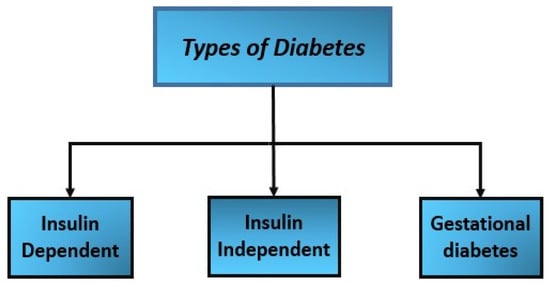
AI Based diabetes prediction system:

Introduction:

Diabetes is a chronic condition that affects the body’s ability to control blood sugar levels. One of the main symptoms of diabetes is high blood glucose levels. The body regulates blood glucose levels through the hormones insulin and glucagon. Usually, proper hormone secretion maintains normal blood sugar levels between 70 and 180 mg/dL. If left untreated or unmanaged, diabetes can lead to long-term complications, such as damage to large and small blood vessels, increasing the risk for cardiovascular disease and kidney, eye, limb, and neurological problems. According to the International Diabetes Federation, there are currently 387 million people living with diabetes worldwide, which is projected to double by 2035. Predicting diabetes early is often a difficult task for medical practitioners. [**Figure 1**](https://www.mdpi.com/2076-3417/13/5/3030#fig_body_display_applsci-13-03030-f001) depicts the three different forms of diabetes that exist. They are as follows:



Several different computational intelligence techniques can be used to predict diabetes complications. The feasibility of these methods can vary depending on the specific application, available data, and computational resources. Here are some of the most commonly used methods and their feasibility for this task:

**Logistic regression (LR):** LR is a classification algorithm used to predict the probability of a binary outcome (e.g., yes or no). It works by fitting a regression model to the data and then applying a sigmoid function to the output to convert it to a probability.

**K-nearest neighbors (KNN):** KNN is a simple classification algorithm that finds the k nearest neighbours of a new data point and assigns it to the most common class among its neighbours.

**Classification and regression trees (CART):** CART is a decision tree algorithm that can be used for both classification and regression tasks. It works by recursively splitting the data into smaller subsets, based on the values of different features, until a stopping criterion is met.

**Random forest (RF):** RF is an ensemble method that combines multiple decision trees to improve the performance and reduce the overfitting of individual trees. It works by creating multiple random samples of the training data and training a decision tree on each sample, then aggregating the results to make predictions.

**Support vector machines (SVM):** SVM is a robust classification algorithm that finds the hyperplane that best separates the classes in the data. Using different kernel functions, it can handle linear and non-linearly separable data.

**XGBoost:** XGBoost is an optimised gradient-boosting algorithm for classification and regression tasks. It works by iteratively adding weak learners to the model and adjusting the weights of misclassified samples to improve overall performance.

**LightGBM:** LightGBM is another optimised gradient boosting algorithm designed to be fast and efficient. It uses a histogram-based algorithm to split the data and reduce memory usage during training.

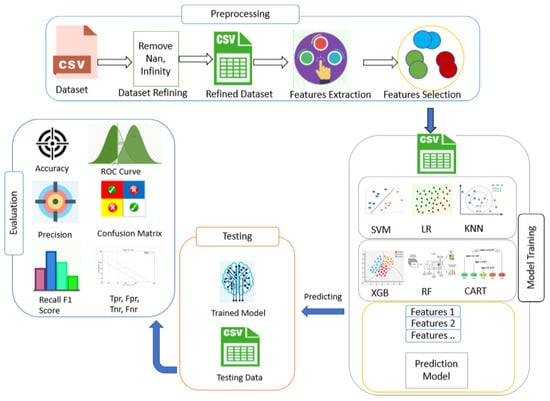
The choice of computational intelligence technique for predicting diabetes complications depends on the specific requirements of the application, available data, and computational resources. While some methods are more interpretable and easier to implement, others can handle more complex relationships in the data and require more computational resources. It is essential to evaluate the performance of different methods on independent datasets and choose the one that best meets the specific application’s needs.

*Problem Statement:*

“The rapid increase in the number of diabetes cases is becoming a global health concern. Early detection and prevention of diabetes are essential to mitigate its impact on individuals and society. However, existing methods for predicting diabetes are not always reliable and may fail to identify high-risk individuals. This paper aims to design a diabetes prediction framework that incorporates computational intelligence techniques to enhance the precision of diabetes predictions and aid in early disease detection and prevention”.

