Scientific Research Publishing

Advances in Bioscience and Biotechnology, 2025, 16(9), 379-409
https://www.scirp.org/journal/abb
ISSN Online: 2156-8502
ISSN Print: 2156-8456

Accurate Classification of Diabetes via PM Generative AI

Philip de Melo, Marie St. Rose

Department of Nursing and Allied Health, Norfolk State University, Norfolk, VA, USA
Email: ferndemelo@gmail.com

How to cite this paper: de Melo, P. and St. Rose, M. (2025) Accurate Classification of Diabetes via PM Generative AI. *Advances in Bioscience and Biotechnology*, **16**, 379-409.
https://doi.org/10.4236/abb.2025.169025

Received: March 5, 2025
Accepted: September 14, 2025
Published: September 17, 2025

Abstract

The recent surge in demand for timely and accurate health information has highlighted the need for more advanced data analysis tools. To reduce the incidence of preventable medical errors, sophisticated IT-driven classification and prediction algorithms are essential. However, extracting meaningful insights from complex biomedical data remains a significant challenge in healthcare transformation. Modern biomedical and health research generates diverse data types, including electronic health records (EHRs), medical imaging, sensor data, and telemedicine inputs, which are often complex, heterogeneous, poorly annotated, and largely unstructured. Traditional statistical learning and data mining methods require extensive preprocessing before developing predictive or clustering models. This process becomes even more challenging when dealing with intricate datasets and limited domain-specific knowledge. Recent advancements in deep learning offer promising end-to-end models capable of handling such complexity. However, these models do not consistently achieve the high levels of accuracy required by healthcare professionals. In this study, we introduce a novel Deep Learning Algorithm combined with a generative AI designed to improve classification accuracy in clinical applications significantly. The algorithm is tailored for seamless integration into hospital workflows and electronic health record systems—an area that is the central focus of our ongoing research. The proposed method combines real-world clinical data with synthetic data generated by Principal Model Generative AI. This approach increased classification accuracy in our experiments from 76% to 95% - 98%.

Keywords

Diabetes, Informatics, PIMA Data Set, Deep Learning, Improved Accuracy, PM Generative AI

1. Introduction

According to the International Diabetes Federation (IDF), diabetes is a widespread chronic condition affecting over 380 million people globally, projected to exceed 600 million in the coming decades. Notably, many of these cases are preventable. Poor blood glucose regulation in diabetic individuals increases the risk of complications such as neuropathy, contributing to higher morbidity and mortality rates. Diabetes also remains a major cause of death due to its strong association with coronary artery disease and stroke.

In 2013, global spending on diabetes care was estimated at a minimum of $550 billion, with projections indicating a rise to over $630 billion by 2035.

Various information technology (IT)-driven solutions have been introduced to address these challenges to enhance blood glucose monitoring and diabetes management. Research shows that IT interventions can improve metabolic control and support the comprehensive care of individuals with chronic diabetes. A literature review in [1] emphasized the potential of technology-based solutions to foster efficient and informative communication between patients and healthcare providers.

In addition, developing end-to-end learning models from complex datasets is essential for advancing diabetes care. However, deep learning algorithms can sometimes lack accuracy. This paper proposes a novel hybrid approach that boosts accuracy by aggregating randomly selected data subsets, thereby significantly enhancing the performance of analytical algorithms.

Technology-driven interventions offer numerous benefits in healthcare, including reduced medical errors, improved research through data generation, and enhanced capacity for continuous quality improvement. Nevertheless, these innovations also come with challenges, such as high implementation and maintenance costs, usability issues for healthcare providers, and the potential for reduced face-to-face patient interaction [2].

Recent studies suggest that IT-based interventions can lead to better glycemic control and more effective diabetes management, though their impact may vary across clinical outcomes. Future research could focus on integrating multiple IT tools into unified systems to improve clinical outcomes and overall diabetes care.

Information technologies also play a crucial role in classifying and diagnosing diabetes using data analytics, machine learning, and artificial intelligence. Key technologies include [3]:

1) Machine Learning & Artificial Intelligence:

Supervised Learning: Algorithms such as decision trees, support vector machines (SVM), random forests, and neural networks classify diabetes based on patient data.

Deep Learning: Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) process large datasets to enhance diagnostic accuracy.

Natural Language Processing (NLP): Extracts insights from electronic health records (EHRs) and medical literature.

2) Big Data and Predictive Analytics: Classify patients based on risk factors (e.g., age, BMI, glucose levels).

Detect patterns for early diagnosis using historical patient records.

3) Wearable Devices and Internet of Things (IoT):

Continuous glucose monitors (CGMs) collect real-time blood sugar data.

Smart insulin pumps adjust insulin delivery based on AI-driven predictions.

Electronic Health Records (EHRs) and Cloud Computing:

Store and analyze large volumes of diabetes-related data.

Cloud-based AI models support real-time diagnosis and continuous monitoring.

Genetic and Biomarker Analysis IT tools analyze genetic data to classify Type 1, Type 2, and gestational diabetes. Omics technologies (genomics, proteomics) assist in precision medicine.

Telemedicine and Mobile Health (mHealth) Mobile apps (e.g., Glucose Buddy, MySugr) help monitor diabetes in real-time. AI-powered chatbots provide diabetes education and lifestyle recommendations.

Utilizing deep learning for diabetes classification poses multiple challenges, such as:

- Limited labeled data: Due to privacy restrictions and difficulties in medical data collection, available datasets tend to be small.
- Class imbalance: Since diabetes cases are less frequent than non-diabetic ones, models may become biased toward the majority class.
- Data inconsistencies: Missing or inaccurate patient records can affect predictive accuracy.
- Overfitting risks: Combining small datasets and complex models may lead to overfitting, reducing the effectiveness of unseen data.
- Lack of generalization: Models trained on a specific population might not perform well across different demographics or regions.
- Integration challenges: Deploying deep learning models in hospitals requires seamless compatibility with existing electronic health record (EHR) systems.
- Real-time processing: Ensuring quick and precise predictions in clinical environments remains a significant hurdle.

This paper introduces an advanced deep learning approach designed to enhance accuracy, as demonstrated by its application to diabetes classification. This approach, applied to PIMA Indian data, largely increased the accuracy of classification.

2. Data Description

Diabetes diagnosis is contingent upon several essential parameters that assess blood glucose levels and related health indicators. The key parameters include:

1) Fasting Blood Glucose (FBG) measures blood sugar levels after a minimum fasting duration of 8 hours. The classifications are as follows: Diabetes: >126 mg/dL (7.0 mmol/L), Prediabetes: 100 - 125 mg/dL (5.6 - 6.9 mmol/L), and Nor-

mal: <100 mg/dL (5.6 mmol/L).

2) The Oral Glucose Tolerance Test (OGTT) evaluates blood sugar levels two hours post-consumption of a 75 g glucose solution. The results are categorized as: Diabetes: >200 mg/dL (11.1 mmol/L), Prediabetes: 140 - 199 mg/dL (7.8 - 11.0 mmol/L), and Normal: <140 mg/dL (7.8 mmol/L).

3) Hemoglobin A1c (HbA1c) reflects the average blood sugar levels over the previous 2 to 3 months, with classifications as follows: Diabetes: >6.5%, Prediabetes: 5.7% - 6.4%, and Normal: <5.7%.

4) The Random Blood Glucose Test assesses blood sugar at any time, regardless of food intake. A diabetes diagnosis is established if the level exceeds 200 mg/dL (11.1 mmol/L) alongside symptoms such as excessive thirst, frequent urination, or unexplained weight loss.

5) Insulin and C-Peptide Levels are instrumental in differentiating between Type 1 and Type 2 diabetes, with low levels of both indicating Type 1 diabetes.

6) Autoimmune Markers for Type 1 Diabetes, including the presence of autoantibodies such as GAD65, IA-2, and ZnT8, can confirm a Type 1 diabetes diagnosis.

7) Ketone Levels, particularly in instances of Diabetic Ketoacidosis (DKA), where elevated urine or blood indicate severe insulin deficiency, are commonly associated with Type 1 diabetes.

8) Body Mass Index (BMI) and Obesity: Increased BMI and obesity are significant risk factors for Type 2 diabetes.

9) Blood Pressure and Cholesterol Levels: Hypertension and dyslipidemia (characterized by high LDL and low HDL) are often observed in individuals with diabetes. This paper will utilize the diabetes data available on Kaggle.

The data set is characterized by the following features depicted by Table 1:

The features of the PIMA data depicted in Table 1 are as follows:

- Pregnancies: Number of times pregnant;
- Glucose: Plasma glucose concentration 2 hours in an oral glucose tolerance test;
- Blood Pressure: Diastolic blood pressure (mm Hg);
- Skin Thickness: Triceps skin fold thickness (mm);
- Insulin: 2-Hour serum insulin (mu U/ml);
- BMI: Body mass index (weight in kg/(height in m)^2);
- Diabetes Pedigree Function: Diabetes pedigree function;
- Age: (years).

Figure 1 shows the number of patients with diabetes and without.

Diabetes triggers a variety of physiological changes in the body, notably altering skin characteristics. One key study [4] aimed to analyze skin thickness in female diabetic patients and evaluate its potential as an indicator of diabetes progression. A one-way ANOVA test was employed to assess the impact of various factors on skin thickness, using a significance threshold of $\alpha \leq 0.05$. Results demonstrated a decrease in skin thickness as diabetes advanced. While insulin levels significantly influenced skin thickness, glucose levels did not show a similar effect.

Table 1. First 20 rows of the data set. The data set consists of 768 rows (patients) and 9 columns (features outcome).

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
5	5	116	74	0	0	25.6	0.201	30	0
6	3	78	50	32	88	31.0	0.248	26	1
7	10	115	0	0	0	35.3	0.134	29	0
8	2	197	70	45	543	30.5	0.158	53	1
9	8	125	96	0	0	0.0	0.232	54	1
10	4	110	92	0	0	37.6	0.191	30	0
11	10	168	74	0	0	38.0	0.537	34	1
12	10	139	80	0	0	27.1	1.441	57	0
13	1	189	60	23	846	30.1	0.398	59	1
14	5	166	72	19	175	25.8	0.587	51	1
15	7	100	0	0	0	30.0	0.484	32	1
16	0	118	84	47	230	45.8	0.551	31	1
17	7	107	74	0	0	29.6	0.254	31	1
18	1	103	30	38	83	43.3	0.183	33	0
19	1	115	70	30	96	34.6	0.529	32	1

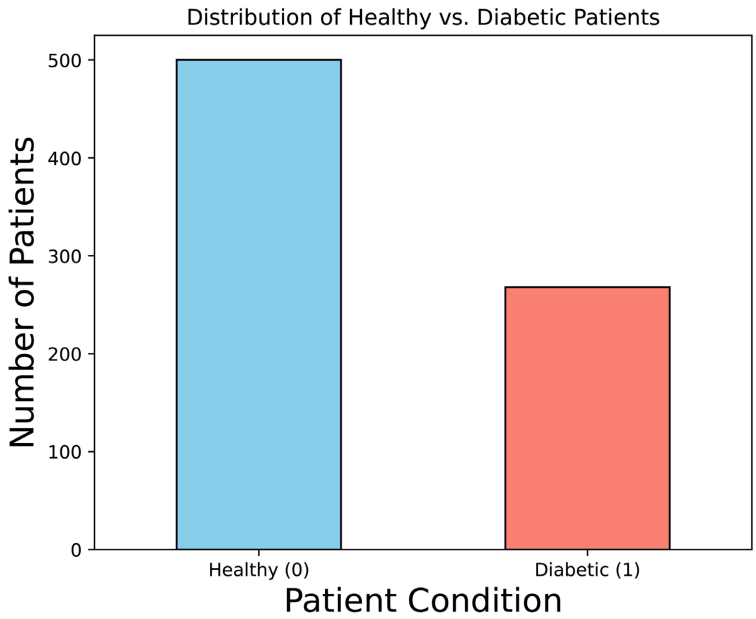

Figure 1. The number of healthy patients (blue) and diabetes patients (orange).

