is 4.
3) **Glucose:** This attribute shows the level plasma glucose concentration or an Oral Glucose Tolerance Test result in 2 hours, and the range of it value is from 0 to 199, where the average is 121.
4) **Blood pressure:** This attribute shows the level of diastolic blood pressure in mm Hg, and the range of it value is from 0 to 122, where the average value is 69.
5) **Skin thickness**: This attribute shows the triceps skin thickness in mm, and the range of it value is from 0 to 99, where the average value is 21.

6) **Insulin:** This attribute shows the level of insulin, in numeric, and the range of it value is from 0 to 846, where the average value is 80.
7) **BMI:** This attribute shows body mass index in Kg/m^2 and the range of it value is from 0 to 67.1, where the average value is 32.
8) **Diabetes pedigree function:** This attribute shows the function scores of the likelihood of diabetes by inheritance, and the range of it value is from 0.078 to 2.42, where the range value is 0.47.
9) **Outcome:** This attribute is a class attribute, it shows the outcome of the prediction using either 0 and 1 or Yes and No, where 1 or Yes indicates diabetes patient is positive, while 0 0r No indicates the diabetes patient is negative.

The PIDD described all the 8 attributes presented in the dataset. However, there exist some strengths and weaknesses associated with the dataset. Table 4 below shows the PIDD strengths and weaknesses.

Table 4 provides and proves that:

1. Only one glucose attribute (Oral Glucose Tolerance Test result) is available in the PIDD, where the patient result of Oral Glucose Tolerance Test result will take minimum of 2 hours, therefore, there is a delay in predicting diabetic patient status.
2. Only patient age is captured in Age attribute, therefore, patient's years with diabetes need to be captured as additional attribute to further predict diabetes types and classes.
3. The PIDD is limited to only 4 risk factors attributes namely: BMI, Diabetes Pedigree Function, Diastolic Blood Pressure and Skin Thickness. Where there are many more important risk factors such as obesity that is not available in PIDD.
4. The PIDD is limited to only one class attribute, therefore no class attribute as outcome of predicting risk factors.
5. The Pregnancy attribute is clearly limiting the PIDD of having male instance.

Table 4. PIDD attributes showing strengths and weaknesses

S_N	PIDD Attributes	Strength	Weakness
1	Age	It contains an integer numeric value. No null attribute value of any instance in the dataset. It is one of the three most important attributes selected during feature selection for prediction of diabetes.	The attribute value range from 21 to 81, therefore, patient of age below 21 cannot be diagnosed using the dataset. Type I diabetes cannot be predicted using the dataset because Type I diabetes is for children below 17 years of age. Age attribute captured only patient age, while patient years of having diabetes is needed to predict diabetes class of patient whenever a patient is tested positive such as class B, class H and class R.
2	BMI	It contains a real numeric value. No null attribute value of any instance in the dataset. It is one of the three most important attributes selected during feature selection for prediction of diabetes.	BMI is a person's weight in kilograms divided by the square of height in meters; therefore, using Weight and Height as two additional attributes is better for the machine learning to compute BMI automatic to avoid using wrong value.
3	Diabetes Pedigree Function	It contains a real numeric value. No null attribute value of any instance in the dataset.	Is a genetic relationship of a patient to relatives with diabetes mellitus history, therefore, this attribute is more of risk factor attributes.
4	Diastolic Blood Pressure	It contains an integer numeric value.	This attribute is one of the least important factors in diabetes prediction. Some instances are having zero (0) value in the PIDD. The attribute is more of risk factor attribute.
5	Glucose	The glucose value in PIDD is an Oral Glucose Tolerance Test result obtained by combination of fasting glucose result and 2 hours blood glucose taken again after patient drink a liquid containing a certain amount of sugary, usually 75 grams.	Oral Glucose Tolerance Test result takes minimum of 2 hours before the glucose value is obtained
6	Pregnancy	This is a number of times that a diabetic patient gets pregnant	This attribute is one of the least important factors in diabetes prediction and this attribute indicated that PIDD contains only female instances.
7	Skin Thickness	This is a skinfold thickness that can be use to predict total amount of body fat of a patient	This attribute is one of the least important factors in diabetes prediction. Some of instances are having zero (0) value in the PIDD and the attribute is more of risk factor attributes.
8	Serum Insulin	This is a hormone that helps move blood sugar, known as glucose, from bloodstream into cells of a patient.	It takes minimum of 2 hours before the Serum Insulin result can be obtained. Some of instances are having zero (0) value in the PIDD, and also the attribute is less important when glucose result is available in predicting diabetes.
9	Outcome	This is a class attribute indicating either a diabetic patient is positive or is negative	PIDD is limited to only one class attribute with two returning values (1 for diabetic patient status is positive and 0 for diabetic patient status is negative).

