DOI : https://www.doi.org/10.62673/JDIU.V13N1.A5 **ISSN : 2077-0111**

Web Based Diabetes Prediction Model through Combined Dataset using Different Machine Learning Algorithms

Md. Tahzib UI Islam[1], Md. Hasanul Hoque[2]

Abstract: The "Web Based Diabetes Prediction Model through Combined Dataset using Different Machine Learning Algorithms" represents an innovative approach to predict both types of diabetes thorough smart web interface. Diabetes is a chronic medical condition characterized by high levels of sugar (glucose) in the blood this occurs because that time our pancreas does not produce enough insulin (Type 1 diabetes), or because the body's cells do not respond properly to insulin (Type 2 diabetes). There is also a condition called gestational diabetes that can develop during pregnancy. It is a leading cause of severe health complications, including blindness, kidney failure, amputations, heart failure, and stroke. Normally while eating, our internal body mechanism turns that food into sugar or glucose. At that point, ones pancreas is supposed to release insulin. Insulin serves as a key to open our cells, to allow the glucose to enter and allow body to use that glucose for energy. But with diabetes, this system does not work properly caused several major instability causing the onset of diabetes. Type-1 and Type-2 diabetes are the most common forms of diabetebs but there are also other kinds such as gestational diabetes, which occurs during pregnancy as well as other forms. This paper focuses on recent developments in machine learning which have made significant impacts in the diagnosis and detection of diabetes. In this study, Machine Learning (ML) techniques are used to predict the presence of diabetes. The proposed method predicts the chances of diabetes and classifies patient's risk level by using different ML algorithm techniques such as Support Vector Machine(SVM), Random Forest Classifier(RF), K-Neighbors Classifier (KNN), Decision Tree Classifier(DT), Gaussian NB, Logistic Regression, Ada-Boost Classifier, Gradient Boosting Classifier(GB), Multi-layer perceptron (MLP), Nearest Centroid Classifier(NCC) and finally applied Voting Classifier for best output result. Two different datasets are combined to train and test the proposed system which have 899 rows with 11 featured columns. The experimental results highlight that the Random Forest Classifier(RF) algorithm yields the highest accuracy at 85.04% compared to other ML algorithms. These findings showcase the potential of machine learning techniques in the early diagnosis and risk

[1] Associate Professor, Dept. of Computer Science & Engineering, Dhaka International University, Dhaka, Bangladesh.

[2] Dept. of Computer Science & Engineering, Dhaka International University, Dhaka, Bangladesh.

assessment of diabetes, which can significantly contribute to improve healthcare outcomes.

***Key Words** : Diabetes, Insulin, Machine Learning, Diagnosis, Detection, Predict, Healthcare outcomes.*

Introduction

Diabetes is known as one of the most critical human diseases in the contemporary world that has a serious impact on quality of life. Diabetes could be a chronic disease that happens either when the body cannot effectively use the insulin it produces or when the pancreas does not produce sufficient insulin. The early methods of forecasting diabetes help in avoiding health damage. However inaccurate diabetes prediction can prove to be lethal. The machine learning algorithm (Machine Learning et al,. 2023) can be very efficient in the prediction of diabetes due to enormous medical data in the healthcare industry. Diabetes affects a large amount of the population and 25% of people with the diagnosis show signs of microvascular diseases, which are diseases that target the finer blood vessels. This is a sign that they had diabetes for approximately 5 years. Many people go undiagnosed for a long time, which affects their health negatively (Cox, M et al., 2023).

This system is developed for public use and it uses resources from diabetes website, health website, and books. The system is developed in Web based format where MySQL tool is used for database management. Scopes of proposed system are:

i. Identify symptoms of diabetes in order to design the proper rules.
ii. Capture the rules as part of expert system within the diabetes detection system.
iii. Implement an online diabetes detection system as a web application.

