GROUP -3

MEMBERS

1. R. LOSHINI
2. M.SHAKTHI
3. V.PRIYADHARSHINI
4. R.NITHYA
5. R.PAVITHRA
6. S.MARIYAMMAL

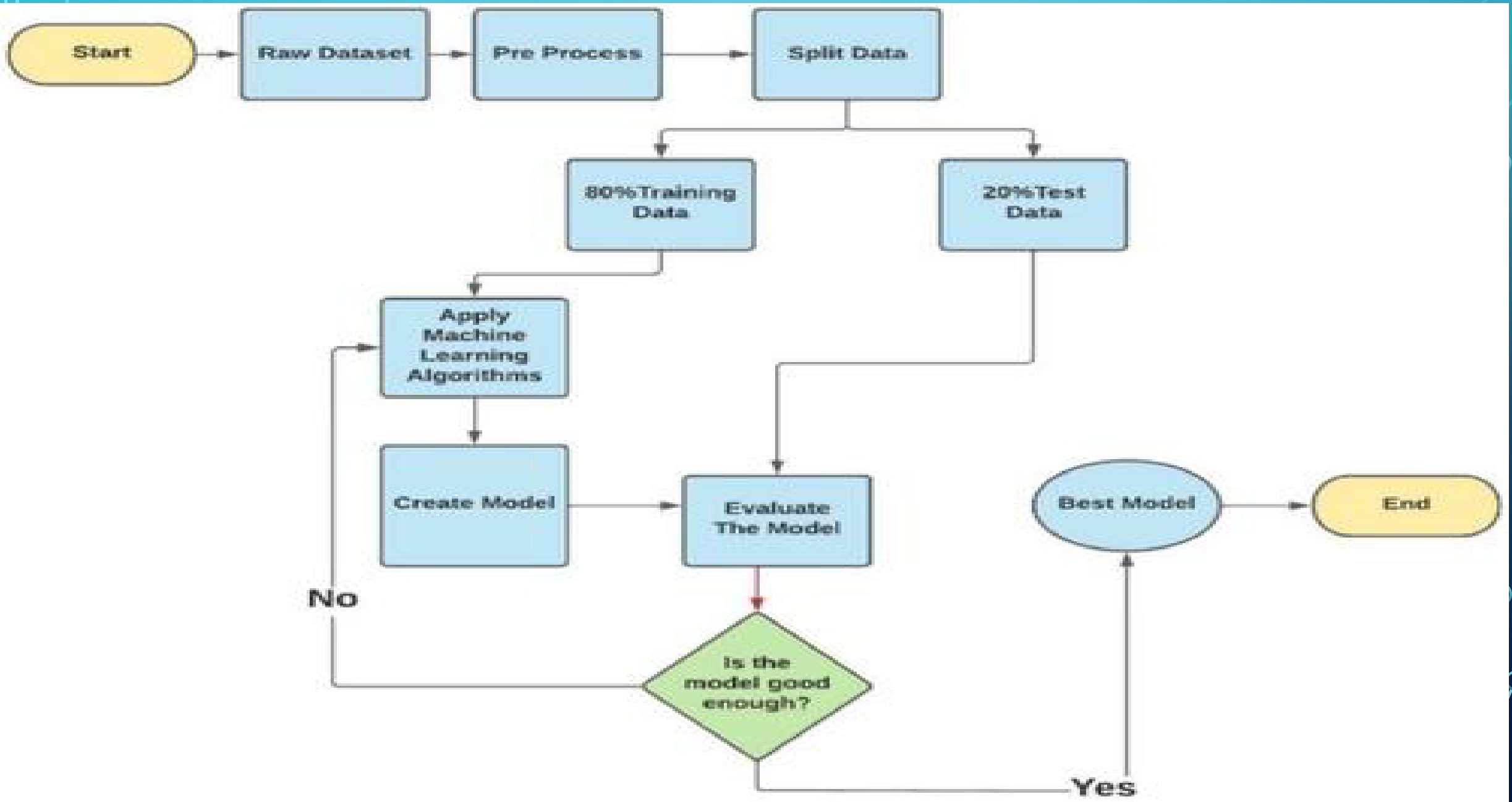
AI-BASED DIABETES PREDICTIONS SYSTEM

INTRODUCTION

- Diabetes is a chronic disease that directly affects the pancreas, and the body is incapable of producing insulin . Insulin is mainly responsible for maintaining the blood glucose level. Many factors, such as excessive body weight, physical inactivity, high blood pressure, and abnormal cholesterol level, can cause a person get affected by diabetes. It can cause many complications, but an increase in urination is one of the most common ones . It can damage the skin, nerves, and eyes, and if not treated early, diabetes can cause kidney failure and diabetic retinopathy ocular disease