These findings suggest that skin thickness may serve as a novel marker for monitoring diabetes progression in women. However, further research is needed to validate this hypothesis. This paper will explore these results in greater detail, with the next step involving the calculation of the correlation matrix (Figure 2).

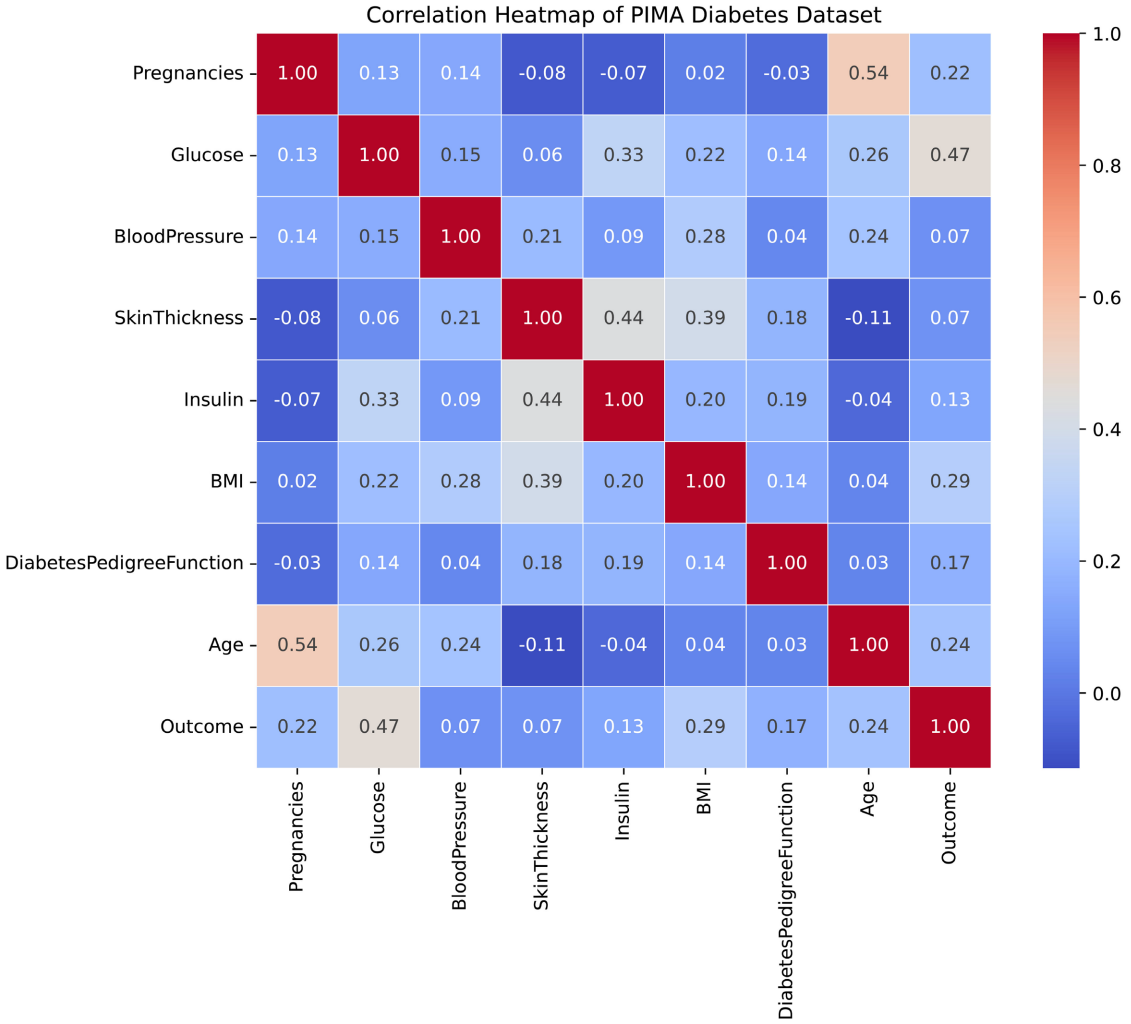

Figure 2. Heatmap shows the correlation between features and outcomes.

Machine learning (ML) plays an increasingly critical role in diabetes research, treatment, and management. By analyzing patient data—such as glucose levels, body mass index (BMI), and family history—ML models can assess an individual's risk of developing diabetes. Techniques like decision trees, neural networks, and support vector machines enable early detection.

ML-based models can also predict blood glucose trends by examining historical data, dietary habits, physical activity, and medication adherence. Continuous glucose monitoring (CGM) devices, enhanced by artificial intelligence, provide real-time alerts to help prevent episodes of hypoglycemia or hyperglycemia [4]. Personalized treatment strategies are greatly improved through ML, which considers patient-specific factors, including lifestyle and medication response.

Moreover, reinforcement learning algorithms assist in optimizing insulin dosages, while deep learning methods analyze retinal images for early detection of diabetic retinopathy. AI-powered tools, such as those developed by Google's DeepMind, support healthcare professionals in diagnosing diabetes-related ocular complications.

Machine learning is crucial in predicting complications such as cardiovascular disease, neuropathy, and kidney disorders, enabling timely interventions. AI-powered applications offer personalized recommendations tailored to individual dietary habits, physical activity, and sleep patterns. Wearable devices like Fitbit and Apple Watch also provide real-time data that supports more effective diabetes management. In addition, automated insulin delivery systems use machine learning to dynamically adjust insulin doses based on glucose level fluctuations, reducing patient burden and improving blood sugar control.

3. Feature Engineering

3.1. Visualization

Feature engineering is a crucial step in data preprocessing for machine learning. It involves identifying and transforming the variables most significantly impacting model outcomes and decision-making. This process converts raw data into meaningful features suitable for machine learning algorithms. Feature engineering includes selecting, extracting, and transforming the most relevant features from a dataset to enhance model performance and accuracy.

Figure 2 presents a cross-correlation heatmap highlighting key patterns and relationships within the data. This visualization aids in improving the model's learning capability by revealing essential interdependence. After data cleaning and labeling, machine learning teams typically conduct thorough exploratory data analysis (EDA) to assess the data quality and fitness for modeling. Visual tools such as histograms, scatter plots, box plots, line graphs, and bar charts are vital in verifying data integrity. These visualizations help scientists detect data patterns, spot anomalies, test hypotheses, and validate assumptions.

The performance of machine learning models is greatly affected by the quality of the features employed in their training. Feature engineering involves a range of techniques that facilitate the generation of new features by combining or transforming existing ones. These approaches are essential.

Exploratory data analysis does not require formal modeling; instead, data science teams can utilize visualizations to gain meaningful insights from the data. The data collection stage involves compiling all relevant information necessary for machine learning, which can be a labor-intensive task due to the often-fragmented nature of data across various sources, such as personal computers, data warehouses, cloud storage, applications, and devices.

Connecting to these diverse data sources can present considerable challenges. Furthermore, the volume of data is increasing at an extraordinary pace, leading to large datasets that demand comprehensive analysis. Additionally, the formats and

types of data can differ significantly depending on their source, complicating the integration of various data types, such as video and tabular data. These figures suggest that four major features determine the outcomes. We will consider both cases when all features are considered (Figure 3) and only four features (Figure 4).

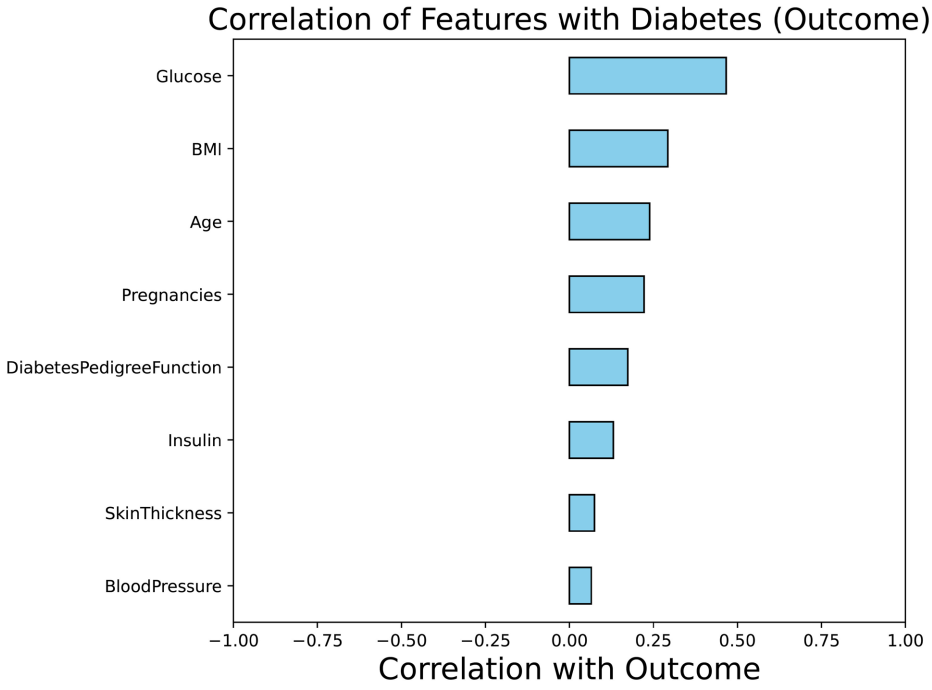

Figure 3. Feature engineering analysis shows the contribution of each feature in the outcome.

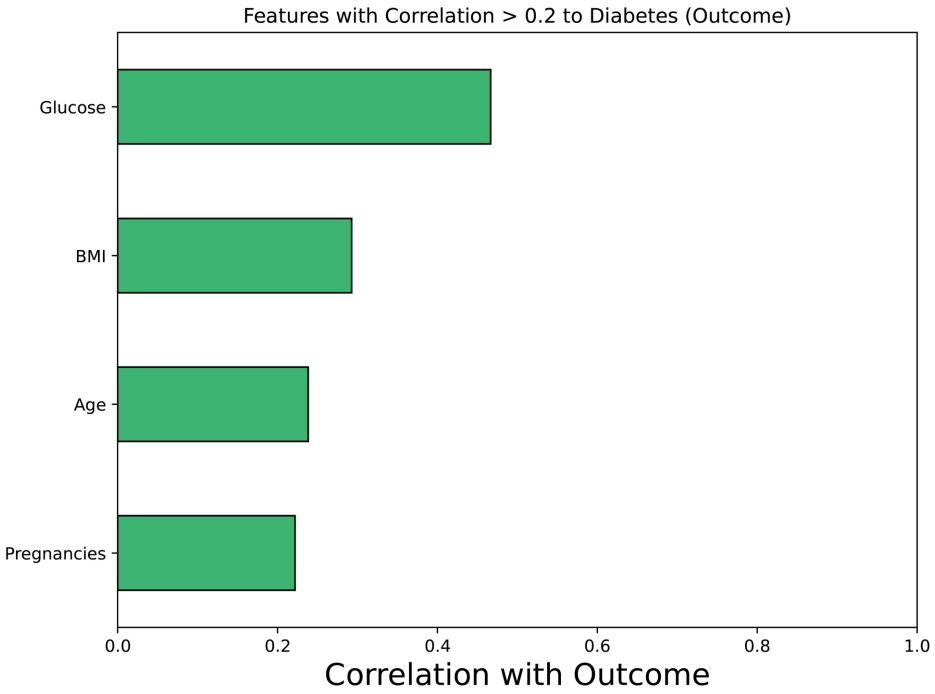

Figure 4. Feature engineering analysis shows the most contribution comes from 4 features.