*Conceptual Description of the Solution:*

This framework aims to develop machine learning paradigms for diabetes prediction and store their results. The technology is used to administer and predict diabetes in patients and perform automated examinations. The data collection process to formulate results is explained step by step, including preprocessing, model training, testing, and evaluations in terms of accuracy, training, and confusion matrix. The affected people were diagnosed with neuromorphic using the random forest approach and evolving alerts.



*Design of the Solution:*

Designing a machine learning solution for diabetic prediction involves several steps:

**Data collection:** The initial step is to gather a large dataset that includes patient demographics, medical history, lab results, and other related information. This data can be obtained from electronic health records (EHRs) or other sources.

**Data preprocessing:** The collected data must be cleaned and preprocessed to eliminate any inconsistencies or missing values. The data must also be transformed and scaled to ensure machine learning algorithms can use it.

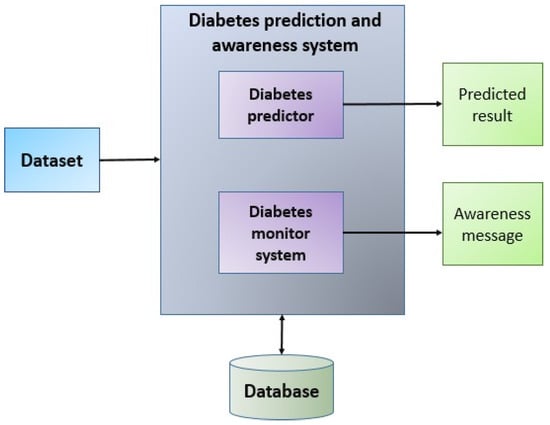
**Feature selection:** The next step is identifying the relevant features that will be used to train the machine learning model. This can include demographic information, lab results, and other related data.

**Model selection:** After the features are selected, the next step is to select the appropriate machine learning model for the task. The model type will depend on the nature of the data and the desired outcome. For example, logistic regression or a decision tree model could be used for classification tasks.

**Model training:** Once the model is selected, it needs to be trained on the collected and preprocessed data. The model will learn from the data and will be able to make predictions on new, unseen data.

**Model evaluation:** After the model is trained, it needs to be evaluated using various metrics, such as accuracy, precision, and recall. The model’s performance can be optimised by fine-tuning the hyperparameters and adjusting the features, if necessary.

**Model deployment:** Once the model is trained and optimised, it can be deployed in a production environment. The model can be integrated with existing systems, such as EHRs, to make predictions about new patients and help with the early detection of diabetes.



1. The system first receives the diabetes data set as input.
2. Based on the given symptoms, the diabetes predictor assists by predicting the presence of diabetes and generates the predicted results.
3. The diabetes monitor device assists in checking blood sugar levels and sends out alerts based on them.
4. The user receives the awareness message to know about their health status.

The use of computational intelligence techniques for the prediction of diabetes complications offers several advantages, including:

**Early detection:** Computational intelligence techniques can identify patterns in the data that may not be apparent to human experts, allowing for the earlier detection of diabetes complications. Early detection can lead to more timely interventions and potentially prevent or delay the onset of complications.

**Personalised treatment:** Predictive models can identify patients at higher risk of developing diabetes complications and tailor their treatment plans accordingly. This can lead to more personalised and effective care.

**Improved patient outcomes:** By identifying patients at a higher risk of developing diabetes complications and providing early interventions, computational intelligence techniques can improve patient outcomes and reduce the overall burden of diabetes-related complications.

**Efficient resource allocation:** By predicting which patients are at a higher risk of developing complications, healthcare resources can be allocated more efficiently, and patients can be prioritised for more intensive care.

**Improved efficiency:** Computational intelligence techniques can process vast amounts of data quickly and efficiently, leading to improved healthcare delivery efficiency.

**Reduced costs:** Early detection and timely interventions can potentially reduce the costs associated with diabetes complications by preventing or delaying the need for more intensive and costly treatments.

**Continuous monitoring:** Wearable devices and other technologies can monitor patients’ health status, allowing for more proactive and personalised care.

Overall, the use of computational intelligence techniques for the prediction of diabetes complications has the potential to improve patient outcomes, reduce costs, and increase the efficiency of healthcare delivery.