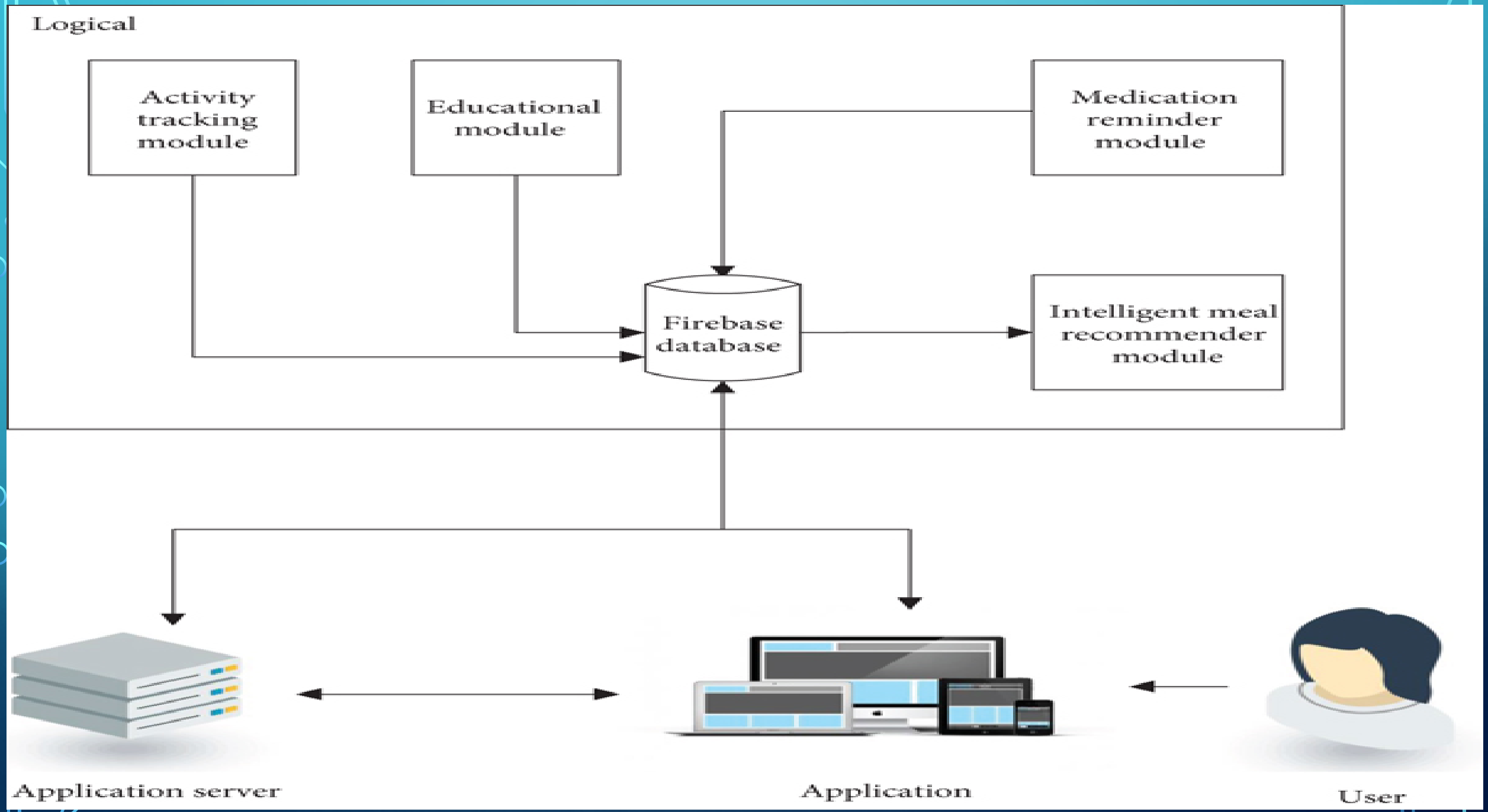
PROPOSED SYSTEM

- the working procedures and implementation of various machine learning techniques to design the proposed automatic diabetes prediction system. Figure 1 shows the different stages of this research work. First, the dataset was collected and preprocessed to remove the necessary discrepancies from the dataset, for example, replacing null instances with mean values, dealing with imbalanced class issues etc. 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.


