**Abstract**

Many factors can cause a person to get affected by diabetes, like excessive body weight, abnormal cholesterol level, family history, physical inactivity, bad food habit etc. People with diabetes for a long time can get several complications like heart disorder, kidney disease, nerve damage, diabetic retinopathy etc. But its risk can be reduced if it is predicted early. We are going to use machine learning classification methods, that is, decision tree, SVM, Random Forest, Logistic Regression, KNN, and various ensemble techniques, to determine which algorithm produces the best prediction results. The explainable AI approach with LIME and SHAP frameworks is implemented to understand how the model predicts the final results.

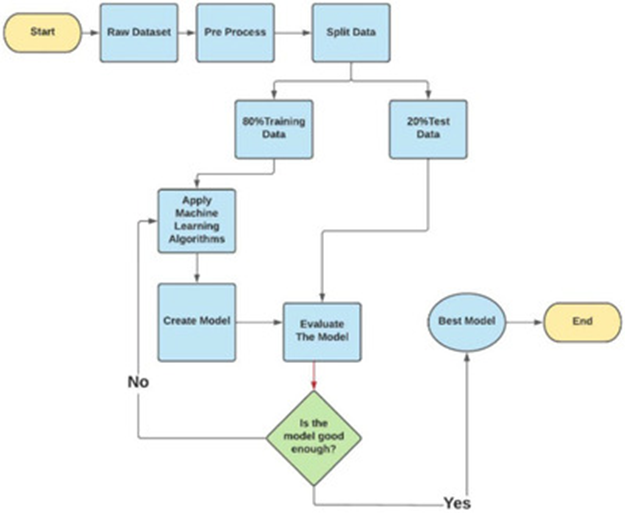
**Introduction**

In this paper, we have employed machine learning and explainable AI techniques to detect diabetes. We have used the holdout validation technique to split the data. In this research paper, we have applied various machine learning‐based classification algorithms, that is, decision tree, logistic regression, KNN, random forest, SVM, and ensemble techniques. Next, the performance of these classifiers has been evaluated in terms of precision, recall, and F1 measure. Finally, the best classifier has been selected as the final model to deploy into an Android smartphone application.

**Proposed system**

This section describes the working procedures and implementation of various machine learning techniques to design the proposed automatic diabetes prediction system. First, the dataset was collected and preprocessed to remove the necessary discrepancies from the dataset. Then the dataset was separated into the training set and test set using the holdout validation technique. Next, different classification algorithms were applied to find the best classification algorithm for this dataset. Finally, the best‐performed prediction model is deployed into the proposed website and smartphone application framework.

**Flow Chart of proposed system**



**Data Set**

The Pima Indian dataset is an open‐source dataset is publicly available for machine learning classification, which has been used in this work along with a private dataset. It contains 768 patients’ data, and 268 of them have developed diabetes. Figure [2](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10107388/figure/htl212039-fig-0002/) shows the ratio of people having diabetes in the Pima Indian dataset. Table [1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10107388/table/htl212039-tbl-0001/) demonstrates the eight features of the open‐source Piman Indian dataset.

[An external file that holds a picture, illustration, etc.
Object name is HTL2-10-1-g012.jpg](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10107388/figure/htl212039-fig-0002/)

TABLE 1

Features of the Pima Indian Dataset

| **Pregnancies** | **Skin thickness** | **Diabetes pedigree function** |
| --- | --- | --- |
| Glucose | Insulin | Age |
| Blood pressure | BMI |  |

