Article

Diabetes Monitoring System in Smart Health Cities Based on Big Data Intelligence

Shadi AlZu'bi [1], Mohammad Elbes [1], Ala Mughaid [2], Noor Bdair [1], Laith Abualigah [3,4,5,*], Agostino Forestiero [6,*] and Raed Abu Zitar [7]

[1] Faculty of Science and IT, Al-Zaytoonah University of Jordan, Amman 11733, Jordan
[2] Department of Information Technology, Faculty of Prince Al-Hussien Bin Abdullah II for IT, The Hashemite University, P.O. Box 330127, Zarqa 13133, Jordan
[3] Hourani Center for Applied Scientific Research, Al-Ahliyya Amman University, Amman 19328, Jordan
[4] Applied Science Research Center, Applied Science Private University, Amman 11931, Jordan
[5] Faculty of Information Technology, Middle East University, Amman 11831, Jordan
[6] Institute for High Performance Computing and Networking, National Research Council of Italy, 87036 Rende, Italy
[7] Sorbonne Center of Artificial Intelligence, Sorbonne University-Abu Dhabi, Abu Dhabi 38044, United Arab Emirates
* Correspondence: aligah.2020@gmail.com (L.A.); agostino.forestiero@icar.cnr.it (A.F.)

Abstract: Diabetes is a metabolic disorder in which the body is unable to properly regulate blood sugar levels. It can occur when the body does not produce enough insulin or when cells become resistant to insulin's effects. There are two main types of diabetes, Type 1 and Type 2, which have different causes and risk factors. Early detection of diabetes allows for early intervention and management of the condition. This can help prevent or delay the development of serious complications associated with diabetes. Early diagnosis also allows for individuals to make lifestyle changes to prevent the progression of the disease. Healthcare systems play a vital role in the management and treatment of diabetes. They provide access to diabetes education, regular check-ups, and necessary medications for individuals with diabetes. They also provide monitoring and management of diabetes-related complications, such as heart disease, kidney failure, and neuropathy. Through early detection, prevention and management programs, healthcare systems can help improve the quality of life and outcomes for people with diabetes. Current initiatives in healthcare systems for diabetes may fail due to lack of access to education and resources for individuals with diabetes. There may also be inadequate follow-up and monitoring for those who have been diagnosed, leading to poor management of the disease and lack of prevention of complications. Additionally, current initiatives may not be tailored to specific cultural or demographic groups, resulting in a lack of effectiveness for certain populations. In this study, we developed a diabetes prediction system using a healthcare framework. The system employs various machine learning methods, such as K-nearest neighbors, decision tree, deep learning, SVM, random forest, AdaBoost and logistic regression. The performance of the system was evaluated using the PIMA Indians Diabetes dataset and achieved a training accuracy of 82% and validation accuracy of 80%.

Keywords: big data intelligence; classification; data science; deep learning; E-health; healthcare analytics; intelligent diagnosis; machine learning; diabetes prediction

1. Introduction

Diabetes is a major concern in different parts of the world, especially in the Middle East and North Africa. As a result of the prevalence of disease-causing factors, which include a sedentary lifestyle and unhealthy diets, diabetes has become a chronic condition affecting an increasing number of patients [1,2]. Intelligent employment of artificial intelligence (AI) techniques has proved significant enhancements in many applications,

such as business [3], agriculture [4,5] social media [6], security applications [7], intelligent transportation [8], supply chain management [9], sustainable cites [10] and many more intelligent applications [11]. Recent advances in the Internet of Things, cloud computing, and artificial intelligence have transformed the traditional healthcare system into an intelligent one. Healthcare services can be greatly improved using the Internet of Things, artificial intelligence, and deep learning. Advanced methods and scientific theories are now generating vast amounts of digital data that can be used to create clinical applications that help diagnose diabetes more efficiently [12].

Diabetes is a disease that causes high blood sugar [13]. The patient's pancreas does not produce suitable amounts of insulin or cannot effectively consume the insulin produced by the pancreas. There are two major types of diabetes, type 1 and type 2 diabetes. Type 2 diabetes is considered the most common form of diabetes by 90% of all diabetic patients. The World Health Organization (WHO) reported that about 1.6 million people yearly die due to diabetes [14].

Predicting and diagnosing diabetes is one of the major challenges in the medical industry and healthcare sector; it depends on several factors, including the chemical examination and the different symptoms and signs that appear on the patient. Diabetes is seen as the world's deadliest disease for human life because the disease has proven to be the precursor to other health problems. Many individuals who suffer from obesity seem to have diabetes. The disease also kills more people yearly than AIDS and breast cancer combined. In all diabetes cases, sugar increases in the bloodstream. This is due to the pancreas not producing enough insulin. Both types 1 and 2 diabetes may be caused by a combination of environmental or genetic factors. It is unclear what those factors are till the current moment. Figure 1 illustrates the diabetes types.
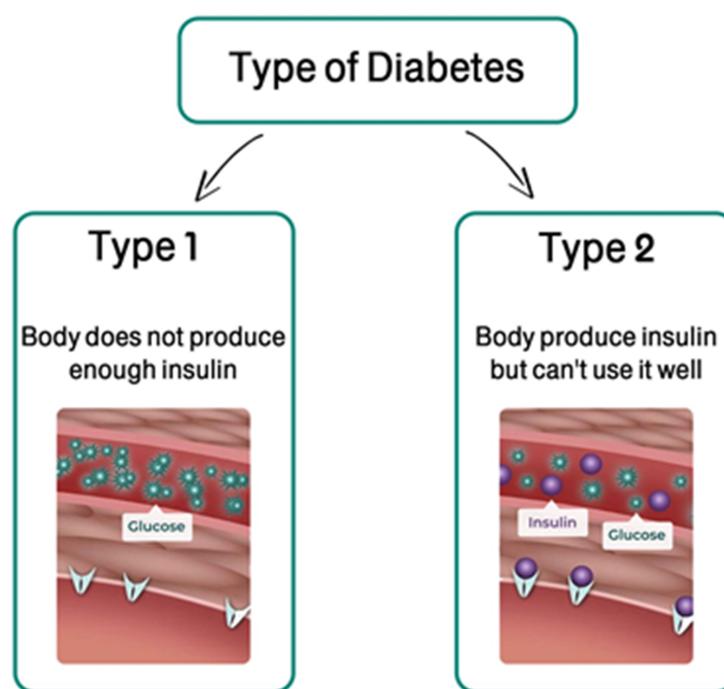
Figure 1. The types of diabetes.

Diabetes is principally caused by obesity or high glucose level [15]. Due to increased glucose levels in the blood, some symptoms arise within the flesh, such as frequent urination and extreme hunger. Normal glucose levels are 70 to 99 mg/dL. Furthermore, a human is considered pre-diabetic if the body glucose concentration is 100–125 mg/dL [16]. Detection of diabetes in its early stages is one of the most critical challenges in the healthcare industry. Diabetes has no permanent cure. Hence, early detection is crucial in the treatment process [17]. Figure 2 illustrates the leading causes of diabetes.

Figure 2. The main causes of diabetes.

The healthcare sector is one of the biggest sectors in the world economy, with a continuous increase in the yearly income and investment [18]. Currently, several clinical procedures, such as diabetic monitoring, blood testing, and pressure monitoring, can be conducted at a remote location in real time [18]. Due to the great advances in wearable devices, fast data acquisition and processing of medical data are possible.