Literature Review

The analysis of related work gives results on various healthcare datasets, where analysis and predictions were carried out using various methods and techniques. Various prediction models have been developed and implemented by various researchers using variants of data mining techniques, machine learning algorithms (Sarwar, M. A., et al., 2018). After analysis and examination of all dataset, methods and techniques a model has proposed which is a smart well-structured and user friendly (American Diabetes Association et al,. 2020). In this study, various data analysis techniques and machine learning algorithms are used to compare the prediction analysis of the Pima Indians Diabetes dataset and Eda Dataset. Here used eleven popular classification algorithm such as Support Vector Machine(SVM), Random Forest Classifier(RF), K-Neighbors Classifier(KNN), Decision Tree Classifier(DT), Gaussian NB, Logistic Regression, Ada-Boost Classifier, Gradient Boosting Classifier(GB), Multi-layer perceptron(MLP), Nearest Centroid Classifier(NCC) and apply Voting Classifier and some classifier performance evaluation metrics such as, confusion matrix, ROC and AUC. The

Random Forest Classifier(RF) algorithm gives the best result among these eleven models in terms of accuracy level (85.04%).

Diabetes which is one of the most chronic diseases (Centers for Disease Control and Prevention et al., 2022) in Bangladesh or elsewhere, the prediction of this in the early stage or even before should be able to detect more easily at early stage and maybe control with a proper diet or a less severe treatment.

The Type-2 diabetes has a much stronger link to family history or lineage than the Type-1 diabetes. So, if a member of a family has Type-2 diabetes it is likely that any member of the family could possess the same thing. So, it has to be eliminated before it gets too complicated. So, the aim of the study is to determine the appropriate classification model or algorithm that gives the best accuracy results ever possible. So, if any algorithm proven to be the best can then be used in the prediction of diabetes to figure out if a person is diabetic or non-diabetic so far. This is to avoid any kind of misconceptions due to the incompetent classification algorithm.

Proposed Model

In this section, it proposed a classifier model after analysis that consists of some parts as follows (Soofi, A et al., 2017). Figure-1 shows the pictorial representation of the proposed model-

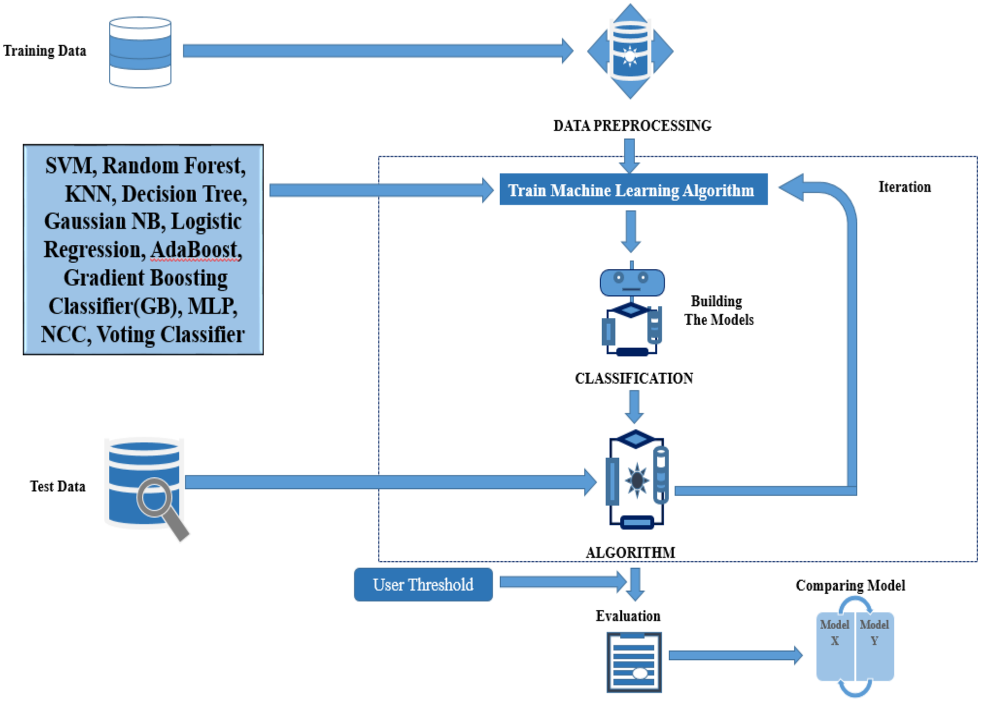

Training Data
DATA PREPROCESSING
SVM, Random Forest, KNN, Decision Tree, Gaussian NB, Logistic Regression, AdaBoost, Gradient Boosting Classifier(GB), MLP, NCC, Voting Classifier
Train Machine Learning Algorithm
Iteration
Building The Models
CLASSIFICATION
Test Data
ALGORITHM
User Threshold
Evaluation
Comparing Model
Model X
Model Y

