Analysis and Future Risk Prediction of Diabetes Mellitus

Project ID:21D211230

Review - III

Group Members

RA1711003030211 Praffulla Kumar Dubey

RA1711003030230 Udbhav Naryani

Supervised By:

Dr. Medhavi Malik

Assistant Professor

Department of Computer Science & Engineering

Faculty of Engineering & Technology

SRM Institute of Science & Technology

May 10, 2021

Table of Contents

1. Abstract
2. Objective
3. Literature Survey
4. Overall Design of the Proposed System
5. Data-set Specification
6. Methodology/Algorithms Used
7. Experimental Results
8. Performance Evaluation
9. Comparison with Existing System
10. Publication Details of Submission of the Work to Journal
11. References

Abstract

The number of patients suffering from diabetes in India are increasing rapidly. One of the major reasons behind this can be that most of the population does not take regular check-ups for diabetes. Hence they are unable to take proper precautions and have a risk of getting diabetes in the future. The current methodologies followed by the doctors are manual and can be time consuming, hence we have come up with a solution for this problem, by providing a model in which the user can enter their health report details and know their future risk of having diabetes and if required they can consult a doctor for the same. In this project, we propose to analyse the Pima Indian diabetes dataset and develop an intelligent system for predicting future risk of diabetes using supervised machine/deep learning algorithms. We would be coding in python using Jupyter Notebook for analysis, training and testing of our model. We would be deploying our prediction model on localhost using Spyder IDE.

Objective

- Analyse and gain insights from the collected data.
- Develop a diabetes prediction model using supervised machine/deep learning algorithms.
- Predict future risk of diabetes.
- Develop a user friendly GUI (Graphical User Interface).

Literature Survey I

Author	Algo.	Conf./Journal & Findings
J.J Khanam, and S. Foo [4]	ANN: 88.6%	**"ICT Express, 2021."** Used Artificial Neural Network for training and testing their model in WEKA tool.
S. Barik et al. [1]	XGB: 74.1%	**"Intelligent and Cloud Computing, 2021."** The data collected was very small and no pre-processing of data-set was performed.
Y. N. and Sri [8]	RF: NA	**"Cluster Computing, 2019."** Used only Random Forest Algorithm on Hadoop Cluster.
S. Lekha and Suchetha M. [5]	CNN: NA	**"IEEE Journal of Biomedical and Health Informatics, 2018".** The data collected was very small and no pre-processing of data-set was performed.
Q. Zou et al. [10]	NN: 76%	**"Frontiers in Genetics, 2018."** MATLAB tool was used for model creation by the authors and we propose to use python.

Table: Literature Survey I

Literature Survey II

Author	Algo.	Conf./Journal & Findings
F. Mercaldo et al. [6]	HT: 76%	**"Procedia Computer Science, 2017."** They used WEKA tool and did not mention the pre-processing steps.
A. Mohebbi et al. [7]	CNN: NA	**"International Conference of IEEE Engineering in Medicine and Biology Society, 2017."** Used Convolution Neural Network small data-set.
J Pradeep et al. [3]	J48: 73.8%	**"Procedia Computer Science, 2015."** They did not mention the data pre-processing techniques and methodology followed.
A. Iyer et al. [2]	NB: 79.5%	**"International Journal of Data Mining Knowledge Management Process, 2015."** Pre-processing was done using WEKA tool and we propose to use Python.
S Nanda et al. [9]	LR: NA	**"Prenatal Diagnosis, 2011."** Used only Logistic Regression for prediction and the accuracy of their model was not mentioned.

Table: Literature Survey II

Overall Design of the Proposed System I

The architectural design for the proposed system is divided into 3 layers:

- **Data Layer:** Pre-process the data and gain knowledge from it.
- **Application Layer:** Create, train and test the diabetes prediction model.
- **Presentation Layer:** Create a GUI (Graphical User Interface) for the user to use the diabetes prediction model.