Start
Raw Dataset
Pre Process
Split Data
80%Training Data
20%Test Data
Apply Machine Learning Algorithms
Create Model
Evaluate The Model
Best Model
End
No
Is the model good enough?
Yes

SYSTEM DESIGN AND DEVELOPMENT

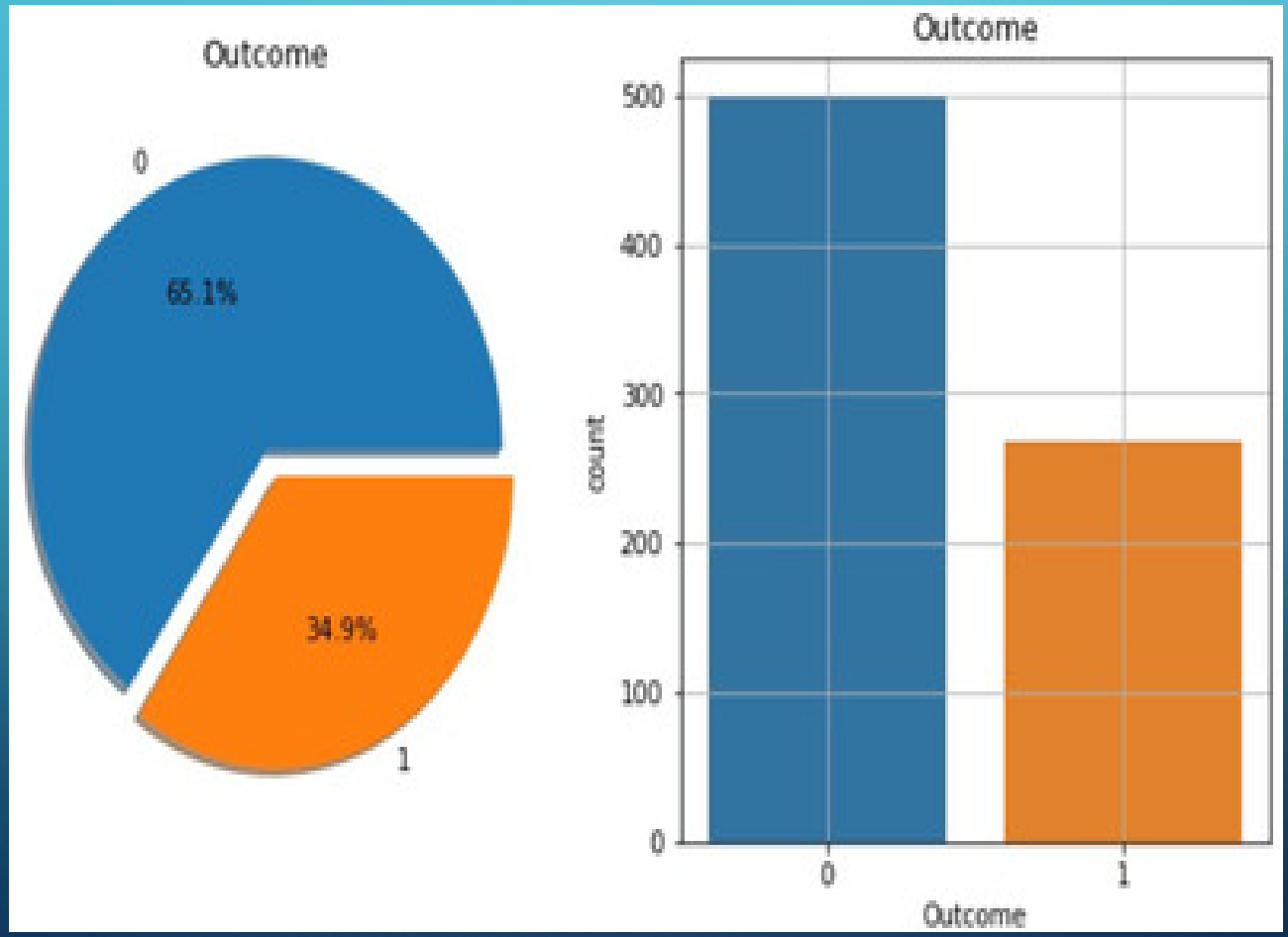
- The system architecture for the Diabetes Management System presented below in Figure is the conceptual model that defines the structure, behavioural interactions, and multiple system views that underpins the system development. It presents the formal descriptions of the systems captured graphically that supports reasoning, and the submodules developed as well as the dataflows between the developed modules.


