Exploring Diabetes Using Advanced Information Technologies

Philip de Melo
American Medical Informatics Association
6218 Georgia Avenue NW,
3077 Washington, DC 20011

Summary

The importance of timely and accurate information about health status increased rapidly. To mitigate inevitable medical errors, a very accurate classification based on advanced IT technologies has to be developed. Gaining knowledge and actionable insights from complex, high-dimensional and heterogeneous biomedical data remains a key challenge in transforming health care. Various types of data have been emerging in modern biomedical research, including electronic health records, imaging, -omics, sensor data and text, which are complex, heterogeneous, poorly annotated and generally unstructured.

Traditional data mining and statistical learning approaches typically need to first perform feature engineering to obtain effective and more robust features from those data, and then build prediction or clustering models on top of them. There are lots of challenges on both steps in a scenario of complicated data and lacking of sufficient domain knowledge. The latest advances in deep learning technologies provide new effective paradigms to obtain end-to-end learning models from complex data. It is also true that current deep learning algorithms do not always give high accurate results. In this paper, a new hybrid method based on the summation of randomly chosen data is developed. It permitted to boost the accuracy from 77% to almost 93%. The next step is to optimize the algorithm by the reduction of the calculations time.

Introduction

According to the International Diabetes Federation (IDF) report, Diabetes is a pervasive chronic disease affects 382 million people worldwide and more than 592 million people will be affected within a generation. However, most of those cases would be preventable.

Diabetic patients with poor blood glucose control have higher mortality and morbidity rates which is related to chronic complications such as neuropathy. Diabetes is a leading cause of death due to increased risk of coronary artery disease and stroke [3].

The estimated total cost of diabetes care in the world was at least USD 548 billion in 2013. This estimation is expected to be more than USD 627 billion for 2035 [4].

Several information technology-based interventions were applied to enhance blood glucose monitoring and diabetes management. Previous evidence demonstrates that information technology can improve diabetes management through better metabolic control and help in the global care of diabetic people with chronic illnesses.

Adaji et al. performed a literature review about the use of information technology to enhance diabetes management. They concluded that promoting a productive and informative interaction between the patient and the care team by using information technology based interventions can lead to improve diabetes care [5].

Information technology-based interventions have some advantages such as reducing medical errors, generating potential data for research, and increasing the ability for continuous improvement. On the other hand, higher cost of initially and maintenance activities, difficulty of using computer and information systems for healthcare providers and spending more time than interacting with a patient are some disadvantages of using information technology in diabetes care.

Recent findings suggest that information technology-based interventions can improve glycemic control in patients with diabetes and lead to better management of diabetes with different effects of intervention on various clinical findings. Combining multiple information technology-based interventions and proposing a comprehensive solution for obtaining better results in various clinical findings lead to better diabetes management may be the suggested future research [6].

This paper is focused on the classification of diabetes using advanced information technology algorithms. The accuracies of these algorithms are compared.

Data Description

We consider the diabetes data available in Kaggle (5 first rows are printed below). This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective is to predict based on diagnostic measurements whether a patient has diabetes.

	a1	**a2**	**a3**	**a4**	**a5**	**a6**	**a7**	**a8**	**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

Table1: First 5 rows of the data set. The data set consists of 768 rows (patients) and 9 columns (features+outcome). The preprocessing procedure included substitution of zeroes by their means or mediums.

Several constraints were placed on the selection of these instances from a larger database. In particular, all patients are females at least 21 years old of Pima Indian heritage. The following features determine the input parameters:

Input:

- a1:Pregnancies-Number of times pregnant
- a2: Glucose-Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- a3: Blood Pressure-Diastolic blood pressure (mm Hg)
- a4: Skin Thickness- Triceps skin fold thickness (mm)
- a5: Insulin- 2-Hour serum insulin (mu U/ml)
- a6: BMI: Body mass index (weight in kg/(height in m)**2)
- a7: Diabetes Pedigree Function: Diabetes pedigree function
- a8: Age- (years)

A particularly interesting attribute used in this study was the Diabetes Pedigree Function. It provided some data on diabetes history in relatives and the genetic relationship of those relatives to the patient. This measure of genetic influence gives an idea of the hereditary risk one might have with the onset of diabetes.

Diabetes is associated with systematic alterations of the body such as skin changes. The main objectives of [7] was to analyze data regarding skin thickness among diabetic women patients, and to evaluate the assumption that skin thickness may be a predictor of diabetic status of women patients. The impacts of study variables on skin thickness were determined based on One Way ANOVA test. Significance was considered at $\alpha \leq 0.05$. Study findings showed that skin thickness decreased as a result of diabetic progression. Skin thickness was significantly influenced by insulin level, but not glucose level. Taken together, the results of the present study showed that skin thickness may be a new predictor of the progression of diabetes in women patients, and further studies are required to establish this assumption. In this paper, we will examine this result. The next step would be the calculation of the correlation matrix.