3.2. Detection and Removal of Outliers

Outlier detection and removal is a technique used to identify and eliminate anomalies from a dataset. The main goal of this process is to improve the accuracy of data representation, which can significantly impact model performance. The extent of this impact varies—some models are highly sensitive to outliers, while others remain largely unaffected.

For instance, linear regression is particularly susceptible to outliers, making it essential to address them before training a model. Various methods can be employed to manage outliers, including:

Removal: This approach involves deleting records containing outliers from the dataset. However, if outliers are present in multiple variables, this can lead to substantial data loss.

Replacing Values: Here, outliers are treated as missing values and replaced with imputed values deemed appropriate.

Capping: This method substitutes extreme values with a predetermined threshold, or a value based on the variable's distribution.

Discretization: This process converts continuous variables into discrete categories by segmenting the range into intervals or bins.

The following boxplots illustrate the selected features. In descriptive statistics, a boxplot (also called a box-and-whisker plot) is a valuable tool for exploration data analysis. It provides a visual representation of data distribution and skewness by displaying quartiles, percentiles, and averages. A boxplot encapsulates the five-number summary of a dataset, consisting of:

Minimum Score: The lowest value in the dataset, excluding outliers, represented by the left whisker's endpoint.

Lower Quartile (Q1): The first quartile, marking the value below which 25% of the data falls.

Median (Q2): The midpoint of the dataset, splitting it into two equal halves. It signifies that half of the values are above and half are below this point.

Upper Quartile (Q3): The third quartile, representing the value below which 75% of the data lies.

Maximum Score: The highest value in the dataset, excluding outliers, shown at the right whisker's endpoint.

Whiskers: These extend from the box, covering the lower and upper 25% of the data, excluding extreme outliers.

Interquartile Range (IQR): The range between the first and third quartiles, representing the middle 50% of the data.

Boxplots efficiently summarize data using a simple visual format, making it easy to identify quartiles, medians, and outliers briefly. Outliers can significantly affect data analysis and machine learning models, leading to biased predictions, misleading conclusions, and distorted statistical measures.

To mitigate these effects, statistical methods such as the Interquartile Range (IQR) are used to quantify dispersion. The IQR measures the spread of the middle 50% of the data and is calculated as the difference between the 75th percentile

(Q3) and the 25th percentile (Q1). The following figures show the boxplots of major features and outliers. Figure 5 shows Q1, Q2, Q3 quartiles and IQR (shadowed). Outliers are beyond the upper bound.

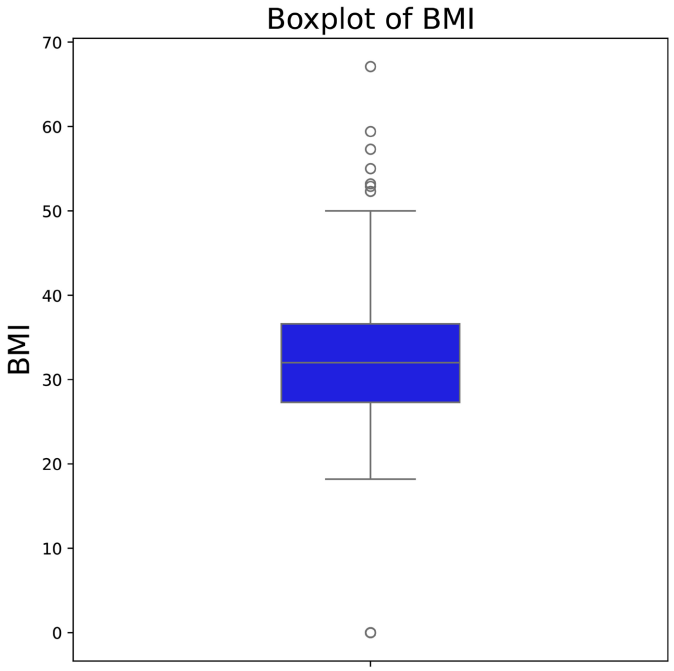

Figure 5. Outliers detected in the BMI records.

Figure 6 shows the boxplot for the Glucose records with outlier at 0 and Figure 7 demonstrates the Boxplot for the Pregnancy records.

Figure 6. Outliers detected in the Glucose records.

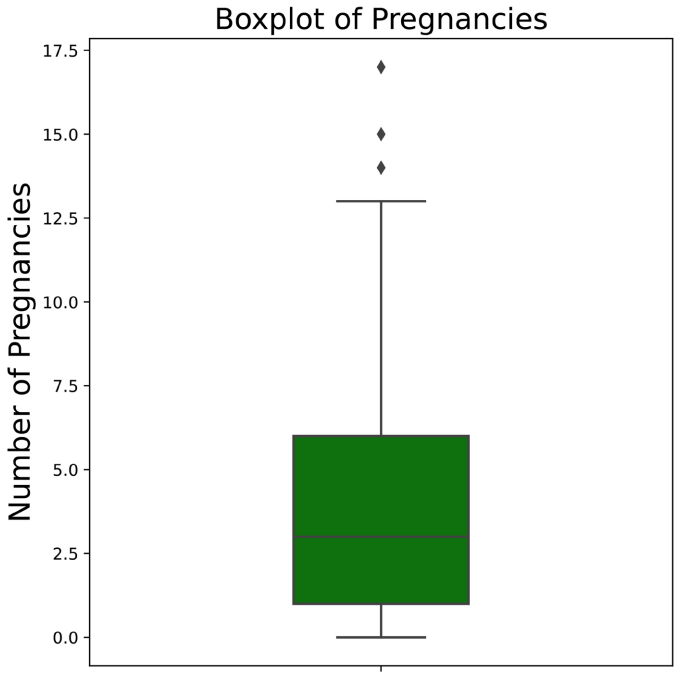

Figure 7. Outliers in the Pregnancies records.

Figure 8 presents the boxplot of the Age data. Although it indicates outliers beyond the upper bound, we will not exclude them in this case, as the cohort includes patients aged 60 - 70.

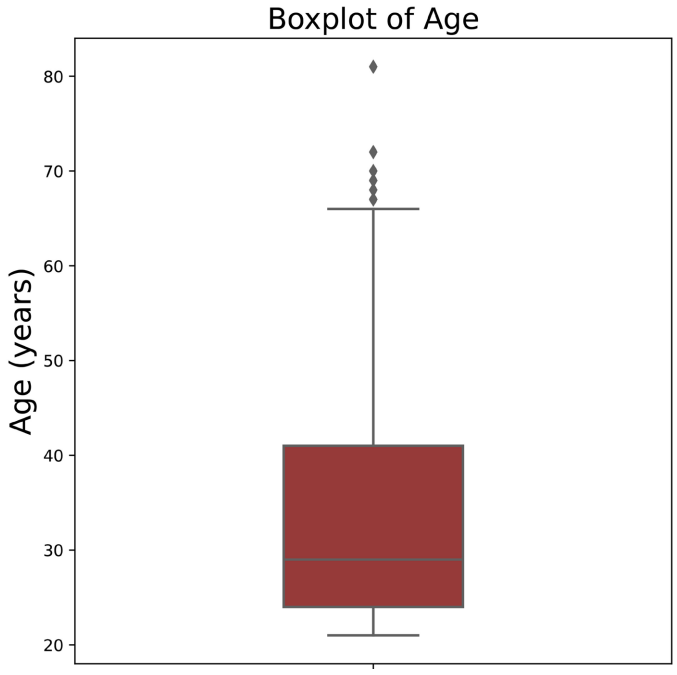

Figure 8. Outliers detected in the Age records, but they were included in the analysis.

The Interquartile Range (IQR) is a statistical measure of dispersion that represents the spread of the middle 50% of a dataset. It is calculated by subtracting the

25th percentile (Q1) from the 75th percentile (Q3). To detect outliers using the IQR method, two boundaries are established:

- Lower Bound: Q1 – 1.5 × IQR
- Upper Bound: Q3 + 1.5 × IQR

These boundaries help identify potential outliers in a dataset. Any data point below the lower bound (Q1 – 1.5 × IQR) is considered an outlier, as it significantly deviates from the rest of the data and may warrant further review or removal. Similarly, any data point above the upper bound (Q3 + 1.5 × IQR) is considered an outlier, as it is substantially higher than most data points and may require special attention.

One of the main advantages of the IQR method is its resilience to skewed data distributions. Since it identifies outliers based on percentiles, it is less sensitive to extreme values. Additionally, the IQR method is straightforward to implement and interpret, providing a clear range within which most data points should fall. This makes it an effective tool for data analysis and quality control.

3.3. Z-Score to Drop Outliers

Outliers can arise from various factors and often result from either genuine variability in the data or errors in data collection, measurement, or recording. Common causes of outliers include:

- Measurement errors: Mistakes in the data collection or measurement processes can lead to outliers.
- Sampling errors: Outliers may occur if there are issues with the sampling process.
- Natural variability: Inherent variability in certain phenomena can lead to extreme values, as some systems naturally exhibit outliers.
- Data entry errors: Human mistakes during data entry can introduce outliers.
- Experimental errors: In experimental settings, anomalies can arise due to uncontrolled factors, equipment malfunctions, or unforeseen events.
- Sampling from multiple populations: When data from different populations with distinct characteristics are combined, outliers can emerge.
- Intentional outliers: Sometimes, outliers are deliberately introduced to test the robustness of statistical methods.

The Z-score is used to standardize variables, giving insight into how far a particular observation is from the meaning. Specifically, the Z-score indicates how many standard deviations a data point is away from the mean. The process of transforming a feature into Z-scores is known as standardization.

If the Z-score of a data point exceeds 3, it suggests the data point is significantly different from the others and could be an outlier. Z-scores can be both positive and negative; the farther the score is from 0, the higher the likelihood that the data point is an outlier. Generally, a Z-score greater than 3 is considered extreme.

For the Z-score method to be effective in identifying outliers, the data should follow a normal distribution. Removing outliers using the Z-score method involves the following steps:

$$Z = \frac{X - \mu}{\sigma}$$

where: X = data point, μ = mean of the dataset, σ = standard deviation of the dataset The procedure includes the following steps:

Choose a threshold (commonly 2 or 3)

A common choice is $Z > 3$ or $Z < -3$, meaning the data point is 3 standard deviations away from the mean. If the dataset is small or you want to be less strict, you might use $Z > 2.5$ or even $Z > 2$. Remove outliers: Any data point with a Z-score beyond the chosen threshold is considered an outlier and can be removed.

The Z-score indicates the number of standard deviations a data point is from the mean. It is calculated as: Z-score removal of outliers requires normal distribution.