Overall Design of the Proposed System II

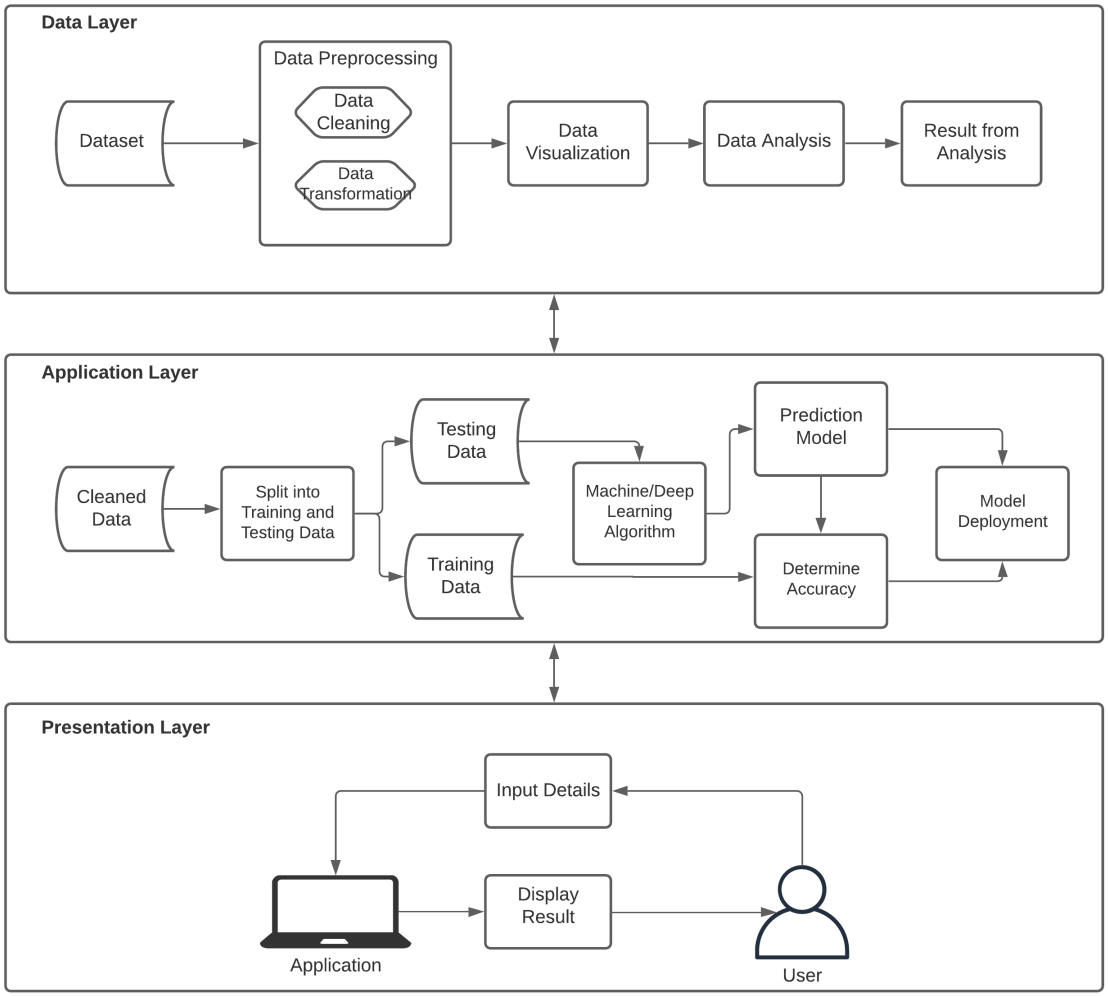

Figure: Architectural Design for Proposed System

Data-set Specification

S.no	Attribute	Description
1.	Pregnancies	Number of times pregnant.
2.	Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test.
3.	BloodPressure	Pressure of circulating blood against the walls of blood vessels (mm Hg).
4.	SkinThickness	Triceps skin fold thickness (mm).
5.	Insulin	Hormone that lowers the level of glucose in blood (mu U/ml).
6.	BMI	Body mass index (weight in kg/(height in m)^2).
7.	DiabetesPedigree Function	A function which scores likelihood of diabetes based on family history.
8.	Age	Age (in years).
9.	Outcome	Class variable (0 or 1), 0 = Non Diabetic 1 = Diabetic
10.	Doctor	Name of Doctor in-charge.
11.	Hospital	Hospital Name.

Table: Data-set Specification

Methodology/Algorithms Used I

- **Methodology:**
  - **Data Cleaning:** Preparing the data-set for analysis and prediction.
    - We dropped columns **"doctor"** and **"hospital"** as they were irrelevant for analysis and prediction.
    - The data-set collected has a lot of null values and they have to be handled properly.
    - We replaced all the null values of a particular attribute by the class median.
  - **Data Visualisation:** Represent information and data graphically.
    - We used **"seaborn"** library to plot **"heatmap"** and **"pairplot"** for the data-set.
  - **Data Analysis:** Extracting useful information from data and making the decision based upon the data analyzed.
    - As glucose levels increase the chances of being diabetic increases.
    - As insulin decreases chances of being diabetic increases.

Methodology/Algorithms Used II

- **Algorithms:**
  - **K-Nearest Algorithm:**
    - It identifies data points that are separated into several classes to predict the classification of a sample point.
    - It uses Euclidean distance in order to find nearest neighbors. As shown in formula below: $d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
    - It classifies new points based on similarity measures.
  - **Support Vector Machine:**
    - It is a linear model for classification as well as regression problems.
    - It is used to solve both linear as well as non-linear problems.
    - Basically, SVM creates a hyper-plane or a line which separates the data into classes.
    - SVM are one of the robust prediction algorithms.
    - It is useful in outlier detection.