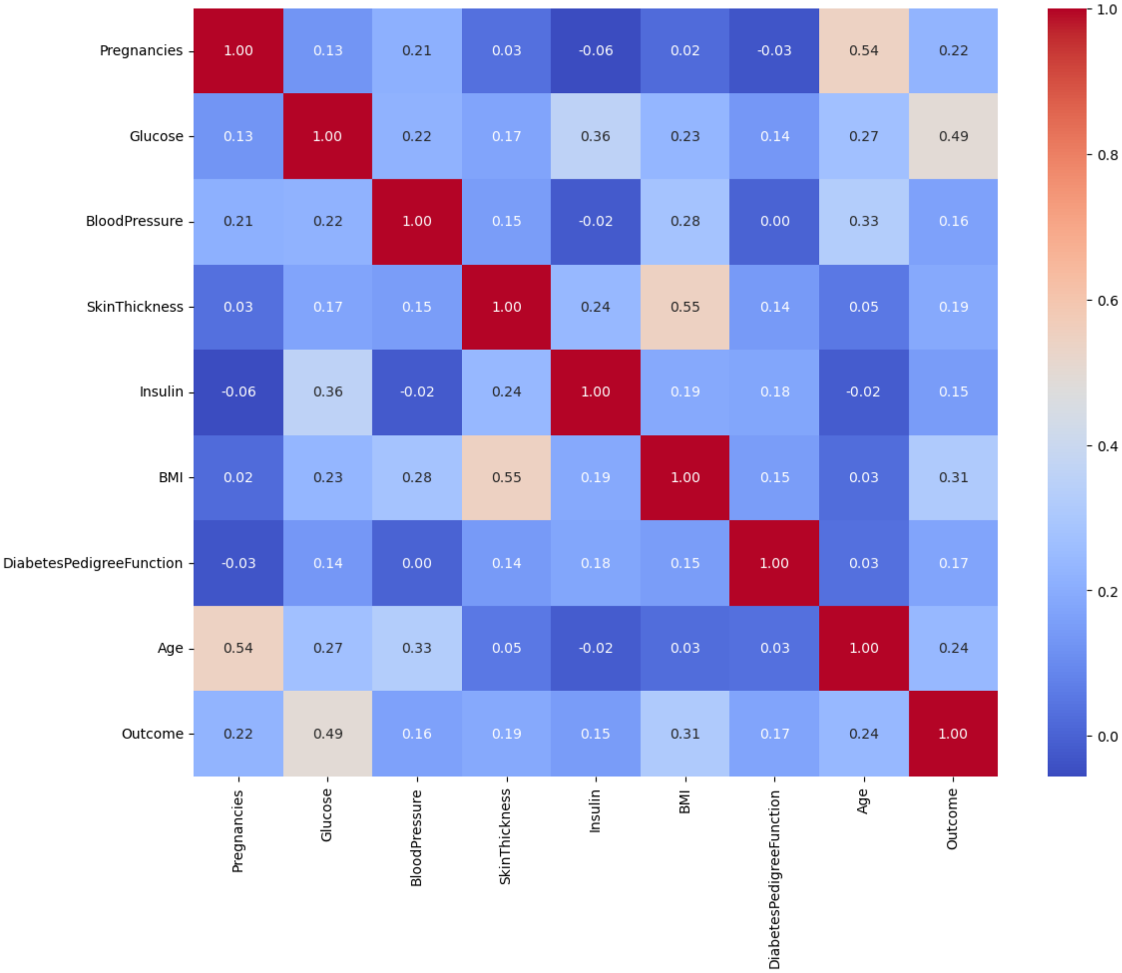

Figure 1: The heatmap shows the correlation between features and outcome

The next Figure demonstrates the number of patients without diabetes (blue) and with diabetes (orange).

Feature engineering

Feature engineering yields the features which mostly contribute to the outcome. It yields features that most accurately describe the collected data with respect to the decision (can be binary decision). Feature engineering is the process of transforming raw data into features that are suitable for machine learning models. In other words, it is the process of selecting, extracting, and transforming the most relevant features from the available data to build more accurate and efficient machine learning models.

The success of machine learning models heavily depends on the quality of the features used to train them. Feature engineering involves a set of techniques that also enable us to create new features by combining

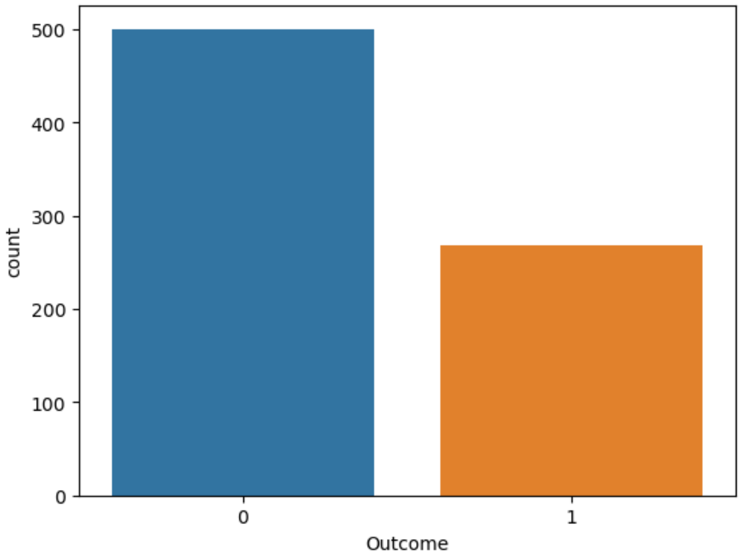

Figure 2: Diagram shows the number of patients with no diabetes (blue) and with diabetes cases.

or transforming the existing ones. These techniques help to highlight the most important patterns and relationships in the data, which in turn helps the machine learning model to learn from the data more effectively. After data is cleaned and labeled, ML teams often explore the data to make sure it is correct and ready for ML.

Visualizations like histograms, scatter plots, box and whisker plots, line plots, and bar charts are all useful tools to confirm data is correct. Additionally, visualizations also help data science teams complete exploratory data analysis. This process uses visualizations to discover patterns, spot anomalies, test a hypothesis, or check assumptions. Exploratory data analysis does not require formal modeling; instead, data science teams can use visualizations to decipher the data. Collecting data is the process of assembling all the data you need for ML. Data collection can be tedious because data resides in many data sources, including on laptops, in data warehouses, in the cloud, inside applications, and on devices. Finding ways to connect to different data sources can be challenging. Data volumes are also increasing exponentially, so there is a lot of data to search through. Additionally, data has vastly different formats and types depending on the source. For example, video data and tabular data are not easy to use together.

The outcome θ can be presented as a map of the sum of features:

$$\sum_{n=1}^{N} F_i(x) \rightarrow \theta$$

If we calculate the cross correlation between each feature and the outcome and if the results is less than a given threshold α, this feature can be dropped.

$$F_i(x) \cap \theta < \alpha$$

Figures 3 and 4 show the results of feature engineering with the threshold is 20

Detection and Removal of Outliers

Outlier detection and removal is a technique for removing outliers from a data set. This method can be used to produce a more accurate data representation. This has an impact on the model's performance.

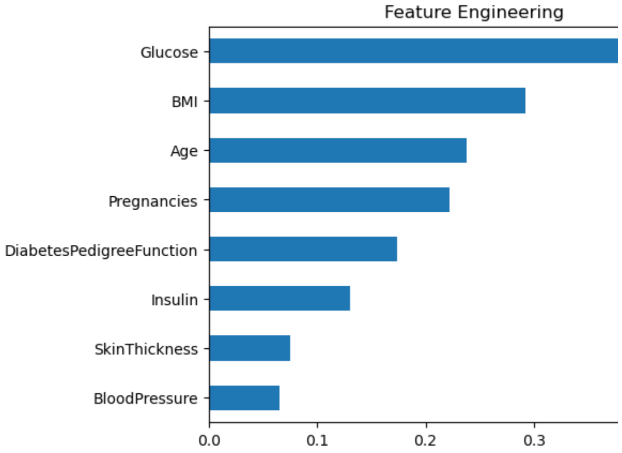

Figure 3: Feature engineering analysis shows the contribution of each feature in the outcome.

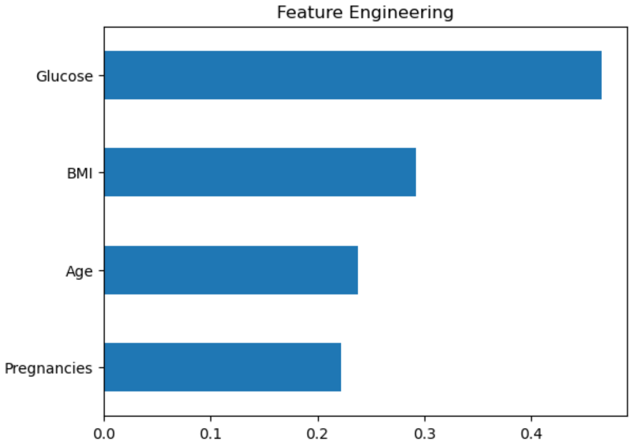

Figure 4: Feature engineering analysis shows the most contribution comes from 4 features.