AI has changed the decision-making process for people in the medical field, as they can diagnose patients early; AI plays a vital role in disease prevention. Electronic health records systems based on AI can be used to collect advanced data reporting about patients [19]. Machine learning (ML) uses previously labeled data to predict the future intelligently, with no explicit programming [20]. It is a field that is comprehensively and inalienably specified with flag handling most notably through data-driven learning systems [21,22] and has rapidly obtained its application in several daily life applications, such as healthcare, business, and many more. ML algorithms are valuable for disease detection [23]. In this work, we present an intelligent system for diabetes prediction using the PIMA Indians Diabetes dataset, and ML algorithms, such as K-nearest neighbors, decision tree classifier, deep learning (multilayer perceptron (MLP)), support vector machine (SVM), random forest classifier, an AdaBoost classifier, and logistic regression.

The motivation for automatic prediction of diabetes in healthcare systems is to improve early detection and intervention of the disease. By utilizing machine learning and other predictive methods, healthcare systems can identify individuals at risk for diabetes before symptoms appear, allowing for early intervention and prevention of serious complications. Additionally, automatic prediction can help improve the efficiency and cost-effectiveness of diabetes management by reducing the need for manual screening and allowing healthcare providers to focus on those individuals most at risk. Furthermore, this approach can also help to reduce health disparities by identifying and targeting high-risk populations that may not have access to traditional screening methods. Overall, the goal is to improve the overall health outcomes for individuals with diabetes and reduce the burden of the disease on the healthcare system.

Healthcare systems for diabetes play a crucial role as a starting point for other health monitoring. Diabetes is a chronic condition that often leads to other serious health complications such as heart disease, kidney failure and blindness. By implementing effective and efficient systems for the management of diabetes, healthcare providers can not only improve the health outcomes for individuals with diabetes, but also identify and intervene in the early stages of other chronic conditions. Additionally, by using diabetes as a starting point, healthcare systems can learn from the methods and strategies that have been successful in managing diabetes and apply them to other chronic conditions. This can help improve the overall health of the population and reduce the burden of chronic disease.

Healthcare systems for diabetes can be generalized to other types of health monitoring by utilizing similar frameworks and methods for early detection and intervention. For example, implementing regular check-ups and screenings for other chronic conditions, such as heart disease or hypertension, can help identify and manage these conditions in their early stages. Additionally, utilizing the same data-collection and analysis techniques, such as machine learning and predictive modeling, can aid in identifying individuals at risk for these conditions. Furthermore, providing education and resources for individuals to manage their overall health can also help in preventing and managing other chronic conditions. By adopting a holistic approach to healthcare that focuses on early detection and intervention, healthcare systems can effectively manage and prevent a variety of chronic conditions.

This paper aims to contribute to the knowledge according to the following points:

- Automatic prediction of diabetes in healthcare systems is to improve early detection and intervention of the disease. Utilizing machine learning and other predictive methods, healthcare systems can identify individuals at risk for diabetes before symptoms appear.
- Help to improve the efficiency and cost-effectiveness of diabetes management, allowing healthcare providers to focus on individuals most at risk, and help to reduce health disparities by identifying and targeting high-risk patients.
- Implement an effective and efficient system for the management of diabetes to improve the overall health of the population and reduce the burden of chronic disease.

The remaining of this manuscript is organized as follows: Section 2 reviews previously conducted research in the field. The proposed methodology is explained in detail in Section 3. Section 4 illustrates the achieved results. The conclusion is presented in Section 5.

2. Related Literature

Many previous related works have been conducted in the field of automatic healthcare processes. In this section, the research topic and the scope of the study are clearly defined, and a literature review using scholars databases is conducted to evaluate the relevance of each source. The rest of this section includes the most recent and related literature. We group the sources based on the themes and concepts they address, and consider the impact of the sources, considering the credibility and reputation of the authors.

Early prediction and diagnosis of diabetes has been investigated by many researchers in the field of machine learning. Those researches gained great popularity recently due to the improved accuracy and efficiency in diabetes predictions. The standing of research conducted in this field lies in the early prediction and diagnosis of diabetes, which allows for better treatment. Early prediction of the disease saves human life, as the disease affects diverse organs of the biosystems, especially the blood veins and nerves.

ML provides effective algorithms to obtain information by building models from previously diagnosed medical data, which could be collected from records of people with diabetes. In 2019, Faruque and Sarker [24] explored some risky aspects connected to diabetes using AI techniques. In 2018, Dagliati et al. [25] used the MOSAIC project, which was funded from EU; they constructed a predictive model set of type 2 diabetes mellitus (T2DM) using a data mining pipeline based on EHRs data of about 1000 diabetic infected people. After processing missing data using the ML algorithm RF and applying necessary techniques to cover class unbalancing, the researchers used logistic regression with stepwise feature selection to predict the onset of neuropathy, retinopathy, or nephropathy at different times. They considered age, gender, time of diagnosis, body mass index (BMI), glycated hemoglobin (HbA1c), smoking habit, and hypertension. They achieved an accuracy of up to 83%.

In 2015, Kandhasamy and Balamurali [26] conducted a comparison study on some ML Algs used in diabetes prediction. Several ML classifiers were employed to detect diabetes. The studies concluded that the best performance (accuracy = 73.82%) was achieved using decision tree J48 classifier. In another scenario, where pre-processing is applied to the

training dataset, the researchers achieved better performance according to the initial study. In 2021, Ravaut et al. established and authorized a ML model based on the population level for predicting type 2 diabetes occurrence long enough before the start of diabetes. They used normally gathered executive health information [16].

In 2019, Alam et al. implemented a neural network and some ML Algs for predicting diabetes. ANN overcame other techniques with accuracy of 75.7%; thus, it could be efficient in assisting doctors with making medical decisions [27]. In 2018, Afzali and Yildiz presented a high-performance diagnostic system with great accuracy by modifying the k-means clustering technique. Initially, a novel clustering method detected uncertain, noisy and inconsistent data records and extracted them from the dataset. Consequently, a diabetes prediction system was implemented using SVM. Employing the proposed Pima system for classifying Indians Diabetes dataset (PID) led to a good accuracy of 99.64% with 10-fold cross validation [28].

Many previous research works have been conducted in the field of health care platforms that deal with big data in the cloud and also in diabetes on the cloud, such as [29–32]. According to the reviewed literature on diabetes prediction, there are several open gaps: lack of generalizability—many studies on diabetes prediction have used small, homogeneous samples, making it difficult to generalize the findings to other populations; limited use of electronic health records (EHRs)—despite the increasing availability of EHRs, many studies on diabetes prediction have not utilized this data source, which could provide a wealth of information for predicting diabetes; and insufficient attention to social determinants of health—social determinants of health, such as poverty and race, have been shown to play a significant role in the development of diabetes, yet many studies on diabetes prediction have not taken these factors into account.

Previously in the literature, there can be noticed a limited consideration of the impact of comorbidities: diabetes is often accompanied by other chronic diseases, yet most studies on diabetes prediction have not considered the impact of comorbidities on diabetes risk. Furthermore, previous studies lack long-term follow-up and display limited use of multifactorial and advanced machine learning methods which could improve the performance of prediction models. Overall, there is a need for more comprehensive, generalizable, and long-term studies that take into account social determinants of health, comorbidities, and advanced machine learning methods, and use electronic health records as a data source.

