ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1760

Design and Implementation of a Hybrid Machine Learning Approach for Improved Diabetes Diagnosis: Comparative Study

Ogu Maris Uchenna[1*]; M. E. Benson-Emenike[2]

[1]M. Sc Student Department of Computer Science Federal University of Technology Owerri, FUTO, Imo State, Nigeria.
[2]Senior Lecturer Department of Computer Science Federal University of Technology Owerri, FUTO, Imo State Nigeria.

Corresponding Author: Ogu Maris Uchenna[1*]

Publication Date: 2026/04/11

Abstract: Diabetes is one of the diseases that are causing a significant increase in the global death rate today. It causes vision loss, heart disease, foot ulcer, and stroke and kidney problems. Because of this, there must be a crucial need to compare some of this machine learning techniques to ascertain the best technique that will be of benefit to built prediction model which can be able to detect diabetes accurately on patients to enable the reduction in the numbers of people who die from the disease. Using machine learning, the research study created a comparative analysis with machine learning techniques to predict Diabetes. In this very research work, six machine learning algorithms were utilized: The Nave Bayes model (NB), Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), and Histogram Gradient Boosting Classifier (HBC). The models were trained in Python programming language and was evaluated on secondary dataset obtained from Kaggle http, the model's capability was evaluated by utilizing the F1 score, accuracy, recall, and R2 score. When compare to other algorithms, Random forest and Histogram Gradient Boosting Classifier had the greatest performance accuracy of 0.97%. The model was deployed on a web application after being saved with the highest accuracy using Python, Django, HTML, CSS, and other IDEs such as VS Code and Jupiter Notebook.

Keywords: *Machine Learning Technique, Algorithm, Diabetes Prediction, Artificial Intelligent, Data Mining, Insulin.*

How to Cite: Ogu Maris Uchenna; M. E. Benson-Emenike (2026) Design and Implementation of a Hybrid Machine Learning Approach for Improved Diabetes Diagnosis: Comparative Study. *International Journal of Innovative Science and Research Technology*, 11(3), 3848-3859. https://doi.org/10.38124/ijisrt/26mar1760

I. INTRODUCTION

Diabetes is a chronic illness or disorder caused by a prolonged elevation of blood glucose levels as a result of inefficient use of the substance by the body. Serious complications from this disease include impaired vision, heart attack, cancer of the leg, coma, stroke, edema illness, chronic renal failure, high blood presure. The number of diabetics worldwide has greatly increased, and the condition is now recognized as a major health concern. Jaiswal, Negi, & Pal, [1]. This illnes is seen as the major cause of death globally. However, data mining tools has been found to help with accurate disease diagnosis and treatment decisions in healthcare while lightening the strain on experts. severial health challeges have been iniciated by this diabetic illnes. Becase of this, it is paramount important to stop it, keep a close watch on it, and disburse knowledge about the diseases. Rastogi & Bansal, [2]. Additionally, the majority of prevalent health diseases, like diabetes and high blood pressure, are brought on by changes in eating life style and an increasing rate of mental stress in the human body. World Health Organization's website offers a statistical summary on diabetes that demonstrates an exponential rise over the recent years. Globally, making us to know that the umerious been with this diabetes diseases has grown dramatically, from a staggering 108 million in 1980 to 422 million in 2014. (Pradhan, Dhaka, & Poonia, [3]. Technological advancements have produced a numerous quantity of data in the healthcare industry. It is feasible to identify changes in human life through making use of diverse data analysis techniques that will enable the early diagnosis of diseases. One such illness that needs to be detected early to prevent further consequences and restrict its progression is diabetes Manikandababu, IndhuLekha, Jenifer, & Theodora, [4]. Many bioinformatics experts have worked to tackle this illness and create instruments and frameworks that could help identify diabetes early. We developed prediction models by making use of series of Machine Learning

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1760

Techniques, including classification algorithms. Decision trees, Support Vector Machines, and Linear Regressions were widely used algorithms Larabi-Marie-Sainte, Aburahmah, Almohaini, & Saba, [5]. By employing big data analytics to analyze enormous datasets, people can find concealed trends and details that can be used to gain insight from the data and create accurate predictions Mujumdar &Vaidehi, [6].

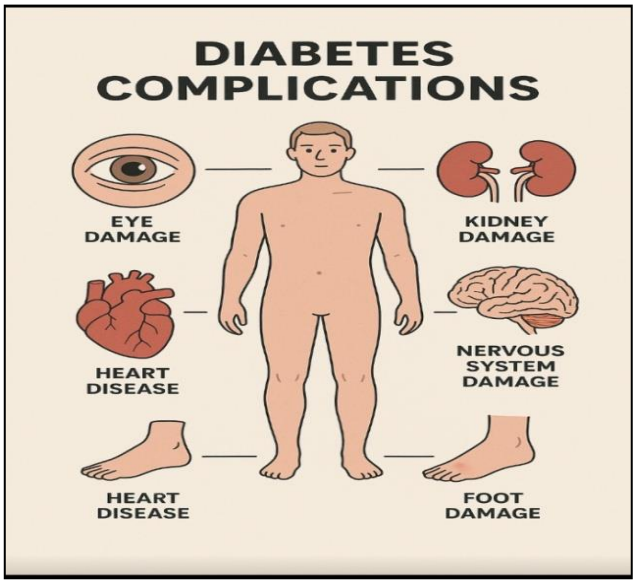

Fig 1 Diabetic Complication
Oputa, R.., & Oputa, P.U. (2024).