Depending on the model, the effect could be large or minimal. For example, linear regression is particularly susceptible to outliers. This procedure should be completed prior to model training. The various methods of handling outliers include:

- Removal: Outlier-containing entries are deleted from the distribution. However, if there are outliers across numerous variables, this strategy may result in a big chunk of the datasheet being missed
- Replacing values: Alternatively, the outliers could be handled as missing values and replaced with suitable imputation.
- Capping: Using an arbitrary value or a value from a variable distribution to replace the maximum and minimum values.
- Discretization : Discretization is the process of converting continuous variables, models, and functions into discrete ones. This is accomplished by constructing a series of continuous intervals (or bins) that span the range of our desired variable/model/function.

Below are the boxplots of the selected features. In descriptive statistics, a box plot or boxplot (also known as a box and whisker plot) is a type of chart often used in explanatory data analysis. Box plots visually show the distribution of numerical data and skewness by displaying the data quartiles (or percentiles) and averages.

Box plots show the five-number summary of a set of data: including the minimum score, first (lower) quartile, median, third (upper) quartile, and maximum score. Definitions:

- Minimum score: Minimum Score The lowest score, excluding outliers (shown at the end of the left whisker).
- Lower Quartile: Twenty-five percent of scores fall below the lower quartile value (also known as the first quartile).
- Median: The median marks the mid-point of the data and is shown by the line that divides the box into two parts (sometimes known as the second quartile). Half the scores are greater than or equal to this value, and half are less.
- Upper Quartile: Seventy-five percent of the scores fall below the upper quartile value (also known as the third quartile). Thus, 25 percent of data are above this value.

- Maximum Score: The highest score, excluding outliers (placed at the end of the right whisker).
- Whiskers: The upper and lower whiskers represent scores outside the middle 50 percent (i.e., the lower 25 percent of scores and the upper 25 percent of scores)
- The Interquartile Range (or IQR):The box plot shows the middle 50 percent of scores (i.e., the range between the 25th and 75th percentile).

Boxplot captures the summary of the data effectively and efficiently with only a simple box and whiskers. Boxplot summarizes sample data using 25th, 50th, and 75th percentiles. One can just get insights(quartiles, median, and outliers) into the dataset by just looking at its boxplot. Outliers can wreak havoc on data analysis and machine learning models. They can lead to incorrect conclusions, biased predictions, and skewed statistical measures. To combat this, we use statistical methods to detect and manage outliers. In this blog post, we will learn the ins and outs of the IQR method.

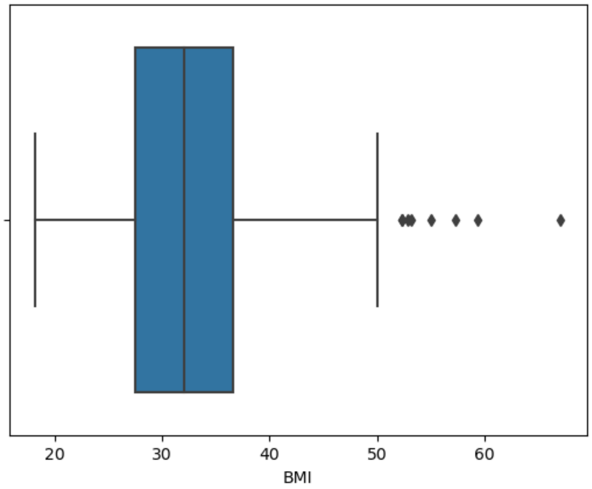

Figure 5: Outliers detected in the BMI records.

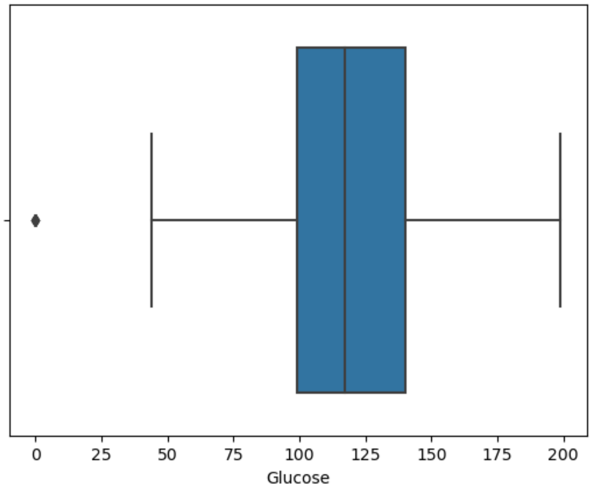

Figure 6: Outliers detected in the Glucose records.

The Interquartile Range, or IQR, is a measure of statistical dispersion. It represents the range within which the middle 50 percent of the data falls. To calculate the IQR, we need to find the difference between the 75th percentile (Q3) and the 25th percentile (Q1).To identify outliers using the IQR method, we need to establish two boundaries:

- Lower Bound: Q1-1.5*IQR
- Upper Bound: Q3+1.5*IQR

These boundaries help us determine which data points might be outliers. Any data point that falls below the lower bound (Q1–1.5 * IQR) is considered an outlier. These values are significantly lower than the majority of the dataset and are potential candidates for removal or further investigation. Conversely, any data point that exceeds the upper bound (Q3 + 1.5 * IQR) is also considered an outlier. These values are much higher than the majority of the dataset and may warrant special attention. One advantage of the IQR method is that it is robust to skewed data distributions. It identifies outliers based on percentiles, making it less sensitive to extreme values. The IQR method is easy to implement and interpret. It provides a clear range within which most data points should fall, making it a valuable tool for data analysis and quality control.

Z-score to eliminate outliers

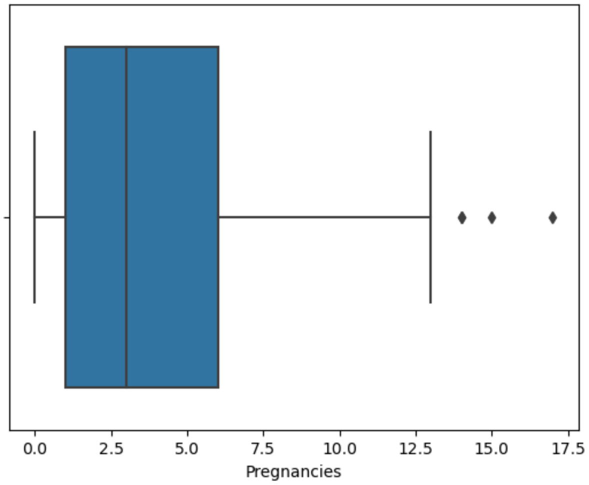

Figure 7: Outliers detected in the Pregnancies records.

