Mapana – Journal of Sciences
2024, Vol. 23, Special Issue 3, 25-54
ISSN 0975-3303|https://doi.org/10.12723/mjs.sp3.2

Effectiveness of Machine Learning Models and Performance Enhancement with Threshold Tuning Method Adopted in Diabetes Prediction

Priyabrata Sahu* and Jibendu Kumar Mantri**

Abstract

Internal organ failure, retinopathy, and neuropathy are a few of the complications of diabetes, a medical condition characterized by persistently high blood sugar levels. By 2040, the WHO predicts that this number will have increased to 642 million, with one in ten individuals suffering from diabetes due to poor diet and inactivity. Using machine learning algorithms to predict diabetes has been the focus of extensive research by a few writers. Most medical data are nonlinear, nonnormal, correlation-based, and complex, making the analysis of diabetic data disease more challenging. In healthcare and medical imaging, machine learning-based algorithms have been prohibited. Prediction of diabetes mellitus at an early stage requires a departure from current practices. Patients can be diagnosed as diabetic or non-diabetic utilising risk stratification algorithms based on machine learning. Utilizing AI and cognitive computing to treat diabetes offers great promise. The objective of this paper is to provide PWDs, clinicians, family members, and carers with a better understanding of the possible uses of current AI advancements in the treatment of PWDs. Automated retinal screening, and patient self-management systems are all paper themes that have the

* Research Scholar, P G Department of Computer Application, MSCBD, University, Odisha, India priyabsahu@gmail.com

** P G Department of Computer Application MSCBD University, Odisha, India jkmantri@gmail.com

potential to revolutionize diabetes care. There are a variety of innovative AI-powered decision-support technologies available currently, including retinal imaging systems, Modelling software for predictions, insulin pumps, apps for smartphones, and more. Millions of PWD could benefit from the enhanced blood glucose management, lower frequency of hypoglycemic episodes, and decreased risk of diabetes-related comorbidities that could be facilitated by AI applications. AI applications enhance the lives of people with disabilities (PWDs), their clinicians, and their family members and carers by enhancing accuracy, efficiency, usability, and satisfaction. In this study, the performance metrics of the top three machine learning models currently available are compared. Four out of the five parameters analyzed, the KNN model fared the best. KNN is superior in terms of accuracy, area under the curve (AUC), precision, and f1 score. Logistic Regression achieves the optimal Recall/Sensitivity of 93%. Our study is highly recommended because it gathers research papers from a range of sources that will be useful to other academics working on various Novel diabetic prediction models.

Keywords—diabetes care, artificial intelligence, retinal imaging, glucose monitoring

I. Introduction

The worldwide prevalence of diabetes has reached epidemic proportions. Diabetes affects 425 million people worldwide as estimated and accounts for 12% of global health expenditures; nonetheless, one out of every two persons with the condition remain undiagnosed and untreated. Obesity and inactivity contribute to the emergence of Type 2 diabetes, which causes individuals to rely on exogenous insulin since their bodies are unable to produce enough. Youths whose bodies are incapable of producing insulin at birth must inject insulin for the remainder of their lives. Diabetes, which is the important reason of kidney failure, and major threat of cardiac arrest and all-cause death, can result in hospitalization, long- term repercussions, and increased costs. Due to a lack of timely, relevant health data, it is difficult to make educated decisions on intensive therapy and rigorous glucose control [1,2]. The impact of technology on the care of diabetes patients appears to be fairly limited, despite the fact that it enables unprecedented and very low-cost access for vital information for a wide variety of individuals and industries. As of June 2018, the index of biological data includes 28 million citations relating to these findings and is rising at 850,000 per year. Over the course of a lifetime, the average person, according to findings, expected to generate around 1 terabyte, or 300 million volumes, of human health data. Approximately 80% of all medical records are disorganized. This category also includes non-clinical data sources, such as device and sensor data, genetic data, and health data related to social determinants [3]. Given that genetic and exogenous data may account for up to 90% of health outcomes, it is vital that people with disabilities and their medical diagnostician basically collect and utilize these collected data for making informed healthcare decisions for PWDs.

AI is now capable of processing massive amounts of data to better serve clients in numerous industries, including healthcare. According to a survey conducted in 2017, 68 percent of developers of mobile health apps believe that diabetes will continue to be the critical health care related area having necessity for related e-health solutions in the future. On the other hand, 61 percent of developers believe that artificial intelligence will be the most disruptive technology that will shape the related sector of digital health. New AI- powered diabetic therapy devices have been approved [4], but there is no systematic study of clinically relevant AI applications for diabetes. This research paper aims to assist PWDs and their primary care physicians, health professionals and their family members, and carers in understanding which important AI advancements may be useful to them right now. Providing access to structured and unstructured health data in real time, AI's rapid development holds enormous potential for enhancing the treatment of those with disabilities. “The process of making computing devices perform tasks which require human intelligence,” describes The Turing Archive for Computing History [6]. The phrase “artificial intelligence” basically refers to a variety of techniques designed to simulate human intelligence and perform reasoning-intensive activities, like the perception of visual or auditory stimuli, data analysis, decision making, and language translation. Using a vast array of AI techniques, cognitive systems enable individuals to solve problems quickly and effectively by drawing from a large pool of data.

Following a brief summary of some of the more prevalent artificial intelligence algorithms used in diabetic diagnoses Section II followed by Existing machine learning models summarized in section III. This research evaluates many of these models in Section IV. The concluding views and insights are included in Section V of the study.

II. Review of Literatures

A study of published papers emphasized the advances in Artificial Intelligence technology over the past few years and a way of supporting PWDs and their physicians in making much more informed decisions. Table 1 summarizes some of the most commonly utilized examples of AI-powered devices for diabetes care and systems found in research. Throughout the last decade, diabetes artificial intelligence (AI) academics and commercial developers have focused on the questions in Table 1.

According to the research so far, numerous artificial intelligence (AI) strategies for diabetes prevention, diagnosis, and treatment are being developed, evaluated, and applied. The multiple no. of published technical research papers outlining various improvements in the required diabetology and Artificial Intelligence has enhanced dramatically over the last several years , rising from 2600 in year 2008 to around further increase by 5500 in year 2013 to continuously increased up to 10,000 in year 2017 [7]. A modern diabetic information base for patients, persons working on research, and physicians requires processing and analyzing millions of patient health records and recent research. These published findings show that AI has great potential for improving diabetes screening, and management [8] because to its capability to rapidly analyze and filter massive amounts of important data into clear ideas for taking decisions.

According to recent studies, AI-based tools are rapidly reshaping care in four key areas: better detection and screening of The effects of diabetes on vision, include diabetic retinopathy (DR) and macular edoema, as well as methods for assessing, reducing, and managing a patient's own risk for these conditions.