➢ *Types of Diabetes:*

In this section, the four major types of diabetes is discussed. Type one (1) Diabetes Mellitus: In past, the words "juvenile diabetes" and "insulin-induced diabetes mellitus" (IDDM) were used. It is unclear what caused it. Young people and adolescents between the ages of 20 are affected by these chronic diseases. Type 1 damage will cause pancreatic cells to become defective Gollapalli et al. [8]. Patients with type 1 diabetes the body stop producing insulin for energy. A patient diagnosed with type 1 diabetes ought to maintain a balanced diet as well regular physical activity.

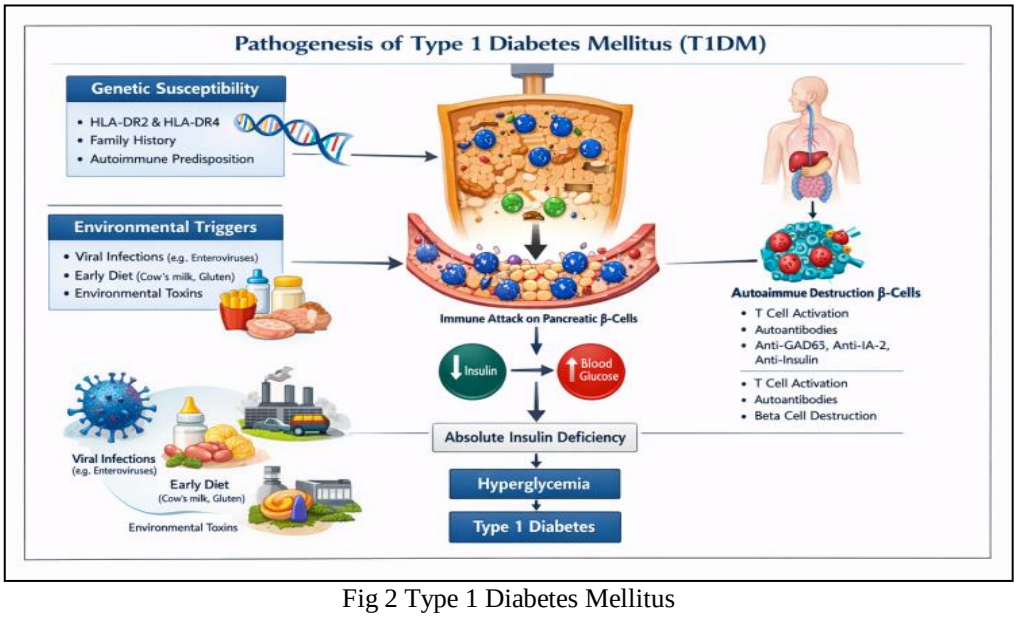

Fig 2 Type 1 Diabetes Mellitus
Dinic et al. (2024)

➢ *Type 2 DM:*

Type 2 diabetes mellitus is mostly caused by insulin resistance, which happens when the body does not use insulin properly and over time, the pancreas cannot produce enough insulin to maintain normal blood glucose levels. As the illness worsens, insulin deficiency could occur. There have been past uses of the terms "adult-induced diabetes" and "non-insulin-based diabetes mellitus." The major causes of this ill health is obesity hypertension and inactivity Ikegami. [7].

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1760

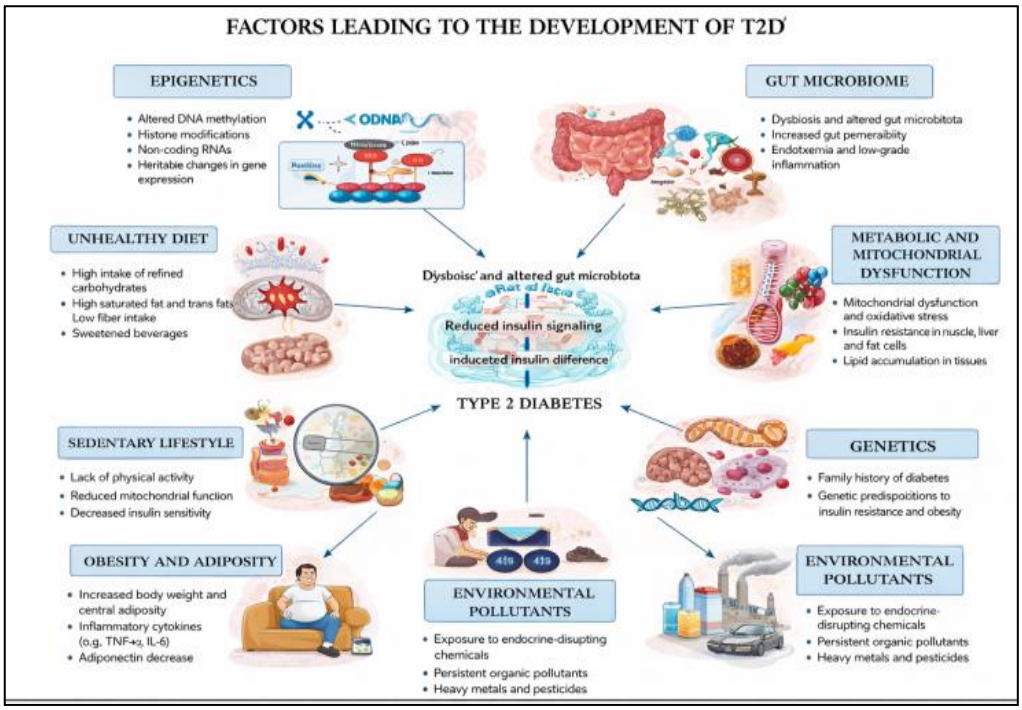

Fig 3 Risk Factors Causing Type 11 Diabetes Mellitus
Singh, S. (2024)

➢ *Gestational Diabetes:*

Higher blood glucose levels cause by the pregnancy hormone in pregnant women without a medical history of diabetes constitutes the third fundamental form of gestational diabetes. Based on the latest diabetic research, around 18% of expectant mothers have the disease. Older women may be more likely to develop gestational diabetes. Overly elevated blood sugar in pregnant women often triggers the third major type of diabetes prior to pregnancy.