Logical
Activity tracking module
Educational module
Medication reminder module
Firebase database
Intelligent meal recommender module
Application server
Application
User

DATA COLLECTION

- The used data in this research consists of two data types, the patient data obtained from an interface provided to the user to input personal details like age, sex, weight, height, and level of activity. The food nutrition data was obtained from the Department of Nutrition and Food Science, University of Ghana, and from the My Fitness database . The diet type of the patient is determined from the obtained data, and calorie needs calculated using the Harris-Benedict's equation .

PERCENTAGE OF PEOPLE HAVING DIABETES IN THE INDIAN DATASET



FEATURES OF PRIVATE DATASET

- Table 1:

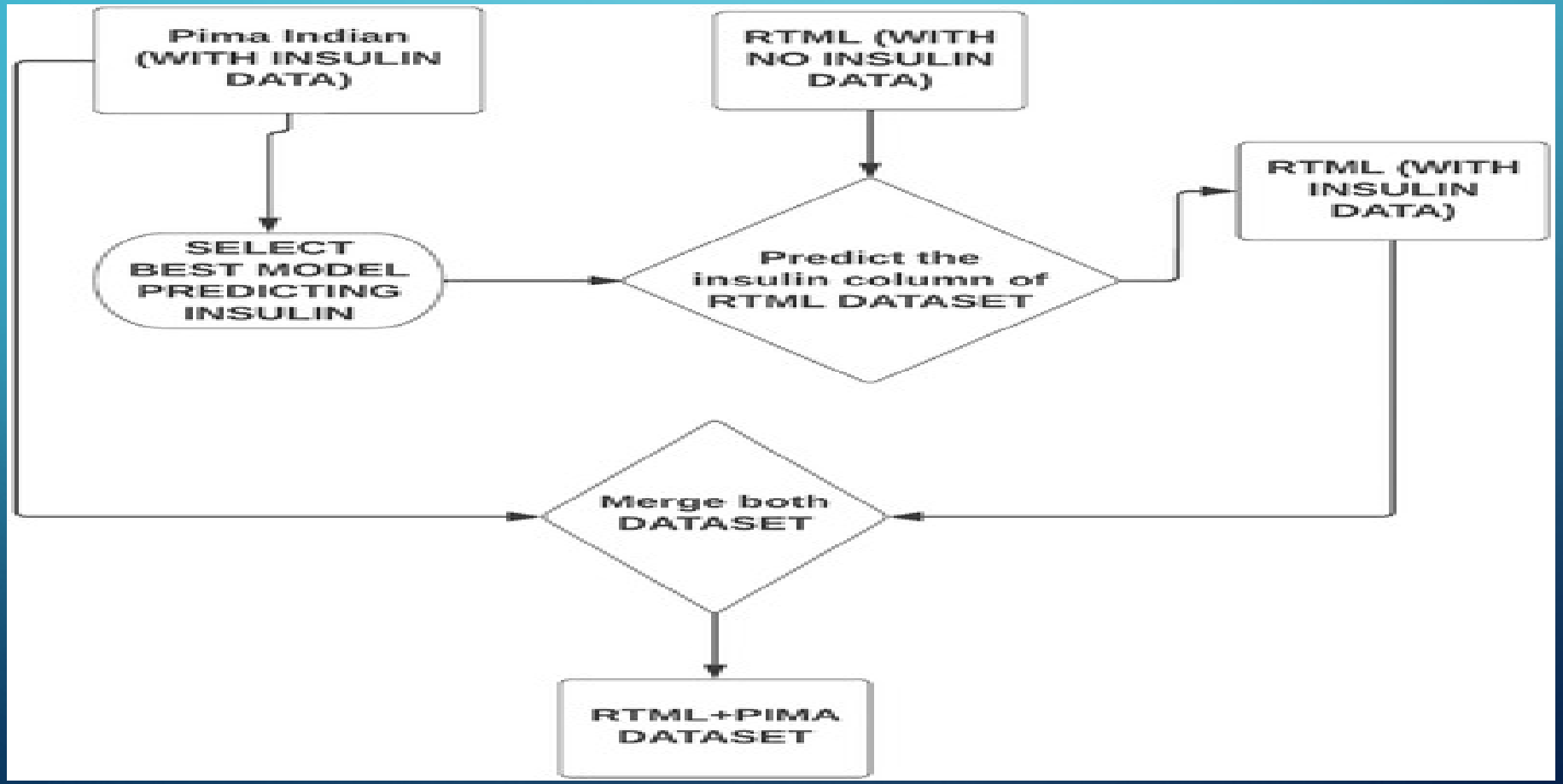
Features	Minimum	Maximum	Average
Pregnancies	0	8	1.61
Glucose (mg/dL)	52.2	274	109.39
Blood pressure (mm Hg)	5.9	115	71.09
Skin thickness (mm)	2.9	23.3	10.78
BMI (kg/m^2)	2.61	41.62	22.69
Age (years)	17	77	27.02