Aside from being the major cause of secondary blindness, DR has a tremendous economic and emotional impact on victims and their loved ones, as well as the healthcare system as a whole. The annual diabetic retinal exam screens and detects treatable retinopathy. Prediction, early detection, and treatment can avoid 98% of DR and macular edoema visual loss [9].However, a shortage of skilled eye care practitioners, as well as other barriers to care, represent significant barriers to undertaking more widespread examinations.

The grading of DR from retinal photos using AI that is based on deep learning has been shown to have a sensitivity and specificity of more than 90%, according to recent studies [10]. The first medical gadget that employs artificial intelligence (AI) to screen diabetic patients for neuropathy was recently given approval for sale by the Food and Drug Administration (FDA) in the United States. The Topcon NW400 retinal camera captures images of the eye, which are then analyzed by an AI algorithm in the IDx-DR (IDx LLC, Coralville, IA) software programme. This application was developed by IDx LLC. A cloud server stores patient retinal photos in order to complete the IDx-DR procedure. Assuming the image quality is sufficient, the software will present the doctor with one of two options: The outcomes have two options: See an eye doctor if diabetic nephropathy is moderate or severe; otherwise, repeat the screening in 12 months. IDx- DR is the first commercially available system that can screen patients without relying on human interpretation of imaging studies or diagnostic data [11]. Patient satisfaction is significantly increased as a result of the streamlined process provided by these automatic systems [12], where in primary care physician offices, non-eye health providers will be able to undertake retinal screening on-site and provide instant regular outcomes, or referrals to an ophthalmologist. To proactively discover and characterize diabetes populations in health care systems, larger physician groups, and health plans, AI "mines" massive volumes of digital and unstructured patient data [13]. find those who are likely to have diabetes and other complications [14], identify proteins and genes related with and predictive of diabetes, and prioritize patients for specialist diabetes disease management system [15].AI has matured to the point that doctors and other medical professionals can utilize it to assist them in making decisions regarding their disabled patients. Using machine learning technologies, doctors can enhance diabetes patients' adherence to treatment and health outcomes [16]. Artificial intelligence (AI)-enabled solutions help with non-invasive diabetes diagnosis [17], precise measurement and monitoring of the degree of diabetic neuropathy [18], and diabetic wound care [19].

For better blood glucose control [20], fewer hypoglycemia episodes [21], and higher satisfaction of patients and reported results [22] are all feasible goals for diabetics and their physicians, thanks to recent breakthroughs in artificial intelligence (AI) such as sensors, pumps, smartphone apps, and more. RT CMG AI devices are approaching the market to help patients with diabetes and their doctors analyze and enhance glycaemic control, prevent hypoglycemia episodes (particularly at night), and increase A1c values, according to research published in 2017.

Reduce hypoglycemia and enhance diabetic self-care with research on the “artificial pancreas,” also called a “Closed Loop System” [23]. Diabetes patients can benefit from using smartphone sensors with ML methods and Logic based on symbols to identify and measure high-level habits of modern life [24]. Diabetes patients and carers can active wound care everyday by using an AI smartphone camera system, which could speed up wound healing, save travel expenditures, and minimize healthcare costs [25]. Researchers discovered that pregnant women with gestational diabetes are open to receiving medical care in the comfort of their own homes utilising AI-enhanced telemedicine. A study incorporated clinical practice suggestions that could be read by a computer, as well as data from a patient's electronic health record and wearable devices measuring glucose, blood pressure, and activity [26]. According to studies, current diabetic applications can assist people with diabetes conveniently keep track of and analyze their data, as well as generate unique, data-driven insights that can be used in everyday life. The best apps now feature vast food information databases, such as the calories and other nutrients in a product after scanning its barcode, restaurant menus and popular meals that can be searched by name, and even the capacity to identify individual foods on a user's plate.

The Guardian Connect from Medtronic is the first artificial intelligence-powered continuous glucose monitoring (CGM) system, and it was approved by the FDA on March 14, 2018, for use in people with diabetes (PWD) aged 14 to 75. Guardian Connect can alert consumers up to 60 minutes before a dangerous decrease in blood sugar levels by using a prediction system. The Guardian Connect system was reliable when used in conjunction with the abdomen-worn Guardian Sensor 3. and sends data on blood glucose levels to a mobile app, and alerted patients of approximately 98.5% of hypoglycaemic events, allowing them to take preventative measures to restore normal blood sugar levels [27].

Caretakers and family members of patients can also access and monitor this data in real time or by text message. Sugar.IQ, a smart diabetes assistant, could be integrated to the Guardian Connect CGM as well. The Sugar.IQ assistant employs Artificial intelligence technology from IBM Watson Health to continuously analyze a user's blood glucose levels in reaction to changes in food, insulin dosage, and daily activities. As 256 Guardian Connect users used Sugar.IQ, they saw statistically significant improvements in blood glucose time-in-range (36 minutes/day), time >180mg/dl (30 minutes/day), and time 70mg/dl (6 minutes/day) when compared to baseline measurements. Sugar.IQ delivered 655 hypoglycemic insights and 699 hyperglycaemic insights for PWDs at the time of the course of 31+ patient-years of use. Because of the large variety in glucose reactions to diet and exercise, the findings underscored the necessity for individualization in Controlling one's own diabetes. Importantly, 231 of 256 users (90%) documented at least two weeks of data, showing a consistent type of interaction with the artificial intelligence based Controlling one's own diabetes application [28].

Another significant hurdle is the difficulty of reproducing AI study results. A basic issue is that academics rarely expose their source code to the public. Only 6% of presenters made their source code public in a survey of around 400 techniques presented at two major Artificial Intelligence conferences over the last few years. One-third did not even bother disclosing the data on which their algorithms were run, and only half cared to give the pseudocode, which is essentially a concise description of the actual source code [29]. Furthermore, even if the original pseudocode is acquired and executed, it may fail to function properly. The training data for required methods, such as the necessary information for training speech-recognition learning systems, can also affect its efficacy in machine learning, a field of AI in which computers learn via experience.

There is still a long way to go before diabetes-related artificial intelligence (AI) software, gadgets, and systems become widespread in the health-care sector. Communication and data sharing between two or more systems is a significant barrier. High initial and recurrent costs, physician engagement use guidelines all hinder adoption and innovation [30].