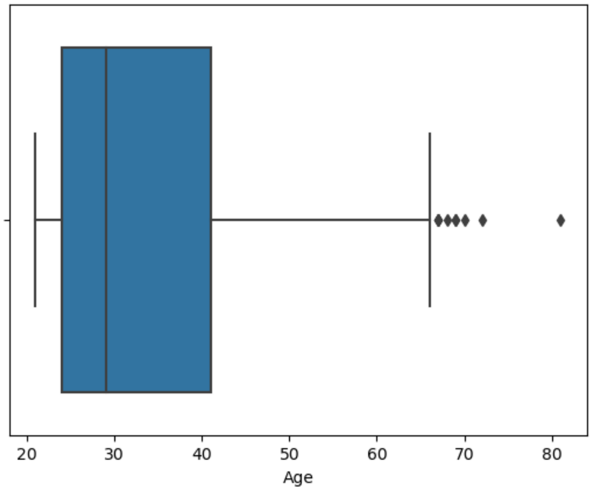

Figure 8: Outliers detected in the Age records, but they were included in the analysis

Z-score is used to standardize the variable, so that just by knowing the value of a particular observation, you get the sense of how far away it is from the mean. More specifically, ‘Z score’ tells how many standard deviations away a data point is from the mean. The process of transforming a feature to its z-scores is called ‘Standardization’. If the z score of a data point is more than 3, it indicates that the data point is quite different from the other data points. Such a data point can be an outlier.

Z-score can be both positive and negative. The farther away from 0, higher the chance of a given data point being an outlier. Typically, Z-score greater than 3 is considered extreme.

$$z_i = \frac{x_i - \bar{x}}{s}$$

Figure 9: Z-scores can be used to estimate the probability of a particular data point occurring within a normal distribution. By converting z-scores to percentiles or using a standard normal distribution table, you can determine the likelihood of a value being above or below a certain threshold. When the population mean and the population standard deviation are unknown, the standard score may be calculated using the sample mean and sample standard deviation (s) as estimates of the population values.

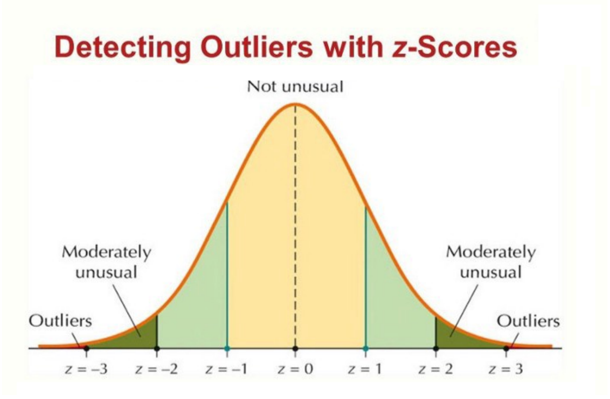

Figure 10: : Z-scores help identify outliers, which are data points significantly different from the rest of the dataset. Typically, data points with z-scores greater than 3 or less than -3 are considered potential outliers and may warrant further investigation.

Z-score and normal distribution

Z-score removal of outliers requires normal distribution. We can check it from diagrams.

The BMI and Glucose data resemble the normal distribution.

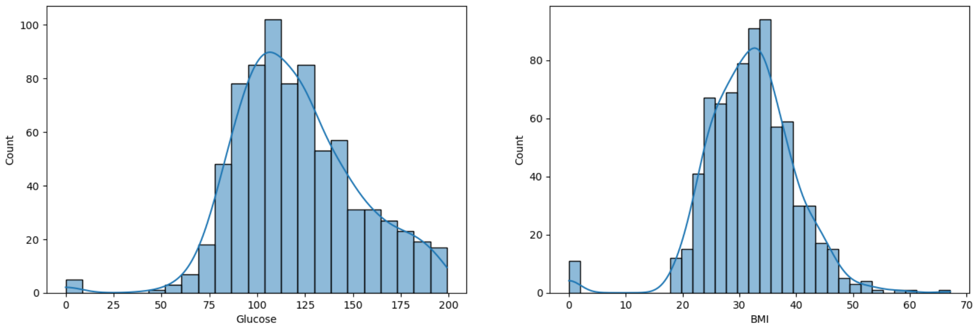

Figure 11: The data distribution for Glucose concentration (11a) and BMI (11bb)

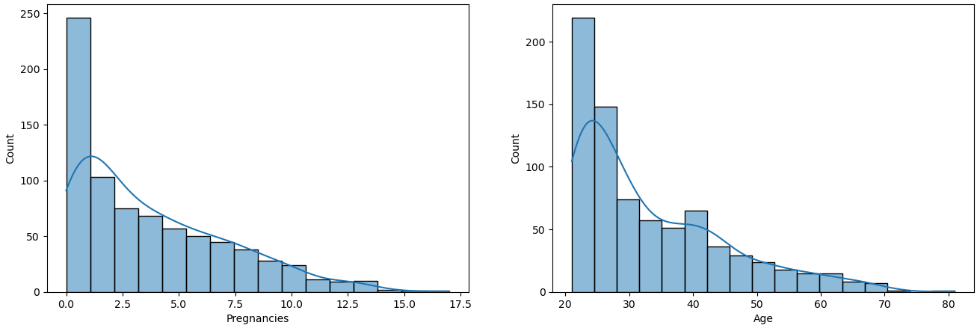

Figure 12: The data distribution for Pregnancies (12a) and Age(12b). The distributions are highly skewed and Z-score cannot be used.

After truncation of outliers, 11a and 11b become

BMI vs. Diabetes

Having obesity makes one more likely to develop diabetes, the condition of having too much glucose (sugar) circulating in the bloodstream. Obesity also causes diabetes to worsen faster.

Here's what happens: Managing the level of glucose in blood is the job of the pancreas. The pancreas creates insulin, which is a hormone that moves glucose out of blood. Normally, insulin transports glucose to muscles to use right away for energy or to the liver, where it's stored for later.

But when one has diabesity, his or her cells resist letting insulin move glucose into them. To make matters worse, the area of the liver where excess glucose is usually stored is filled with fat.