Figure-1: Diagram of Proposed model.

Dataset

In this research, a combined and obviously new dataset have been used works for patients both male and female. The dataset contains 899 instances and their associated 11 different attributes. Table-1 demonstrates the description of the dataset and the corresponding attributes. The 11 attributes are Gender, Age, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Hypertension, Smoking History and Outcome. The target or dependent variable is the outcome attribute, consisting of two binary values ("1" depicts tested positive and "0" depicts tested negative). The remaining attributes are considering the independent features variables (Modified Dataset et al,. 2023). In dataset gender and smoking history are categorical and Age, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Hypertension are numerical datatype. The resultant feature outcome is considered as boolean type which depicts the existence of diabetes in a patients body. The following table Table-1 shows the details of the datatypes of used dataset.

Table-1: The attribute of dataset part.

Attribute	Description	Type
Gender	Male and Female	Categorical
Age	Age (years)	Numerical
Glucose	Plasma glucose concentration 2 h in an oral glucose tolerance test.	Numerical
Blood Pressure	Diastolic blood pressure (mm Hg).	Numerical
Skin Thickness	Triceps skinfold thickness (mm).	Numerical
Insulin	2 hours serum insulin (ml)	Numerical
BMI	Body mass index (kg/m^2)	Numerical
Diabetes Pedigree Function	Diabetes pedigree function	Numerical
Hypertension	The gauge uses a unit of measurement (mmHg)	Numerical
Smoking History	Never/Current	Categorical
Outcome	Diabetes diagnose results (1 means tested positive and 0 means tested negative)	Boolean

Preprocessing

The primary function of this part is used to peprocess data that existed in the data set. The preprocessing output helps to build the best machine learning model that can provide better accuracy. The preprocessing performs various functions: delete the outlier values, filling the missing data, and do scaling. For example, in dataset, 579 instances were classified as tested positive and 320 instances are tested negative Figure-2.

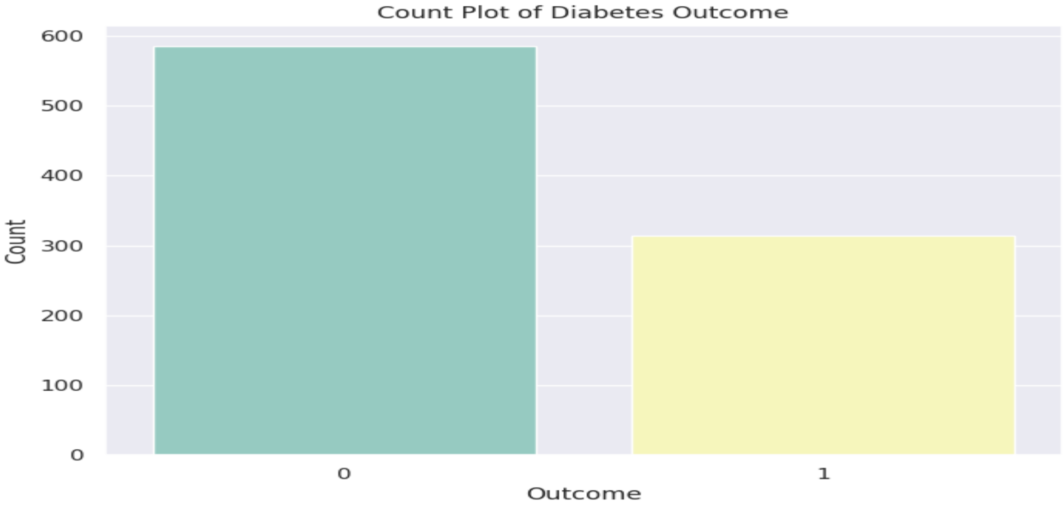


Figure-2 : count plot on dataset

Missing Data Identification

Table-2 explains the missing values of each attribute in the datasets. To overcome missing values it replaced by the corresponding mean value. So, the importance of the current process will reflect on the results (accuracy and performance) of the following parts.