Methodology/Algorithms Used III

- **Logistic Regression:**
  - It is used when target variable is categorical for example: to predict tumor (if 1 then true else if 0 then false). It is used for predictive analysis.
  - It can be used for classification algorithms.
  - Logistic Regression uses logistic function to model a binary dependent variable as defined below: $f(x) = L/(1 + e^{-k(x-x_0)})$ where,
    - x_0 is the x value of sigmoid midpoint
    - L is curves maximum learning value.
    - k is the steepness of the curve or the logistic growth rate.
- **Decision Tree:**
  - It is a classifier in the structure of tree.
  - Internal nodes of the decision tree represent features of data-set, branch of tree represents decision rules and leaf nodes of the tree represents the outcome.
  - It is a graphical representation of all possible solution to a problem based in the conditions given.

Methodology/Algorithms Used IV

- **Naive Bayes:**
  - It is mainly used for solving classification problems.
  - It predicts the outcome on the basis of the probability of the data and hence it is called as **Probabilistic Classifier**. Naive stands for strong assumption and it uses Bayes theorem which is given as: $P(A|B) = P(B|A)P(A)/P(B)$ where,
    - $P(A|B)$ is **Posterior Probability**: Probability of hypothesis A on the observed event B.
    - $P(B|A)$ is **Likelihood Probability**: Probability of evidence given that probability of hypothesis is true.
    - $P(A)$ is **Prior Probability**: Probability of hypothesis before observing the evidence.
    - $P(B)$ is **Marginal Probability**: Probability of evidence.
- **Random Forest:**
  - Random forest is a classifier that contains number of decision trees on various subset of given data-set and takes average to improve the accuracy of prediction. The greater number of tree the greater is the accuracy and it prevents overfitting.

Methodology/Algorithms Used V

- **Perceptron:**
  - It is an algorithm used for binary classifiers in Supervised Learning.
  - Perceptron is a single layered neural network.
  - They consist of four main parts:
    - input values
    - weights
    - bias
    - activation function (It is the function which outputs a small value for small inputs, and a larger value if its inputs exceed a threshold)
    - net sum

- **Multi-layer Perceptron:**
  - It classifies data-sets that are not linearly separable.
  - It has the same input and output layers of Perceptron but in addition to that it has multiple hidden layers.
  - MLP form the basis for all neural networks and they have greatly improved the results of classification as well as regression problems.

Methodology/Algorithms Used VI

- **Hybrid Model:**
  - Hybrid Models are an ensemble of different machine learning algorithms.
  - They can be used for taking advantage of efficiency of different models on particular sets of data which will in turn improve the accuracy of overall prediction model.
  - We used stacking to design our hybrid model.
    - In base layer we will be using algorithms which gave good accuracy after testing.
    - In meta layer we used Logistic Regression as our problem is classification based.