With nowhere to be stored, the glucose remains in the bloodstream. So the pancreas creates even more insulin trying to accomplish the job of moving glucose out of the blood. It's trying to push against the resistance created by the fat. The pancreas becomes overworked, and as a result, it wears out. It starts producing less insulin. Diabetes develops and then quickly worsens if the fat resistance remains." If you have obesity, you're about six times more likely to develop Type 2 diabetes than those at a healthy weight. But not everyone with obesity automatically gets diabetes. Other factors are likely at play, too, including:

- Family history
- Diet
- Exercise
- Stress
- Gut health

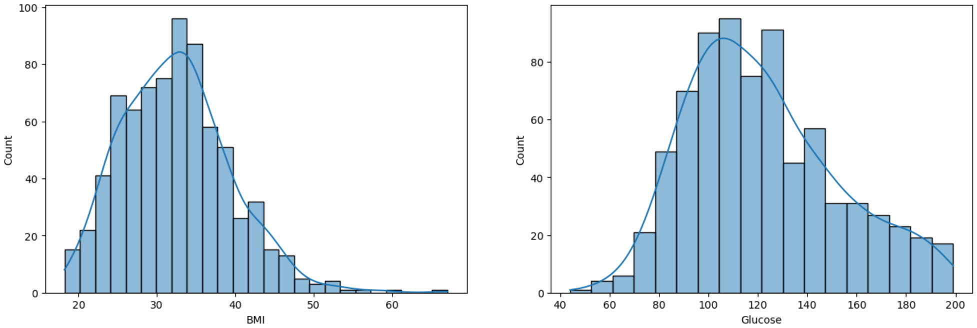

Figure 13: The data distribution for BMI and Glucose represent slightly skewed normal distributions to be used using z-score for outliers removal.

It may be that some people with obesity can produce more insulin without overtaxing the pancreas. Others might be limited in insulin production, making it more likely that obesity will lead to diabesity.

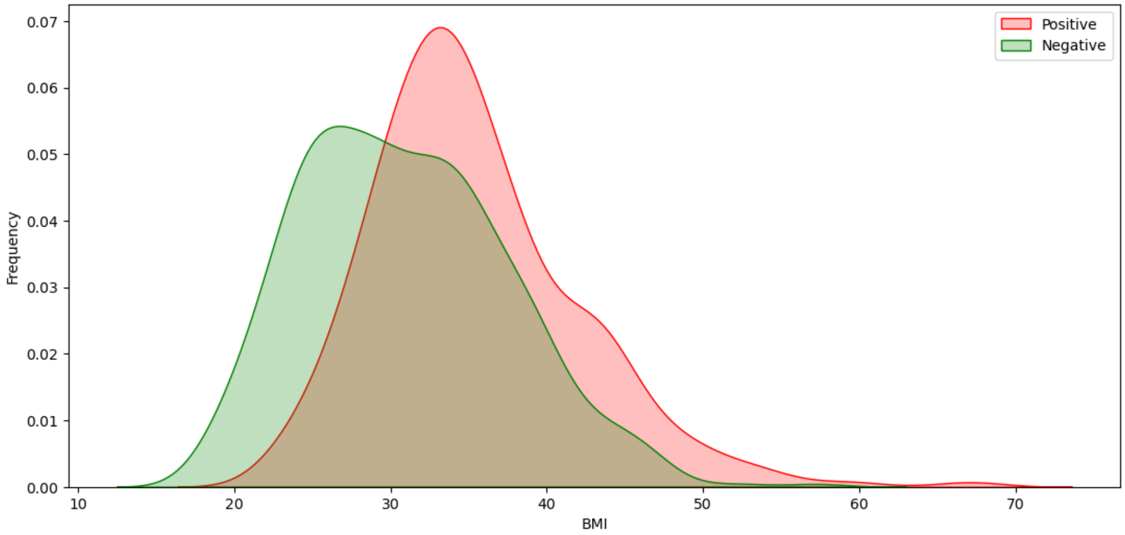

Figure 14: BMI near 30 threatens the development of diabetes.

Glucose vs. Diabetes

Glucose comes from the Greek word for "sweet." It's a type of sugar you get from foods you eat, and your body uses it for energy. As it travels through your bloodstream to your cells, it's called blood glucose or blood sugar.

Insulin is a hormone that moves glucose from your blood into your cells for energy and storage. If you've been diagnosed with diabetes, you have higher-than-normal levels of glucose in your blood. Either you don't have enough insulin to move it through your bloodstream, or your cells don't respond to insulin as well as they should.

High blood glucose for a long period can damage your kidneys, eyes, and other organs. The glucose molecules in your bloodstream mainly come from foods that are rich in carbohydrates, including bread, potatoes, and fruit. As you eat, food travels down your esophagus to your stomach. There, acids and enzymes break it down into tiny pieces. During that process, glucose is released.

From there, it goes into your intestines where it's absorbed and passes into your bloodstream. Insulin then helps glucose get inside your cells. Your blood sugar level should naturally rise after you eat. Then, it dips a few hours later as insulin moves glucose into your cells. Between meals, a healthy blood sugar level is less than 100 milligrams per deciliter (mg/dL). This is called your fasting blood sugar level. But if you have diabetes, your body can't turn food into energy the way it should.

There are two types of diabetes:

In type 1 diabetes, your body doesn't have enough insulin. Your immune system attacks and destroys cells in your pancreas, where insulin is made, by mistake. In type 2 diabetes, your cells don't respond well to insulin. As a result, your pancreas needs to make more and more insulin to move glucose into your cells. Over time, this can damage your pancreas so it can't make enough insulin to meet your body's needs. Without enough insulin, glucose stays in your bloodstream. A level above 200 mg/dL 2 hours after a meal or above 125 mg/dL during fasting is high blood glucose, also called hyperglycemia. Too much glucose in your bloodstream for a long period can damage the vessels that carry oxygen-rich blood to your organs. High blood sugar can increase your risk for:

- Heart disease, heart attack, and stroke
- Kidney disease
- Nerve damage
- Stress
- An eye disease called retinopathy

If you live with diabetes, it's important to test your blood sugar often. Exercise, being careful about the foods you eat, and medicine can all help keep your blood glucose at a healthy level and prevent complications

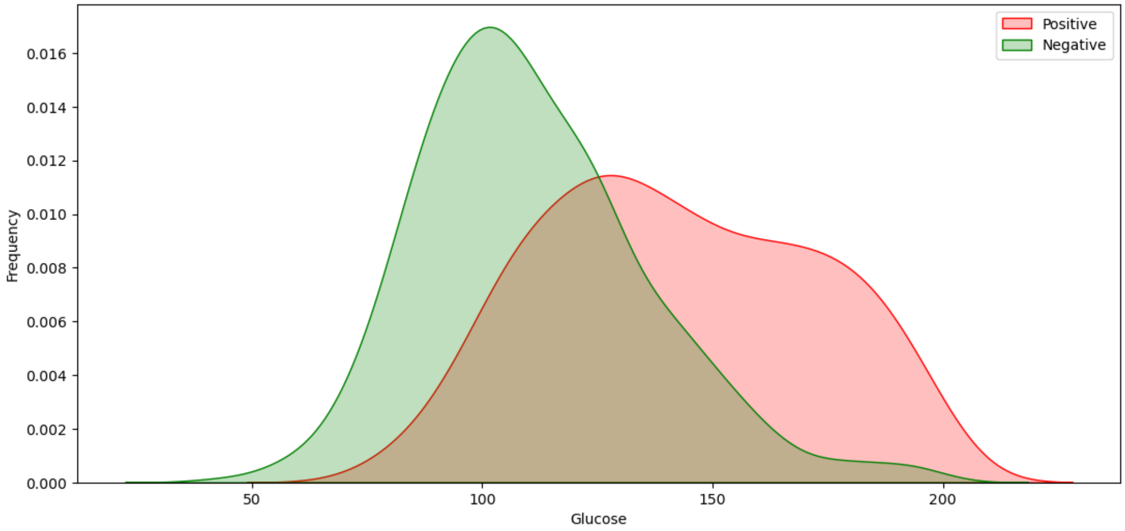

Figure 15: Glucose levels over 130 pose higher diabetes risk development.