- *Pre-Gestational:*

Pregnancy causes diabetes before insulin-dependent diabetes manifests itself. A man with pre-diabetes seems more inclined to score two in these kinds of situations or assessments.

Diabetes mellitus is a chronic health condition that has become a major public health concern due to its rapidly increasing prevalence across the world. The disease affects millions of individuals and poses serious long-term complications if not detected and managed early. In response to this challenge, several computational and machine learning-based models have been developed to predict the occurrence of diabetes using selected patient attributes. However, many of these models demonstrate inconsistent performance, limited predictive reliability, and reduced effectiveness when applied in real clinical settings.

Consequently, the need for a more accurate, reliable, and clinically applicable diabetes prediction model remains critical. This study aims to evaluate and compare the performance of multiple machine learning algorithms—namely Hierarchical Gradient Classifier (HGC), Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), and Naïve Bayes (NB) in predicting diabetes. By analyzing their predictive capabilities, the study seeks to identify the most effective algorithm and highlight the strengths and limitations associated with each approach.

The overarching goal of this research is to support informed algorithm selection for diabetes prediction by prioritizing high classification accuracy, strong generalization ability, and clear feature interpretability. Furthermore, the study investigates the potential clinical usefulness of the proposed predictive models, thereby contributing to the development of efficient and data-driven decision-support systems for diabetes diagnosis and management. Hence, in this study the data set were derived from kaggle. The specified objectives of this study include:

- ✓ To compute and compare the models using appropriate classes of evaluation metrics.
- ✓ To build a massive Machine Learning Model which will be able to predict diabetes risk factor accurately on patients by using the best ML techniques
- ✓ To examine the activeness of the Machine Learning Model developed on diabetes prediction using (8) eight relevant features in the dataset.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1760

Table 1 Summary of Related Works

AUTHORS/YEARYE	TECHNIQUES	PREDICTION ACCURACY	APPLICATION
Chou et al. (2023)	two-class l r, two-class N N, two-class D J, or two-class B D T,	0.991, with two-class boosted decision tree,	They applied MS, ML Studio to train models of various types of N N.
Sarkar and Pawar (2023)	ANN), CATBoost (CGB), XGBoost (XGB), XGBoost-histogram (XGB-h), and LightGBM (LGBM).	XGB-h model outperforms other ML methods in (AUC-ROC)	Mobile app and a website.
Tasin et al. (2022)	SMOTE, SVM, DT, RF, LR, XGBoost, KNN, ADASYN. The explainable AI approach with the LIME and SHAP	with 81% accuracy, 0.81 F1 coefficient,	a website and an Android smart phone app
Menon et al. (2023)	(LDA), (ASV-RF), (PSO)	Results show that the proposed method is more accurate.	M L, for Smart phones, sensors, & devices E-Health App.
Ginting et al. (2023)	RFA, bivariate and multivariate statistical methods with a 5% significance level,	With an accuracy rate of more than 80%. and accuracy rate of 84%/	an early detection system for type 2 diabetes risk factors app.
Islam et al. (2023)	CDSS, KNN DT, NB RF, SVM, and HBGB.	The accuracy of DT and HBGB was greater than 90%.	a web-based user interface to obtain decision support.
Datta et al. (2023)	RF, AdaBoost, and Gboost are combined with LR, DT, and KNN	97+% accuracy	Application of soft and hard. Ensemble learning models
Wee et al. (2023)	Lab-based and invasive test measurements.	Better prediction as the accuracy improves	To develop a cost-effective high-performance solution, anthropometric measurements.

II. MATERIALS AND METHODS

This section presents the methodological implementation for developing the prediction model. The steps involved in designing this system are shown in fig. 4. Also, the tools and technologies used for developing this system are presented the steps are thus explained below.