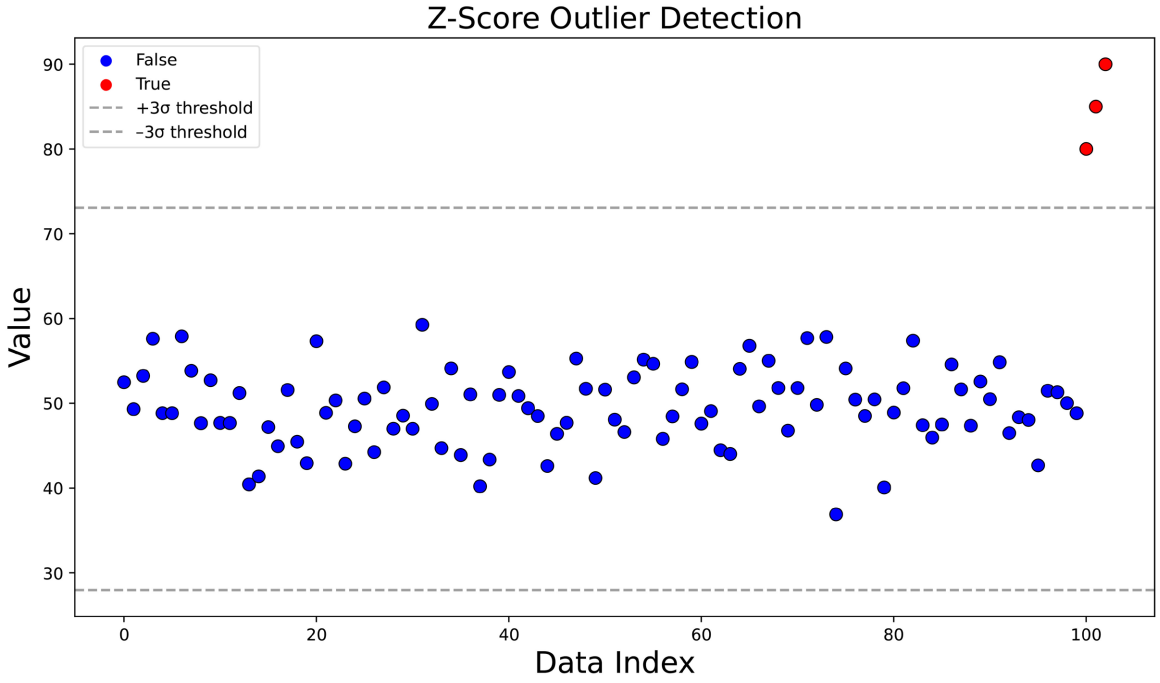

Figure 9. Z-scores are useful for detecting outliers—data points that deviate significantly from the rest of the dataset. Generally, data points with z-scores exceeding 3 or falling below -3 are considered potential outliers and may require further analysis.

Figure 9 shows the application of z-scoring to data sets which after the z-transform becomes normally distributed. While Z-scores can be computed for any distribution, they are particularly useful when the data is normally distributed, because:

- We can then use standard normal distribution tables.
- Probabilities and percentiles become meaningful.

Z-scores don't require a normal distribution but are most useful when the data is (approximately) normal. Figure 10 shows the detection of outliers using z-scores. Figure 11, Figure 12 show the feature distributions: Figure 11 represents skewed normal distributions and can be used to eliminate outliers, while Figure

12 does not represent normal distribution and cannot use a-scores to eliminate outliers.

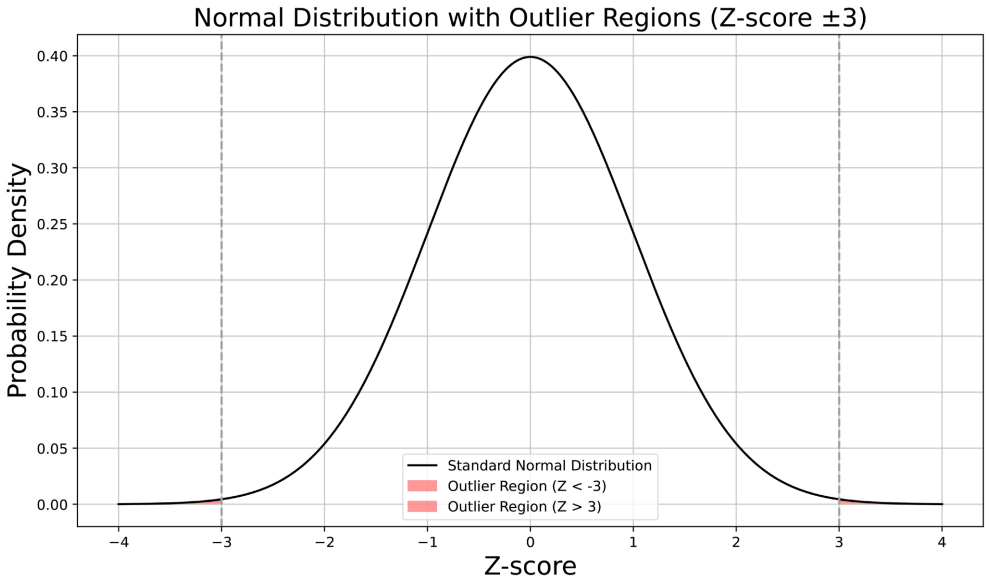

Figure 10. Moderately unusual and evident outliers.

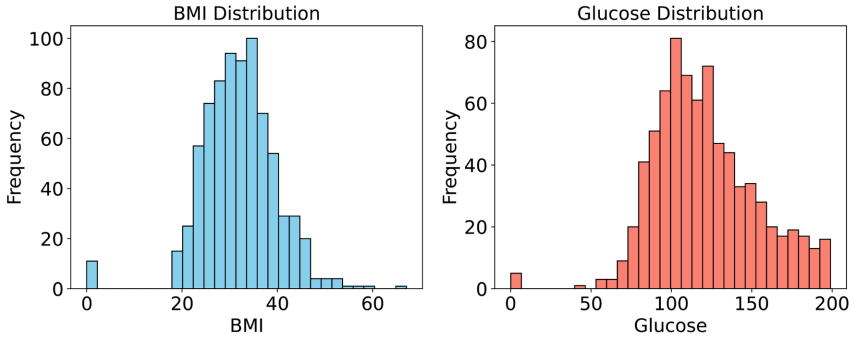

Figure 11. The data distribution for BMI and Glucose represent slightly skewed normal distributions and can be used in z-scoring for outliers' removal.

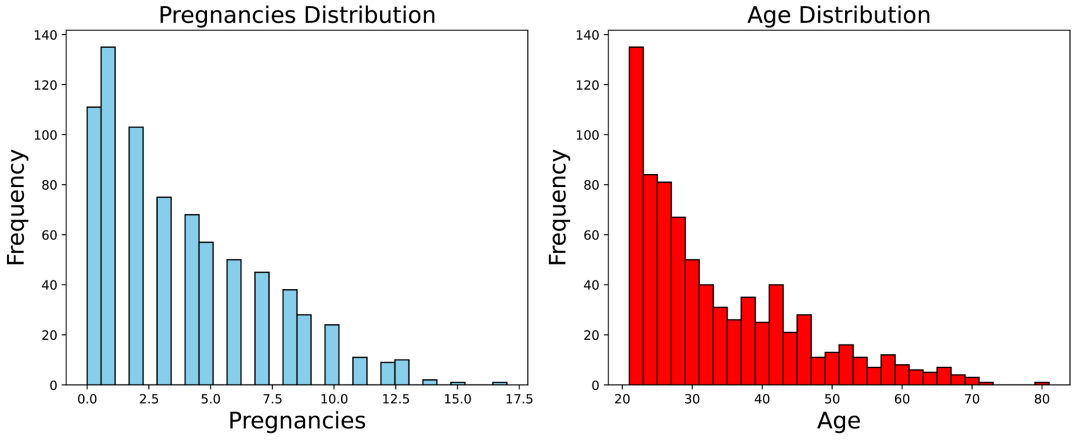

Figure 12. The data distribution for Pregnancies and Age cannot be used for z-scoring in outliers removal.

Figure 13 shows the BMI and Glucose features after outliers have been removed using z-scores. It is important to emphasize that outliers can disproportionately impact models, particularly those sensitive to extreme values (e.g., linear regression, k-means clustering). In classification tasks, outliers can lead to misclassification by distorting decision boundaries. A model trained on data with outliers may struggle to generalize effectively to unseen data. Tree-based methods, such as Random Forest and XGBoost, exhibit reduced sensitivity to outliers. However, in the context of machine learning applied to public health or healthcare, it is crucial to remove outliers to avoid degrading model accuracy.

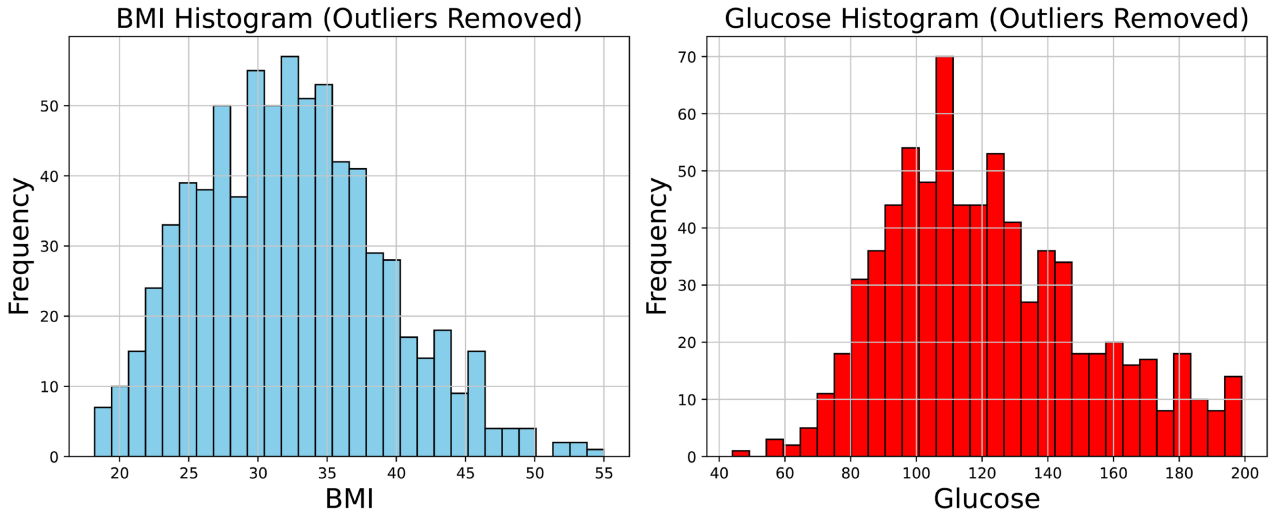

Figure 13. The data distribution for BMI and Glucose after elimination of outliers.

4. Diabetes Features against Outcomes

4.1. BMI vs. Diabetes

Having obesity significantly increases the likelihood of developing diabetes, a condition characterized by excessive glucose (sugar) in the bloodstream. It also accelerates the progression of diabetes. A high BMI (in the overweight or obese range) raises the risk of insulin resistance, which can lead to type 2 diabetes. Excess fat, particularly around the abdomen, disrupts insulin function, making it harder for cells to absorb glucose from the blood [5].

Research indicates that individuals with a BMI of 30 or higher face a much greater risk of developing type 2 diabetes compared to those with a normal BMI (18.5 - 24.9). Those in the overweight range (BMI 25 - 29.9) are more susceptible to prediabetes, a state where blood sugar levels are elevated but not yet high enough for a diabetes diagnosis. Losing just 5% - 10% of body weight can significantly lower this risk.

Here's how it works: The pancreas regulates blood glucose levels by producing insulin, a hormone responsible for moving glucose out of the bloodstream. Under normal conditions, insulin helps transport glucose to muscles for immediate energy use or stores it in the liver for future needs.

However, in cases of diabesity (a combination of obesity and diabetes), cells become resistant to insulin, preventing glucose from entering. Additionally, the liver's glucose storage becomes saturated with fat. With nowhere else to go, glucose remains in the bloodstream, prompting the pancreas to produce more insulin to counteract this resistance.

Over time, the pancreas becomes overworked and starts producing less insulin, leading to diabetes, which can quickly worsen if insulin resistance persists.