DATASET PROCESSING

- 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.

Regression model	RMSE
XGB	0.36
SVR	0.45
GPR	0.43

- # WORKING STEPS OF PREDICTING INSULIN OF THE DATASET



MODEL SELECTION

- Logistic regression:
- Logistic Regression (LR) is a subset of generalized linear models which deals with the analysis of binary data, which seeks out the best-fitting model for describing the connection between dependent and independent predictors .When it comes to predicting sickness or health status, the LR model is most commonly used .Based on the risk factors given, the LR model can calculate the likelihood of an individual acquiring diabetes disease .

- Extra Gradient boosting
- The Extreme Gradient Boosting is an improved supervised algorithm proposed by Chen and Guestrin based on the Gradient Boosting Decision Tree algorithm XGBoost can be used to solve problems for regression and classification, which has been chosen to be used by data scientists because of its high execution speed and the high accuracy that it supplies . The XGBoost objective function includes its loss function and regularization term, which can help to prevent overfitting by smoothing the final learned weights to obtain an optimal solution . The loss function controls the ability of the prediction, which determines the deviation between predicted label and the actual label . The regularization term controls the complexity of the model and it can also handle the overfitting issue . XGBoost can also optimize the loss function using first-order and second-order gradient statistics.

Random forest:

RF is one of the most common uses of classifier integration.

The steps are as follows

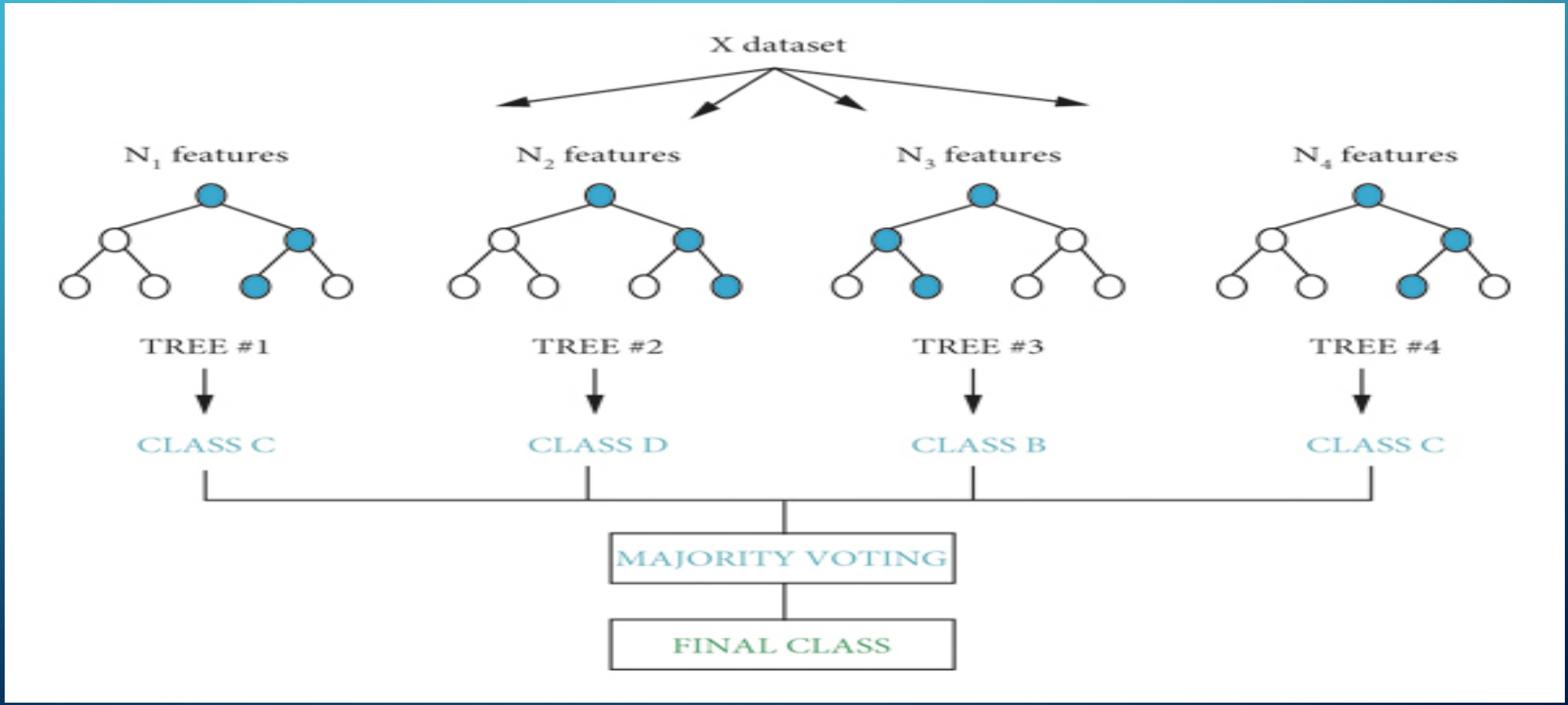
(i)Extracting some samples from the training set as a training subset using the bootstrap method, which is a self-help sampling approach.