3. Methodology

3.1. The Proposed System

Diabetes is caused by the pancreas not producing enough insulin or by the body's inability to consume the insulin produced. According to reports generated by WHO, the percentage of people affected by diabetes in the US is very high. Continuous glucose monitors (CGMs) are becoming an increasingly popular way to monitor blood glucose. The method works by inserting a right needle wearable device into the skin that monitors blood sugar levels (tissue levels). The CGM reports blood glucose trends and provides alerts to stop hypoglycemia and supply the best glycemic monitoring [33]. The data are connected to the smartphone via Bluetooth, and then the information is connected to the cloud computing (internet or network) server. For people with diabetes, if the blood sugar levels rise, the system immediately generates various alerts. If the percentage of diabetes exceeds the permissible limit, especially for diabetic patients, the emergency will be called directly. This work illustrates how machine learning is implemented to predict diabetes. Our dataset contains 768 records with 9 features, and the output is the probability of whether the patient has diabetes or not. Figure 3 illustrates the proposed intelligent health system for predicting diabetes in the cloud of health cities. The remote sensing of the patient's current status is transferred automatically into the intelligent healthcare system and stored in an EHR for further processing. It can be noticed from this figure that the patients data are processed in the cloud and trained according to pretrained models, and the required action could be activated automatically based on the results obtained from the

cloud processing. Figure 4 shows the proposed diabetes prediction/monitoring system in smart health cities that is running in the cloud.
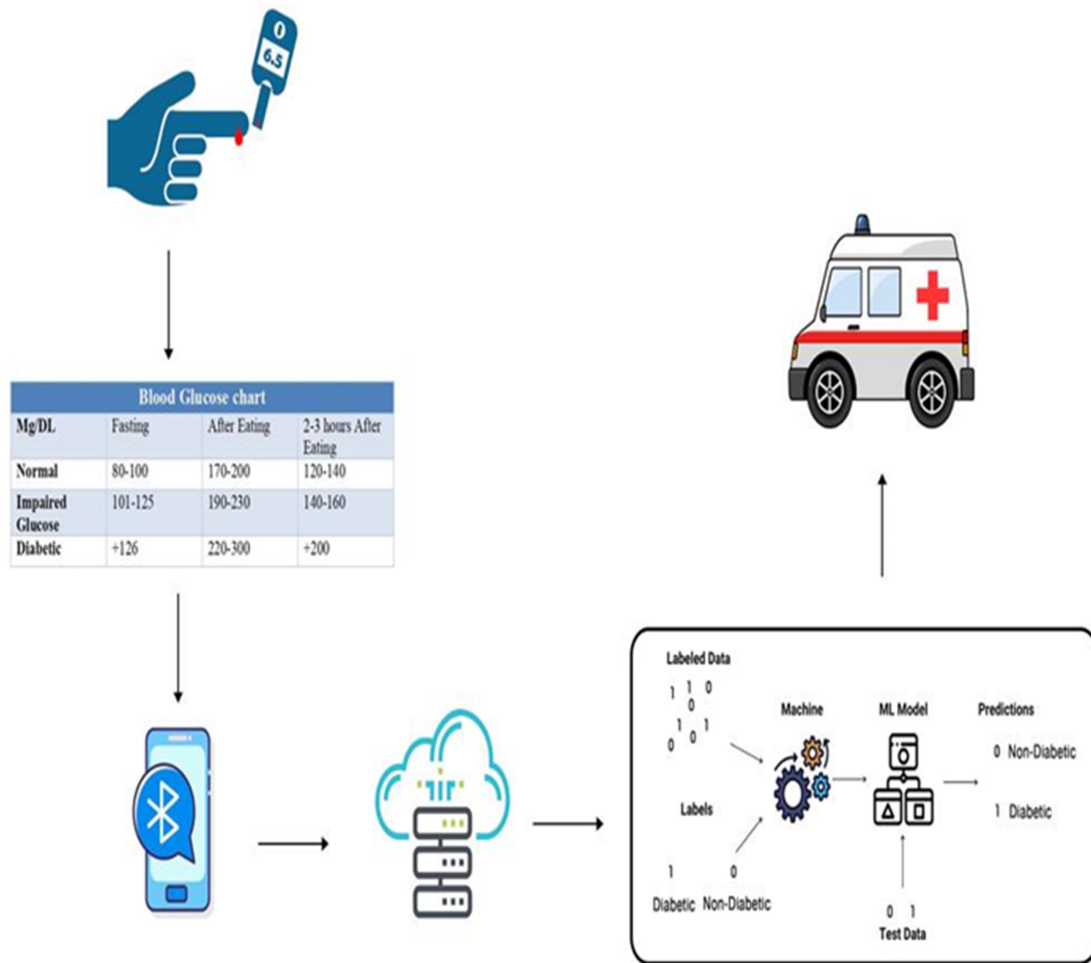

Figure 3. Cloud computing architecture related to diabetes healthcare.

The integration of cloud computing and Bluetooth-connected devices can enhance healthcare monitoring by providing remote and real-time data collection, storage, and analysis. With cloud computing, data collected from Bluetooth devices, such as glucose monitors, blood pressure cuffs, and activity trackers, can be transmitted and stored securely on remote servers for easy access by healthcare providers. This allows for more efficient and frequent monitoring of patients and can assist in the early detection and intervention of any health issues. Additionally, cloud computing can enable the use of advanced analytics and machine learning algorithms to identify patterns and make predictions, improving the overall management of chronic conditions, such as diabetes.

Notifications in healthcare monitoring systems are set to trigger when certain thresholds are met. They are typically based on established clinical guidelines and are specific to the condition being monitored. For example, notifications may be sent when blood sugar levels exceed a certain level, indicating a need for immediate intervention. These thresholds are fixed, as they are based on the expert's knowledge and evidence-based research and are intended to provide objective measures for emergency response. The notifications are used to alert healthcare providers, who will then determine if an ambulance should be called.
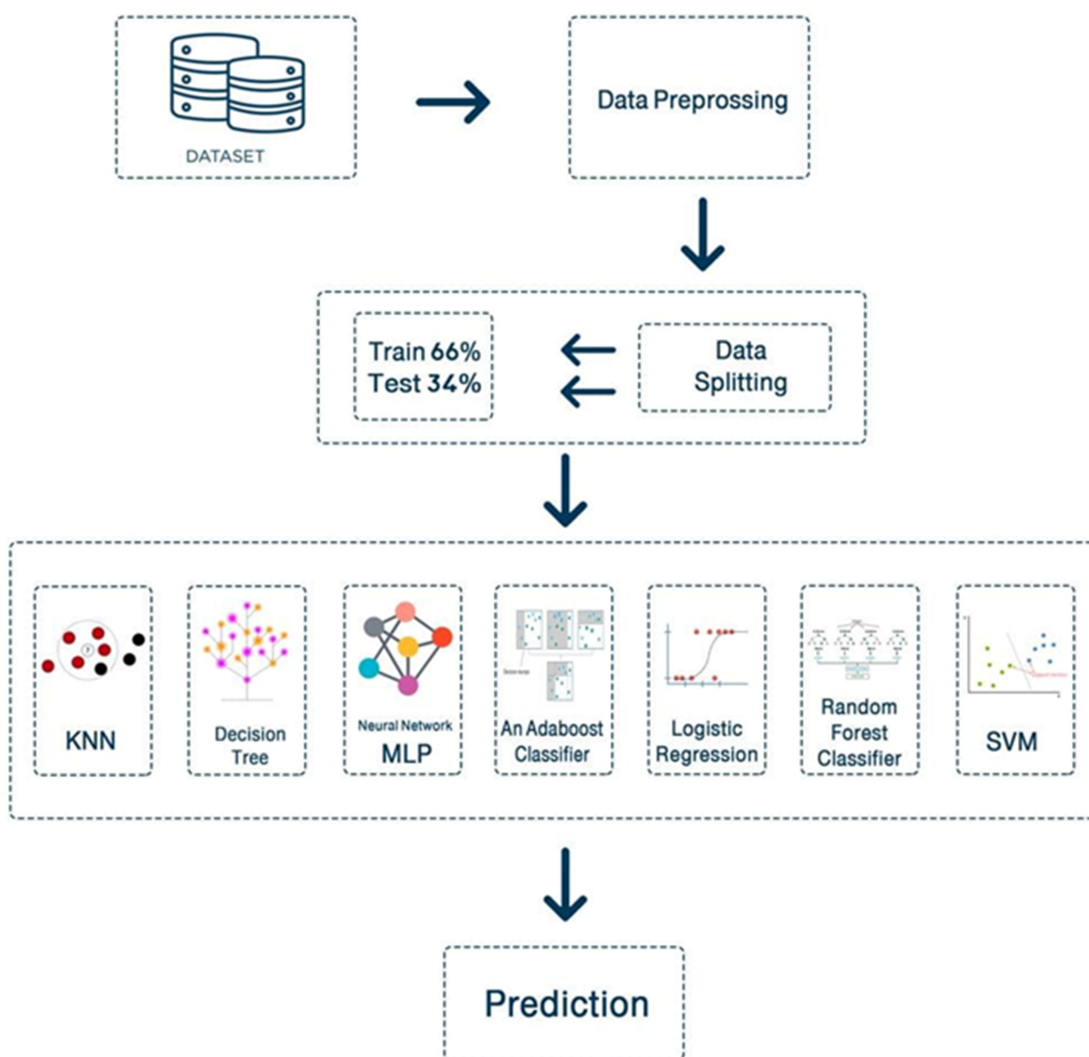
Figure 4. The proposed diabetes prediction/monitoring system in intelligent health cities.