Individuals with obesity are approximately six times more likely to develop type 2 diabetes than those at a healthy weight. However, not everyone with obesity will necessarily develop diabetes. Other contributing factors include:

- Family history;
- Diet;
- Physical activity;
- Stress levels;
- Gut health.

Some people with obesity may produce enough insulin without overburdening the pancreas, while others may have a limited insulin production capacity, making them more susceptible to diabesity.

Figure 14 demonstrates the cross plot between BMI values of patients with diabetes (pink color) and healthy ones (green color).

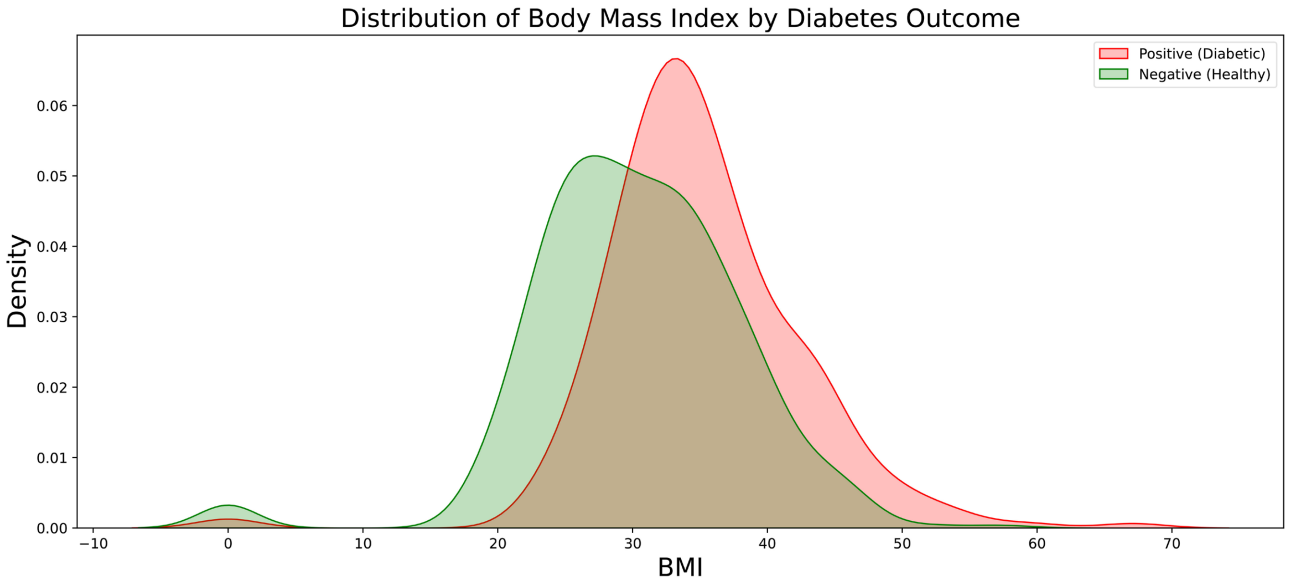

Figure 14. BMI near 30 threatens the development of diabetes.

4.2. Glucose vs. Diabetes

The term "glucose" comes from the Greek word meaning "sweet." It is a type of sugar derived from the foods you consume, serving as a primary energy source for your body. When glucose circulates in your bloodstream to reach your cells, it is known as blood sugar. Insulin, a hormone, helps transfer glucose into cells for energy and storage. In the case of diabetes, blood glucose levels become higher

than normal. This can happen either because your body doesn't produce enough insulin to regulate glucose or because your cells don't respond effectively to insulin.

Prolonged high blood sugar can damage the heart and blood vessels, increasing the risk of heart disease, high blood pressure, and stroke. It can also lead to kidney disease or failure (diabetic nephropathy), damage to the retina (diabetic retinopathy), causing vision loss or blindness, and nerve damage (diabetic neuropathy), which can result in pain, tingling, or numbness, especially in the feet. Additionally, high blood sugar raises the risk of cognitive decline and Alzheimer's disease, contributes to fatty liver disease, and worsens insulin resistance.

The glucose in your bloodstream primarily comes from carbohydrate-rich foods, such as bread, potatoes, and fruit. As shown in Figure 15, when glucose levels exceed 125 - 130 mg/dL, the risk of developing diabetes significantly increases.

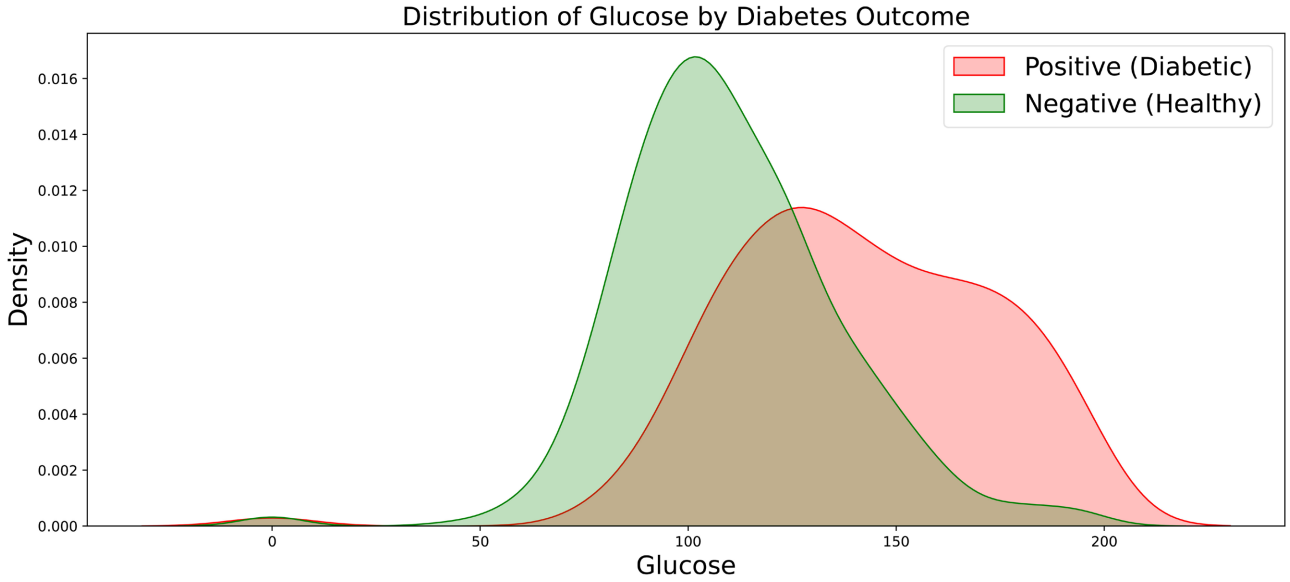

Figure 15. Glucose vs. diabetes shows numbers with elevated risks of diabetes.

As we eat, food travels down the esophagus into the stomach, where acids and enzymes break it down into smaller components, releasing glucose in the process. This glucose then moves to the intestines, where it is absorbed into the bloodstream. Insulin plays a crucial role in helping glucose enter cells. After eating, blood sugar levels naturally rise and gradually decrease a few hours later as insulin facilitates glucose absorption.

Hyperglycemia, or high blood glucose, occurs when blood sugar levels exceed 200 mg/dL two hours after eating or 125 mg/dL while fasting. For patients with diabetes, regular blood sugar testing is essential. Exercise, a balanced diet, and medication can help maintain healthy blood glucose levels and prevent complications. Figure 16 presents a violin plot showing that higher glucose levels are associated with diabetes.

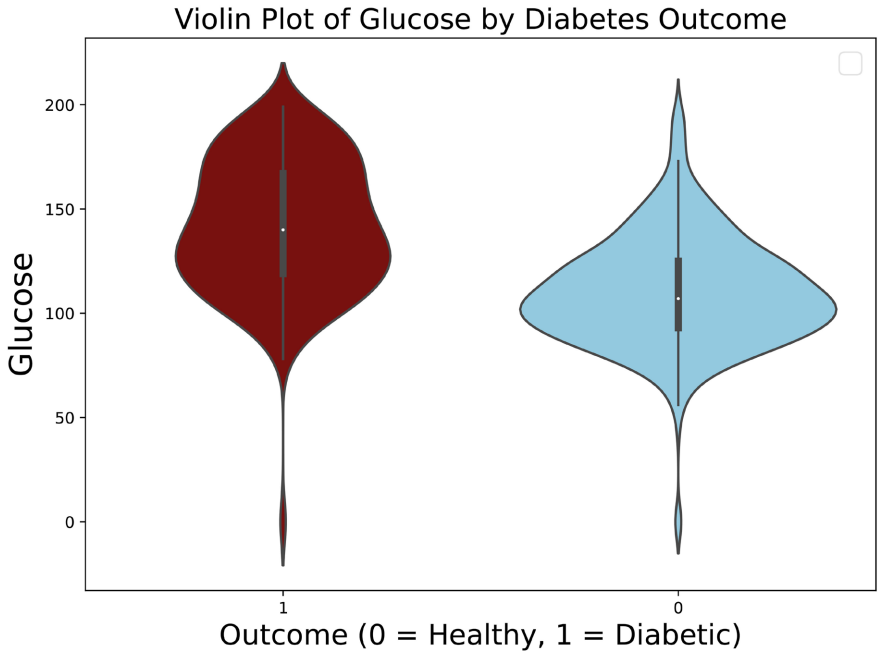

Figure 16. Violin plot shows the glucose values vs. the outcomes.

Chronic high blood glucose can lead to nerve damage (diabetic neuropathy), causing pain, tingling, or numbness, especially in the feet. It also increases the risk of cognitive decline and Alzheimer's disease, contributes to fatty liver disease, and worsens insulin resistance. The glucose in the bloodstream primarily comes from carbohydrate-rich foods, such as bread, potatoes, and fruit. **Figure 15** illustrates that when glucose levels exceed 125 - 130 mg/dL, the risk of diabetes increases significantly.

4.3. Age vs. Diabetes

Diabetes is mostly diagnosed in individuals over the age of 45. The risk increases with age due to factors such as reduced insulin sensitivity, weight gain, and decreased physical activity. However, an increasing number of younger people are being diagnosed with diabetes, largely due to lifestyle changes. More than 90% of individuals with diabetes have Type 2.

For middle-aged individuals, it's essential to focus on weight management, blood sugar control, and regular monitoring for complications.

For older adults, the risk of complications like heart disease, nerve damage, and kidney problems rises, necessitating careful management and monitoring.

Diabetes can go undiagnosed for years, as symptoms like excessive thirst, blurred vision, and tingling in the hands and feet may develop gradually and go unnoticed.

Middle age marks a significant rise in diabetes diagnoses. Approximately 15% of middle-aged Americans are diagnosed with Type 2 diabetes, which is nearly five times the rate of those under 45. Incidence increases even more as individuals age. Nearly 25% of older adults in the U.S. have been diagnosed with Type 2 diabetes, with undiagnosed cases possibly accounting for an additional 5%. This

means that over one in four senior Americans live with Type 2 diabetes.

Age is a significant risk factor for Type 2 diabetes. The older a patient is, the more likely they are to develop the condition. This is also true for preteens and teenagers, as the rates of diabetes in this age group have increased sharply in recent years. Type 2 diabetes is a disease caused by a combination of genetic factors and lifestyle choices. Figure 17 displays a cross-plot of age and outcomes, with diabetes shown in red and non-diabetes in green.