Table-2: The missing values in each attribute of the dataset.

Attribute	Numbers of missing values
Gender	0
Age	0
Glucose	0
Blood Pressure	0
Skin Thickness	0
Insulin	0
BMI	0
Diabetes Pedigree Function	0
Hypertension	0
Smoking History	0
Outcome	0

Normalization

Normalizing the data in the range [0-1] helps to perform feature scaling, which boosted the time processing of ML algorithms. It achieves the target of this process by increasing the processing time of proposed Model.

Feature Extraction

To enhance the quality of the data, feature extraction is essential in the classification model. Dataset correlation approach is a widely effective method used for determining the most relevant features. Figure-3 shows the relationship between input and output features after preprocessing. It shows that glucose and outcome are highly correlated which have a greater correlation coefficient value.

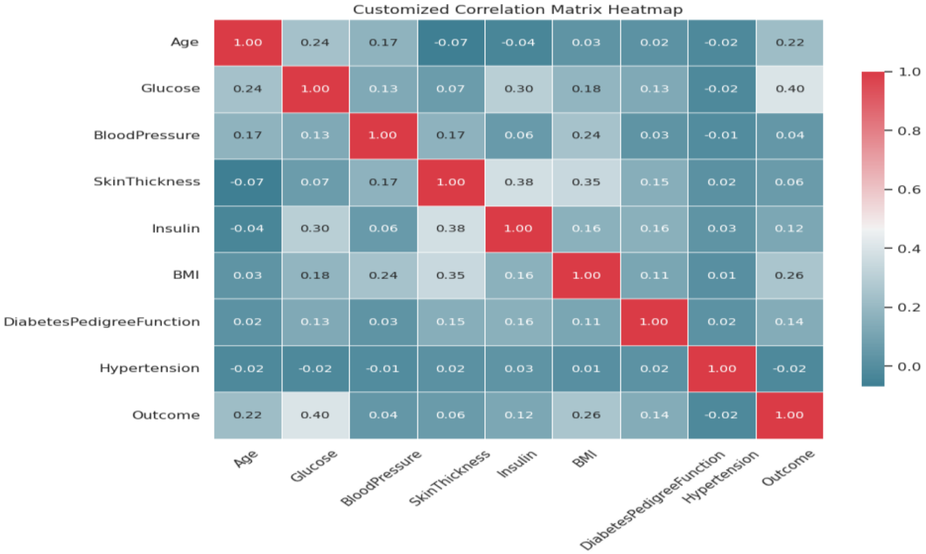


Figure-3: The correlation matrix.

Training

After preprocessing and data filtering, dataset is ready to use for model training. In model, dataset is divided for training and testing where 70% of data from the dataset is used for training the model and 30% data for testing purpose.

Methodology

The classification model depends on the above parts to receive the training data/testing data. The data is processed using 11 different machine learning algorithms: Support Vector Machine(SVM), Random Forest Classifier(RF), K-Neighbors Classifier(KNN),Decision Tree Classifier(DT), Gaussian NB, Logistic Regression, Ada-Boost Classifier, Gradient Boosting Classifier(GB), Multi-layer perceptron(MLP),Nearest Centroid Classifier(NCC) and applied Voting Classifier. The output refers that RF can get the best result for classification data based on the best parameters that enroll from the ML algorithms. The following Figure-4 depicts the flowchart diagram (Scikit-Learn documentation et al,. 2022) of the proposed model.