Age vs. Diabetes

More than 9 out of 10 people with diabetes have type 2. It used to be called adult-onset diabetes because it was rarely diagnosed in children.

Age is a big risk factor for type 2. The older you are, the more likely you are to have it. That also holds true for preteens and teenagers, whose diabetes rates have climbed sharply in recent years. Type 2 is a disease caused by a mix of your genes and your lifestyle. Being overweight, having high blood pressure, and not exercising all raise your chances for type 2.

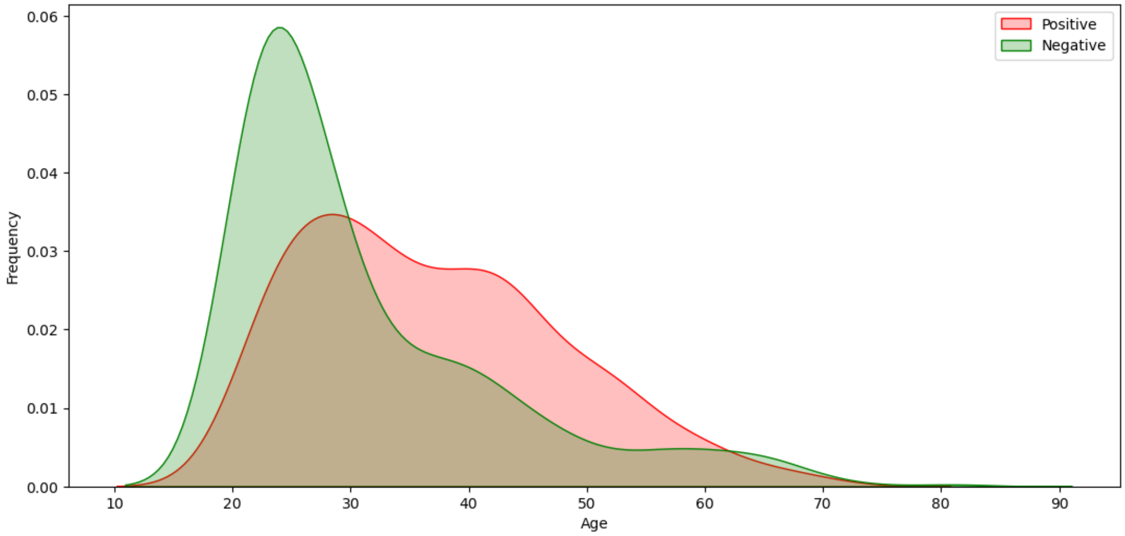

Figure 16: Glucose levels over 130 pose higher diabetes risk development.

You can have diabetes for years and not know it. Symptoms like thirst, blurry eyesight, and tingling hands and feet may come on slowly without your noticing.

Middle age is when diabetes diagnoses really start to spike. An estimated 14% of Americans ages 45 to 64, or 11 million people, are diagnosed with type 2. That's almost five times the rate for those 18 to 44. Diabetes rates jump even higher at the onset of your senior years. Almost 25% of Americans 65 and older have been diagnosed with type 2. Undiagnosed cases may account for another 4.7%. That means more than 1 of every 4 oldest Americans lives with type 2 diabetes.

The disease also is affecting ever more teens and even children. In 2002, 8 out of 100,000 adolescents were diagnosed with type 2 between the ages of 10 to 14. A decade later, the rate was 50% higher, or 12 per 100,000 youths.

Researchers believe childhood obesity and lack of exercise are among the reasons behind that trend. Doctors now screen kids as young as 10 for diabetes if they're overweight or have other risk factors for the disease, Compared to those who were diagnosed later, research found that people who had type 2 before they turned 40 were more likely to have:

- Quicker damage to insulin-making cells called beta cells
- More complications, mainly because they live with the disease longer
- Shorter life spans

As you age, you're more likely to have multiple medical conditions, including high blood pressure and high cholesterol. That can make it harder for you to keep your diabetes under control.

In turn, diabetes can lead to other health problems such as heart disease.

Low blood sugar, called hypoglycemia, is more common in older adults with diabetes. Symptoms such as dizziness, confusion, and weakness might worsen as you age [8].

Pregnancies vs. Diabetes

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.

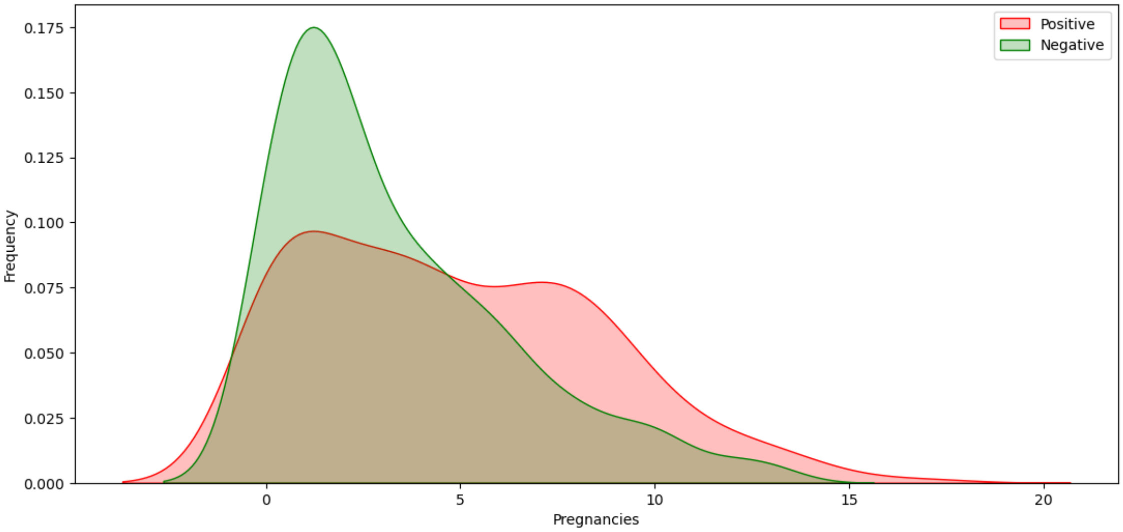

Figure 17: The number of pregnancies over 5 poses the higher risk of diabetes in women.