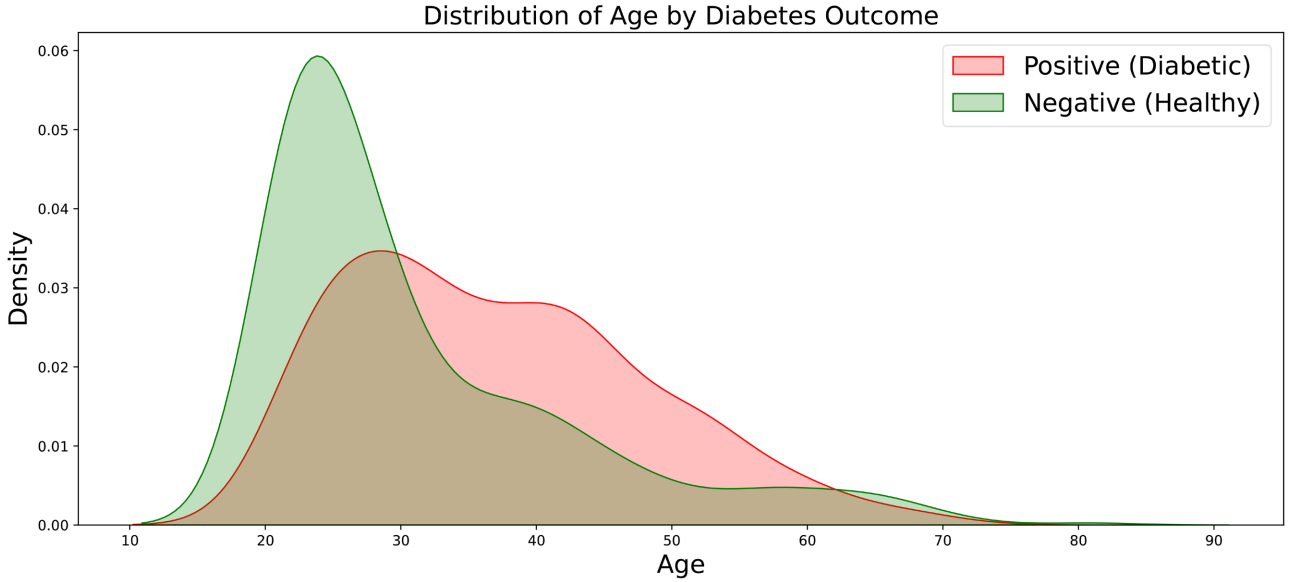

Figure 17. Age over 30 causes a higher risk of diabetes.

Being overweight, having high blood pressure, and leading a sedentary lifestyle all increase the risk of Type 2 diabetes. Managing diabetes at different ages involves different approaches:

Younger individuals: Focus on diet, exercise, and insulin management (for Type 1).

Middle-aged individuals: Emphasize weight management, blood sugar control, and monitoring for complications.

Older adults: The risk of complications such as heart disease, nerve damage, and kidney issues increase, requiring careful management.

Diabetes can go undiagnosed for years. Symptoms such as excessive thirst, blurry vision, and tingling in the hands and feet may develop gradually and be overlooked.

Middle age is when the number of diabetes diagnoses starts to rise significantly. Approximately 15% of middle-aged Americans are diagnosed with Type 2 diabetes, nearly five times the rate among those under 45. The rate increases even further as individuals enter their senior years. Nearly 25% of older adults in the U.S. have been diagnosed with Type 2, with undiagnosed cases possibly accounting for an additional 5%. This means that more than one in four senior Americans lives with Type 2 diabetes [6].

The disease also is affecting ever more teens and even children. Researchers believe childhood obesity and lack of exercise are among the reasons behind that trend.

4.4. Pregnancies vs. Diabetes

Pregnancy can have a significant impact on diabetes, and diabetes can also affect pregnancy. There are two main scenarios to consider:

Gestational Diabetes Mellitus (GDM)—This type of diabetes develops during pregnancy, usually in the second or third trimester, and often resolves after childbirth. It occurs when the body cannot produce enough insulin to meet the increased needs of pregnancy.

Pre-existing Diabetes (Type 1 or Type 2)—Women who have diabetes before pregnancy need to manage their blood sugar levels carefully to prevent complications for both them and the baby.

Gestational diabetes is diabetes that a woman can develop during pregnancy. When you have diabetes, your body cannot use the sugars and starches (carbohydrates) it takes in as food to make energy. As a result, your body collects extra sugar in your blood.

We don't know all the causes of gestational diabetes. Some with gestational diabetes are overweight before getting pregnant or have diabetes in the family. From 1 in 50 to 1 in 20 pregnant women has gestational diabetes. It is more common in Native American, Alaskan Native, Hispanic, Asian, and Black women, but it is found in white women, too. Figure 18 shows the Pregnancies data distribution for patients with diabetes (red) and without (green).

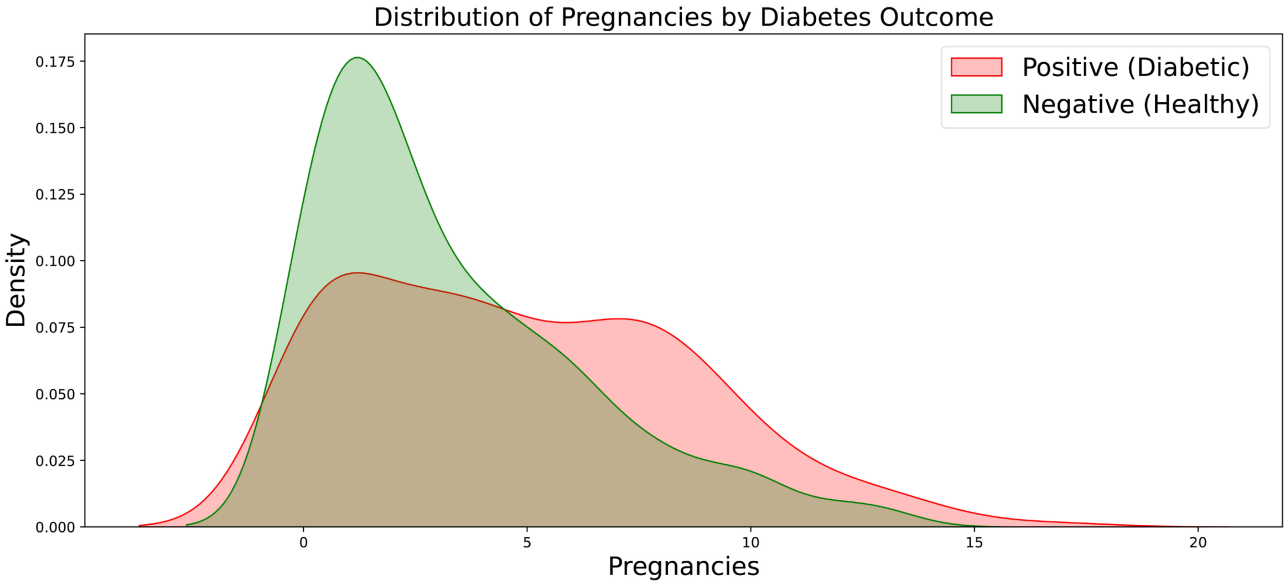

Figure 18. Higher number of pregnancies indicate risks of diabetes. the KDE plot for Pregnancies showing non-zero values for values < 0 is a common side effect of kernel density estimation as it smooths the data using Gaussian kernels).

5. Deep Learning Analysis of Diabetes Data

Classification is the process of finding or discovering a model or function that

helps separate the data into multiple categorical classes, *i.e.*, discrete values. In classification, data is categorized under different labels according to some parameters given in the input, and then the labels are predicted for the data.

In a classification task, we predict discrete class labels using independent features. In the classification task, we should find a decision boundary that can separate the different classes in the target variable. For more information on advanced regression/classification algorithms, we send the reader to [7]. The derived mapping function could be demonstrated as "IF-THEN" rules. The classification process deals with problems where the data can be divided into binary or multiple discrete labels. For example, suppose we want to predict the possibility of winning a match by Team A based on some parameters recorded earlier. Then, there would be two labels: Yes and No.

Regression is finding a model or function for distinguishing the data into continuous real values instead of using classes or discrete values. It can also identify the distribution movement that depends on historical data. Because a regression predictive model predicts a quantity, the skill of the model must be reported as an error in those predictions.

Deep learning, or hierarchical learning, is a subset of machine learning in AI that mimics the brain's computing abilities and decision-making patterns. In contrast to task-based algorithms, deep learning systems learn from data representations. It can learn from unstructured or unlabeled data. A neural network with multiple hidden layers and nodes in each hidden layer is known as a deep learning system or a deep neural network. *i.e.* depth of the neural network. Essentially, every neural network with more than three layers, that is, including the Input Layer and Output Layer can be considered a Deep Learning Model.

Deep learning is the field of artificial intelligence (AI) that teaches computers to process data in a way inspired by the human brain. Deep learning models can recognize data patterns like complex pictures, text, and sounds to produce accurate insights and predictions. Neural networks power deep learning. It consists of interconnected nodes or neurons in a layered structure. The nodes process data in a coordinated and adaptive system.

Neural networks form the foundation of deep learning systems. They exchange feedback on generated outputs, learn from mistakes, and improve continuously. While deep learning models are powerful, simpler neural networks are often favored for basic machine learning (ML) tasks due to their lower development costs and modest computational requirements. These simpler models are particularly feasible for smaller projects, enabling organizations to develop internal applications for tasks such as data visualization and pattern recognition cost-effectively [8].

In contrast, deep learning systems have a broad range of practical applications. Their capacity to learn from data, extract complex patterns, and automatically develop features enables them to deliver state-of-the-art performance. Common use cases include natural language processing (NLP), autonomous driving, and speech

recognition [9].

However, training and developing deep learning systems require substantial computational resources and funding. As a result, many organizations opt to use pre-trained models offered as fully managed services, which can be customized for specific applications.

To better understand neural networks, consider them as a series of algorithms inspired by the structure and function of the human brain. These networks are designed to recognize patterns in data, interpreting sensory inputs by labeling and clustering raw information. Their ability to learn and adapt over time makes them essential for image and speech recognition tasks.

This learning path provides a comprehensive introduction to neural networks, preparing you to explore more advanced applications.

Figure 19 represents a simple neural network and a deep learning neural network architecture with an activation function.

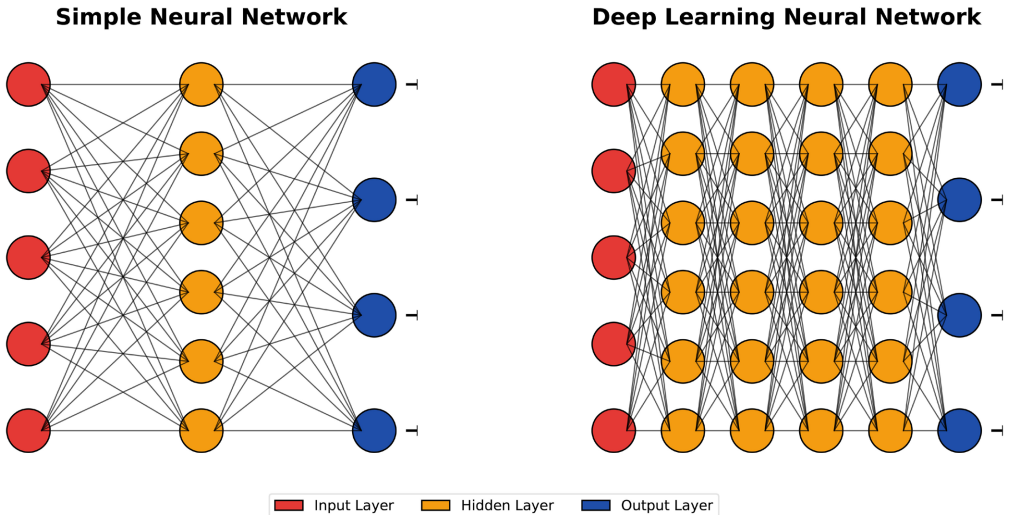

Figure 19. A simple neural network and a deep learning neural network characterized by deep forward and backpropagation.