(ii)A number of features are randomly picked from the feature set for the training subset as the basis for splitting each node of the Decision Tree.

(iii)Repeat steps (i)-(ii) to generate a large number of training subsets and Decision Trees, which are then combined to build a Random Forest.

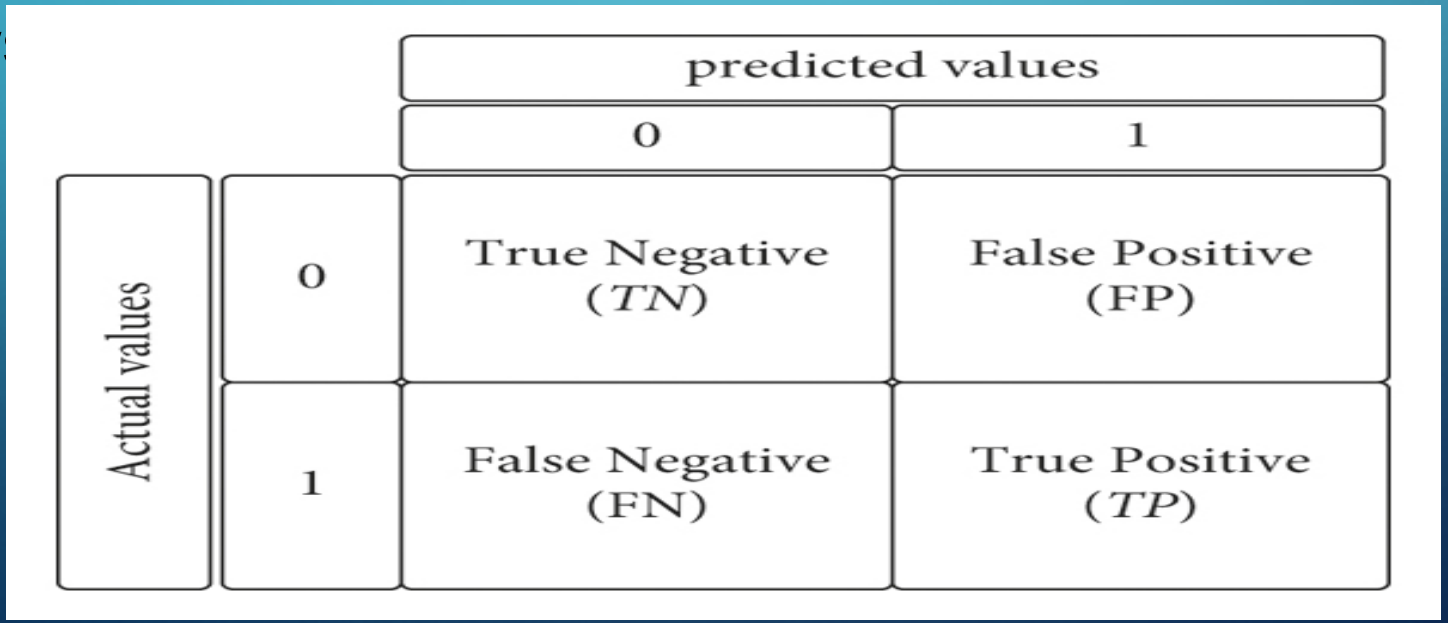
(iv)The test set's samples are fed into the Random Forest, where each Decision Tree makes a choice based on the data. After receiving the findings, the results are voted on using a voting technique to determine the sample categorization results.