Experimental Results I

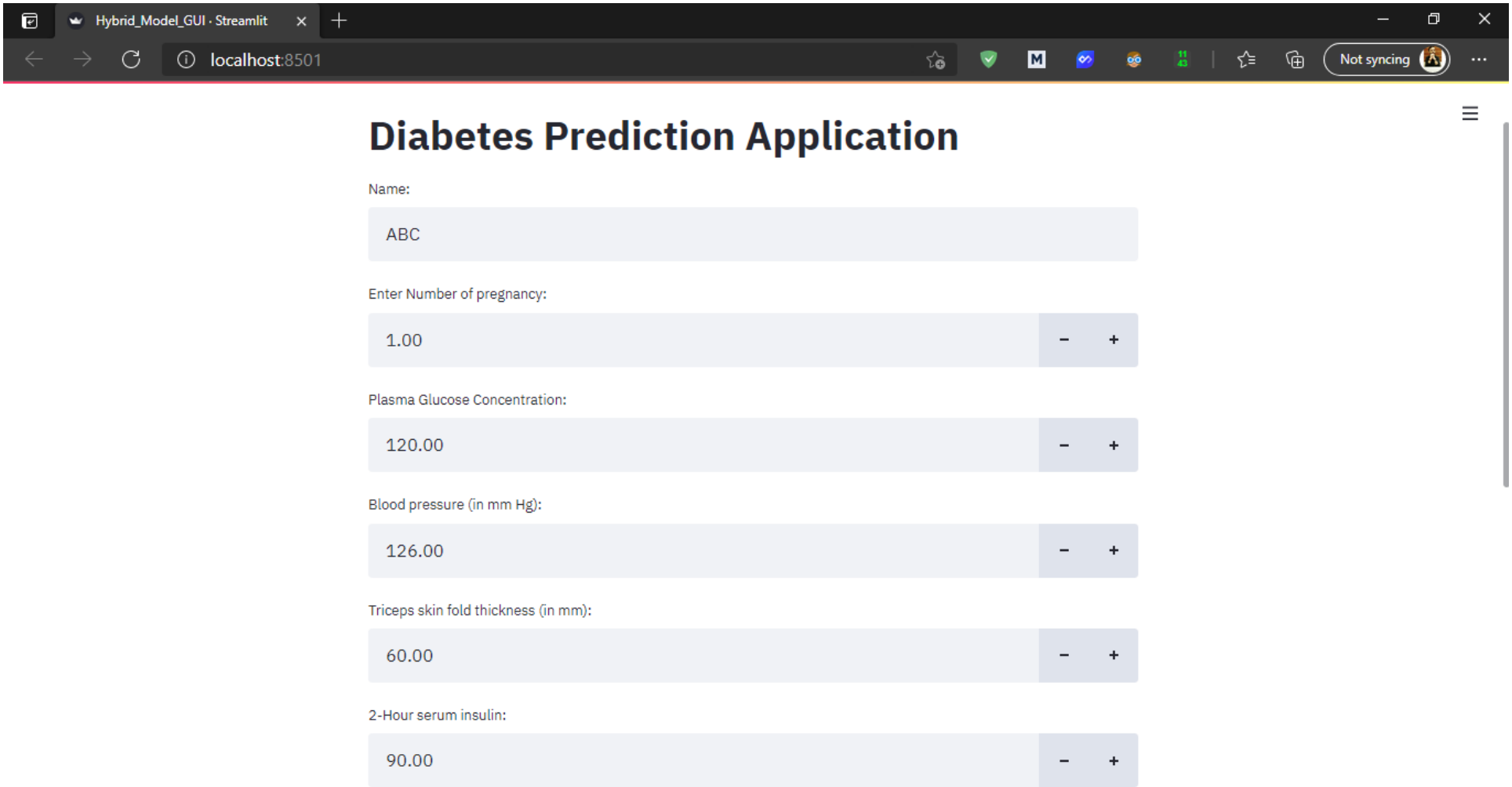

Figure: Experimental Results I

Experimental Results II

Figure: Experimental Results II

Experimental Results III

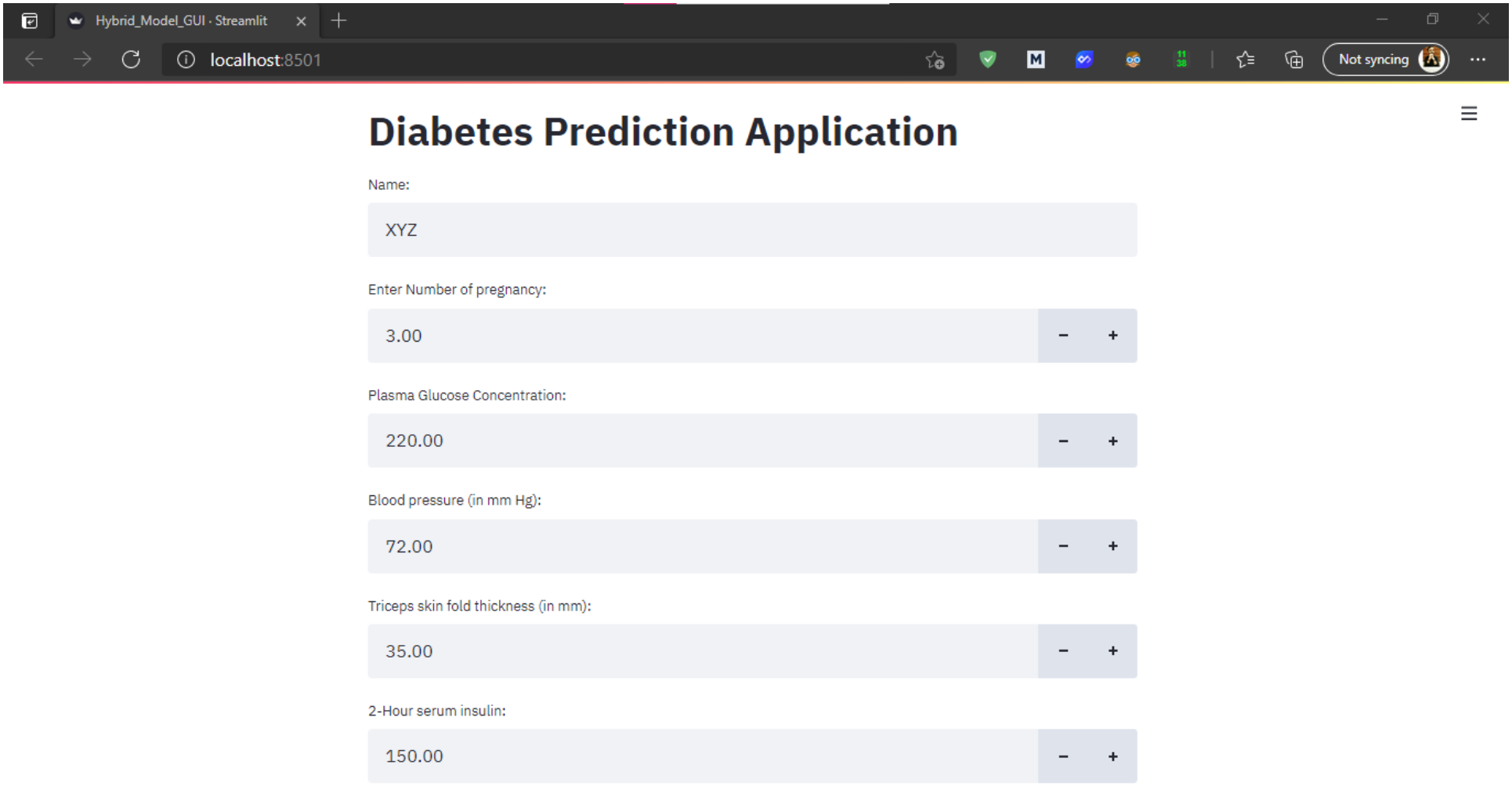