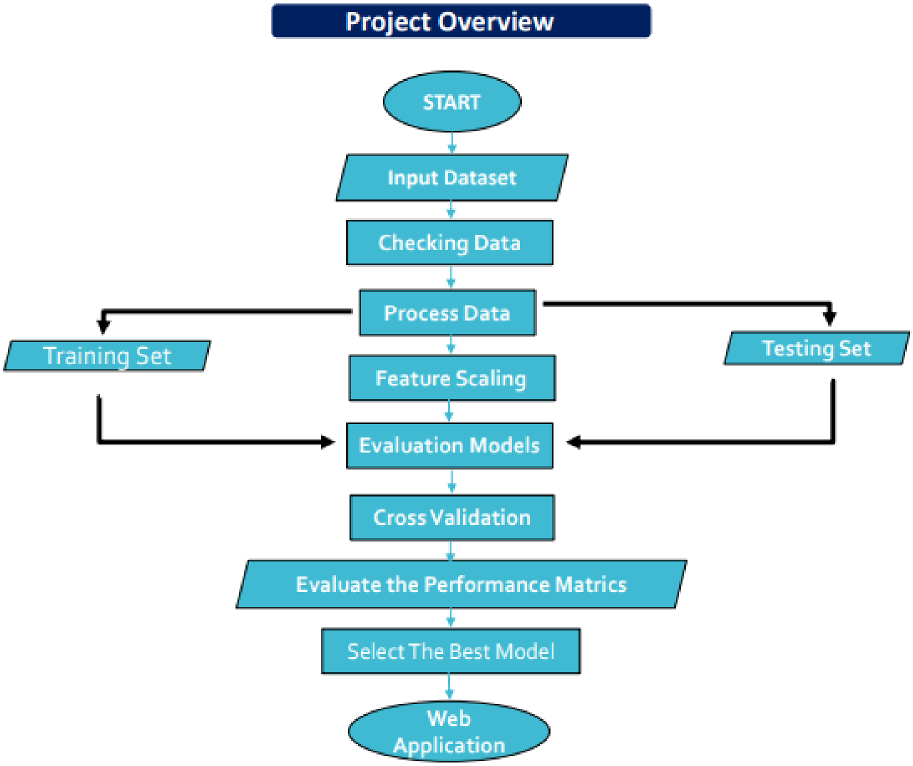


Figure-4: The proposed project diagram.

Result and Discussion

This section discusses on model's results based on some measurements equations: 1–4 such as accuracy, precision, recall, and f1-score. The confusion matrix considers the main factor for computing these measurements. The measures are given below-

	Predicted No (0)	Predicted Yes (1)
Actual No (0)	TN	FP
Actual Yes (1)	FN	TP

$$\text{Accuracy} = \frac{\text{True Positive+True Negative}}{\text{True Positive+True Negative+False negative+False Positive}} \quad \ldots \quad (1)$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive+False negative}} \quad \ldots \quad (2)$$

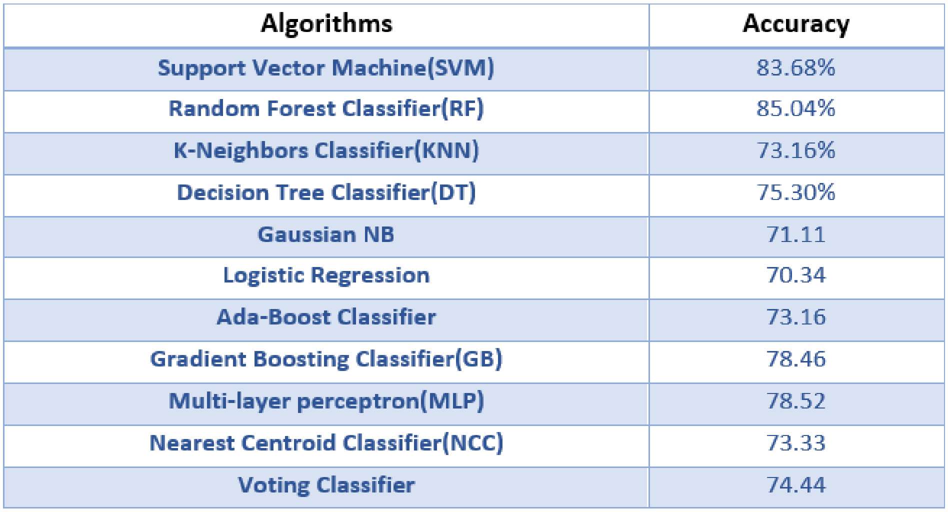
$$\text{Precession} = \frac{\text{True Positive}}{\text{True Positive+False Positive}} \quad \ldots \quad (3)$$

$$\text{F1-SCOR} = \frac{2 \times Precession \times Recall}{Precession \times Recall} \quad \ldots \quad (4)$$