Pregnancy leads to substantial maternal metabolic and lifestyle alterations, including decreased insulin sensitivity, an accumulation of body fat, body fat central redistribution, decreased physical activity, and increased calorie intake, which are associated with higher risk of GDM. However, it is still unclear whether repeated exposure to these metabolic and lifestyle changes influences the development of GDM in future pregnancies. Various studies have investigated the association between parity and GDM, with conflicting results. However, none of these studies took abortion or stillbirth into consideration. Spontaneous miscarriage occurs in approximately 15% of clinically recognized pregnancies.

A total of 7,008 subjects from the Healthy Baby Cohort study were included in the study reported in [9]. The number of pregnancies was classified into three categories: 1, 2, or $>$ pregnancies. GDM was diagnosed using International Association of Diabetes and Pregnancy Study Groups criteria. Multivariate logistic regression models were used. In the fully adjusted model, women with $>$ 3 pregnancies had a 1.27-fold (95% confidence interval, 1.05–1.54) higher risk of GDM. Among women $>$ 30 years old, 2 and $>$ 3 pregnancies were associated with a higher risk of GDM 1.32; Among women with a pre-pregnancy BMI $<$ 24 kg/m2, $>$ 3 pregnancies were associated with a 1.35-fold (95% CI, 1.09–1.67) higher risk of GDM.

Deep Learning Analysis of Diabetes Data

Let us consider a neural network. Neural networks are a series of algorithms modeled after the human brain, designed to recognize patterns in data. They interpret sensory data through machine perception, labeling, and clustering raw input. These networks can learn and improve over time, making them essential for tasks like image and speech recognition. This learning path gives you a comprehensive introduction to the topic so you can springboard into more advanced applications for neural networks.

Neural networks are capable of learning and identifying patterns directly from data without pre-defined rules. These networks are built from several key components:

- Neurons: The basic units that receive inputs, each neuron is governed by a threshold and an activation function.

- Connections: Links between neurons that carry information, regulated by weights and biases.
- Weights and Biases: These parameters determine the strength and influence of connections.
- Propagation Functions: Mechanisms that help process and transfer data across layers of neurons.
- Learning Rule: The method that adjusts weights and biases over time to improve accuracy

An N-layer neural network can be presented as a nested function

$$y = f_N(f_{N-1}........f_1(x))$$

In this expression,

$$f_i(z) = \psi_i(z)(W_i z + b_i)$$

where 'i' is called the layer index and can span from 1 to any number of layers, the function ψ is called an activation function it is fixed and usually nonlinear chosen by the data explored before the learning. W is a matrix and b is the bias vector for each layer.

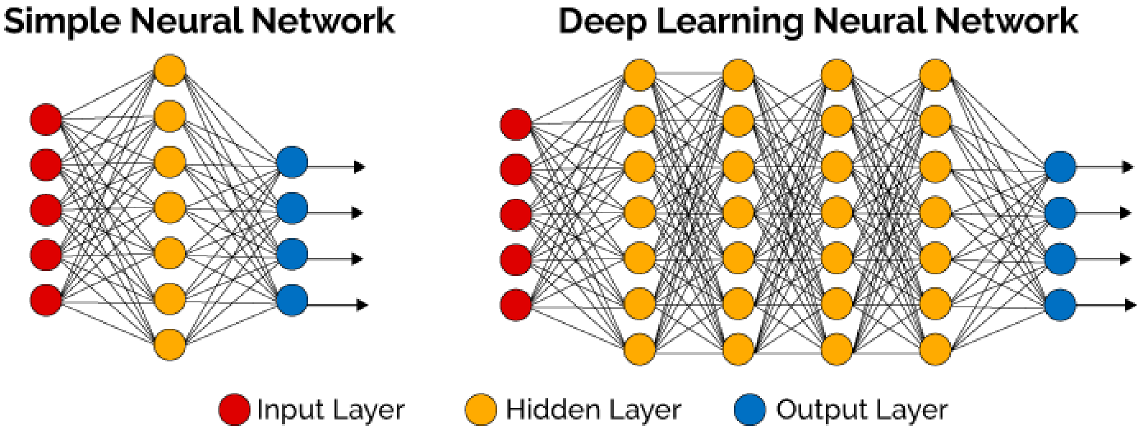

Figure 18: A simple and deep learning networks.

Deep learning, also known as hierarchical learning, is a subset of machine learning in artificial intelligence that can mimic the computing capabilities of the human brain and create patterns similar to those used by the brain for making decisions. 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 multiple nodes in each hidden layer is known as a deep learning system or a deep neural network. Deep learning is the development of deep learning algorithms that can be used to train and predict output from complex data. The word "deep" in Deep Learning refers to the number of hidden layers 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. A neural network is the underlying technology in deep learning. It consists of interconnected nodes or neurons in a layered structure. The nodes process data in a coordinated and adaptive system. They exchange feedback on generated output, learn from mistakes, and improve continuously. Thus, artificial neural networks are the core of a deep learning system. We often use simple neural networks for machine learning (ML) tasks due to their low-cost development and accessible computational demands. Organizations can internally develop applications that use simple neural networks. They're more feasible for smaller projects because they have limited computational requirements. If a company needs to visualize data or recognize patterns, neural networks provide a cost-effective way of creating these functions.

On the other hand, deep learning systems have a wide range of practical uses. Their ability to learn from data, extract patterns, and develop features allows them to offer state-of-the-art performance. For example, you can use deep learning models in natural language processing (NLP), autonomous driving, and speech recognition.

However, you need extensive resources and funding to train and self-develop a deep learning system. Instead, organizations prefer using pretrained deep learning systems as a fully managed service they can customize for their applications.

Classification vs Regression.

Classification is the process of finding or discovering a model or function that helps in separating 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 are supposed to predict discrete target variables(class labels) using independent features. In the classification task, we are supposed to find a decision boundary that can separate the different classes in the target variable. The derived mapping function could be demonstrated in the form of "IF-THEN" rules. The classification process deals with problems where the data can be divided into binary or multiple discrete labels. Let's take an example, suppose we want to predict the possibility of the winning of a match by Team A on the basis of some parameters recorded earlier. Then there would be two labels Yes and No.

Classification is the process of finding or discovering a model or function that helps in separating 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 are supposed to predict discrete target variables(class labels) using independent features. In the classification task, we are supposed to 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 [10]. The derived mapping function could be demonstrated in the form of "IF-THEN" rules. The classification process deals with problems where the data can be divided into binary or multiple discrete labels. Let's take an example, suppose we want to predict the possibility of the winning of a match by Team A on the basis of some parameters recorded earlier. Then there would be two labels Yes and No.

Regression is the process of 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 depending on the historical data. Because a regression predictive model predicts a quantity, therefore, the skill of the model must be reported as an error in those predictions.