Neural networks learn and identify patterns directly from data without relying on predefined rules. They consist of several key components:

- Neurons: The basic processing units that receive input. Each neuron operates based on a threshold and an activation function.
- Connections: Pathways between neurons that transmit information, modulated by weights and biases.
- Weights and Biases: Parameters defining each connection's strength and influence.
- Propagation Functions: Mechanisms that process and pass data through layers of neurons.

Learning Rule: The method used to adjust weights and biases over time to improve performance.

An L-layer neural network can be defined as a nested function:

$$a^{(L)} = f^{(L)}\left(W^{(L)} f\left(W^{(L-1)} f\left(\cdots f\left(W^{(1)} x + b^{(1)}\right)\cdots\right) + b^{(L-1)}\right) + b^{(L)}\right) \tag{1}$$

where we used $f^{(L)}$ to allow for possibly different activation functions in different layers. This nested expression explicitly shows the feedforward computation as a nested function of x applying each layer in sequence.

We minimize the loss function for regression:

$$L = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_i - y_i\right)^2 \tag{2}$$

and classification problems:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log\left(\hat{y}_i\right) + \left(1 - y_i\right)\log\left(1 - \hat{y}_i\right)\right] \tag{3}$$

Weights and biases are updated using the following expressions:

$$W^l := W^l - \alpha\frac{\partial L}{\partial W^l} \tag{4}$$

$$b^l := b^l - \alpha\frac{\partial L}{\partial b^l} \tag{5}$$

The whole procedure consists of the following steps:

1) Compute the loss function

2) Perform forward propagation to compute activations.

3) Compute gradients using backpropagation.

4) Update weights and biases using gradient descent.

If we want to solve regression or classification problem, the last layer of a neural network usually contains only one unit. If the activation function of the last unit is linear then the neural network is a regression model if the activation function is a logistic function the neural network is a binary classification model. The results of the diabetes data (accuracy) are shown in Figure 20.

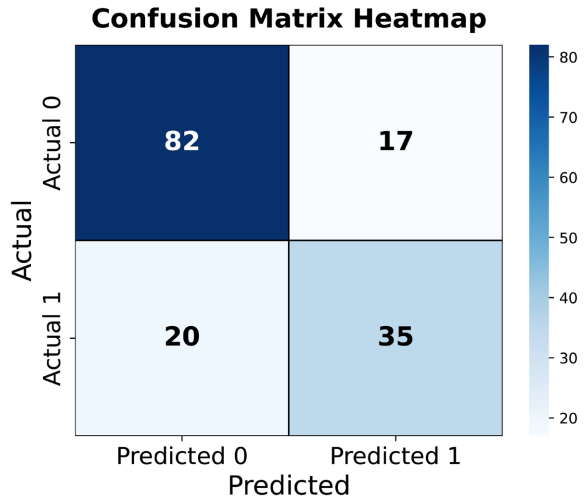

Figure 20. Confusion matrix for conventional deep learning.

Table 2 depicts the Classification report of a conventional deep-learning algorithm

applied to the diabetes data. The accuracy of the algorithm is 0.7597402597402597 which is quite typical for classification approaches (Figure 21).

Table 2. Classification report for a conventional deep learning algorithm. The accuracy is 76%.

Outcome	Precision	Recall	f1-score	Support
0 (negative)	0.80	0.83	0.82	99
1 (Positive)	0.67	0.64	0.65	55
Accuracy		76%		154
Macro avg	0.74	0.73	0.74	154
Weighted avg	0.76	0.76	0.76	154

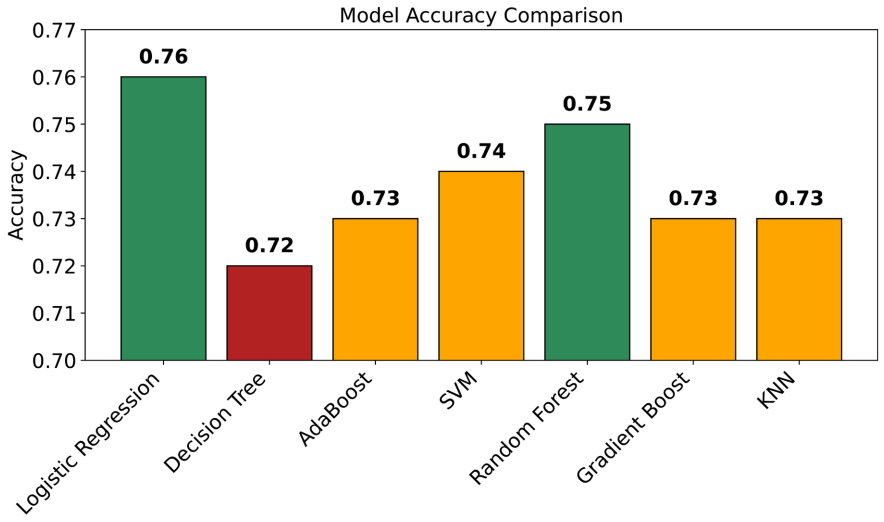

Figure 21. Accuracy of other methods applied to the data.

The confusion matrix is given by Figure 20:

Metric Description:

Precision: Proportion of positive identifications that were correct;

Recall: Proportion of actual positives that were correctly identified;

F1-score: Harmonic mean of precision and recall;

Support: Number of actual occurrences of the class in the dataset.

Deep learning has various challenges:

Data Challenges

- Data Scarcity: Many deep learning models require vast amounts of labeled data, which can be expensive and time-consuming to obtain.
- Data Quality: Noisy, biased, or imbalanced datasets can lead to poor model performance.

Computational Challenges

- High Computational Cost: Training large models requires powerful GPUs/TPUs, which are expensive.
- Energy Consumption: Deep learning models consume significant energy, rais-

ing sustainability concerns.

Let us apply other methods to the diabetes data. In supervised learning, the model is trained on labeled data, meaning the input data comes with corresponding output labels. The goal is to learn a mapping from input to outputs.

We will apply the following supervised learning algorithms:

1) Logistic Regression (for classification tasks);

2) Decision Trees;

3) Random Forests;

4) Support Vector Machines (SVM);

5) KNeighborsClassifier.

KNN is a simple, supervised machine learning (ML) algorithm that can be used for classification or regression tasks - and is also frequently used in missing value imputation. It is based on the idea that the observations closest to a given data point are the most "similar" observations in a data set, and we can therefore classify unforeseen points based on the values of the closest existing points. By choosing K, the user can select the number of nearby observations to use in the algorithm [10]. The accuracy of all known methods is in the interval (72% - 26%) (**Figure 21**).

In the next section, we will consider an optimized deep learning algorithm that improves the accuracy to up to 95% and higher with the same architecture and training conditions.

6. Generative Artificial Intelligence Algorithms

6.1. Clinical Significance of Generative AIs

The literature shows the effectiveness of generative AI in clinical decision-making, as well as highlighting its challenges. For example, AI systems provided an effective method for the management of gastrointestinal disease including early detection and diagnosis, and the author suggested using various methods of generative AI for different gastrointestinal health information in future research [11]. Furthermore, AI has already demonstrated great capability in breast cancer management and patient outcomes [12] and [13]. For example, AI systems have shown promise in mammography screening and radiotherapy treatment [13]. In addition to the benefits of AI, the researchers noted some of its challenges such as patient confidentiality, problems with moral principles, and regulation issues [13]. They emphasized the need to conduct specific studies that can prove the accuracy of AI techniques in healthcare delivery [12].

Despite these challenges, AI has the potential of making a dramatic change in the delivery of healthcare services by empowering health practitioners to provide the best patient care [14]. Similarly, AI should be held in high regard because of its ability to identify not only specific diseases but all diseases and healthcare professionals should pay attention to its effectiveness [15]. Other studies highlighted the potential of Generative AI into clinical practice, specifically in the prevention of cardiovascular disease [16].

Moreover, AI has proven to be effective in clinical decision-making despite limitations. Similarly, AI techniques have the potential of accurately classifying diabetes to improve the health status of patients.

6.2. Principal Model Generative Artificial Intelligence Algorithm (PM GenAI)

In this paper, we developed a new generative AI algorithm that consists of the following steps: 1) PM GenAI explores the data and generates new data by sampling from the posterior distribution of the model weights. This allows it to create diverse outputs based on the learned distributions, rather than relying on a fixed, deterministic model, 2) Uncertainty Quantification: By modeling uncertainty in parameters, PM GenAI produces a distribution of outputs instead of a single prediction. This is particularly useful for tasks that benefit from multiple plausible outcomes, such as data generation or decision-making under uncertainty, 3) Design the neural network architecture and specify prior distributions for the model parameters. Use techniques such as variational inference or Markov Chain Monte Carlo (MCMC) to approximate the posterior distribution of the weights, 4) Once the posterior is approximated (typically via backpropagation), sample weights from the distribution and perform feedforward passes through the network to generate predictions or synthetic data points that reflect the learned uncertainty.

The algorithm divides the data into test and training. In the training data set it defines the trend and statistics: Mean vector (μ), the center of the distribution, covariance matrix (Σ): the shape, spread, and orientation. The PDF of a Gaussian component is:

$$P\left(x\middle|\mu_k,\Sigma_k\right)=\frac{1}{\left(2\pi\right)^{1/d}\left|\Sigma_k\right|^{1/2}}\exp\left(-\frac{1}{2}\left(x-\mu_k\right)^{\mathrm{T}}\Sigma_k^{-1}\left(x-\mu_k\right)\right) \tag{6}$$

The covariance matrix is:

$$\Sigma_k=\frac{1}{N_k}\sum_{i=1}^{N}\gamma_{ik}\left(x_i-\mu_k\right)\left(x_i-\mu_k\right)^{\mathrm{T}} \tag{7}$$

In this expression:

$$N_k=\sum_{i=1}^{N}\gamma_{ik} \tag{8}$$

PM GenAI generates augmented data that resembles real data, which is especially valuable in scenarios with limited labeled samples largely improving the quality of data sets. By incorporating uncertainty, the algorithm also enables models to be more robust to outliers and noise, ultimately improving generalization. In summary, using PM GenAI in deep learning for data generation enables the incorporation of uncertainty into models, generate realistic synthetic data, and make more reliable and accurate predictions.

6.3. The Choice of Hyperparameters

Learning Rate**:** Controls how much the model updates its weight during training.

A high learning rate may lead to convergence issues, while a low one can make training slow (we use Adam or Adagrad optimizers to automatically adjust the learning rate).

Batch Size**:** Defines the number of training samples used in one forward and backward pass. Common choices include 32, 64, and 128. We use 128 and then split it into mini batches.

Number of Epochs**:** Specifies how many times the entire dataset is passed through the neural network. We usually do not use more than 100 epochs.

Optimizer (e.g., SGD, Adam, RMSprop): Determines how the model updates weights during training. We usually use an Adam optimizer.

Loss Function (e.g., MSE, Cross-Entropy): Measures the difference between predicted and actual outputs. We use Cross-Entropy because we solve the classification problem. MSE us used in regression algorithms,

Number of Layers and Neurons per Layer**:** Defines the depth and complexity of the network. We start with 8 neurons (the number of features). It is highly advisable this way of building the neural network to prevent overfitting if many neurons are selected.