It used a combined dataset from (Dataset: Pima et al,. 2023) and (Dataset: Eda et al,. 2023) which (Modified Dataset et al,. 2023) having 899 rows and 11 columns. In one dataset there have both numerical data and categorical data which is unlike in another dataset. After processing with all required steps it applied 11 different popular ML algorithms to get the best accuracy for the model.

After analyzing the results it came to the conclusion that the Random Forest Classifier(RF) (85.04%) has performed as the most efficient method to analyze the dataset using means of splitting it into training and testing datasets. Table-3 shows the performance result of different algorithms.

Table-3: The performance measure of all classification methods.



Algorithms	Accuracy
Support Vector Machine(SVM)	83.68%
Random Forest Classifier(RF)	85.04%
K-Neighbors Classifier(KNN)	73.16%
Decision Tree Classifier(DT)	75.30%
Gaussian NB	71.11
Logistic Regression	70.34
Ada-Boost Classifier	73.16
Gradient Boosting Classifier(GB)	78.46
Multi-layer perceptron(MLP)	78.52
Nearest Centroid Classifier(NCC)	73.33
Voting Classifier	74.44

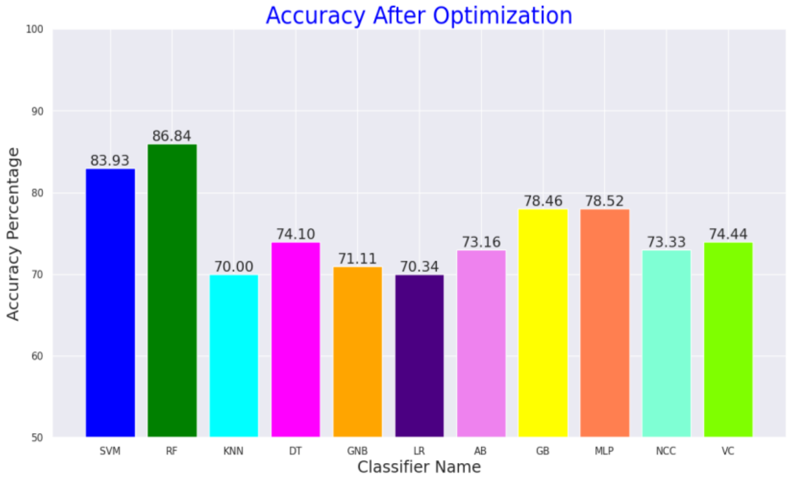


Figure-5: Model accuracy comparison.

Figure-5 depicts the graphical representation of the comparison based on accuracy among all algorithms.

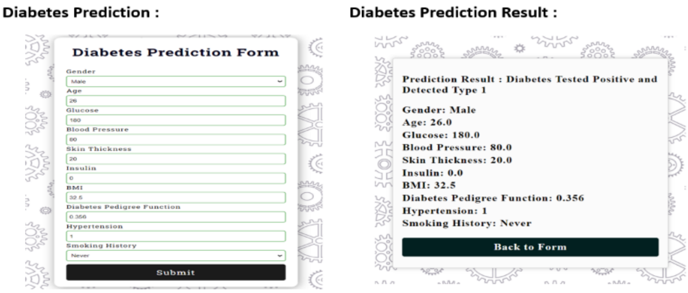
Table-4: Result comparison accuracy table(Hossain et al,. 2023)(Llaha, O et al., 2013) (Yassin, Aet al., 2019).