III. Methodologies Used in Diabetes Care

A range of AI strategies are being used by researchers basically make analysis of a lot of information which is necessary to make sense of it all and evaluated [15]. AI encompasses a large array of increasingly complex techniques referred to as machine learning, and cognitive computing. Computer learning professionals frequently "train" AI systems with massive amounts of related data and techniques that allow the device for analyzing and further learning from relationships. Understanding, thinking, interacting, and learning are all within the capabilities of intelligent AI systems. These such systems quickly process and further deeply analyze structured and unstructured input. They provide natural dialogue, visualization, and cooperation [16]. Cognitive AI acquires knowledge via gathering and analyzing system-wide feedback. The end output consists of information, tools, and technology that patients with diabetes and their physicians may employ to save time, boost efficiency, improve clinical decision making, empower diabetic persons, and maybe improve diabetic patient and clinician satisfaction. This study reveals that artificial intelligence is becoming more prevalent in the treatment of diabetes, which has the potential to alter the lives of millions of people. Although a more comprehensive examination of artificial intelligence is outside the scope of this Research paper, this study reveals that AI has the potential to alter the lives of millions of people. General-purpose artificial intelligence techniques presented in the paper and their practical usefulness in diabetes management are listed in table 1.

A multilayer perceptron: A multilayer perceptron is made up of neurons in the input, output, and hidden layers. All of the neurons in one layer are coupled to all of the neurons in the layer below them. It learns using the "backpropagation" method. Its main advantage is that it can simulate complex nonlinear interactions. However, they have limitations in terms of the various number of parameters that must be calculated without method of convolution and are unable to compare well to the performance of competing deep models. Patient Self-Management Tools and Prediction Models are among its applications.

Convolutional Neural Network: The convolution layer has neurons that, like filters, analyze tiny patches of the input image at a time and convolute throughout the whole image while sharing parameters. These neurons are part of the convolution layer. It does most of its learning through a mechanism known as "back propagation." Each layer of the CNN is responsible for determining whether or not certain features are present in the given region, with subsequent layers of the algorithm being able to identify increasingly complex features. It is capable of modelling complex nonlinear interactions, making it excellent for image, audio, and video processing. Its drawbacks are that it requires a large quantity of data to train, that it is computationally costly, and that many more parameters require precise modification at the time of training the model. This method is suitable for retinal screening.

Random Forest: It's like having a bunch of decision trees generated for you. Each tree's root nodes and branches are determined by analyzing a different set of attributes at random. Its basic advantages include being easy to set up and giving satisfactory results in most cases. Useful for resolving classification and regression issues. It's quite good at weighing the significance of various features. When the sample size is large enough, it is immune to outliers and prevents overfitting. However, this method can only be used with discrete results, as continuous results require further categorization before being understood. Common uses include diagnostic aids, such as patient self-management instruments and prediction models, as well as screening for retinal diseases and other forms of eye disease.

Fuzzy_logic or fuzzy_system: It simply basically provides a method of probability value between 0 and 1 for membership in a particular class rather than a deterministic judgement (0 or 1). Its advantages include the fact that it resembles human reasoning, has a high level of interpretability, is easy to alter rules, and does not require large amounts of data. However, it necessitates a good curation of rules. Its applications include retinal screening, sensors, and an artificial pancreas.

Table-1: Reviews of Some machine learning methods for Diabetes Predictions

Author, date	Proposed Research purpose and Learning Model Used	Outcomes of the proposed study	Application of the Model used and its performance
Gulshan V. 2016 [10].	Deep CNN Model for Developing as well as Validating a DL - Deep Learning System to Identify Diabetic Retinopathy used for Retinal Fundus Images	93.9% for Messidor-2,ROC: 0.991 0.990 (0.986–0.995) for Mess Sensitivity: 97.5% for EyePACS-1, idor-2 , (95% CI, 0.988– 0.993) for EyePACS-1,	Deep machine learning detected referable DR with excellent sensitivity and specificity.
Keel S. 2018 [12]	Screening Model adopted to a New AI-based Diabetic Retinopathy for Endocrinology Outpatient Services: Feasibility and Patient Acceptability (Deep CNN)	Sensitivity: 92.3%, Specificity: 93.7%, 96% were happy or extremely satisfied with the automated model.	AI-based DR screening is practical, accurate, and well-accepted by endocrinologist outpatients.
Shankara charya. 2012 [13]	Mixture of expert system based on MLP model used for Indian Artificial Diagnosis Aid for Pre-Diabetes and Type 2 Diabetes Based on Artificial Intelligence	Excellent result found as Sensitivity is 99.5% and Specificity is 99.07%. Furthermore, Accuracy is 99.36%	The suggested technique for diagnosing high-volume diabetic testing in people with prediabetes, diabetes, and those without the condition works well.

Corey KE. 2016 [14].	Development as well as rechecking of an EMR method to Detect "Nonalcoholic Fatty Liver Disease" (NAFLD). (LR with adaptive LASSO)	"91% Sensitivity: 51%, PPV: 89%, NPV: 56%, AUC: 0.85 (compared to 0.75 using ICD-9 billing codes alone, P<0.0001)".	NAFLD classification algorithms surpass ICD-9 billing data. An easy-to-implement strategy produces EMR-based NAFLD cohorts across institutions.
Vyas R. 2016 [15]	Biomedical Text Mining , Support Vector Machine -SVM , and Network Analysis of Diabetes Mellitus Protein-Protein Interactions (SVM)	The results are Accuracy is 78.20%, Precision is 68.26% and AUC is 0.788	This integrated strategy may discover illness-related proteins, and research of proteomics .
Lo´pez B. 2018 [16]	Models used are (Random Forest, k-NN) for Random Forest SNP Relevance Learning for prediction of Type 2 related Diabetes Risk	RF surpassed SVM and LR in risk prediction of basic accuracy and relevance relevant stability (AUC: 0.89).	RF helps clinicians detect type 2 diabetes- related SNPs.
Shu T. 2017 [17].	A Comprehensive Study Extractors related to texture features to Identify Diabetes Mellitus in Face Areas (k-NN, SVM with 8 image extractor methods)	The excellent extractor for texture features, Picture Grayscale paired with SVM, Sensitivity is 99.64%, Specificity is 98.26%.	The Picture Grayscale Histogram uses face and tongue traits to accurately detect diabetes without blood samples.
Wang L. 2017 [19].	Stacked Two-Stage SVM-Based Classification of Diabetic Foot Ulcer Images. The model used are The SVM is used in two stages, the first of which employs conditional random fields to do basic linear iterative clustering.	Sensitivity is 73.3% Specificity is 94.6%	Computer-based solutions are effective for smartphone-based picture analysis of diabetic wounds and wound healing state.