Figure: Experimental Results III

Experimental Results IV

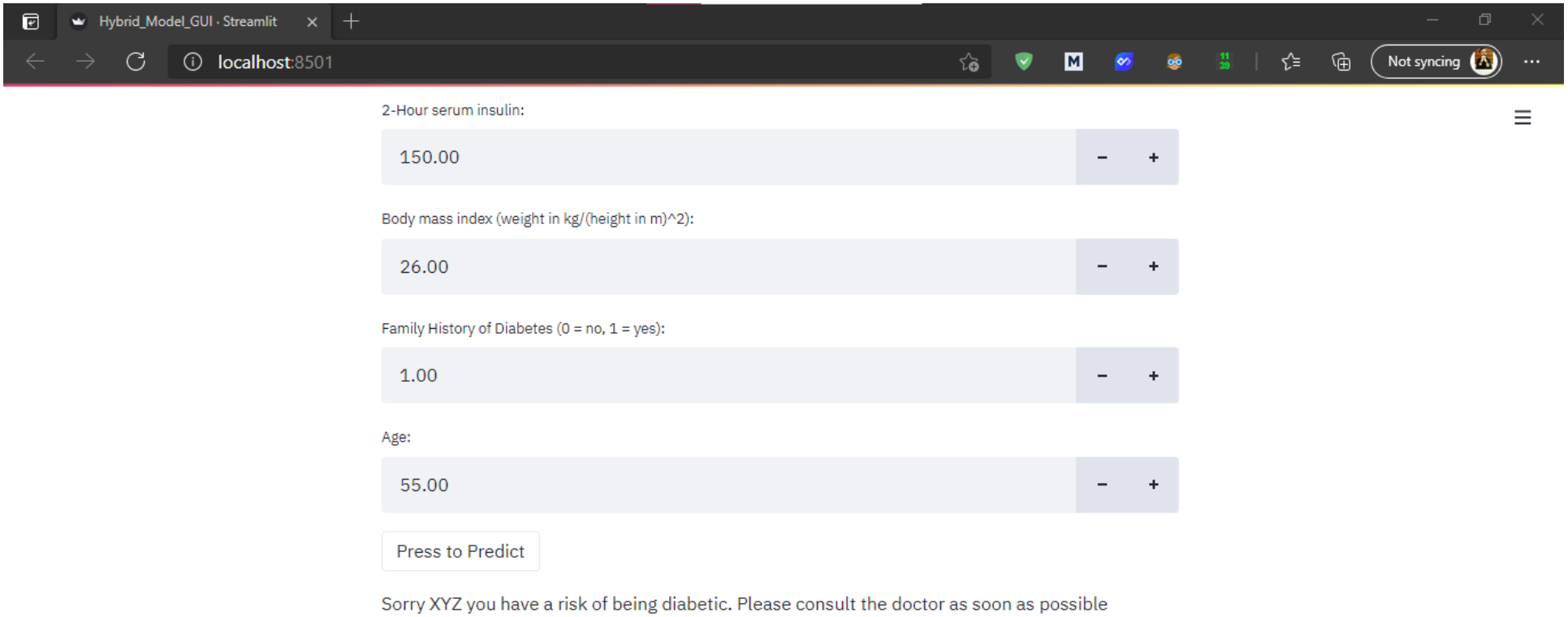

Figure: Experimental Results IV

Performance Evaluation

The performance of the various models trained and used in our project is shown in the graph below:

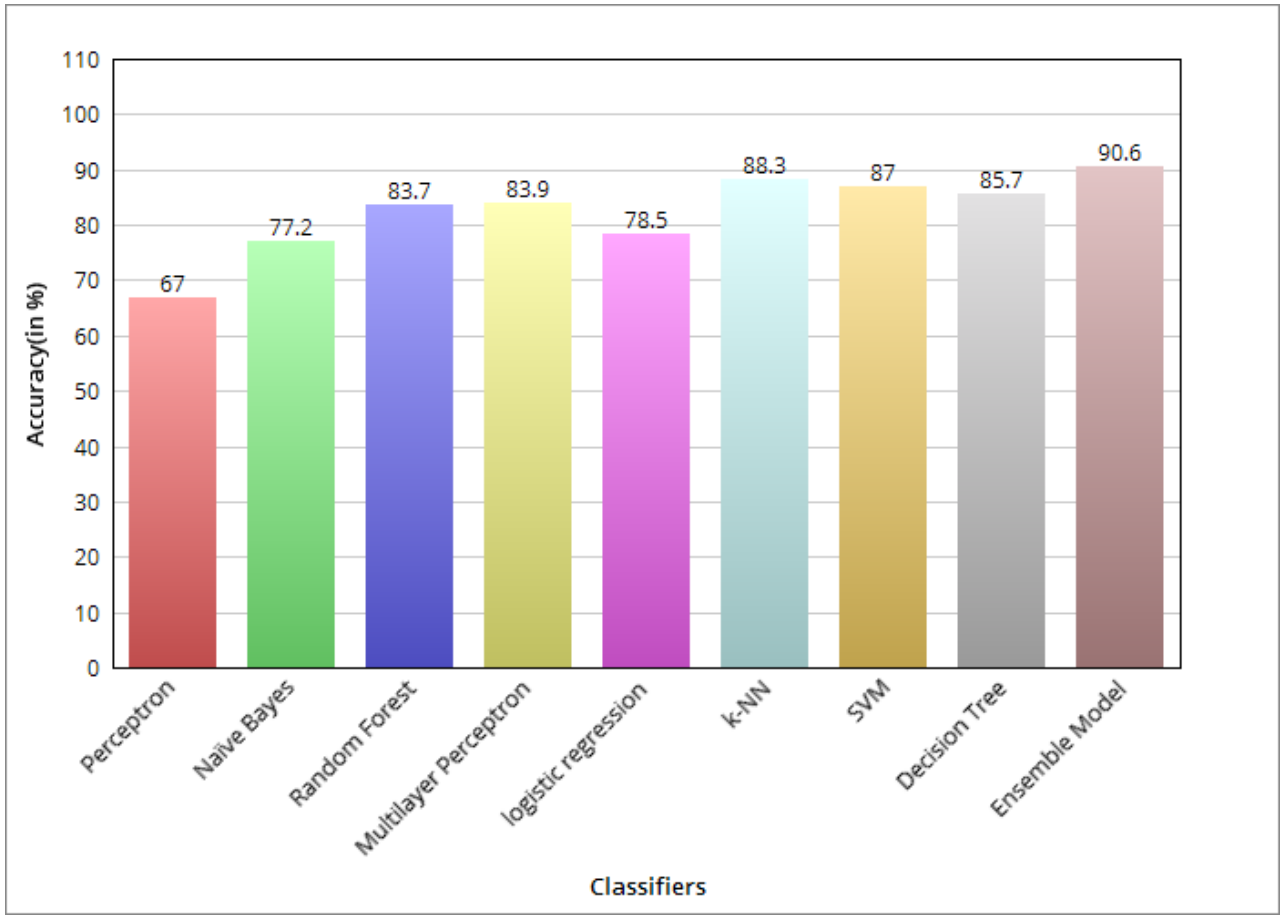

Figure: Performance Evaluation

Comparison with Existing System

The comparison of system with existing system (our base paper [4]) is shown in the graph below:

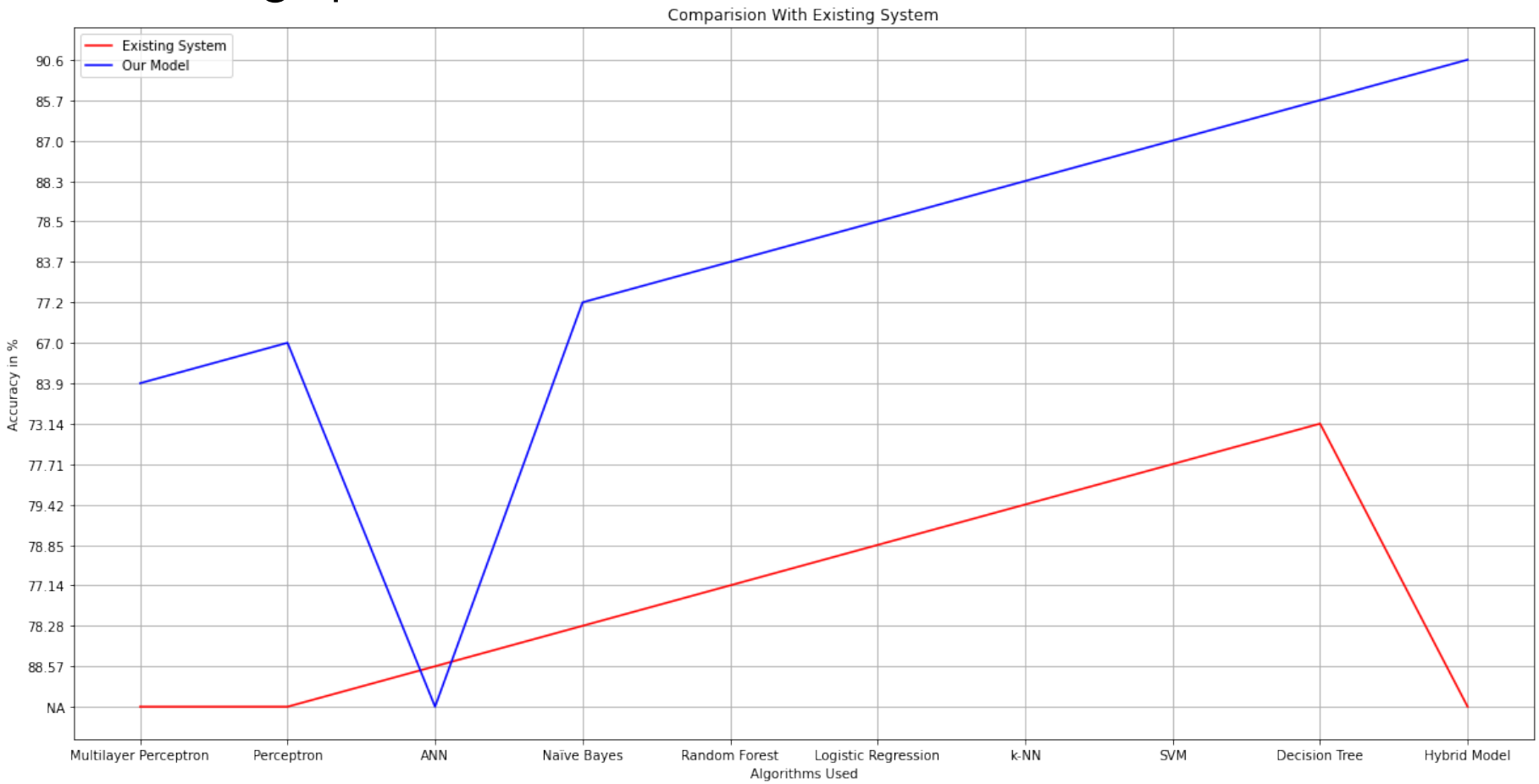

Figure: Performance Evaluation

Publication Details of Submission of the Work to Journal

Submitted our paper in the **Scopus Indexed Journal "Computer Methods and Programs in Biomedicine"**.