3.2. The Dataset

The database contains eight features, and the first one is pregnancy; at least 5% of pregnant women have diabetes during pregnancy. This is called gestational diabetes. The second feature is glucose. Glucose tolerance declines steadily with age, leading to an increase in the incidence of type 2 diabetes. The third feature is the patient's blood pressure, as it was found that having high blood pressure would raise the risk of diabetes. The fourth feature is skin thickness, as one of the complications of diabetes is in the skin; scleroderma causes an increase in the thickness of the skin on the back of the hands and difficulty moving the fingers. The fifth is 2-hour insulin; the existence of insulin resistance, a combination of a small amount of insulin and its low level of effectiveness, leads to a deviation from the acceptable glucose level. One of the most important features is BMI, which is a more substantial risk factor for diabetes than genetics, and weight loss can prevent or even reverse diabetes. The seventh feature is diabetes pedigree, which appears in the examination of blood sugar levels. The last feature is age.

Some preprocessing steps must be performed before starting any classification process. The preprocessed dataset are then split into 70–30 percent for training and testing respectively. Before performing the classification process, several parameters have to be optimized for achieving a valid results, such as, batch size, epochs, optimizer, augmentation, and normalization before being input to the pertained TL models used.

There may be no specific age for the onset of type 2 diabetes, but the age factor dramatically increases the chances of developing the disease. The dataset contains 768 labels for diabetes, where a label 1 means the patient has diabetes and 0 otherwise. Of these 768 records, 500 are labeled as 0 and 268 as 1. Table 1 illustrates a sample from the collected dataset. Figure 5 shows the total outcome of diabetic and non-diabetic patients.

**Table 1.** A sample from the database.

Pregnancies	6	1	8	1	0
Glucose	6	1	8	1	0
Blood Pressure	148	85	183	89	137
Skin Thickness	72	66	64	66	40
Insulin	35	29	0	23	35
BMI	33.6	26.6	23.3	28.1	43.1
Diabetes Pedigree	0.6	0.3	0.6	0.1	2.2
Age	50	31	32	21	33
Outcome	1	0	1	0	1
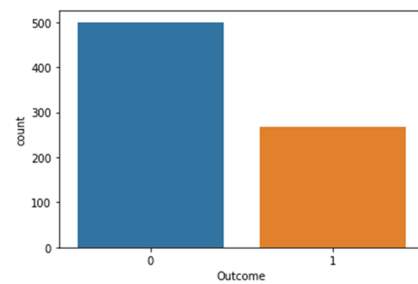
Figure 5. The total outcome of diabetic and non-diabetic patients.

*3.3. Data Visualization*

The correlation between each column is visualized using heat-map; from the output, the lighter colors indicate more correlation. The correlation is shown for each pair of features, such as age and pregnancies, or BMI and skin thickness, etc. Figure 6 illustrates the data visualization using heat-map.
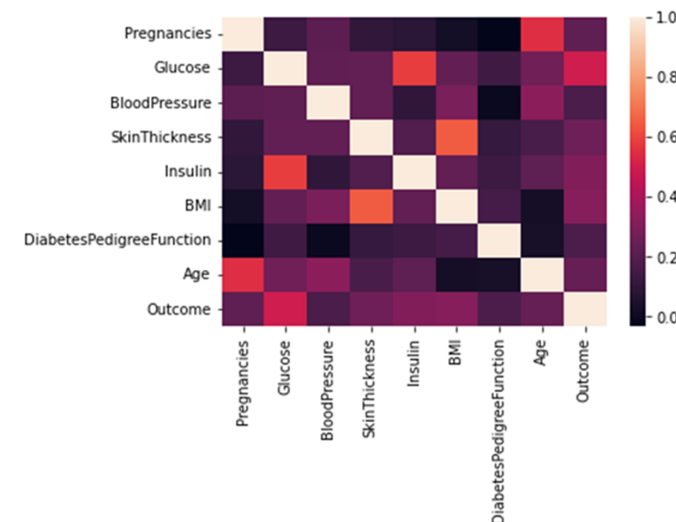

Figure 6. Data visualization using heat-map.

Figure 7 shows that the most significant feature is the glucose level as shown in the figure. Each feature is presented as a binary digit for predicting targets or not. Figure 7 illustrates the most important extracted feature.
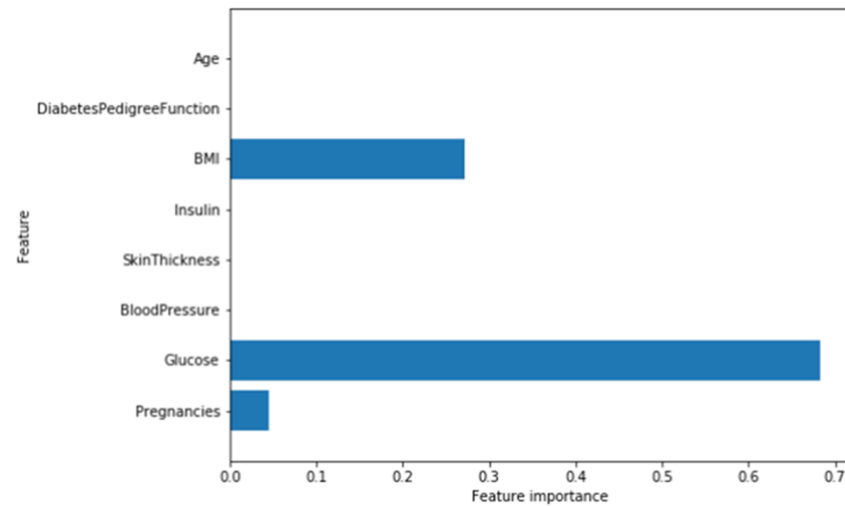
Figure 7. The most important extracted feature.