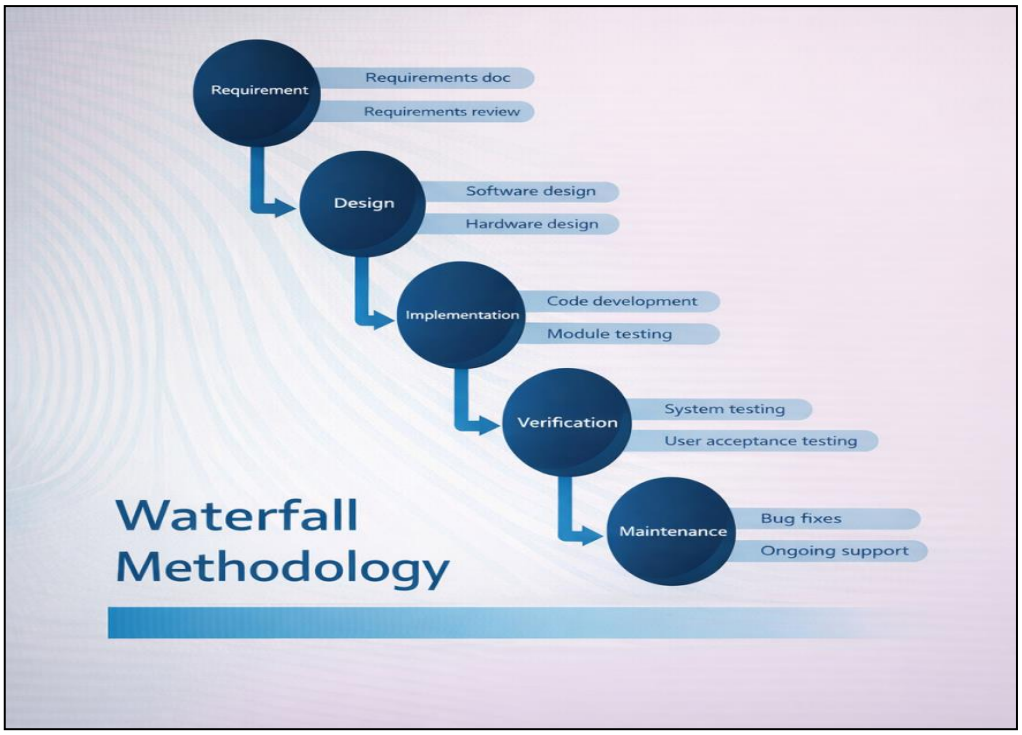

Fig 5 Water Fall Methodology
Khan, S. M. A. (2023)

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1760

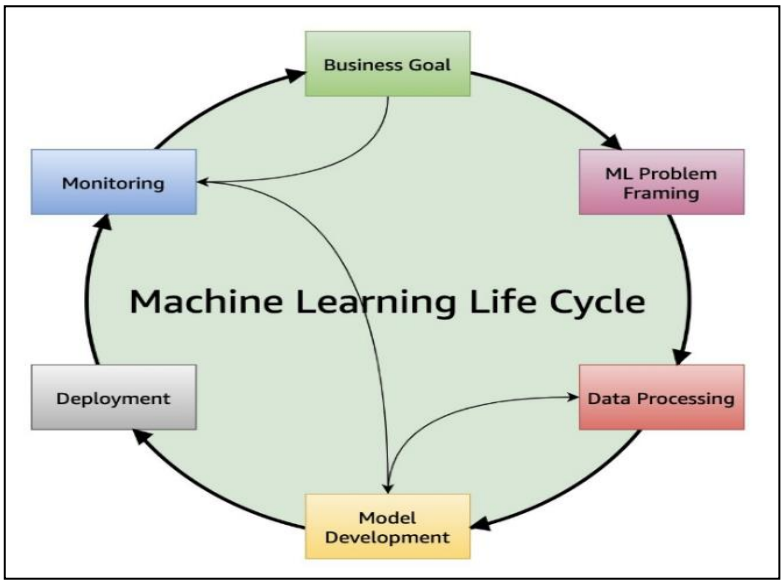

Fig 6 Steps in Machine Learning Life Cycle
Analytics vidhya. (2024)

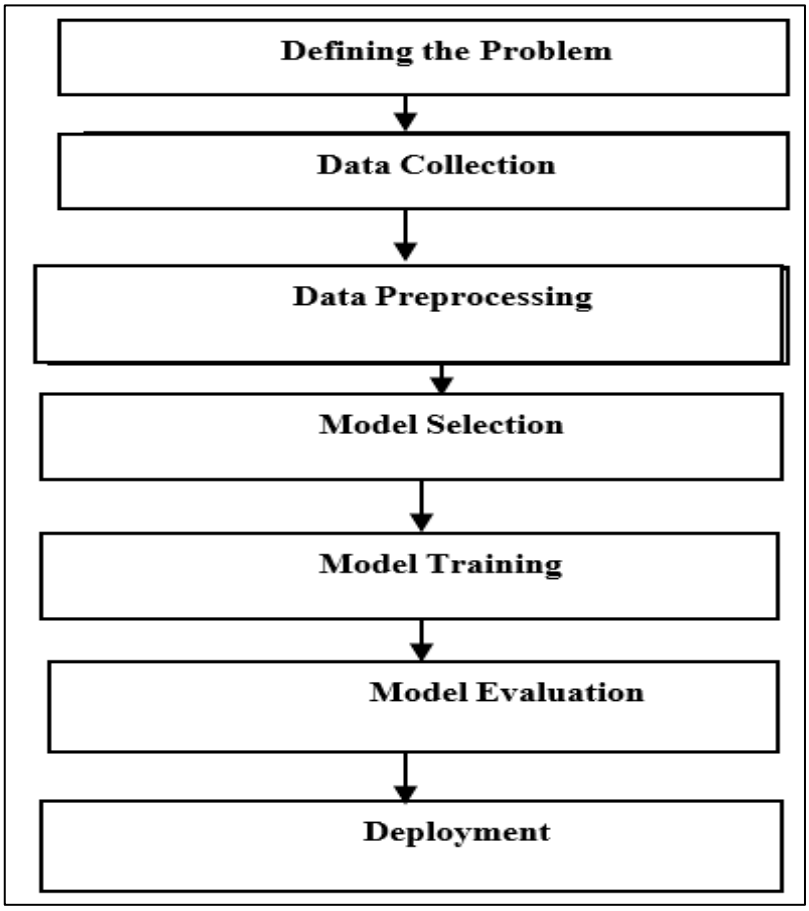

Fig 4 Defining the Problem

The first step in solving any problem is to properly define the problem. In this scenario, the problem identified was to make a comparison study of six (6) Machine Learning Techniques for the predict of diabetes mellitus timely and accurately on patients. Hence, to perform this, the necessary input and output must be considered.