Algorithms	Paper-1	Paper-2	Paper-3	Proposed
Support Vector Machine	77.4%	77.02%	60%	83.68%
Random Forest	75.6%	74.03%	91%	85.04%
K-Nearest Neighbors	75%	76.62%	90%	73.16%
AdaBoost	N/A	N/A	93%	73.16%
Multi-layer perceptron(MLP)	N/A	N/A	N/A	78.52%
Logistic Regression	77.2%	76.62%	96%	70.34%
Gradient Boosting Classifier(GB)	N/A	N/A	N/A	78.46%
GaussianNB	76.3%	75.32%	93%	71.11%
Decision Tree	N/A	72.73%	86%	75.30%
Nearest Centroid Classifier(NCC)	N/A	N/A	N/A	73.33%
Voting Classifier	N/A	N/A	N/A	74.44%

In this comparison its been noticed that it got comparatively a low accuracy from other papers in some cases but there is a difference than other paper that it used a combined dataset (Modified Dataset et al. 2023) and rest of papers used Dataset Pima (Dataset: Pima et al. 2023) and Dataset EDA (Dataset: Eda et al., 2023).

User Interface

To make more efficient it a well structure and user friendly web interface has been introduced in this study for diabetes prediction which not only detect the diabetes but also both Type-1 and Type-2 forms. This additional feature in this proposed model gave a completely new dimension of this study which created a huge difference with other implemented models showed below in Figure-5, Figure-6 and Figure-7-



Diabetes Prediction :

Diabetes Prediction Form

Gender: Male

Age: 26

Glucose: 180

Blood Pressure: 80

Skin Thickness: 20

Insulin: 0

BMI: 32.5

Diabetes Pedigree Function: 0.356

Hypertension: 1

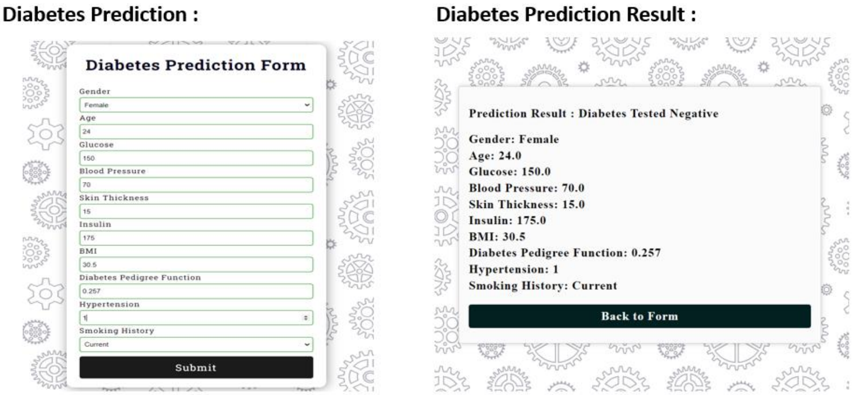
Smoking History: Never

Diabetes Prediction Result :

Prediction Result : Diabetes Tested Positive and Detected Type 1

Gender: Male
Age: 26.0
Glucose: 180.0
Blood Pressure: 80.0
Skin Thickness: 20.0
Insulin: 0.0
BMI: 32.5
Diabetes Pedigree Function: 0.356
Hypertension: 1
Smoking History: Never

Figer-6: Prediction form of project web app for diabetes tested positive.



Diabetes Prediction :

Diabetes Prediction Form

Gender: Female
Age: 24
Glucose: 150
Blood Pressure: 70
Skin Thickness: 15
Insulin: 175
BMI: 30.5
Diabetes Pedigree Function: 0.257
Hypertension: 1
Smoking History: Current

Diabetes Prediction Result :

Prediction Result : Diabetes Tested Negative

Gender: Female
Age: 24.0
Glucose: 150.0
Blood Pressure: 70.0
Skin Thickness: 15.0
Insulin: 175.0
BMI: 30.5
Diabetes Pedigree Function: 0.257
Hypertension: 1
Smoking History: Current

Figure-7: Prediction form of project web app for diabetes tested negative.