Scatter plots are usually used to plot data points on vertical and horizontal axes with the attempt to show how much one variable is related to another or to look for a relationship between different variables, where the better the correlation, the tighter the points will hug the line. Figure 8 illustrates scatter plots visualizing glucose vs. BMI vs. pregnancies.
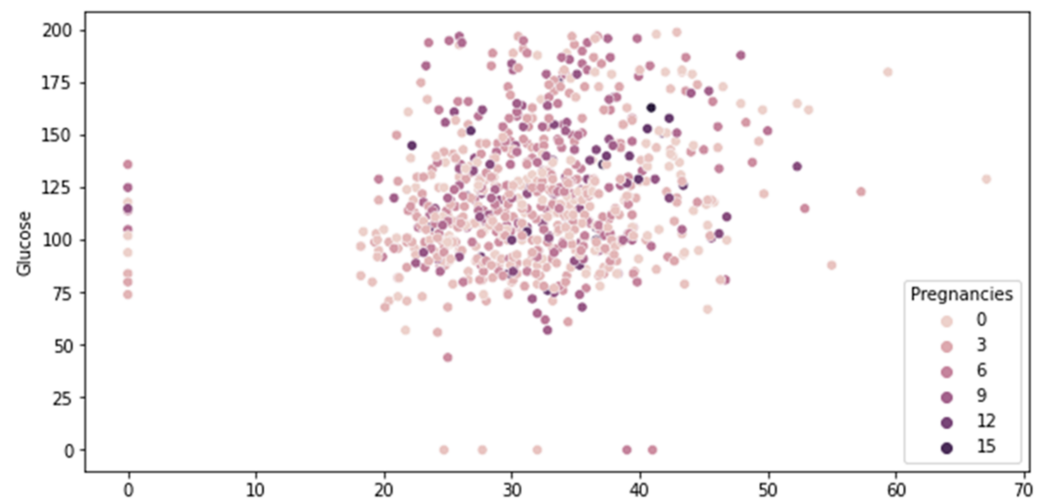
Figure 8. Visualizing glucose vs. BMI vs. pregnancies using scatter plots.

3.4. Proposed Methods

In this research, we utilize the continuous glucose monitor (CGM), which is connected to the smart phone via Bluetooth. Then, the collected information is uploaded to a server for further processing. In the proposed system, we implement 7 ML techniques to predict diabetes, implemented in Python programming language. The following machine learning algorithms are supervised:

- Logistic regression;
- Decision trees (DT);
- Support vector machine (SVM);
- Multi-layer perceptron (MLP);

- AdaBoost;
- Random forest.

The k-nearest neighbor (KNN) is unsupervised:

Supervised learning algorithms are used when there are labeled data available and the goal is to train a model to make predictions or classify data based on that labeled information. Unsupervised learning algorithms are used when there are no labeled data available and the goal is to find patterns or structure in the data without any prior knowledge.

3.4.1. K-Nearest Neighbors (KNN)

KNN uses proximity to make predictions by grouping individual data points. Model building contains training data storing. Then, KNN locates the nearest points in the training data and groups the records as nearest neighbors.

3.4.2. Decision Tree Classifier (DT)

DT is used in many applications including data mining, statistics, and ML. DT guarantees a 100% accuracy on training subset, but the results may not be accepted on the test set. We improved the testing accuracy by implementing different max depth, and minimizing the overfitting by preventing the deepness of the tree, which led to a low training accuracy, but an improvement on the test set.

3.4.3. Random Forest Classifier (RF)

RF belongs to the supervised ML techniques; it is fast and guileless, and very effective in a diversity of applications. The hidden secret behind the effectiveness of RF is the employment of numerous humble decision trees in the training phase and the popular ballot in the classification phase. The model is trained with typical parameters using the training data, and kept in the rcf file. Then, the proposed model performance is evaluated using the test dataset.

3.4.4. Support Vector Machine (SVM)

SVM is a supervised ML technique which is used frequently in classification. It makes a discriminative classifier formally by the separation of hyper-planes.

3.4.5. Deep Learning for Diabetes Prediction

Deep learning (DL) is an advanced ML which emulates the procedure of human brains acquiring knowledge. DL plays a significant role in data science, which implements predictive and statistics modeling. Multilayer perceptron (MLP) in deep learning is a feed forward ANN that produces outputs directly from inputs without any convolutional feedback. MLP is categorized by numerous layers of connected input nodes as a layer graph between inputs and outputs. MLP accuracy is not convincing because of the scaling of data.

3.4.6. An AdaBoost Classifier

AdaBoost starts by correcting a classifier on the initial dataset, then it corrects more copies of the estimator in the targeted dataset with attuned weights for the wrongly classified cases. Moreover, consequent estimators emphasize more on interesting circumstances.

3.4.7. Logistic Regression

Logistic regression is a supervised ML algorithm used to predict target label probability. The label environment or reliant variable is dichotomous, which leads to binary classes possibilities.

4. Experimental Results and Analysis

### 4.1. Evaluation Metrics and Model Evaluation

Machine learning help the physicians understand and treat diabetic disorders. The performance evaluation used in this research is calculating the accuracy. After performing the training process, accuracy is measured using the score(X_train,Y_train), which measures the accuracy of the model against the testing data (how well the model explains the data it was trained with).

Performance evaluation is an important step in machine learning applications, as it allows researchers to measure the effectiveness of the model. It can be done using a variety of metrics, such as accuracy, precision, recall, F1 score, and ROC AUC. However, not all metrics are suitable for all applications, and researchers may choose to use specific metrics based on the research goals and characteristics of the data. For example, accuracy may be appropriate for balanced datasets, while the F1 score may be more suitable for imbalanced datasets. Additionally, it is not practical to use all metrics, as it can be time consuming and may not provide clear results. Researchers must carefully select the appropriate evaluation metrics for their specific application to ensure the effectiveness of the model.

4.2. Experiments

The major target of the proposed system is to predict diabetes and implement the algorithm with the highest accuracy to predict the existence of diabetes in the early stage.