Mauseth R. 2015 [20].	The Fuzzy Logic Controller related to Artificial Pancreas System was Improved Via testing of Pizza and Exercise.	FLC v2.1 improved mean blood glucose after pizza, activity testing, , and euglycemic time and hyperglycemia	Experimenting When it is retested for mean, high, and normal glucose in blood , the AP system and dosage matrix changes enhanced FLC performance.
Ling SH. 2012 [21].	Natural Hypoglycemia "Particle Swarm Optimized Fuzzy Reasoning" (Fuzzy Logic Controller systems (FLC))	Improved night-time hypoglycemic episode detection 85.7%, 79.8%. Detection of Hypoglycemic episodes detection, , 55.1% specificity.	The technology detects type 1 diabetic hypoglycemia episodes noninvasively.
Herrero P. 2015 [22].	Run-To-Run Control and Case-Based Reasoning Advanced Insulin Bolus Adviser	CBR improved mean blood glucose value in adults and adolescents, eliminating hypoglycemia ,CBR alone was not actually able to perform it in the adolescent population.	The suggested device improves bolus calculator performance while maintaining its simplicity.
DeJour nett L. 2016 [23].	Knowledge-based system model used in Silico Testing	The average time in control, 70–140mg/dl, and hyperglycemic ranges was 94.2%, 97.8%, and 2.1%, respectively.	AI-adopted closed-loop glucose controllers can outperform ICU-based PID/MPC controllers.

Cvetkovic´ B. 2016 [24].	Smartphone Activity Recognition for Diabetics (Ensemble of models like SVM, RF, Jrip, and Bagging algorithms), symbolic rules)	MCAT yields best results. Accuracy obtained is 83.4% and value related to F- score obtained is 0.82	Sensors based on smartphones employing ML and symbolic reasoning may detect and Putting a number on a luxurious lifestyle behaviours of diabetes persons for making better exercise alternatives.
Wang L. 2015 [25].	Smartphone Diabetes Wound Assessment System. The model used are for detection of image boundaries and for mean-shift segmentation method is used and for the purpose of Colour segmentation ,K-means is used	Simulated image visual assessment Matthews Correlation: 0.736	Camera system of the smartphone lets diabetes persons and carers participate in regular wound care.
Rigla M. 2018 [26].	Smart Mobile Telemedicine for GDM Management (Mobile telemedicine system)	Telemedicine patients had reduced BP and comparable metabolic and perinatal outcomes.	GDM patients might benefit from AI- augmented telemedicine.

Support vector machine (SVM): It is a way for describing outcomes using a binary scale (While approaches for multi class State vector machine exist), it is not often employed for such tasks. In order for it to function properly, data are first added to a space with several dimensions, and then the optimal hyperplane for dividing two classes is calculated (based on the margin, which is optimizes the margin (distance from plane to related data points) between the plane and neighboring data points. It is not susceptible to overfitting and performs exceptionally well at nonlinear decision boundaries. It is difficult to understand and does not scale well to vast volumes of data.

This makes it less useful. It can be used in patient self- management systems, retinal screening, decision assistance, and prediction models, among other uses.

Logistic regression: This is essentially a categorization method for binary outcomes. Based on the attributes, it predicts the likelihood of a result (0 or 1). Maximum likelihood estimation is used to learn the model coefficients. It finds the best line or hyperplane to represent the data points. It is simple to install, efficient, and scalable. It is supported by the vast majority of standard software. It can calculate the likelihood of the outcome, which is useful. However, its limitations are that it only allows for binary categorization. It is vulnerable to outliers and necessitates the processing of nonlinear characteristics. It is beneficial to prediction models.

Natural language processing (NLP): It refers to technology and methods used in computing for handling, analyzing, and inferring human languages. It is critical in the development of intelligent machines and human-computer interactions. It can analyze and interpreting free text of data, such as electronic medical notes. Various numbers of human-annotated records are frequently used in training. It can be applied to prediction models.

The algorithm of K-nearest neighbors:

It divides input data into classifications based on method of k-nearest neighbors (KNN). Furthermore, It does not make any particular assumptions about the underlying distribution. It applies to both regression and classification related problems. It is straightforward to understand and apply. It necessitates a large amount of computer power. It is, however, susceptible to outliers and regional data. Ophthalmic monitoring, decision assistance, modeling techniques, and self - regulation and self-systems are among of the most common uses.

IV. Diabetes Data - Model Comparisons

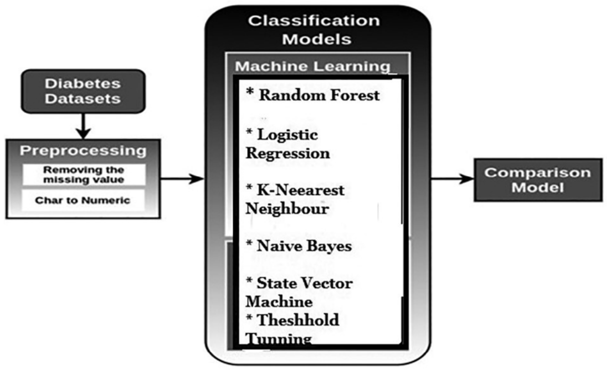

Figure 4. 1. Proposed Research Model

Dataset and Models used: The proposed methodology adopted for finding out effectiveness of different machine learning models and performance enhancement without and with threshold cleaning for diabetes prediction is as shown in figure 4.1 Utilization of a dataset is the first step in this paper. The utilized dataset originates from the NIDDK study, which strives to comprehend and cure the most chronic, expensive, and debilitating diseases. The basic purpose is to determine whether a patient has diabetes based on diagnostic parameters extracted from the dataset. The following step is to develop a model capable of properly predicting whether the patients in the dataset have diabetes. Further the datasets contain multiple medical predictor variables and one target variable (Outcome). Among the predictor variables are the patient's age, BMI, insulin level, and number of pregnancies.

Pregnancies, Glucose, Blood Pressure (BP), Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, and Outcome("Class variable") are examples of variables used in data sets(either 0 or 1). 268 of the 768 numbers are 1, with the remainder being 0.

Data Cleaning and Analysis:

a. Carry out an analysis that is descriptive: Gain an understanding of the variables and the values that relate to them. In the columns that follow, a value of zero is illogical and therefore denotes the absence of data, as follows: glucose levels, blood pressure (BP), skin_thickness, insulin_levels, and BMI.

b. Investigate these factors using histograms as a visual tool. Apply the appropriate treatment to the missing values.

c. This dataset contains variables of the integer and float datatypes respectively. Produce a count (or frequency) plot that outlines the data categories as well as the number of variables.

Exploratory Data Analysis:

a. Age, Insulin, Diabetes Pedigree Function and Pregnancies are rightly skewed.

b. Zero values in blood pressure, BMI, Insulin and Glucose clearly stand out in the plot.

c. After removing zeros for non-zero expected columns, we see that except Insulin which is highly right skewed, all other are near to gaussian distribution.

d. Except for Insulin, for rest of other non-zero columns, we can take mean value.

e. For Insulin, we took median value to fill.

f. In Data type count plot, we can see that there are 2 int type columns and 7 float types. (As shown in table 4.1)

Plot data types

Exploratory Data Analysis: The steps involved are.

a. Determine whether or not the data are evenly distributed by graphing the count of outcomes against their value. Describe what you've discovered and make plans for the next step to take.

b. Construct scatter plots between the two variables in order to better comprehend their correlations. Describe what you have discovered.

c. Carry out an investigation of correlations. Utilize a heat map to have a visual understanding of it.