Dropout Rate**:** Controls the percentage of neurons randomly dropped during training to prevent overfitting (we use dropout rate 0.3-0.4 to prevent over or underfitting)

Weight Initialization: Determines how weights are initialized before training (e.g., Xavier, He initialization). Choosing the right weight initialization method depends on the activation function and network depth. We use Xavier initialization to work well with classification problems. Proper weight initialization leads to more stable training and improved model performance.

Activation Functions (e.g., ReLU, Sigmoid): Define how neurons process inputs. We use ReLU for hidden layers and Sigmoid for the output layer typical to classification problems.

L1/L2 Regularization (Weight Decay): Helps prevent overfitting by penalizing large weights. It is implemented in optimizers like AdamW and SGD.

K-fold cross-validation. We used K-fold cross-validation as a resampling technique to evaluate the performance of the method. It evaluates how well a model generalizes unseen data by splitting the dataset into multiple subsets. K-fold cross-validation includes:

1) The dataset is divided into K equally sized folds (subsets).

2) The model is trained on K-1 folds and tested on the remaining one.

3) The process is repeated K times, with each fold serving as the test set once.

4) The final model performance is obtained by averaging the results across all K iterations.

5) Stratified K-Fold ensures class distribution is preserved in each fold (useful for imbalanced datasets and can be used in several applications including public health).

Statistical analysis of the results shows that the comparison of previous methods

(Figure 21) and the PM GenAI accuracy yields:

$$P \sim 0.004 < 0.05$$

We may conclude that PM GenAI significantly improves the performance of the classification algorithms for diabetes.

6.4. PM GenAI Confusion Matrix and Accuracy

Figure 22 depicts the Confusion matrix generated by PM GenAI. Table 3 is a complete classification report demonstrating 97% accuracy.

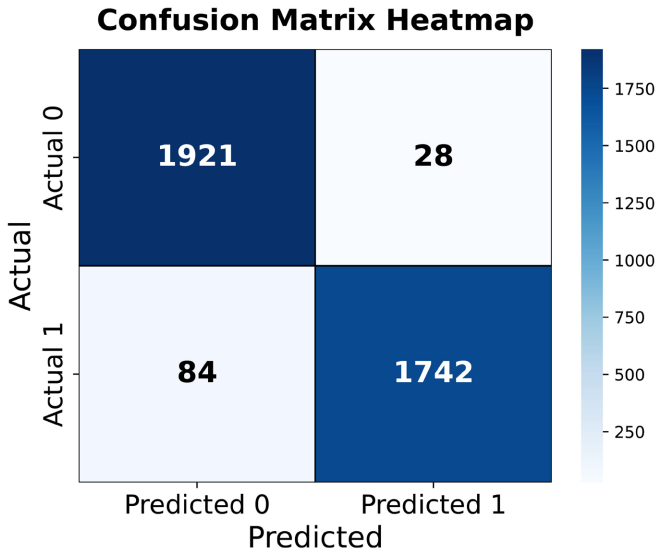

Figure 22. Confusion matrix shows large improvement of accuracy compared with other methods.

Table 3. Classification report for a PM GenAI algorithm. The accuracy is 97%.

Outcome	Precision	Recall	f1-score	Support
0 (negative)	0.95	0.98	0.97	1949
1 (Positive)	0.98	0.95	0.97	1826
Accuracy		97%		3775
Macro avg	0.97	0.96	0.97	3775
Weighted avg	0.97	0.97	0.97	3775

7. Conclusions

The abundance of biomedical data presents significant opportunities and challenges in healthcare research. A crucial aspect is identifying relationships among diverse data points to develop reliable medical tools using data-driven techniques and machine learning. Previous studies have integrated multiple data sources to achieve this and create comprehensive knowledge bases for predictive analysis and discovery. While existing models show considerable potential, machine learning-based predictive tools have yet to gain widespread adoption in the medical field.

Research has demonstrated that optimized algorithms, such as Support Vector Machines (SVM), play a vital role in accurately diagnosing various diseases [17].

Despite these advancements, deep learning approaches remain underutilized in addressing a wide range of healthcare and medical challenges despite their significant potential. Deep learning offers several advantages, including superior performance, an end-to-end learning framework with built-in feature extraction, and the ability to process complex, multi-modal data. However, for these methods to be effectively implemented in healthcare, researchers must address challenges associated with medical data, which is often sparse, noisy, heterogeneous, and time dependent [18].

Additionally, improved methodologies and tools are needed to facilitate the seamless integration of deep learning into healthcare workflows and clinical decision-making.

In this study, we initially utilized a Kaggle dataset provided by the National Institute of Diabetes and Digestive and Kidney Diseases. This dataset aims to predict diabetes risk in patients based on specific diagnostic indicators. The data were curated with constraints, including only Pima Indian women aged 21 and above. It contains multiple predictor variables, including the number of pregnancies, body mass index (BMI), insulin levels, age, and other relevant factors, alongside the target outcome variable.

The results of this study are auspicious, achieving an accuracy rate of approximately 95% - 98% (see also [12]). [19] compared different methods and found that deep learning gives the best results. [20] published a paper on projections of global mortality and burden of disease focusing on diabetes. [21] [22] considered the data mining approach to prediction, while [22] and [23] discussed the risks of disease development using ML approaches.

The proposed method offers several advantages:

1) It reduces variance and enhances predictive accuracy.

2) It performs well on large datasets, effectively handling missing and noisy data.

3) It prioritizes features based on their impact on decision-making, enabling analysts to focus on critical variables.

4) It efficiently processes high-dimensional data, making it highly suitable for real-world decision support applications.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Adaji, A., Schattner, P. and Jones, K. (2008) The Use of Information Technology to Enhance Diabetes Management in Primary Care: A Literature Review. *Journal of Innovation in Health Informatics*, **16**, 229-237. https://doi.org/10.14236/jhi.v16i3.698

[2] Barnett, A.H., Eff, C., Leslie, R.D.G. and Pyke, D.A. (1981) Diabetes in Identical Twins. A Study of 200 Pairs. *Diabetologia*, **20**, 87-93.

https://doi.org/10.1007/bf00262007

[3] Redondo, M.J., Jeffrey, J., Fain, P.R., Eisenbarth, G.S. and Orban, T. (2008) Concordance for Islet Autoimmunity among Monozygotic Twins. *New England Journal of Medicine*, **359**, 2849-2850. https://doi.org/10.1056/nejmc0805398

[4] Tol, A. and Baghbanian, A. (2012) The Introduction of Self-Management in Type 2 Diabetes Care: A Narrative Review. *Journal of Education and Health Promotion*, **1**, 35. https://doi.org/10.4103/2277-9531.102048

[5] Cho, N.H., Whiting, D., Guariguata, L., Montoya, P.A., Forouhi, N., Hambleton, I., *et al.*, (2013) IDF Diabetes Atlas. 6th Edition, International Diabetes Federation.

[6] Riazi, H., Larijani, B., Langarizadeh, M. and Shahmoradi, L. (2015) Managing Diabetes Mellitus Using Information Technology: A Systematic Review. *Journal of Diabetes & Metabolic Disorders*, **14**, Article No. 49. https://doi.org/10.1186/s40200-015-0174-x

[7] Funjan, K.I. (2020) Skin Thickness Can Predict the Progress of Diabetes Type 2: A New Medical Hypothesis. *EC Diabetes and Metabolic Research*, **4**, 8-12.

[8] Helmer, J. (2024) How Age Relates to Type 2 Diabetes. Google Publication. https://www.webmd.com/diabetes/diabetes-link-age

[9] Liu, B., Song, L., Zhang, L., Wang, L., Wu, M., Xu, S., *et al.* (2020) Higher Numbers of Pregnancies Associated with an Increased Prevalence of Gestational Diabetes Mellitus: Results from the Healthy Baby Cohort Study. *Journal of Epidemiology*, **30**, 208-212. https://doi.org/10.2188/jea.je20180245

[10] De Melo, P. (2024) Public Health Informatics and Technology. Library of Congress, Washington DC.

[11] Lee, K. and Kim, E.S. (2024) Generative Artificial Intelligence in the Early Diagnosis of Gastrointestinal Disease. *Applied Sciences*, **14**, Article 11219. https://doi.org/10.3390/app142311219

[12] Ahn, J.S., Shin, S., Yang, S., Park, E.K., Kim, K.H., Cho, S.I., *et al.* (2023) Artificial Intelligence in Breast Cancer Diagnosis and Personalized Medicine. *Journal of Breast Cancer*, **26**, 405-435. https://doi.org/10.4048/jbc.2023.26.e45

[13] AlSamhori, J.F., AlSamhori, A.R.F., Duncan, L.A., Qalajo, A., Alshahwan, H.F., Alabbadi, M., *et al.* (2024) Artificial Intelligence for Breast Cancer: Implications for Diagnosis and Management. *Journal of Medicine, Surgery, and Public Health*, **3**, Article 100120. https://doi.org/10.1016/j.glmedi.2024.100120

[14] Aamir, A., Iqbal, A., Jawed, F., Ashfaque, F., Hafsa, H., Anas, Z., *et al.* (2024) Exploring the Current and Prospective Role of Artificial Intelligence in Disease Diagnosis. *Annals of Medicine & Surgery*, **86**, 943-949. https://doi.org/10.1097/ms9.0000000000001700

[15] Kaur, S., Singla, J., Nkenyereye, L., Jha, S., Prashar, D., Joshi, G.P., *et al.* (2020) Medical Diagnostic Systems Using Artificial Intelligence (AI) Algorithms: Principles and Perspectives. *IEEE Access*, **8**, 228049-228069. https://doi.org/10.1109/access.2020.3042273

[16] Parsa, S., Somani, S., Dudum, R., Jain, S.S. and Rodriguez, F. (2024) Artificial Intelligence in Cardiovascular Disease Prevention: Is It Ready for Prime Time? *Current Atherosclerosis Reports*, **26**, 263-272. https://doi.org/10.1007/s11883-024-01210-w

[17] De Melo, P. and Davtyan, M. (2023) High Accuracy Classification of Populations with Breast Cancer: SVM Approach. *Cancer Research Journal*, **11**, 94-104.

[18] de Melo, P. (2025) Augmented and Synthetic Data in Artificial Intelligence. *International Journal of Artificial Intelligence & Applications*, **16**, 93-108.

https://doi.org/10.5121/ijaia.2025.16307

[19] Naz, H. and Ahuja, S. (2020) Deep Learning Approach for Diabetes Prediction Using PIMA Indian Dataset. *Journal of Diabetes & Metabolic Disorders*, **19**, 391-403. https://doi.org/10.1007/s40200-020-00520-5

[20] Mathers, C.D. and Loncar, D. (2006) Projections of Global Mortality and Burden of Disease from 2002 to 2030. *PLOS Medicine*, **3**, e442. https://doi.org/10.1371/journal.pmed.0030442

[21] Swapna, G., Vinayakumar, R. and Soman, K.P. (2018) Diabetes Detection Using Deep Learning Algorithms. *ICT Express*, **4**, 243-246. https://doi.org/10.1016/j.icte.2018.10.005

[22] Wu, H., Yang, S., Huang, Z., He, J. and Wang, X. (2018) Type 2 Diabetes Mellitus Prediction Model Based on Data Mining. *Informatics in Medicine Unlocked*, **10**, 100-107. https://doi.org/10.1016/j.imu.2017.12.006

[23] The Emerging Risk Factors Collaboration, (2010) Diabetes Mellitus, Fasting Blood Glucose Concentration, and Risk of Vascular Disease: A Collaborative Meta-Analysis of 102 Prospective Studies. *The Lancet*, **375**, 2215-2222. https://doi.org/10.1016/s0140-6736(10)60484-9