RANDOM FOREST



EVALUATION METRICS

- The confusion matrix is considered as a great tool to show the results summary of a model with the classification issues .In the classification, the prediction can be one of four special cases as follow



		predicted values	
		0	1
Actual values	0	True Negative (*TN*)	False Positive (FP)
	1	False Negative (FN)	True Positive (*TP*)

- The following metrics are used to evaluate the proposed model .
- **Accuracy (Acc)** is the percentage of the correct predictions that a classifier has made compared with the actual values of the target in the testing phase.

$$ACC=(TP+TN)/(TP+TN+FP+FN)*100\%$$

- **Precision (Pre)** is the percentage of instances that a classifier has labelled as positive with respect to the total predictive positives (the exactness of a classifier).

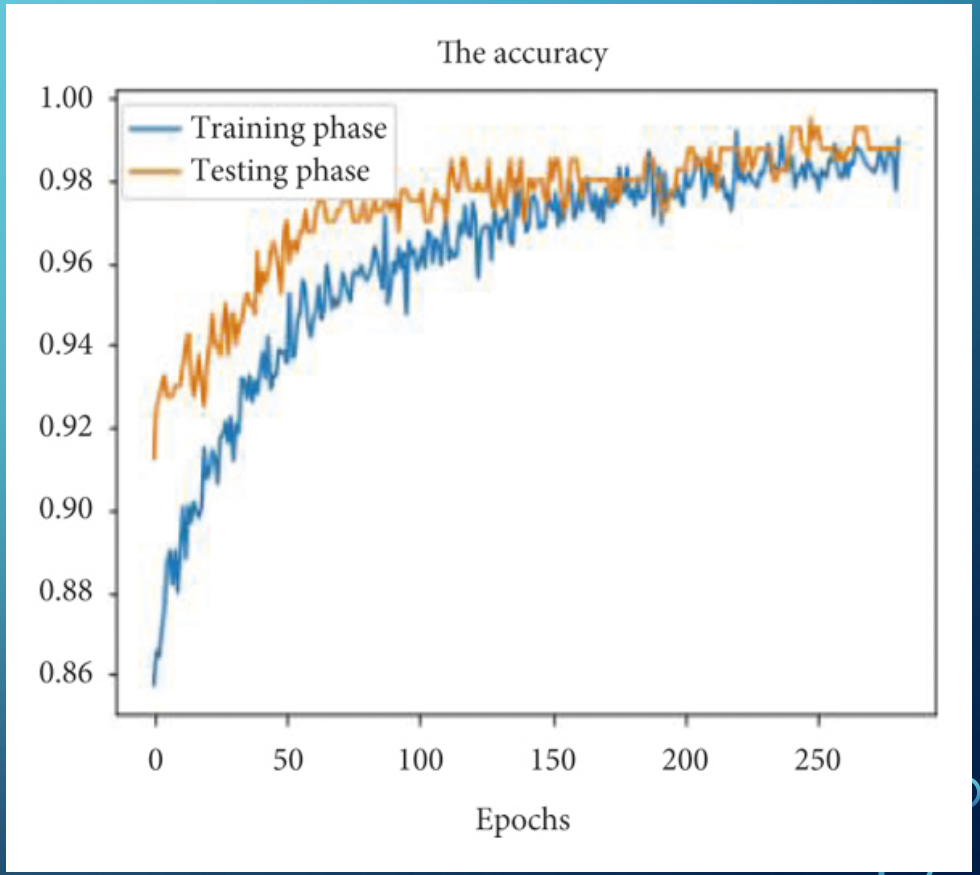
$$Pre=TP/(TP+FP)*100\%$$

- **F1-score** shows the harmonic mean of precision and recall.