- *Data Collection:*

The next stage is to them collect quality data to profile a solution to the problem identified. The dataset was derived using diabetes detection data set derived from this link; https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1760

➢ *Data Preprocessing*

Next step is to preprocess the data. The data set was preprocessed to modify unnecessary data and replaced it with numerical values to enable computer to understand it for training, among the 9 features of the dataset, it was separated into 8 predicting features and 1 was the target feature that which is to be predicted.

➢ *Model Selection*

The dataset was trained using 6 machine learning algorithms like HGC, SVM, RF, DT, LR, NB.

➢ *Model Training*

Six machine learning algorithms were trained using 80% of dataset while 20% was used for testing the trained models.

➢ *Model Evaluation*

The model was evaluated using the six standard evaluation matrices as stated in the results section.

➢ *Deployments*

The tools and technologies used for the deployment of this system includes: Jupiter Notebook, Visual Studio Code, Python, and the Django framework were the primary tools deployed in the research study. Using Python and the Django framework, the Random Forest classifier and Histogram Gradient Boosting Classifier ware deployed to produce a real-time predictive app that predicts diabetes.

III. DESIGN AND RESULTS

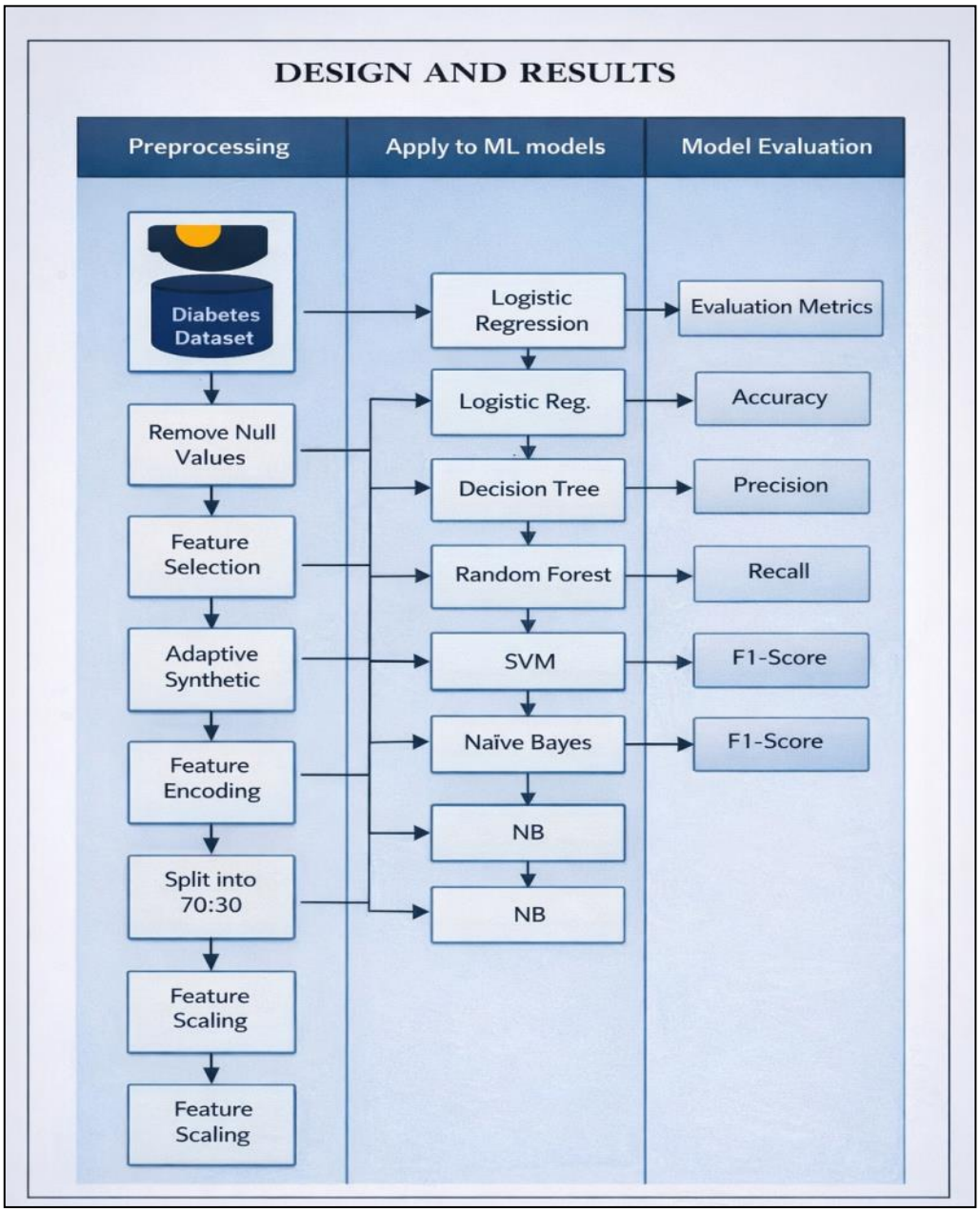

Fig 7 The Application of Machine Learning Models and Evaluation Models

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1760

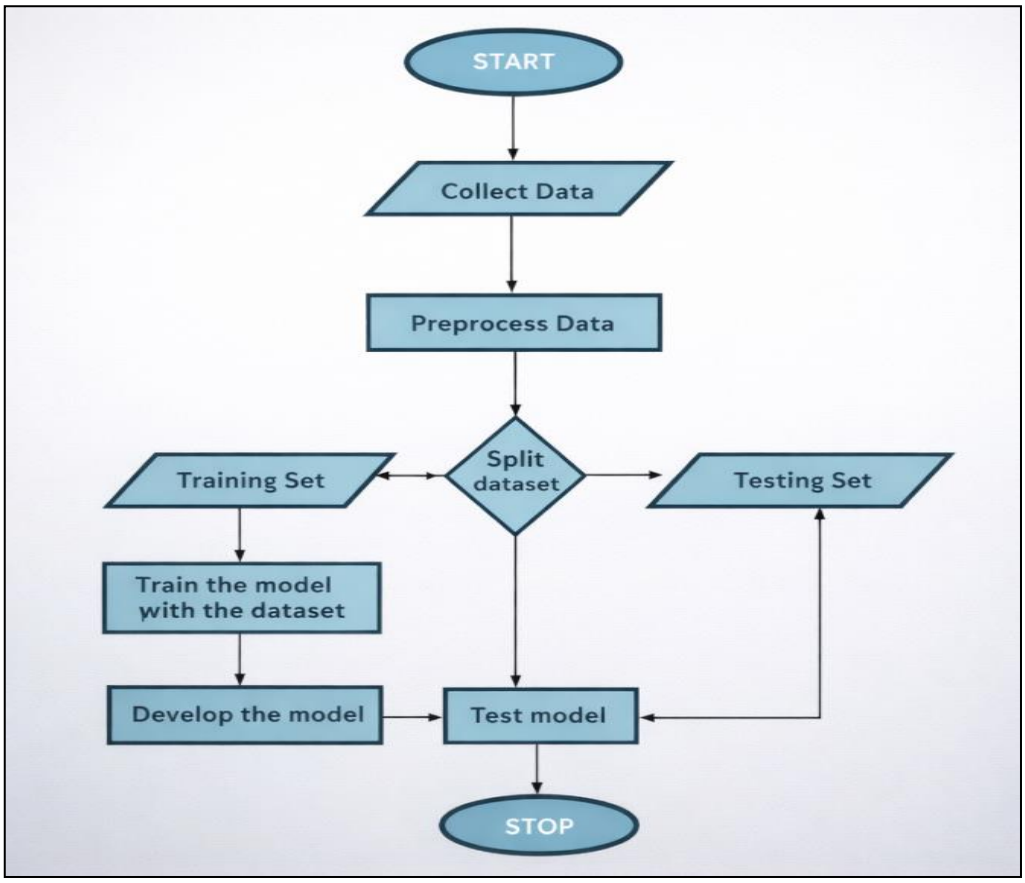

Fig 8 Flowchart of the New System.