4. Conclusion

There are many issues bedeviling research in the field of diabetes using Machine Learning. Part of the issues related to diabetes datasets is the lack of enough attributes or even the diabetes datasets that are available online for free download. This issue limits the way diabetes related researches progress in the field of Machine Learning. There exist some researches that played the use of the Pima Indian Diabetes Dataset (PIDD) where Type II diabetes are mostly considered using the available attributes presented in the dataset. Type I diabetes, classes of diabetes mellitus and diabetes risk factors as well requires additional experimentation in the field of Machine Learning. Applying machine learning techniques to medical dataset is a trending research area, because there are too many healthcare diseases that require urgent result in investigation with accurate and efficient prediction [4]. As it is discussed in the literature review, the majority of the diabetes datasets have only one class attribute. Therefore, the outcome of the prediction is always one. And for the non-class attributes, the attributes are limited to predicting only the status of the diabetic patient and no provision for predicting diabetes types, classes and risk factors. This paper identified some gaps that require attention from the research community. The identified gaps are:

1. To enhance the existing diabetes dataset attributes by providing the required additional attributes for diabetes prediction.
2. To identify attributes required for glucose prediction, diabetes types, classes prediction, and risk factors prediction.
3. To use a machine learning techniques to develop a model that can predict diabetes status, types of diabetes, classes of diabetes and diabetes risk factors.

References

[1] DelVecchio, A. (2019). Health informatics. https://searchhealthit.techtarget.com/definition/health-informatics.

[2] Azhar, F. (2020). Data Mining in Healthcare: Benefits, Techniques, and Prospects https://www.way2smile.ae/blog/data-mining-in-healthcare/.

[3] Chaves, L. & Marques, G. (2021) Data Mining Techniques for Early Diagnosis of Diabetes: A Comparative Study.

[4] Yusuf, A. B., Dima, R. M., & Aina, S. K. (2021). Optimized Breast Cancer Classification using Feature Selection and Outliers Detection. Journal of the Nigerian Society of Physical Sciences, 298-307.

[5] Hina, S., Shaikh, A., & AbulSattar, A. (2017). Analyzing Diabetes Datasets using Data Mining. Journal of Basic & Applied Sciences, 13, 466-471.

[6] Peker, M., Özkaraca, O., & Şaşar, A. (2018). Use of Orange Data Mining Toolbox for Data Analysis in Clinical Decision Making: The Diagnosis of Diabetes Disease.

[7] World Health Organization (2021) Diabetes. https://www.who.int/news-room/fact-sheets/detail/diabetes.

[8] Saeedi, P.; Petersohn, I.; Salpea, P.; Malanda, B.; Karuranga, S.; Unwin, N.; Colagiuri, S.; Guariguata, L.; Motala, A.A.; & Ogurtsova, K. (2019) Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition. Diabetes Res. Clin. Pract.

[9] Khanam, J.J. & Foo, S.Y. (2021) A comparison of machine learning algorithms for diabetes prediction, ICT Express.

[10] Manimaran, R., & Vanitha, M. (2017) Prediction of Diabetes Disease Using Classification Data Mining Techniques. International Journal of Engineering and Technology, https://www.researchgate.net/publication/331672855.

[11] Alshammari1, R., Atiyah, N., Daghistani, T., & Alshammari, A. (2020) Improving Accuracy for Diabetes Mellitus Prediction by Using Deepnet. Public Health Informatics * ISSN 1947-2579 * http://ojphi.org * 12(1):e11.

[12] Breault, J. L. (2011). "Data Mining Diabetic Databases: Are Rough Sets a Useful Addition?

[13] Parthiban, G., Rajesh, A., & Srivatsa, S.K. (2011). "Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method", International Journal of Computer Applications, 24(3).

[14] Padmaja, P. (2008) "Characteristic evaluation of diabetes data using clustering techniques", IJCSNS International Journal of Computer Science and Network Security, 8(11).