Observations

It is an imbalanced dataset where positive outcomes are almost half of the negative outcomes. While creating model, we need to balance the outcomes either by oversampling the minority class or under sampling of majority class.

Other workaround could be to do a weighted computation while training the model.

Pair plot analysis

a. BMI and Skin thickness have a positive correlation.

b. Insulin and Glucose have a positive correlation.

c. The rest of the other fields are uncorrelated or very weakly correlated.

Table 4.1: Exploratory Data Analysis:

Para meter	Pregn ancies	Glucose	Blood Pressure	Skin Thick ness	Insulin	BMI	Diabetes Pedigree Function	Age	Out-come
count	768	768	768	768	768	768	768	768	768
mean	3.84	120.89	69.1	20.53	79.79	31.99	0.47	33.24	0.34
std	3.36	31.97	19.35	15.95	115.24	7.88	0.33	11.76	0.47
min	0	0	0	0	0	0	0.078	21	0
25%	1	99	62	0	0	27.3	0.24	24	0
50%	3	117	72	23	30.5	32	0.37	29	0
75%	6	140.25	80	32	127.25	36.6	0.62	41	1
max	17	199	122	99	846	67.1	2.42	81	1

Another workaround could be to do a weighted computationwhile training the model.

a. BMI and Skin thickness have a positive correlation.

b. Insulin and Glucose have a positive correlation.

c. The rest of the other fields are uncorrelated or very weakly correlated.

Correlation Analysis

a. There is no strong correlation between either of the two fields.

b. The BMI-Skin thickness and Insulin-Glucose are the highest correlated in the set, but they are moderatelycorrelated.

c. Outcome is moderately correlated to Glucose.

Checking Data balance and Correlation analysis

Creating, Tuning and Comparing Models: 1. Devise strategies for model building. It is necessary to decide regarding the appropriate validation framework.

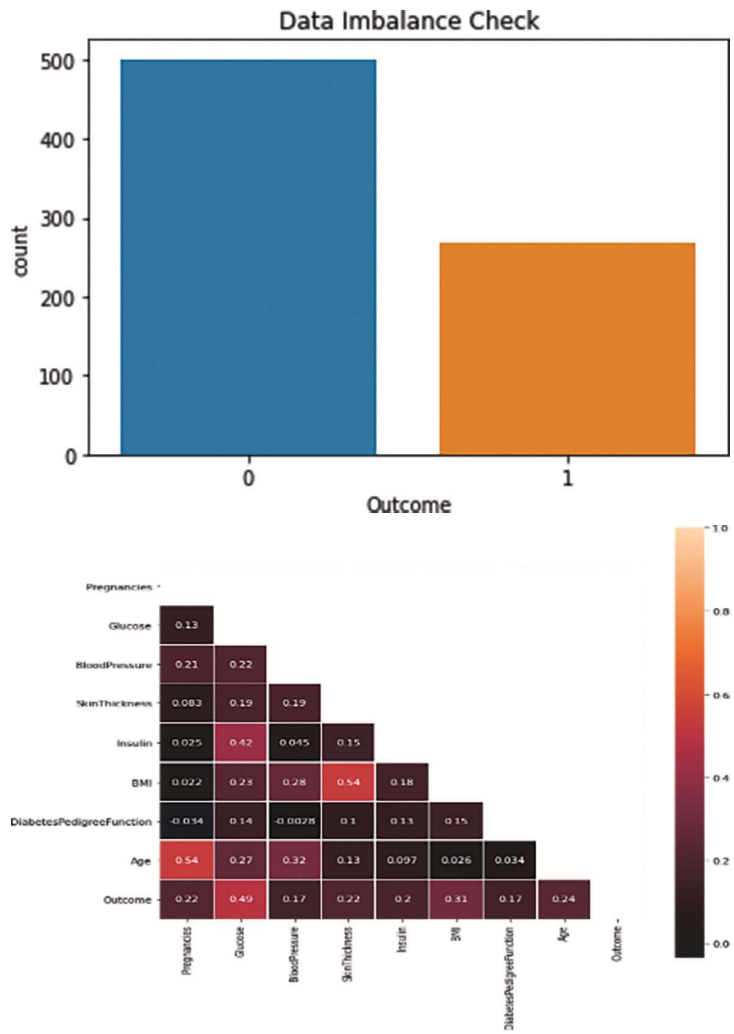

Figure 4.3: Checking Data balance and Correlation analysis. Describe the mental process that you are using.

2. To construct a model, select an appropriate classification algorithm and run it. Analyze the performance of the KNN algorithm in comparison to the various models.

Observations

This is a binary data classification problem were depending on the all the features, the model has to decide whether a person has diabetes or not. We have several ways to build the model for binary/multi-class classification. Few of them are listed below:

1.Logistics Regression, 2. Naive Bayes classification.3 Stochastic Gradient Descent 4. K-Nearest Neighbors 5. DT, 6. RF, 7. SVM.

We are going to build four models and compare their performance on test and train dataset. We will tune the models if there is need to tune. Here we will basically make use of K-Fold Cross Validation to validate the models. We will plot all the models' stats together and compare their performance. Ofall the tuned models, we will pick the best model. The step- by-step procedure can be followed below:

a. Common Terminology
b. Data Scaling & Splitting
c. Logistic Regression
d. Naive Bayes Classification
e. Random Forest
f. K Nearest Neighbours
g. Putting it all together: ROC AUC Curves, Model Comparison.

From Model Comparison, we find that KNN is the most stable classifier. All the parameters are quite good. It has the best accuracy, AUC, precision and f1_score of all the models.

If we are looking for a highly sensitive model, we can take logistic regression model, which has the highest recall.

Classification Terminology

Precision: The level of precision can be estimated by dividing the number of confirmed results by the number of results that were not false positives. It is a measure of how accurately a model can forecast the positive class. Accuracy, or positive predictive value, is another name for precision.

Precision = True Positives / (True Positives + False Positives) What proportion of actual positives were identified correctly is called **Recall/Sensitivity/True Positive Rate (TPR)**

In medical terminology, Sensitivity quantifies the frequency with which a test correctly produces a positive result for individuals with the ailment being tested for. A test with a high level of sensitivity will identify virtually all people whohave the ailment and will result in a negligible number of false-negative findings. (For instance, a test with a sensitivity of 90% will correctly produce a positive result for 90% of people who have the condition, but it will return a false- negative result for 10% of people who have the disease and should have tested positive.)