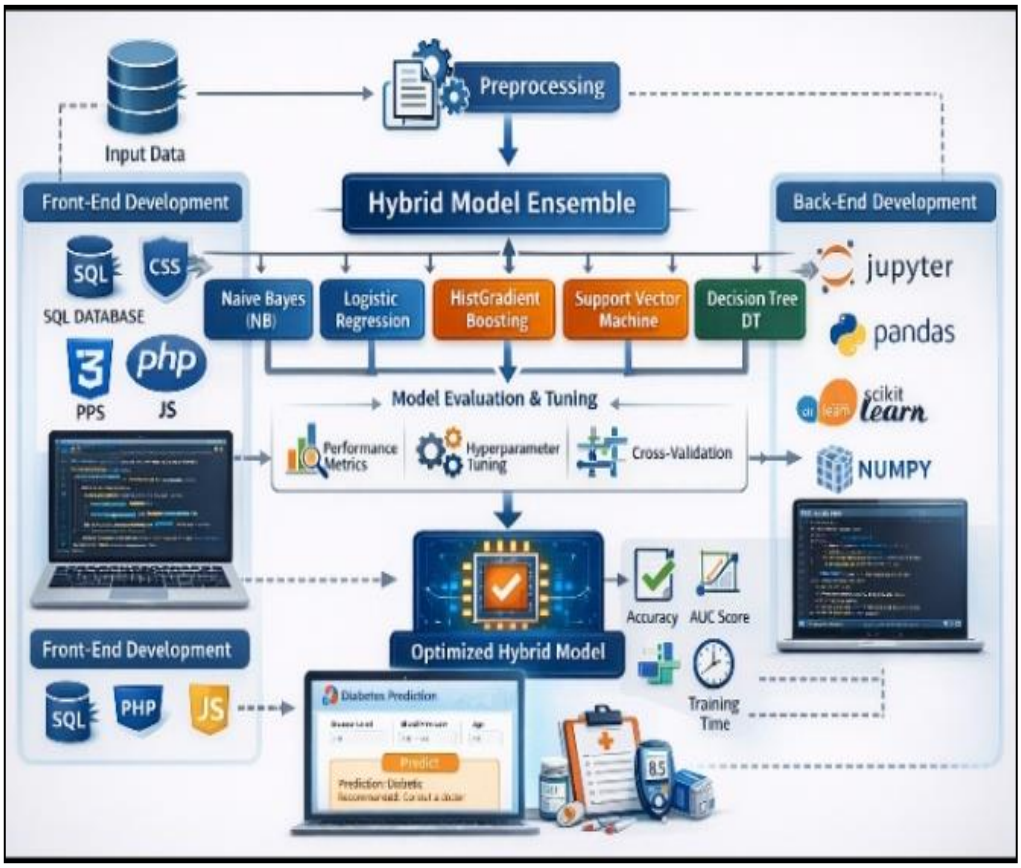

Fig 9 Architecture of the New System

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1760

IV. RESULTS

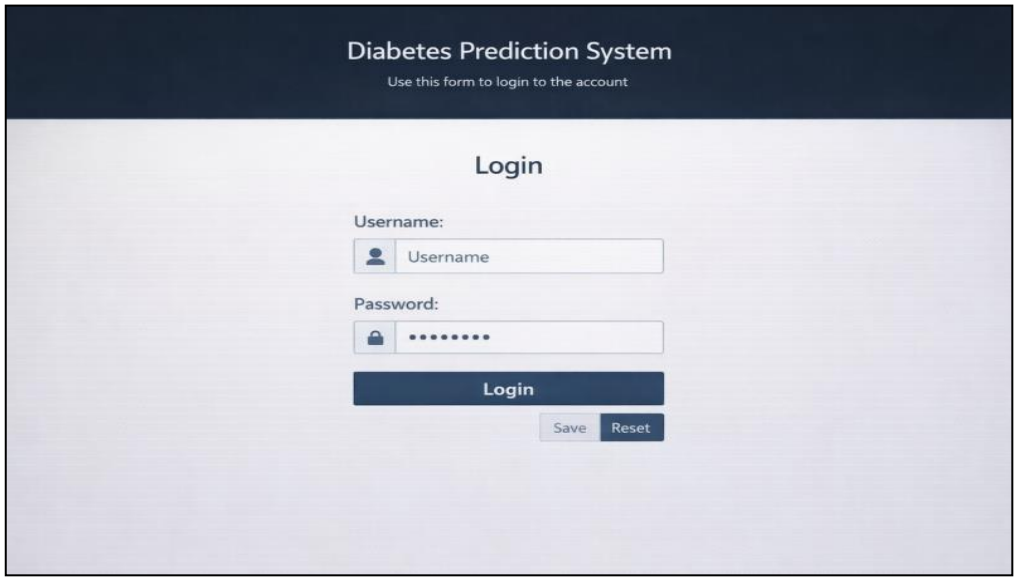

Fig 10 Login Page of The Diaetes Predictiom System

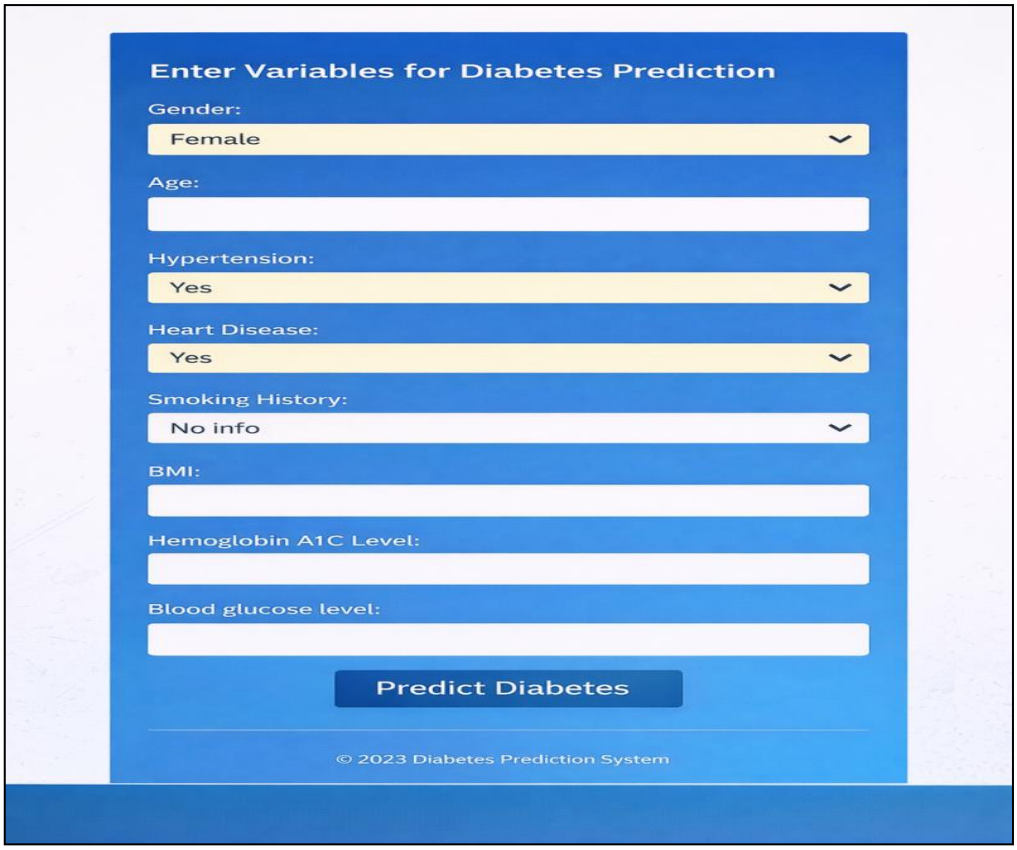

Fig 11 Home Page of the Diabetes Prediction new System.