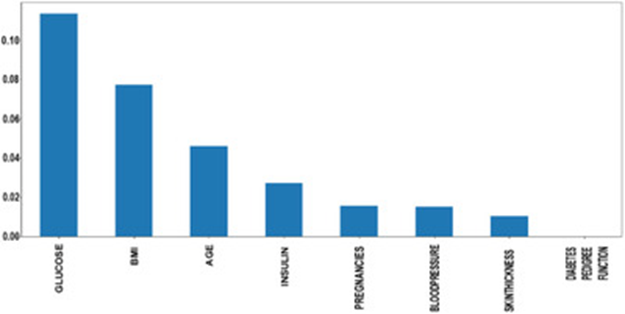
TABLE 2

Features of the RTML private dataset

| **Features** | **Minimum** | **Maximum** | **Average** |
| --- | --- | --- | --- |
| Pregnancies | 0 | 8 | 1.61 |
| Glucose (mg/dL) | 52.2 | 274 | 109.39 |
| Blood pressure (mm Hg) | 5.9 | 115 | 71.09 |
| Skin thickness (mm) | 2.9 | 23.3 | 10.78 |
| BMI (kg/m2) | 2.61 | 41.62 | 22.69 |
| Age (years) | 17 | 77 | 27.02 |

**Dataset preprocessing**

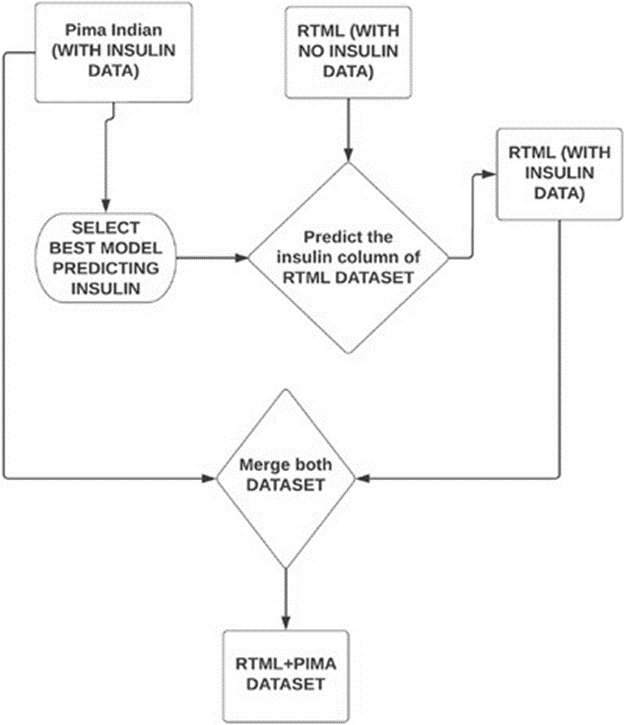
In the merged dataset, we discovered a few exceptional zero values. For example, skin thickness and Body Mass Index (BMI) cannot be zero. The zero value has been replaced by its corresponding mean value. The training and test dataset has been separated using the holdout validation technique, where 80% is the training data and 20% is the test data.



we computed the root mean square error (RMSE) of various regression frameworks as

RMSE=∑Ni=1(Predictedi−Actuali)2N−−−−−−−−−−−−−−−−−−−−−−−−√

**Working steps insulin of the RTML dataset**

****

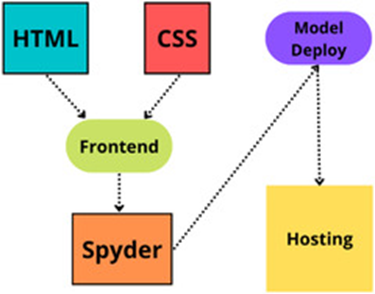
The data has been scaled to the same range using the following equation:

**Xscaled=X−XminXmax−Xmin**

. **Deployment of the prediction system**

The proposed machine learning‐based diabetes prediction system has been deployed into a website and smartphone application framework to work instantaneously on real data.

**Web application:** We have used HTML and CSS for the frontend part of the proposed website. After that, we finalized the machine learning model XGBoost with ADASYN, as it provided the best performance. The model deployment has been done with Spyder, a Python environment platform that works with Anaconda.



**Android smartphone application:** To demonstrate the automatic diabetes forecasting system in real time, we also designed an Android smartphone application to test its performance. Android Studio is used for the frontend part of this application. We employed Java as the necessary coding language. After that, the model has been implemented in Android Studio using the pickle package. While developing the API, we used Heroku to host our model on the corresponding hosting server.

**RESULT**

In this paper, an automatic diabetes prediction system using various machine learning approaches has been proposed.