The proposed system utilises a trained dataset for predicting the likelihood of an individual developing diabetes. The diabetes prediction and awareness system generates predictions and sends out health-related alerts to users.

The diabetic predictor uses the information in [**Table 1**](https://www.mdpi.com/2076-3417/13/5/3030#table_body_display_applsci-13-03030-t001) to make its computations and provide its results.

**Table 1.** Blood glucose level chart.



*Validation Prototype:*

The proposed system utilises a dataset to predict the likelihood of an individual developing diabetes. The system utilises the Iterative Dichotomiser 3 algorithm to create decision trees and to aid in monitoring the patient’s health by providing results from various checkups for fasting and postprandial blood sugar levels. The system generates awareness messages based on the patient’s blood sugar levels, as shown in [**Table 1**](https://www.mdpi.com/2076-3417/13/5/3030#table_body_display_applsci-13-03030-t001). This table displays the normal blood sugar levels and the type of diabetes associated with them.

Incomplete or missing data is a common issue in medical datasets and can cause problems for machine learning models, leading to biased or inaccurate results [[**27**](https://www.mdpi.com/2076-3417/13/5/3030#B27-applsci-13-03030)]. Handling missing values in medical datasets before training a model is an important step in the data preprocessing stage. There are several approaches to handling missing values in medical datasets, including:

**Imputation:** This involves replacing missing values with a substituted value, such as the mean or median value of the corresponding feature, or using more complex techniques, like regression models.

**Deletion:** This involves removing the samples or features with missing values. However, this approach can result in a significant loss of data.

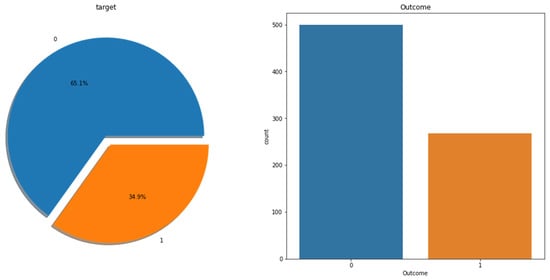
**Using specialised algorithms:** Some machine learning algorithms, such as decision trees or random forests, can handle missing values directly during the training process.

It is important to choose an appropriate strategy based on the specific characteristics of the dataset and the requirements of the problem at hand. Proper handling of missing values can help to improve the accuracy and reliability of the machine learning model and reduce the potential for biases or inaccuracies in the results.

## Results and Performance Evaluation:

This dataset contains information on 768 individuals with reports of symptoms associated with diabetes. The data was collected through a questionnaire given to individuals who had recently been diagnosed with diabetes or had symptoms but were not yet diagnosed. Missing values were handled by ignoring incomplete data, resulting in a dataset of 500 cases. Within this dataset, there were 186 negative cases (indicating no diabetes diagnosis) and 314 positive cases (indicating a diabetes diagnosis).

The pie chart in [**Figure 4**](https://www.mdpi.com/2076-3417/13/5/3030#fig_body_display_applsci-13-03030-f004) illustrates the distribution of the diabetic and non-diabetic populations in the dataset. The blue portion of the chart represents the non-diabetic population, with a value of 0, and the orange portion represents the diabetic population, with a value of 1. The bar chart in [**Figure 4**](https://www.mdpi.com/2076-3417/13/5/3030#fig_body_display_applsci-13-03030-f004) illustrates how the diabetic and non-diabetic population is affected by different parameters, as shown in [**Table 2**](https://www.mdpi.com/2076-3417/13/5/3030#table_body_display_applsci-13-03030-t002), which lists the dataset’s attributes.



**Figure 4.** Correlation matrix graph of the data set.

**Table 2.** Dataset description.



[**Table 3**](https://www.mdpi.com/2076-3417/13/5/3030#table_body_display_applsci-13-03030-t003) shows a detailed description of the dataset and its properties.

**Table 3.** Attributes and their feature types.

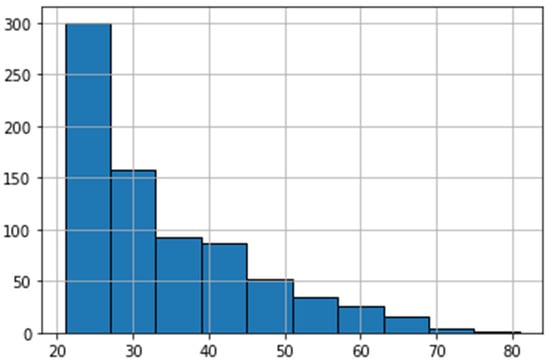


The sample dataset with values is shown in [**Table 4**](https://www.mdpi.com/2076-3417/13/5/3030#table_body_display_applsci-13-03030-t004). The output class variable (0 or 1) shows the prediction of the diabetic.