In this study, we used K-nearest neighbors, decision tree classifier, deep learning (multilayer perceptron (MLP)), support vector machine (SVM), random forest classifier, an AdaBoost classifier and logistic regression. All these algorithms were applied on the PIMA Indian dataset. Of the records, 80% in the dataset were used for training; the remaining 20% was used for testing. Table 2 lists the obtained results, and the results are visualized in Figure 9.
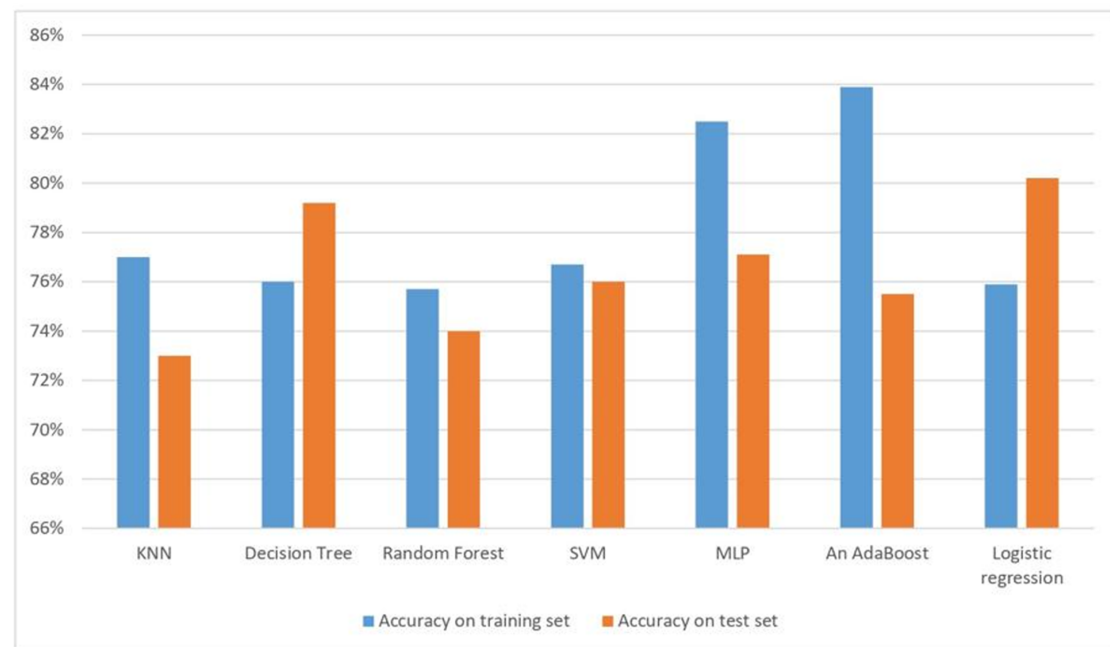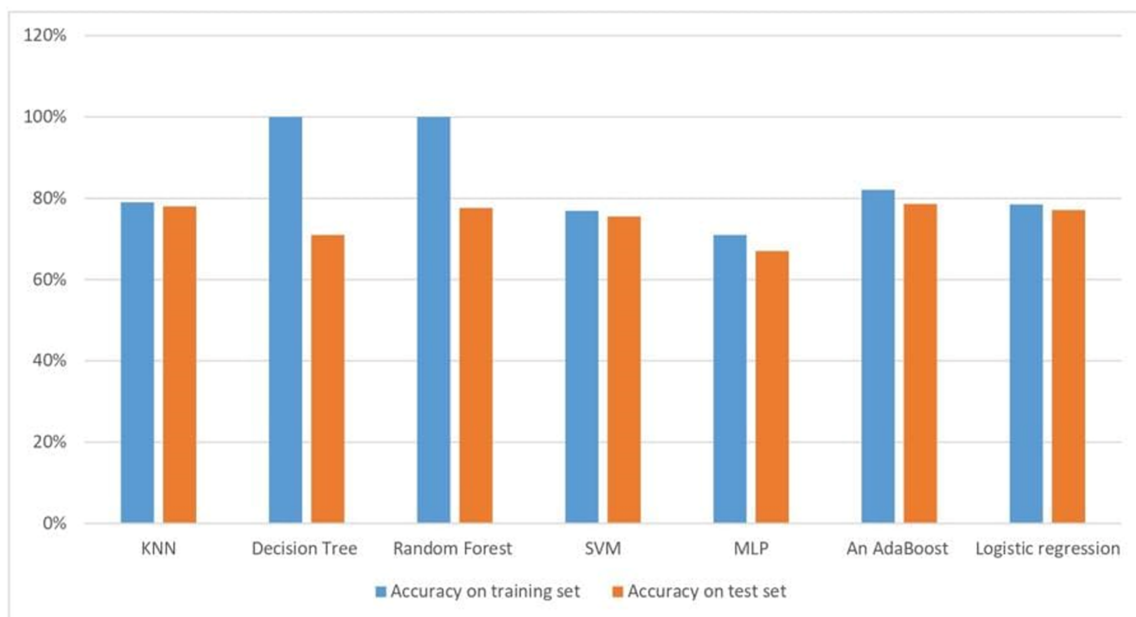

Figure 9. The classification algorithms accuracy using 80% training and 20% testing dataset.

Table 2. The classification algorithms accuracy using 80% training and 20% testing dataset.

Classification Algorithms	Training Accuracy	Testing Accuracy
KNN	77%	73%
Decision Tree	76%	79.2%
Random Forest	75.7%	74%
SVM	76.7%	76%
MLP	82.5%	77.1%
An AdaBoost	83.9%	75.5%
Logistic regression	75.9%	80.2%

The performance evaluation used in this research is calculating the accuracy. Compared to previous research, the result of the accuracy of the algorithms in this proposed system is very competitive. The results of the proposed research suggest that an AdaBoost classifier is more suitable for the detection of diabetes in healthcare.

The KNN accuracy for diabetes prediction is 79% for training accuracy and 78% for testing accuracy, the training accuracy using DT is 100%. When training a deep learning model to predict diabetes, the training accuracy is 71% and testing accuracy is 67%. For RF, the training accuracy is 100% and testing accuracy is 77.6%. The SVM has a training accuracy of 76.9% and testing accuracy of 75.5%. All results are presented in Table 3, and visualized in Figure 10.

Figure 10. The first result for classification algorithms accuracy.

The training accuracy with DT is 100%, but the testing accuracy is not impressive. After setting the max_depth=3, this leads to a lower training accuracy, but testing accuracy is improved; the training accuracy becomes 77% and the testing accuracy becomes 74%. As we mentioned before, scaling the data negatively affected the MLP accuracy. Therefore, we re-scaled the data to achieve significant requirements for predicting diabetes accurately; the training accuracy becomes 82% and the testing accuracy becomes 80%. However, if iterations are increased, more alpha and other parameter will be added to the model weights, this leads to a training accuracy of 79% and a testing accuracy of 79%. RF has a training accuracy of 100% and testing accuracy of 77.6%. Setting the max_depth to 2, a

lower training accuracy is achieved, but testing accuracy is improved; the training accuracy is 76.2% and the testing accuracy is 75%. The proposed system proved the best two models are the deep learning model after scaling and an AdaBoost classifier by the same training accuracy of 82% and testing accuracy of 80% for the DL model. After scaling the data, the training accuracy is 78.6% for the AdaBoost classifier. Table 4 presents the final result for classification algorithms accuracy, visualized in Figure 11.

Table 3. The first result for classification algorithms accuracy.

Classification Algorithms	Training Accuracy	Testing Accuracy
KNN	79%	78%
Decision Tree	100%	71%
Random Forest	100%	77.6%
SVM	76.9%	75.5%
MLP	71%	67%
An AdaBoost	82.1%	78.6%
Logistic regression	78.5%	77.1%

Figure 11. The final result for classification algorithms accuracy.

Table 4. The final result for classification algorithms accuracy.

Classification Algorithms	Training Accuracy	Testing Accuracy
KNN	79%	78%
Decision Tree	77%	74%
Random Forest	76.2%	75%
SVM	76.9	75.5%
MLP	82%	80%
An AdaBoost	82%	78.6%
Logistic regression	78.5%	77.1%

The maximum depth of the tree is defined by max_depth. Here, a depth three is employed to improve the proposed decision tree. Over-fitted DTs are achieved due to the zero value, which is set by default. max_depth in RF denotes each tree depth in the forest, more splits are accordingly related to the deeper trees, and extra information from data is released.