Recall or Sensitivity or True Positive Rate (TPR)
= True_Positives / (True_Positives + False_Negatives)

Specificity/True Negative Rate: The capacity of a test to provide a negative result in individuals who do not have the disease that is being tested for in the first place is one of the aspects that are evaluated by this criterion. A test has a high level of specificity if it can rule out almost everyone who doesnot have the disease while at the same time producing relatively few false positive results. (For instance, a test with a specificity of 90% will correctly provide a negative result for 90% of individuals who do not have the disease. However, the same test will correctly return a positive result, also known as a false-positive, for 10% of individuals who do not have the condition and should have tested negative.)

Specificity or True Negative Rate
= True_Negatives /(True_Negatives + False_Positive) False Positive Rate=1−Specificity.

Inverted specificity =
False alarm rate = False_Positive Rate
= False_ Positives / (False_Positives + True_Negatives)
For any test, there is usually a trade-off between TPR and FPR. “

ROC-AUC Curve
It is basically a graph of the false_ positive_ rate (x-axis) versus the true_positive_rate (y-axis) for various candidate threshold values

ranging from 0.0 to 1.0. In other words, it shows basically the false alarm rate against the hit rate.

Precision - Recall Curve

Reviewing both precision and recall is important when there is an imbalance between the two classes' observations. Thereare numerous instances of no event (class 0) and few instancesof an event (class 1).

Due to the vast number of class 0 cases, we are often less concerned with the model's ability to properly forecast class 0instances, i.e., high true negatives.

Important to the computation of precision and recall is the factthat the calculations do not employ genuine negatives. It is only concerned with the accurate forecast of class 1, the minority class.

A precision-recall plot or curve is a depiction of the precision(y-axis) and recall (x-axis) at various thresholds. It isconceptually like the receiver operating characteristic (ROC)curve. F-Measure summarizes model skill for a certainprobability threshold (e.g., 0.5), whereas the area under thecurve shows model skill across thresholds, like ROC AUC. **F1 Score=** 2∗precision∗recall/ (precision + recall)

Scaling and splitting the data: Tunning by AUC_ROC Threshold

The geometric mean of TPR and FPR is an ideal value that ismaximum for every given TPR and FPR. If our goal is to construct a model that forecasts both sides, this threshold value could be picked as the best one.

Table 4.2: The optimum threshold to Test classification Report With tuned threshold.

	precision	recall	f1-score	support
0	0.94	0.66	0.77	96
1	0.62	0.93	0.74	58
Accuracy			0.76	154
Macro avg	0.78	0.79	0.76	154
Weighted avg	0.82	0.76	0.76	154

Table 4.3: Test classification Report Without tuned threshold.

	Precision	Recall	F1-score	Support
0	0.83	0.79	0.81	96
1	0.68	0.72	0.7	58
Accuracy			0.77	154
Macro avg	0.75	0.76	0.75	154
Weighted avg	0.77	0.77	0.77	154

Table 4.4: Train Classification Report

	Precision	Recall	F1-score	Support
0	0.79	0.83	0.81	404
1	0.64	0.59	0.61	210
Accuracy			0.74	614
Macro avg	0.72	0.71	0.71	614
Weighted avg	0.74	0.74	0.74	614

Random Forest

Table 4.5: Train Classification Report

	Precision	Recall	F1-score	Support
0	1	1	1	404
1	1	1	1	210
Accuracy			1	614
Macro avg	1	1	1	614
Weighted avg	1	1	1	614

Table 4.6: Test Classification Report

	Precision	Recall	F1-score	Support
0	0.81	0.86	0.83	96
1	0.75	0.66	0.7	58
Accuracy			0.79	154
Macro avg	0.78	0.76	0.77	154
Weighted avg	0.78	0.79	0.78	154

Observation

Given the appearance of the training data, we can conclude that Random Forest has been overfit. It may have very good responses due to overfitting, but it is ultimately not a suitable model. We will adjust the parmas accordingly. We will modify the Cost parameter and determine whether cost function produces comparable training and test data accuracy.

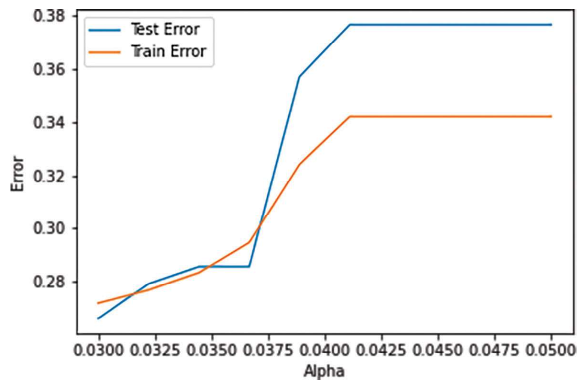

Figure 4.4: Random Forest report

Table 4.6: Test Classification Report after Tuning

	Precision	Recall	F1-score	Support
0	0.86	0.77	0.81	96
1	0.68	0.79	0.73	58
Accuracy			0.78	154
Macro avg	0.77	0.78	0.77	154
Weighted avg	0.79	0.78	0.78	154

Table 4.7: Train Classification Report after Tuning

	Precision	Recall	F1-score	Support
0	0..87	0.77	0.82	404
1	0.64	0.78	0.78	210
Accuracy			0.78	614
Macro avg	0.76	0.78	0.76	614
Weighted avg	0.79	0.78	0.78	614

K-Nearest Neighbor

We're collecting rmse, error rate, and accuracy data for a variety of nearest neighbors. We'll plot them all to observe ournearest neighbors. We can see that error rate and rmse produce the same plot, whereas accuracy produces a mirror copy of theother two.