[15] Rajesh, K. & Sangeetha, V. (2012). Application of Data Mining Methods and Techniques for Diabetes Diagnosis. International Journal of Engineering and Innovative Technology (IJEIT), 2(3).

[16] Rahim, S.S. (2016). Automatic Screening and Classification of Diabetic Retinopathy Eye Fundus Images. Unpublished PhD Thesis. Coventry: Coventry University.

[17] Neilesh, B. & Gandhi, K. (2014) Diabetes prediction using feature selection and classification. Int. J. Adv. Eng. Res. Dev.

[18] Vijayan, V. & Anjali, C. (2015) Prediction and Diagnosis of Diabetes Mellitus - A Machine Learning Approach. IEEE.

[19] Miss, S.J., & Megha, B. (2016) detection and prediction of diabetes mellitus using back-propagation neural network. IEEE.

[20] Mohebbi, A., Tinna, A.B., Alexander, J.R., Henrik, B., Marco, F., & Morten, M. (2017). A deep learning approach to adherence detection for type 2 diabetics. IEEE.

[21] Francesco, M., Nardone, V., & Santone, A. (2017) Diabetes mellitus affected patients classification and diagnosis through machine learning techniques. Sci. Direct;112:2519-28.

[22] Maham, J., Hammad, A., Mehreen, A., Khawar, K., Raheel, N. (2017) An expert system for diabetes prediction using auto-tuned multi-layer perceptron. In: IEEE, vol. 2017 intelligent systems Conference (IntelliSys). London: IEEE.

[23] Wenqian, C., Shuyu, C., Hancui, Z., Tianshu, W. (2017) A hybrid prediction model for type 2 diabetes using K-means and decision tree. In: 8th IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS Beijing IEEE.

[24] Mangrulkar R.S. (2017) Retinal image classification technique for diabetes identification. Int. Conf. Comput. Methodol. Commun. ICCMC Erode IEEE.

[25] Sidong, W., Xuejiao, Z., & Chunyan, M. (2018) A comprehensive exploration to the machine learning techniques for diabetes identification. IEEE 4th world forum internet of things WF-IoT IEEE.

[26] Ashiquzzaman, A. (2018) Reduction of overfitting in diabetes prediction using deep learning neural network. IT Converge. Secure. 2017 Lect. Notes Electr. Eng, vol. 449. Springer Singap.

[27] Deepti, S., & Dilip, S.S. (2018) Prediction of diabetes using classification algorithms. Sci. Direct.

[28] Han, W., Shengqi, Y., Zhangqin, H., Jian, H., & Xiaoyi, W. (2018) Type 2 diabetes mellitus prediction model based on data mining. Sci. Direct.

[29] Safial, I.A., & Islam M. (2019) Diabetes prediction: a deep learning approach. Int. J. Inf. Eng. Electron. Bus, vol. 11.

[30] Ayon, S.I & Islam, M. (2019) "Diabetes Prediction: A Deep Learning Approach", International Journal of Information Engineering and Electronic Business (IJIEEB), Vol.11, No.2.

[31] Naz, H., & Ahuja, S. (2020) Deep learning approach for diabetes prediction using PIMA Indian dataset. Journal of Diabetes & Metabolic Disorders.

[32] Bhoia, S.K, Pandab, S.K., Jenaa, K.K., Abhisekhc, P.A., Sahood, K.S., Samae, N.U., Pradhan, S.S., & Sahooa, R.R. (2021) Prediction of Diabetes in Females of Pima Indian Heritage: A Complete Supervised Learning Approach. Turkish Journal of Computer and Mathematics Education. Vol.12 No.10 3074-3084.

[33] Islam, M., Rahman, J., Roy , D.C., Maniruzzaman, M. (2020) Automated detection and classification of diabetes disease based on Bangladesh demographic and health survey data, 2011 using machine learning approach. Diabetes and Metabolic Syndrome Clinical Research and Reviews https://www.researchgate.net/publication/339846671.

[34] Alpan, K., & İlgi, G.S. (2020) Classification of Diabetes Dataset with Data Mining Techniques by Using WEKA Approach. 978-1-7281-9090-7, IEEE.

[35] Anwar, F., & Ul-Ain, Q., & Ejaz, M., & Mosavi, A. (2020). A comparative analysis on diagnosis of diabetes mellitus using different approaches -A survey. Informatics in Medicine Unlocked. 21. 100482.