$$F1\text{-}score=2*TP/2*TP+FN+FP*100\%$$

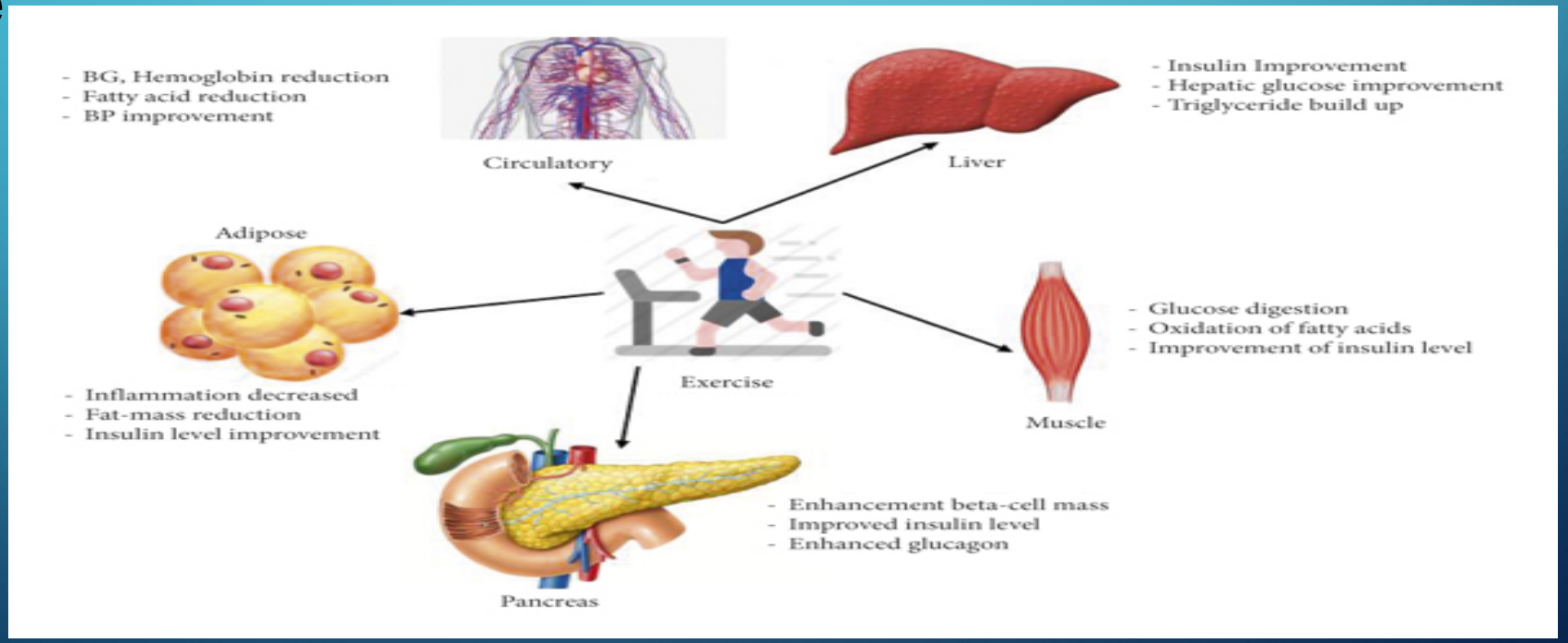
ITERATIVE IMPROVEMENT

- The behavior of the accuracy is shown in Figure. where the blue line represents the training phase, and the orange one represents the testing phase resulting in the best values of the accuracy, 99.0% and 99.75%, respectively.



PREVENTION MEASURES

- they studied the metabolic effect on tissues of diabetic patients and found very significant improvements in individuals performing regular exercise. Moser et al. have also highlighted the significance of regular exercise in improving the functionality of various organs of the body, as shown in Figure
- OBJ



CONCLUSION

- an approach to assist the healthcare domain. The primary objective of this study is twofold. First, we proposed an MLP-based algorithm for diabetes classification and deep learning based LSTM for diabetes prediction. Second, we proposed an IOT-based hypothetical real-time diabetic monitoring system.

As future work, we plan to implement the android application for the proposed hypothetical diabetic monitoring system with the proposed classification and prediction approaches. Genetic algorithms can also be explored with the proposed prediction mechanism for better monitoring