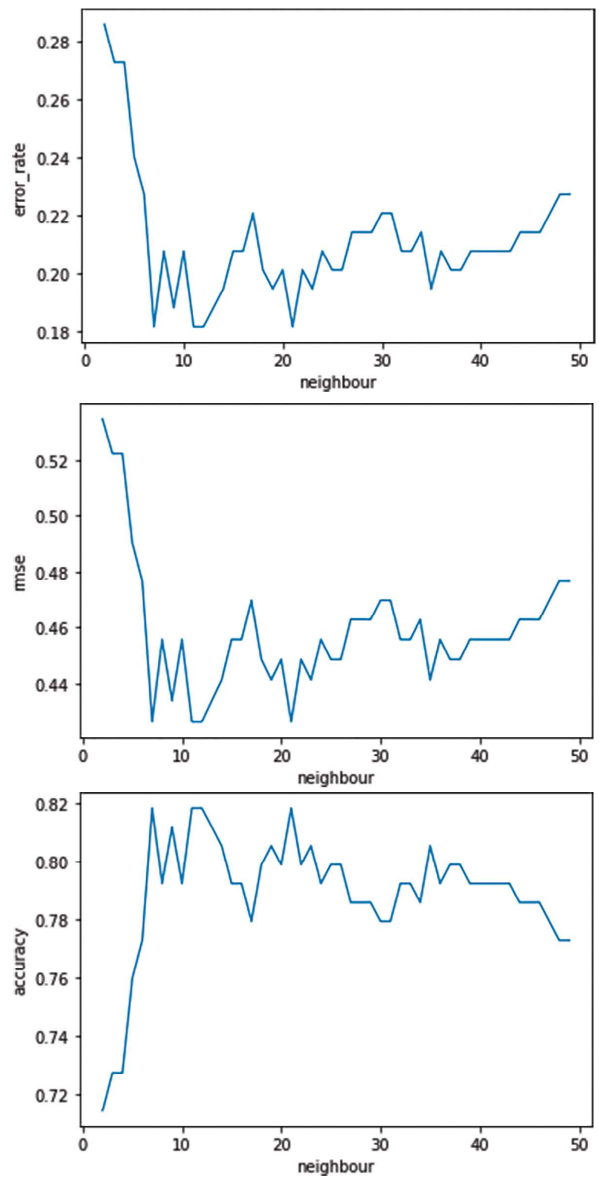

Figure 4.5: K-Nearest Neighbor reports

Model Comparison Data

Modelling:

In this process the purpose of data modelling is to develop a report on categorization by doing analyses of sensitivity, specificity, AUC (ROC curve), and other metrics. While doing this the proper care should be taken that it should be as descriptive as possible when explaining the values that have been used for these parameters.

Observations

We have plotted AUC-ROC curve of all the models together for comparison purposes. Also, I have plotted all the model stats together, tuned, and non-tuned version. We can compare the performance by looking at the plots. Of that, we have selected 3 best tuned models and plotted again to show what model is best of the lot. AUC ROC Curve of all the models put together.

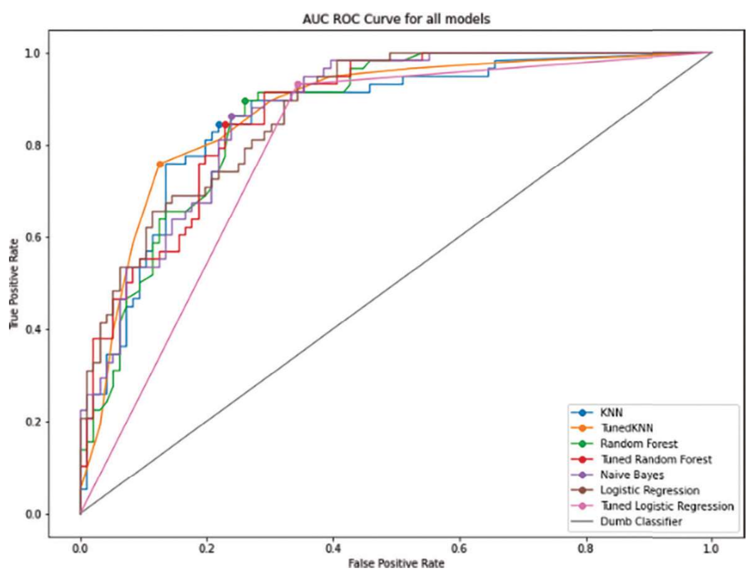

Figure 4.6: Model performance comparisons

Comparing model parameters

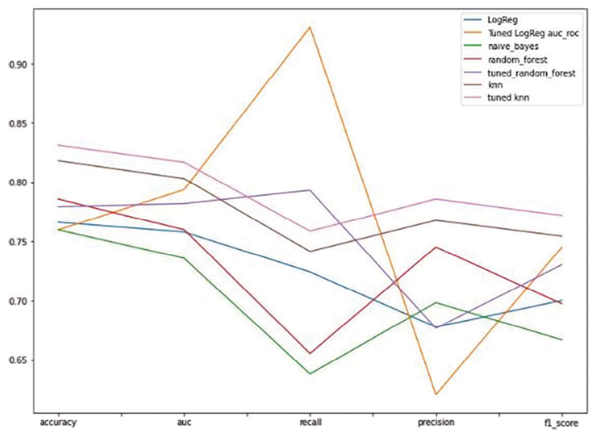

Figure 4.7: Model Parameter comparisons.

Observations

We have plotted the various parameters of the models. All models' recall and precision are in the other direction, as evidenced by a single glance. The overfit Random Forest hadcomparable features to KNN, however after tuning, its recall increased but its precision decreased.

In terms of overall parameter values, KNN is the most accurate predictor of all models. We have also tuned in to KNN. Through adjustment, the model's performance has improved across all parameters. We have also tuned Logisticregression. Following adjustment, TPR has increased while accuracy has decreased. If we want a model with a high Sensitivity, we can choose the Logistic Regression model. We can choose KNN for overall superior performance.

Final Model:

We have chosen the top three models for performance evaluation and analysis, and we compared their performanceparameter values. Based on these results obtained, we canconclude that the KNN model performs best in four of the five parameters, making it one of the best models. KNN is the bestin terms of accuracy, AUC, precision, and f1 score. At 93%, logistic regression has the highest Recall/Sensitivity.

V. CONCLUSION

According to this overview of recently published high-impact and clinically significant research papers, diabetes is attracting big health care method-based corporations and start-ups using breakthrough Artificial Intelligence technology and methods to answer PWDs' frequent issues. Numerous applications have been approved and commercialized in recent years. PWDs may benefit from continuous monitoring and real-time feedback.

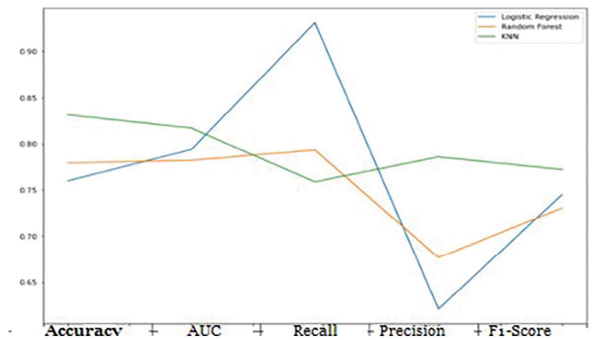

Figure 4.8: Final model

Table 4.8: Final Model Comparisons

Metrics	Accuracy	AUC	Recall	Precision	F1-Score
LR	0.759	0.793	0.931	0.620	0.744
RF	0.779	0.781	0.793	0.676	0.730
KNN	0.831	0.816	0.758	0.785	0.771

This can help uncover important trends and lead to individualinsights that increase patient and clinician involvement, as well as confidence and blood glucose management, according to research. For diabetic data analysis, we used a binary data classification problem in which the model must predict whether a person has diabetes or not based on all of the features. There are several methods for building the model for binary/multi-class classification. Only a few are classified using Logistics regression and naive Bayes. Gradient DescentStochastic. We built four models and compared their performance on test and train datasets: K-Nearest Neighbours(KNN), Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM). We discovered that the KNN model is the best for four of the five parameters, making it the top model. KNN is the most precise, has the highest AUC, and has the highest f1 score. The logistic regression has the highest recall/sensitivity at 93%. Our research is highly recommendedbecause it is made up of research articles from various sourcesthat can help other academics working on alternative diabeticprediction models.