4.3. Analysis and Discussion

Diabetes mellitus is considered one of the risky diseases. Obesity, age, sedentary lifestyle, living way, inherited diabetes, poor diet, high blood pressure, and many more habits are the main causes of diabetes. In this study, we used KNN, DT classifier, deep learning using MLP, SVM, RF, AdaBoost classifier and logistic regression classification algorithms as supervised learning. Unsupervised learning methods are mostly used for dimensionality lowering. Undue attributes in the diabetes dataset are hinder the accuracy of the classifier.We can have the integration of supervised and unsupervised learning for the best prediction of diabetes. Table 5 reviews the summary of major findings of diabetes prediction from the literature. The results of the proposed research suggest that the AdaBoost classifier is more suitable for the detection of diabetes in healthcare.

There are several limitations of using machine learning (ML) in diabetes prediction in health-monitoring work:

1. Data bias: ML models are only as good as the data they are trained on, and if the data are biased, the model will also be biased. This can lead to inaccurate predictions and poor performance, particularly for underrepresented groups.
2. Lack of interpretability: Many ML models, particularly deep learning models, are considered "black boxes" and can be difficult to interpret. This can make it challenging for researchers and healthcare providers to understand how the model is making its predictions and identify potential errors.
3. Limited generalizability: ML models are trained on specific datasets, and may not perform well on new or unseen data. This can be a significant limitation in healthcare, where data are often collected from diverse populations with unique characteristics.
4. Lack of domain knowledge: Many machine learning algorithms and models are complex and require a deep understanding of the domain and data to be used effectively.
5. Limited data availability: In some cases, there may not be enough data available to train effective ML models, which can be a significant limitation in healthcare, where data are often collected from diverse populations with unique characteristics.
6. Lack of explainability and transparency: ML models are often considered "black boxes" which can be difficult for healthcare practitioners and patients to understand the decision-making process and the factors that influence the predictions.

Overall, while ML can be a powerful tool for diabetes prediction, it is important to consider these limitations and use appropriate techniques to address them.

Using machine learning (ML) in diabetes prediction in health monitoring can make significant contributions to society by improving early detection and intervention of the disease, reducing healthcare costs, and improving the overall health outcomes for individuals with diabetes. ML-based models can assist healthcare providers in identifying individuals at risk for diabetes and in monitoring the progression of the disease, allowing for more efficient and cost-effective management. Additionally, ML can also help to reduce health disparities by identifying high-risk populations that may not have access to traditional screening methods. For technology transfer, researchers and industry partners need to work together to develop and test the models in real-world settings, identify the challenges and limitations of the models, and adapt the models to meet the specific needs of the target population. Furthermore, regulatory bodies, such as FDA, should be involved to ensure the safety and efficacy of the developed products.

As specialists in the field guarantee, the prediction life cycle is based on the amount of data used, and as mentioned earlier, this research could be a part of a huge project that helps in managing the healthcare sector based on AI. It could be efficient for a long time when more data are collected using the proposed system.

There are several limitations of using AI in the automation of the healthcare sector, including the following:

- Data quality and availability: The quality and availability of data are a major challenge for AI in healthcare. Health data are often siloed and not in a format that can be easily analyzed by AI algorithms.
- Privacy and security: The handling of sensitive health data is a concern for privacy and security. There is a risk of data breaches and unauthorized access to patient information.
- Ethical and legal issues: The use of AI in healthcare raises ethical and legal questions, such as issues related to bias, accountability, and informed consent.
- Lack of regulatory framework: There is currently a lack of a clear regulatory framework for AI in healthcare, which makes it difficult to ensure the safety and efficacy of AI-powered medical devices and systems.
- Technical limitations: AI algorithms have a limited ability to handle complex medical conditions and perform well when there are limited data available. The development of AI algorithms for healthcare also requires domain-specific expertise.
- Resistance to change: The healthcare sector is often slow to adopt new technologies, and there may be resistance to the use of AI from healthcare professionals and patients.

Despite these limitations, AI has the potential to significantly improve the efficiency, accuracy, and cost-effectiveness of healthcare delivery. However, careful consideration of these limitations is necessary to ensure the responsible and ethical use of AI in healthcare.

Overall, ML-based systems for diabetes prediction have the potential to improve the lives of millions of people, but technology transfer is an essential step to ensure that these systems are implemented in a way that maximizes their potential benefits.

Finally, the following summarizes the contribution of this work to society:

Automatic prediction of diabetes in healthcare systems using machine learning and predictive methods can improve early detection and intervention of the disease, leading to better management and cost-effectiveness. By identifying individuals at risk before symptoms appear, healthcare providers can focus their efforts on high-risk patients, reducing health disparities. Implementing an efficient and effective system for diabetes management can improve overall population health and reduce the burden of chronic disease.

Table 5. Summary from previously conducted research in diabetes prediction.

Reference No.	Experimented Algorithms	Best Algorithm
[17]	SVM NB KNN DT	DT
[2]	SVM NB KNN ANN	ANN
[18]	ANN RF K-means	ANN
[19]	SVM Vs kmeans And SVM on PID dataset	K-Mean SVM

5. Conclusions and Future Work

Early prediction of illness can be controlled and save human lives, as illness affects completely different organs of the physical body that therefore damage an oversized variety of the body's systems, such as the blood veins and nerves. The prosperous detection of diabetes may be an important medical issue for physicians and researchers.

In this article, we utilized machine learning to predict diabetes. Diabetes prediction can be useful for diabetic patients. We used the PIMA Indians Diabetes dataset to predict diabetes mellitus effectively. The algorithms used in this research are K-nearest neighbors, decision tree classifier, deep learning (multilayer perceptron (MLP)), support vector machine (SVM), random forest, AdaBoost classifier and logistic regression. Our experiments showed promising results compared to the techniques implemented in previous research. The proposed system proved the best two models, the deep learning model after scaling and AdaBoost classifier, with the same accuracy: the training set is 82%, and the test set is 80% for the deep learning model after scaling, and the accuracy of the training set is 78.6% for the AdaBoost classifier. The results of the proposed research suggest that the AdaBoost classifier is more suitable for the detection of diabetes in healthcare.