**Table 4.** Dataset values.

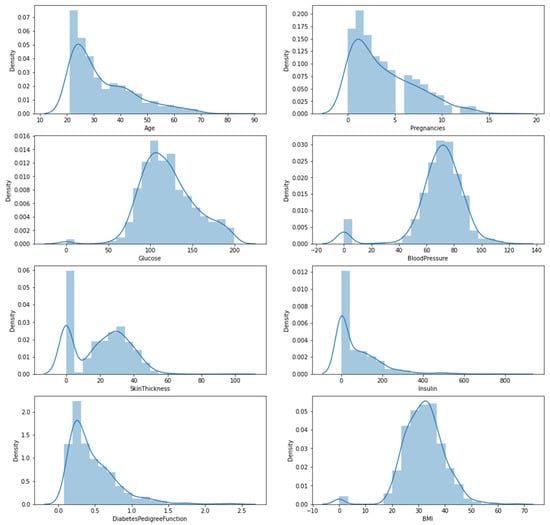


[**Figure 5**](https://www.mdpi.com/2076-3417/13/5/3030#fig_body_display_applsci-13-03030-f005) shows the age variables of the sample size dataset. Total patient records, age-wise, reflect the number of patients with age-wise segmentation.



**Figure 5.** Representation of age-wise data.

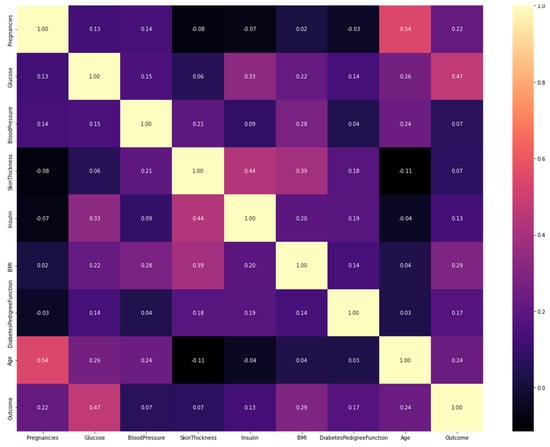
The distribution of the outcome variable in the data was examined and visualised in [**Figure 6**](https://www.mdpi.com/2076-3417/13/5/3030#fig_body_display_applsci-13-03030-f006). All variables are plotted with density.



**Figure 6.** Histogram and density graphs of all variables.

The train-test split is a technique used to evaluate the performance of a machine learning model. It is used for classification, regression, and supervised learning strategy. This technique divides a dataset into two subsets: a training set and a test set. The training set is used to fit the model, while the test set is used to evaluate the model’s performance. The test set is fed the model’s input, and the predicted output is compared to the actual output. This technique allows for assessing the model’s ability to generalise to new, unseen data. In this case, the training dataset accounts for 80% of the data and the test dataset accounts for 20% of the data.

Research often uses a number of parameters to evaluate the performance of diabetes prediction algorithms, one of which is the confusion matrix. A confusion matrix is a tool to visualise a classification algorithm’s performance. It displays the number of true positive, true negative, false positive, and false negative predictions. [**Figure 7**](https://www.mdpi.com/2076-3417/13/5/3030#fig_body_display_applsci-13-03030-f007) shows the confusion matrix of our classifier, which is used to evaluate different performance measures. This technique has been used in our study’s comparative analysis of multiple algorithms.



**Figure 7.** Confusion matrix.

In Outlier Observation Analysis, we specified several conditions to determine the most accurate results. A collection of data preparation methods was used to improve the classifier’s performance. To reduce bias towards normal examples, we oversampled certain data instances, regardless of class, to expose the specific attributes hidden within outlier occurrences. A function was used to calculate the risk of developing diabetes based on family history; the greater the function, the greater the risk of diabetes. The outcome of whether the individual had diabetes (1 = yes, 0 = no) was also considered.