REFERENCES

[1]. International Diabetes Federation (IDF). IDF diabetes atlas,7th edition. Brussels, Belgium: International Diabetes Federation, 2015.

[2]. Centres for Disease Control and Prevention. Estimates of diabetes and its burden in the United States. Washington, DC: US Department of Health and Human Services, 2017: 1–20.

[3]. Lewis K. "Humans and systems: creating natural interfaces to augment human ability".https://www.ibm.com/blogs/internetof- things/creating-natural-interfaces/ Accessed July 18, 2018.

[4]. Yeung S, Downing NL, Fei-Fei L, Milstein A.," Bedside computer vision—moving artificial intelligence from driver assistance to patient safety". N Engl J Med 2018;378:1271– 1273.

[5]. Young K. "Newly approved software uses AI to improve diabetic retinopathy". https://www.jwatch.org/fw114063/2018/04/12 /newly- approved-software-uses-ai-detect-diabetic Accessed July 18, 2018.

[6]. "Turing Archive for the History of Computing". www. alanturing.net /turing_archive/archive/index/archive index. html Accessed July 18, 2018.

[7]. Pfaffl MW. "A new mathematical model for relative quantification in real-time RT-PCR". Nucleic Acids Res 2001;29: 45e.

[8]. Lo´pez B,Martin C,Vin˜as PH. Special section on artificial intelligence for diabetes. ArtifIntell Med2018;85:27-28.

[9]. Vijan S, Hofer TP, Hayward RA. "Cost-utility analysis of screening intervals for diabetic retinopathy in patients with type 2 diabetes mellitus". JAMA 2000;283:889–896.

[10]. Gulshan V, Peng L, Coram M, et al. "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs". JAMA 2016;316:2402–2410.

[11]. "US Food and Drug Administration. FDA permits marketing

of artificial intelligence-based device to detect certain diabetes-related eye problems". https://www.fda.gov/ News Events/ Newsroom/Press Announcements/ucm604357.htm Accessed July 18, 2018.

[12]. Keel S, Lee PY, Scheetz J, et al. "Feasibility and patient acceptability of a novel artificial intelligence-based screening model for diabetic retinopathy at endocrinology outpatient services: a pilot study. Sci Rep 2018;8:4330.

[13]. Shankaracharya, Odedra D, Samanta S, Vidyarthi AS. "Computational intelligence-based diagnosis tool for the detection of prediabetes and type 2 diabetes in India". Rev Diabetes Stud 2012;9:55–62.

[14]. Corey KE, Kartoun U, Zheng H, Shaw SY. "Development and validation of an algorithm to identify non-alcoholic fatty liver diseasein the electronic medical record". Dig Dis Sci 2016;61:913–919.

[15]. Vyas R, Bapat S, Jain E, Karthikeyan M, Tambe S, Kulkarni BD. "Building, and analysis of protein-protein interactions related to diabetes mellitus using support vector machine, biomedical text mining and network analysis". Comput Biol Chem 2016;65:37–44.

[16]. Lo´pez B, Torrent-Fontbona F, Vin˜as R, Ferna´ndez-Real JM. "Single nucleotide polymorphism relevance learning with random forests for type 2 diabetes risk prediction". Artif Intell Med 2018;85:43–49.

[17]. Shu T, Zhang B, Yan Tang Y. "An extensive analysis of various texture feature extractors to detect diabetes mellitus using facial specific regions". Comput Biol Med 2017;83: 69–83.

[18]. Katigari MR, Ayatollahi H, Malek M, Haghighi MK. "Fuzzy expert system for diagnosing diabetic neuropathy". World J Diabetes 2017;8:80–88.

[19]. Wang L, Pedersen PC, Agu E, Strong DiM, Tulu B. "Area determination of diabetic foot ulcer images using a cascaded two-wtage SVM-based classification." IEEE Trans Biomed Eng 2017;64:2098–2109.

[20]. Mauseth R, Lord SM, Hirsch IB, Kircher RC, Matheson DP, "Greenbaum CJ. Stress testing of an artificial pancreas system with pizza and exercise leads to improvements in the system's

fuzzy logiccontroller". J Diabetes Sci Technol 2015;9:1253–1259.

[21]. Ling SH, Nguyen HT. "Natural occurrence of nocturnalhypoglycemia detection using hybrid particle swarm optimized fuzzy reasoning model". Artif Intell Med 2012;55: 177–184.

[22]. Herrero P, Pesl P, Reddy M, Oliver N, Georgiou P, Toumazou C. "Advanced insulin bolus advisor based on run-torun control and case- based reasoning". IEEE J Biomed Health Inform 2015;19:1087–1096.

[23]. DeJournett L, DeJournett J. "In silico testing of an artificial intelligence-based artificial pancreas designed for use in the intensivecare unit setting". J Diabetes Sci Technol 2016;10: 1360–1371.

[24]. Cvetkovic´ B, Janko V, Romero AE, Kafalı O¨, Stathis K, Lusˇtrek M. "Activity recognition for diabetic patients using a smartphone". JMed Syst 2016;40:256.

[25]. Wang L, Pedersen PC, Strong DM, Tulu B, Agu E, IgnotzR. "Smartphone-based wound assessment system for patients with diabetes". IEEE Trans Biomed Eng 2015;62:477– 488.

[26]. Rigla M, Martı´nez-Sarriegui I, Garcı´a-Sa´ez G, Pons B, Hernando ME. Gestational diabetes management using smart mobile telemedicine. J Diabetes Sci Technol 2018; 12:260–264.

[27]. ClinicalTrails.gov. "Adult accuracy study of the Elite 3 Glucose Sensor" (E3). https://www.clinicaltrials.gov/ct2/show/study/NCT02246582?term= NCT02246582&rank=1& sect=X0123456 Accessed July 18, 2018.

[28]. Neemuchwala H. Sugar. "IQ improves time-in-range by36mins/day, time >180 by 30mins/day, time <70 by 6mins/day". Presented at: American Diabetes Association 78th Scientific Session, June 22–26,2018, Orlando, FL.

[29]. Hutson TM. "Missing data hinder replication of artificial intelligence studies". www.sciencemag.org/news/2018/02/missing-data-hinder- replication-artificial-intelligence-studies? Accessed July 18, 2018.

[30]. Benson T. "Principles of health interoperability" HL7 and SNOMED, 2nd edition. London: Springer Verlag, 2012.