Conclusion and Future Scope

The aim of this study has focused on using of machine learning in the detection of diabetes using 11 different machine learning algorithms. Random Forest Classifier(RF) has got the highest accuracy among all other popular ML algorithms. The study found that machine learning could be very useful tool to detect diabetes. It can help to control and reduce the affection rate of diabetes in under developed populated country like Bangladesh. Though the model did not reach the expected goal and has also many future scopes to enhance the prediction percentage result. It needs to modify and customize the model to improve accuracy and precision of diabetes prediction with this dataset compared to existing models. The web interface also has a great chance to improve with more attractive UI and make more user friendly to make easy use for the illiterate people also. There is a great scope to convert this web interface in app version in future. To increase the prediction rate it needs to analyze more datasets and more features to be considered.

References :

1. Cox, M. E., & Edelman, D. (2009). Tests for screening and diagnosis of type 2 diabetes. In Clinical Diabetes, 27(4), 132-138.

   Roth, R. A. (2011). Nutrition & Diet Therapy. Delmar Cengage Learning.
2. Centers for Disease Control and Prevention. (2022). Diabetes Basics. Retrieved from https://www.cdc.gov/diabetes/basics/index.html.
3. American Diabetes Association. (2020). Diagnosis. Retrieved from https://www.diabetes.org/a1c/diagnosis.
4. Sarwar, M. A., Kamal, N., Hamid, W., & Shah, M. A. (2018). Prediction of Diabetes Using Machine Learning Algorithms in Healthcare. In 2018 24th International Conference on Automation and Computing (ICAC) (pp. 1-6): 10.23919/IConAC.2018.8748992.
5. Machine Learning. (2023). Retrieved from https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained.

6. Soofi, A., & Awan, A. (2017). Classification Techniques in Machine Learning: Applications and Issues. Journal of Basic & Applied Sciences, 13, 459-465. 10.6000/1927-5129.2017.13.76.
7. Scikit-Learn documentation. (2022). Retrieved from https://www.scikit-learn.org/0.21/index.html.
8. Rezaei, F., Abbasitabar, M., Mirzaei, S., et al. (2022). Improve data classification performance in diagnosing diabetes using the Binary Exchange Market Algorithm. J Big Data, 9(1), 43.: 10.1186/s40537-022-00598-z.
9. Tigga, N. P., & Garg, S. (2020). Prediction of Type 2 Diabetes using Machine Learning Classification Methods. Procedia Computer Science, 167, 706-716: 10.1016/j.procs.2020.03.336.
10. Yassin, A. A.(2019) Title of the article. Journal Name, Volume(Issue), Page numbers. URL: [https://www.iasj.net/iasj/journl/260/issues] (Last accessed: 2-09-2023; Time: 1.20pm).
11. Llaha, O., & Rista, A. (2013). Title of the article. Journal Name, Volume(Issue), Page numbers. URL: [https://www.ceur-WS.org/Vol-2872/paper13pdf] (Last accessed: 24-08-2023; Time: 2.20pm).
12. Hossain, Alishah. (2023). URL: [https://www.sciencedirect.com] (Last accessed: 25-08-2023; Time: 1.22pm).
13. Varma, S. (2023). Title of the article. International Journal of Engineering & Research & Technology. URL: [https://www.ijert.org] (Last accessed: 27-08-2023; Time: 2.33am).
14. Mujumdar, A. (2019). Title of the conference paper. International Conference on Recent Trends in Advanced Computing. URL: [https://www.sciencedirect.com] (Last accessed: 27-08-2023; Time: 6.30pm).
15. Dataset: Pima. (2023). URL: [https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset] (Last accessed: 07-07-2023; Time: 7.00am).
16. Another Dataset. (2023). URL: [https://www.kaggle.com/code/tumpanjawat/diabetes-eda-random-forest-hp] (Last accessed: 08-07-202; Time: 9.43am).
17. Modified Dataset. (2023). URL:[https://drive.google.com/file/d/1JXv0T7XOrMePHKXS_6ObhrESy3jehTgk/view?usp=sharing] (Last accessed: 08-07-202; Time: 8.55pm).