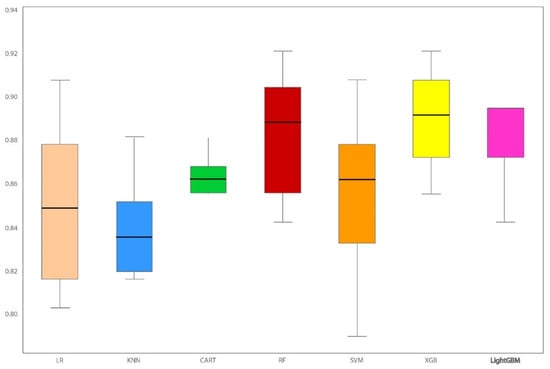
Using the data collected, prediction models for the risk of developing diabetes were created to potentially prevent the disease in the future. The accuracy of these models is presented in [**Table 5**](https://www.mdpi.com/2076-3417/13/5/3030#table_body_display_applsci-13-03030-t005), where it can be seen that the XGBoost classifier had the highest accuracy of 89%. This indicates that the XGBoost classifier can correctly predict diabetes risk 89% of the time when using the input data.

**Table 5.** Performance comparison of the generated prediction models.



The experimental results presented in [**Table 5**](https://www.mdpi.com/2076-3417/13/5/3030#table_body_display_applsci-13-03030-t005) indicate little difference in performance between the single models used (LR, KNN, CART, RF, SVM, XGBoost, and LightGBM methods). On the test dataset, the XGBoost model achieved 89% accuracy in predicting the occurrence of diabetes. In comparison, the LightGBM model achieved 88% accuracy, considered a standard statistical analysis method at the time.

This study used a large dataset and ensemble machine learning approaches to create the prediction models, which is different from previous works. A data-driven feature selection method was used to develop predictors that effectively identified the different classes in the dataset. The statistics for the different classifiers for the accuracy of the proposed system showed the highest values for the dataset, as shown in [**Figure 8**](https://www.mdpi.com/2076-3417/13/5/3030#fig_body_display_applsci-13-03030-f008). Furthermore, by adjusting the number of iterations used to train the models, the study also demonstrated the effect of accumulated medical data on prediction accuracy. This approach aims to improve the accuracy of diabetes prediction and early identification.



**Figure 8.** Statistical analysis of seven classifiers with accuracy.

The trade-off between using ensemble techniques, like XGBoost, and single models, like logistic regression, KNN, or SVM, is valid. While ensemble techniques often achieve better performance, single models can be more easily interpretable and have lower complexity. This is an important consideration for applications in which model interpretability is a priority, such as in medical or healthcare settings where the decisions made by the model may have a significant impact on patient outcomes. Regarding unsupervised learning methods, hierarchical clustering is a valid suggestion for clustering categorical data. Hierarchical clustering is a bottom-up approach that builds a hierarchy of clusters, with the option to stop at a particular level of granularity to obtain a desired number of clusters. This method can be particularly useful when the number of clusters is unknown or when the data is challenging to interpret. The paper “Estimating the Optimal Number of Clusters in Categorical Data Clustering by Silhouette Coefficient” is an excellent reference for the author to discuss in the context of unsupervised learning methods. The paper proposes a method for determining the optimal number of clusters in categorical data clustering based on the silhouette coefficient, which measures the quality of the clustering solution [[**28**](https://www.mdpi.com/2076-3417/13/5/3030#B28-applsci-13-03030)].

## 5. Comparison with Other Techniques

The paper “Prediction of Diabetes Complications using Computational Intelligence Techniques” uses several machine learning techniques to predict the likelihood of diabetes complications. Overall, the paper’s approach of using ANNs, SVMs, and DTs for diabetes complication prediction is consistent with previous studies. The authors showed that the ANN model outperforms the other models in terms of accuracy and sensitivity, consistent with previous studies showing that ANNs are an effective technique for diabetes complication prediction. However, the choice of machine learning technique depends on several factors, such as the size and complexity of the dataset, the type of outcome variable, and the need for interpretability. Therefore, it is important to choose the appropriate technique based on the specific requirements of the study.

## Conclusions:

This study aimed to build a prediction model for diabetes complications using a classification data mining approach. The classification technique’s effectiveness in constructing the best rule-based model for the prediction goal was evaluated. The process of discovering valuable and previously unknown information from large databases is known as data mining. This paper provided a comprehensive classification of the commonly used diabetes prediction techniques based on a literature review on data mining-based diabetes diagnosis, classification, and prediction techniques. Additionally, the paper proposed a Disease Influence Measure based method to improve performance, resulting in an accuracy of 89%.