Author Contributions: Conceptualization, S.A. and M.E.; methodology, S.A., M.E., A.M., N.B., L.A., A.F. and R.A.Z.; software, M.E.; validation, XS.A. and M.E.; formal analysis, L.A., A.F. and R.A.Z.; investigation, S.A., M.E., A.M., N.B., L.A., A.F. and R.A.Z.; resources, S.A.; data curation, S.A.; writing—original draft preparation, S.A., M.E., A.M., N.B., L.A., A.F. and R.A.Z.; writing—review and editing, S.A., M.E., A.M., N.B., L.A., A.F. and R.A.Z.; visualization, M.E.; supervision, S.A. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement: Data is available from the authors upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. What is Type 1 Diabetes? A Mayo Clinic Expert Explains. Available online: https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc (accessed on 31 August 2022).
2. Why Is Diabetes Research so Important? Available online: https://ocrc.net/why-is-diabetes-research-so-important-2/ (accessed on 31 August 2022).
3. Kanan, T.; Mughaid, A.; Al-Shalabi, R.; Al-Ayyoub, M.; Elbes, M.; Sadaqa, O. Business intelligence using deep learning techniques for social media contents. *Clust. Comput.* **2022**, 1–12.. [CrossRef]
4. Aqel, D.; Al-Zubi, S.; Mughaid, A.; Jararweh, Y. Extreme learning machine for plant diseases classification: A sustainable approach for smart agriculture. *Clust. Comput.* **2022**, *25*, 2303–2303. [CrossRef]
5. AlZu'bi, S.; Hawashin, B.; Mujahed, M.; Jararweh, Y.; Gupta, B.B. An efficient employment of internet of multimedia things in smart and future agriculture. *Multimed. Tools Appl.* **2019**, *78*, 29581–29605. [CrossRef]
6. Mughaid, A.; Al-Zu'bi, S.; Al Arjan, A.; Al-Amrat, R.; Alajmi, R.; Zitar, R.A.; Abualigah, L. An intelligent cybersecurity system for detecting fake news in social media websites. *Soft Comput.* **2022**, *26*, 5577–5591. [CrossRef]
7. Mughaid, A.; AlZu'bi, S.; Hnaif, A.; Taamneh, S.; Alnajjar, A.; Elsoud, E.A. An intelligent cyber security phishing detection system using deep learning techniques. *Clust. Comput.* **2022**, *25*, 3819–3828. [CrossRef]
8. AlZu'bi, S.; Jararweh, Y. Data fusion in autonomous vehicles research, literature tracing from imaginary idea to smart surrounding community. In Proceedings of the 2020 Fifth International Conference on Fog and Mobile Edge Computing (FMEC), Paris, France, 20–23 April 2020; pp. 306–311.
9. AlZu'bi, S.; Aqel, D.; Lafi, M. An intelligent system for blood donation process optimization-smart techniques for minimizing blood wastages. *Clust. Comput.* **2022**, *25*, 3617–3627. [CrossRef]
10. Elbes, M.; Alrawashdeh, T.; Almaita, E.; AlZu'bi, S.; Jararweh, Y. A platform for power management based on indoor localization in smart buildings using long short-term neural networks. *Trans. Emerg. Telecommun. Technol.* **2022**, *33*, e3867. [CrossRef]
11. AlZu'bi, S.; Alsmirat, M.; Al-Ayyoub, M.; Jararweh, Y. Artificial intelligence enabling water desalination sustainability optimization. In Proceedings of the 2019 7th International Renewable and Sustainable Energy Conference (IRSEC), Agadir, Morocco, 27–30 November 2019; pp. 1–4.
12. Garg, A.; Mago, V. Role of machine learning in medical research: A survey. *Comput. Sci. Rev.* **2021**, *40*, 100370. [CrossRef]
13. Karlsson, K.; Kjellmer, I. The outcome of diabetic pregnancies in relation to the mother's blood sugar level. *Am. J. Obstet. Gynecol.* **1972**, *112*, 213–220. [CrossRef]
14. Diabetes. Available online: https://www.who.int/health-topics/diabetes#tab=tab_1 (accessed on 31 August 2022).
15. Dutta, D.; Paul, D.; Ghosh, P. Analysing feature importances for diabetes prediction using machine learning. In Proceedings of the 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, Canada, 1–3 November 2018; pp. 924–928.

16. Ravaut, M.; Harish, V.; Sadeghi, H.; Leung, K.K.; Volkovs, M.; Kornas, K.; Watson, T.; Poutanen, T.; Rosella, L.C. Development and Validation of a Machine Learning Model Using Administrative Health Data to Predict Onset of Type 2 Diabetes. *JAMA Netw. Open* **2021**, *4*, e2111315. [CrossRef]

17. Khanam, J.J.; Foo, S.Y. A comparison of machine learning algorithms for diabetes prediction. *ICT Express* **2021**, *7*, 432–439. [CrossRef]

18. Jagadeeswari, V.; Subramaniyaswamy, V.; Logesh, R.; Vijayakumar, V. A study on medical Internet of Things and Big Data in personalized healthcare system. *Health Inf. Sci. Syst.* **2018**, *6*, 1–20. [CrossRef]

19. Singh, R.; VIT, V. Artificial Intelligence in Patients Care. Available online: (accessed on). [CrossRef]

20. AI vs. ML–What's the Difference Between Artificial Intelligence and Machine Learning? Available online: https://www.freecodecamp.org/news/ai-vs-ml-whats-the-difference/ (accessed on 31 August 2022).

21. Polat, K.; Güneş, S.; Arslan, A. A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine. *Expert Syst. Appl.* **2008**, *34*, 482–487. [CrossRef]

22. Duda, R.O.; Hart, P.E. *Pattern Classification*; John Wiley & Sons: Hoboken, NJ, USA, 2006.

23. Ahmed, U.; Issa, G.F.; Khan, M.A.; Aftab, S.; Khan, M.F.; Said, R.A.; Ghazal, T.M.; Ahmad, M. Prediction of diabetes empowered with fused machine learning. *IEEE Access* **2022**, *10*, 8529–8538. [CrossRef]

24. Faruque, M.F.; Asaduzzaman; Sarker, I.H. Performance analysis of machine learning techniques to predict diabetes mellitus. In Proceedings of the 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox'sBazar, Bangladesh, 7–9 February 2019; pp. 1–4.

25. Dagliati, A.; Marini, S.; Sacchi, L.; Cogni, G.; Teliti, M.; Tibollo, V.; De Cata, P.; Chiovato, L.; Bellazzi, R. Machine learning methods to predict diabetes complications. *J. Diabetes Sci. Technol.* **2018**, *12*, 295–302. [CrossRef]

26. Khamees, M.; Jarrar, Y.; Al-Qirim, T.; Mahmoud, I.S.; Hatmal, M.M.M.; Alshaer, W.; Lee, S.J. No impact of soluble epoxide hydrolase rs4149243, rs2234914 and rs751142 genetic variants on the development of type II diabetes and its hypertensive complication among Jordanian patients. *Int. J. Clin. Pract.* **2021**, *75*, e14036. [CrossRef]

27. Alam, T.M.; Iqbal, M.A.; Ali, Y.; Wahab, A.; Ijaz, S.; Baig, T.I.; Hussain, A.; Malik, M.A.; Raza, M.M.; Ibrar, S.; et al. A model for early prediction of diabetes. *Informatics Med. Unlocked* **2019**, *16*, 100204. [CrossRef]

28. Afzali, S.; Yildiz, O. An effective sample preparation method for diabetes prediction. *Int. Arab J. Inf. Technol.* **2018**, *15*, 968–973.

29. Kong, X.; Peng, R.; Dai, H.; Li, Y.; Lu, Y.; Sun, X.; Zheng, B.; Wang, Y.; Zhao, Z.; Liang, S.; et al. Disease-specific data processing: An intelligent digital platform for diabetes based on model prediction and data analysis utilizing big data technology. *Front. Public Health* **2022**, *10*. [CrossRef]

30. Bahmani, A.; Alavi, A.; Buergel, T.; Upadhyayula, S.; Wang, Q.; Ananthakrishnan, S.K.; Alavi, A.; Celis, D.; Gillespie, D.; Young, G.; et al. A scalable, secure, and interoperable platform for deep data-driven health management. *Nat. Commun.* **2021**, *12*, 5757. [CrossRef]

31. Peixoto, H.; Ramos, V.; Marques, C.; Machado, J. Predicting Diabetes Disease for healthy smart cities. *EAI Endorsed Trans. Smart Cities* **2022**, *6*, e1. [CrossRef]

32. Naseem, A.; Habib, R.; Naz, T.; Atif, M.; Arif, M.; Allaoua Chelloug, S. Novel Internet of Things Based Approach Towards Diabetes Prediction Using Deep Learning Models. *Front. Public Health* **2022**, *10*, 914106. [CrossRef]

33. Mian, Z.; Hermayer, K.L.; Jenkins, A. Continuous glucose monitoring: Review of an innovation in diabetes management. *Am. J. Med. Sci.* **2019**, *358*, 332–339. [CrossRef]