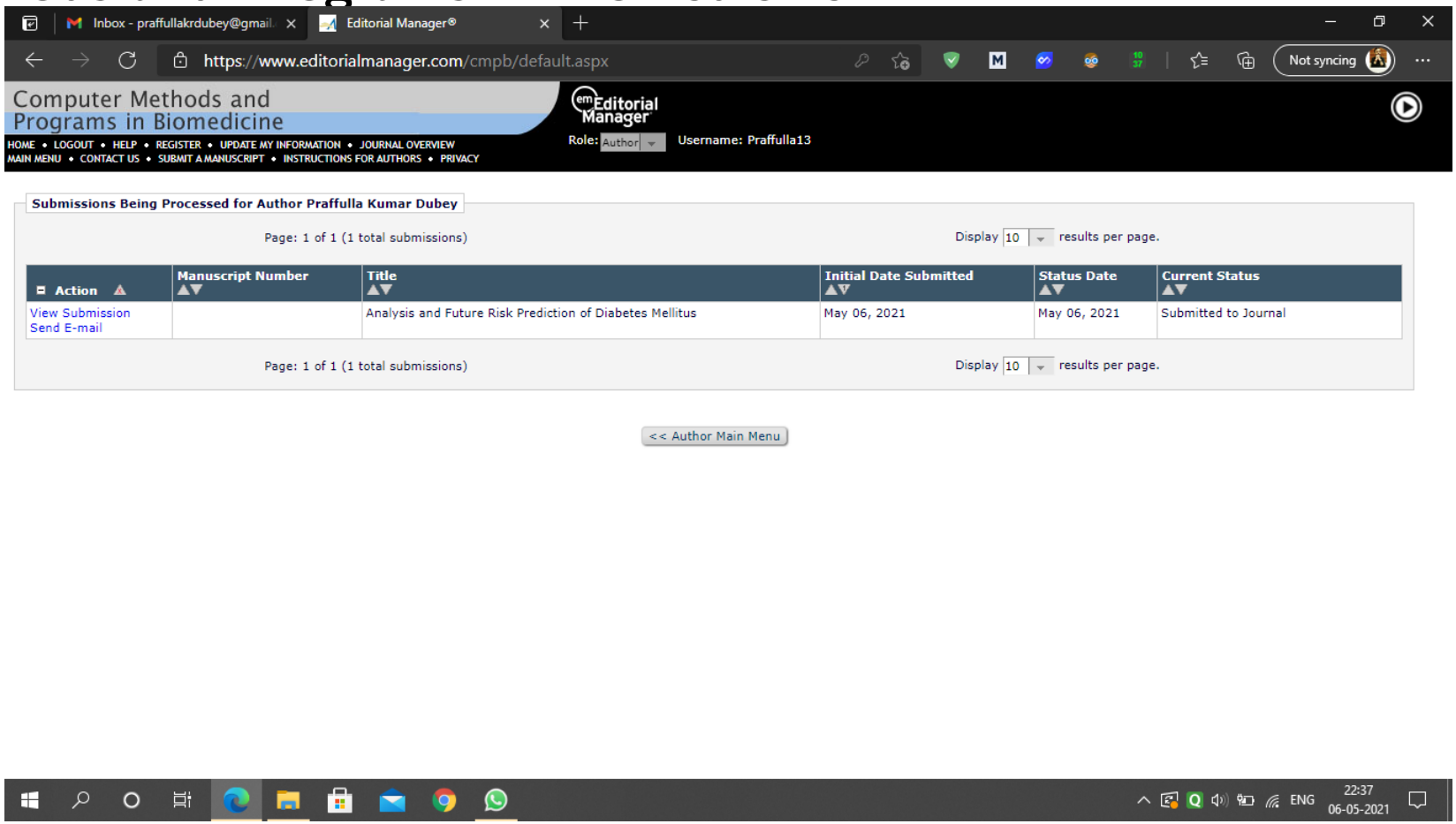

Figure: Publication Details

References I

S. Barik, S. Mohanty, S. Mohanty, and D. Singh.
Analysis of Prediction Accuracy of Diabetes Using Classifier and Hybrid Machine Learning Techniques, pages 399–409.
01 2021.

A. Iyer, s. Jeyalatha, and R. Sumbaly.
Diagnosis of diabetes using classification mining techniques.
International Journal of Data Mining Knowledge Management Process, 5:1–14, 02 2015.

J. p. Kandhasamy and S. Balamurali.
Performance analysis of classifier models to predict diabetes mellitus.
Procedia Computer Science, 47:45–51, 12 2015.

J. J. Khanam and S. Foo.
A comparison of machine learning algorithms for diabetes prediction.
ICT Express, 02 2021.

References II


S. Lekha and S. M.
Real-time non-invasive detection and classification of diabetes using modified convolution neural network.
IEEE Journal of Biomedical and Health Informatics, 22(5):1630–1636, 2018.


F. Mercaldo, V. Nardone, and A. Santone.
Diabetes mellitus affected patients classification and diagnosis through machine learning techniques.
Procedia Computer Science, 112:2519–2528, 12 2017.


A. Mohebbi, T. Aradóttir, A. Johansen, H. Bengtsson, M. Fraccaro, and M. Mørup.
A deep learning approach to adherence detection for type 2 diabetics.
In *Proceedings of 2017 39th Annual International Conference of the Ieee Engineering in Medicine and Biology Society*, 2017 39th Annual

References III

International Conference of the Ieee Engineering in Medicine and Biology Society (embc), pages 2896–9, United States, 2017. IEEE. 2017 39th Annual International Conference of the Ieee Engineering in Medicine and Biology Society, EMBC 2017 ; Conference date: 11-07-2017 Through 15-07-2017.

Y. N. and K. SriPreethaa.
Diabetes prediction in healthcare systems using machine learning algorithms on hadoop cluster.
Cluster Computing, 22, 01 2019.

S. Nanda, M. Savvidou, A. Syngelaki, R. Akolekar, and K. Nicolaides.
Prediction of gestational diabetes mellitus by maternal factors and biomarkers at 11 to 13 weeks.
Prenatal Diagnosis, 31(2):135 – 141, Feb. 2011.

References IV


Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang.
Predicting diabetes mellitus with machine learning techniques.
Frontiers in Genetics, 9:515, 2018.