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. One of the most used activation function is a ReLu function given by the following formula:

$$ReLU(x) = \begin{cases} 0, & \text{if } x \leq 0 \\ x, & \text{if } x > 0. \end{cases}$$

Let us consider the following algorithm. We will randomly split the data into batches of size M. For each batch, we will run the deep learning algorithm. Our algorithm, we call HDLA (Hybrid Deep Learning Algorithm) constructs many individual batches at training. Predictions from all batches are pooled to make the final prediction; the mode of the classes for classification or the mean prediction for regression. As we use a collection of results to make a final prediction, we refer the algorithm to as hybrid technique.

$$F_i(z) = \frac{1}{M} \sum_{j=1}^{M} \psi_i(z)(W_i^{(j)} z + b_i^{(j)})$$

where M is the number of batches.

Results

In a fully connected Deep neural network, there is an input layer and one or more hidden layers connected one after the other. Each neuron receives input from the previous layer neurons or the input layer. The output of one neuron becomes the input to other neurons in the next layer of the network, and this process continues until the final layer produces the output of the network. The layers of the neural network transform the input data through a series of nonlinear transformations, allowing the network to learn complex representations of the input data. The deep learning algorithm yields continuous numbers and an excellent tool for regression. To estimate accuracy, we create a table:

Number of samples randomly chosen	1	2	3	4	5	6	7	8	9	10
True values	1	0	0	1	0	1	0	1	1	0
Predicted values	.65	.14	0.12	0.45	0.27	0.78	0.36	0.71	0.37	0.08

Regression mode of the deep learning

In order to make a classification results, we can introduce an additional condition:

$$f(x) = \begin{cases} 1, & \text{if } f(x) \geq 0.5 \\ 0, & \text{if } f(x) < 0.5. \end{cases}$$

Number of samples randomly chosen	1	2	3	4	5	6	7	8	9	10
True values	0	0	0	1	0	0	0	1	0	0
Predicted values	0	1	0	1	0	0	0	1	1	0

The table above demonstrates the classification mode of the deep learning. The overall accuracy of the classification model of the deep learning was 76.77%

Number of samples randomly chosen	1	2	3	4	5	6	7	8	9	10	11	12	13
True values	0	0	0	1	0	0	0	1	0	0	1	0	0
Predicted values	0.24	0,16	0.24	0.78	0.16	0.33	0.14	0.66	0,23	0.15	0.57	0.44	0.27

The result of the HDLA approach in the regression mode. In order to run the HDLA in the classification mode, we use the following approach: To calculate the result, we randomly choose batches in the data sets and features and do voring for 1 or 1 depending on the intermediate results.

$$F_i(z) = \frac{1}{M} \sum_{k=1}^{Q} \sum_{j=1}^{M} \psi_i(z)(W_{i,k}^{(j)} z + b_{i,k}^{(j)}),$$

where summation is carried out over randomly taken features (Q stands for te number of features taken into the analysis). The accuracy of classification mode of the hybrid deep learning was largely increased. The overall accuracy of the classification model of the hybrid deep learning was 92.76%

Number of samples randomly chosen	1	2	3	4	5	6	7	8	9	10	11	12	13
True values	0	0	0	1	0	0	0	1	0	0	1	0	0
Predicted values	0	0	0	1	0	0	0	1	0	0	1	0	1

Conclusion

The large availability of biomedical data brings tremendous opportunities and challenges to health care research. In particular, exploring the associations among all the different pieces of information in these data sets is a fundamental problem to develop reliable medical tools based on data-driven approaches and machine learning. To this aim, previous works tried to link multiple data sources to build joint knowledge bases that could be used for predictive analysis and discovery. Although existing models demonstrate great promises, predictive tools based on machine learning techniques have not been widely applied in medicine. [11] showed that the use of optimized algorithms (SVM) can ba indispensable tools in accurate diagnostics of various diseases.

However, deep learning approaches have not been extensively evaluated for a broad range of health care and medical problems that could benefit from its capabilities. There are many aspects of deep learning that could be helpful in health care, such as its superior performance, end-to-end learning scheme with integrated feature learning, capability of handling complex and multi-modality data and so on. To accelerate these efforts, the deep learning research field as a whole must address several challenges relating to the characteristics of health care data (i.e. sparse, noisy, heterogeneous, time-dependent) as need for improved methods and tools that enable deep learning to interface with health care information workflows and clinical decision support.

In this work, we used data set from kaggle.com. This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

At the beginning, we completed data preprocessing including feature engineering. We have selected 4 features and performed data analysis (deep learning) on bot 4 and 8 features. We achieved about 77% accuracy. However, the hybrid algorithm demonstrated in this paper, can lift the accuracy to about 92%.

References

1.Barnett AH, Eff C, Leslie RD, Pyke DA. Diabetes in identical twins. A study of 200 pairs. Diabetologia. 1981;20(2):87-93. doi:10.1007/BF00262007 [PubMed] [CrossRef]

2.Redondo MJ, Jeffrey J, Fain PR, Eisenbarth GS, Orban T. Concordance for islet autoimmunity among monozygotic twins. N Engl J Med. 2008;359(26):2849-2850. doi:10.1056/NEJMc0805398 [PubMed] [CrossRef]

3.Baghbanian A, Tol A. The introduction of self-management in Type 2 Diabetes care: a narrative review. J Educ Health Promot. 2012;1:35. doi: 10.4103/2277-9531.102048. [DOI] [PMC free article] [PubMed] [Google Scholar]

4.Cho NH, Whiting D, Guariguata L, Montoya PA, Forouhi N, Hambleton I, et al. IDF Diabetes Atlas. 6th edn. International Diabetes Federation; 2013.

5.Adaji A, Schattner P, Jones K. The use of information technology to enhance diabetes management in primary care: a literature review. Inform Primary Care. 2008;16:229–37. doi: 10.14236/jhi.v16i3.698. [DOI] [PubMed] [Google Scholar]

6. H Riazi, B Larijani, M Langarizadeh,, L Shahmoradi, Managing diabetes mellitus using information technology: a systematic review, J Diabetes Metab Disord., 2015 Jun 3;14:49. doi: 10.1186/s40200-015-0174-x

7. Khaled Ibrahim Funjan, 2020, Skin Thickness can Predict the Progress of Diabetes Type 2: A New Medical Hypothesis.

8. J. Helmer, 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, Cao Z, Wang Y. Higher Numbers of Pregnancies Associated With an Increased Prevalence of Gestational Diabetes Mellitus: Results From the Healthy Baby Cohort Study. J Epidemiol. 2020 May 5;30(5):208-212. doi: 10.2188/jea.JE20180245. Epub 2019 Apr 20. PMID: 31006717; PMCID: PMC7153959.

10. P. de Melo, Public Health Informatics and Technology, Washington, DC, 2024, ISBN 979-8894090962

11. P. de Melo and M. Davtyan, High Accuracy Classification of Populations with Breast Cancer: SVM Approach, Cancer Research Journal, Cancer Research Journal DOI:10.11648/j.crj.20231103.13