Table 2 Comparison Table of ML Models Performance

LR	VALUES
TP	18, 129
TN	1,048
FP	163
FN	660
AUC CURVE	0.96

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1760

Table 3 Comparative Analysis of CM & AUC Curves

DT	VALUES
TP	17,798
TN	1,270
FP	494
FN	438
AUC CURVE	0.86

HGBC	VALUES
TP 18269	
TN	1,181
FP	23
FN	527
AUC CURVE	0.98

LR	VALUES
TP	18, 129
TN	1,048
FP	163
FN	660
AUC CURVE	0.96

NB	VALUES
TP	17,020
TN	1088
FP	1,272
FN	624
AUC CURVE	0.92

SVM	VALUES
TP	18,270
TN	975
FP	22
FN	733
AUC CURVE	0.96

RF	VALUES
T P	18,212
T N	1,174
F P	80
F N	536
AUC curve	0.96

Table 4 Comparing the Parameters Values of the Old &New System

S/N	Parameters for comparison	Values old system	Values mew system
	Software development model	not specified	Machine learning approach
	Uses of application	none	1 web application
	Development tools	python	HTML, CSS, Django, Python, and Jupiter note book.
	Machine learning models.	RF, NB, RF, SVM.	RF, LR, HGBC, DT, N B SVM.
	Accuracy	82.46%	0.9724%

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1760

Table 5 Evaluating the Parameters of the New System

S/N	Evaluated parameters	values
1	The software development model	Machine learning approach
2	Number of application (s)	A single web application
3	Development tools	Html, CSS, Django, Python, and Jupiter note book.
4	Machine learning models.	The machine learning models are presented in tale 1
5	best fit machine learning model	Hist. gradient boosting classifiers
6	Accuracy of the best fit machine learning model	0.9724

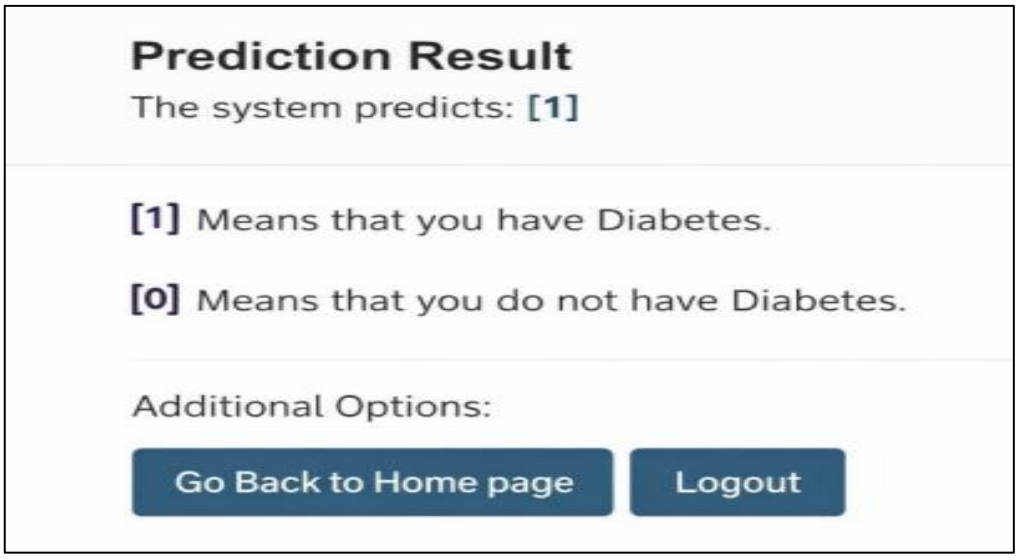

Fig 12 Diabetes Prediction Result Page

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1760

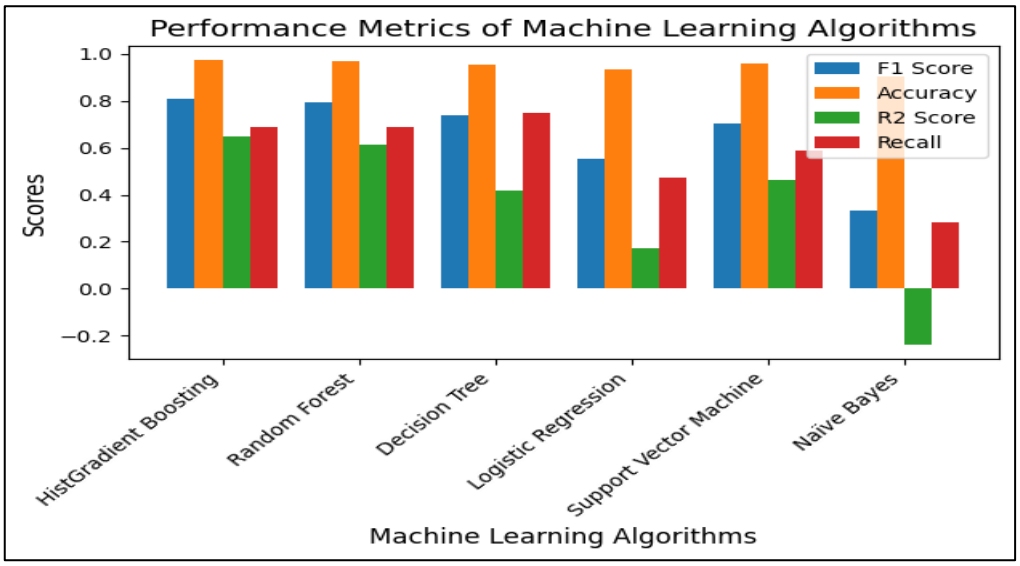

Fig 13 Overall Performance Matrices Comparison of ML Algorithms: Confusion Matrix Results.

V. CONTRIBUTION TO KNOWLEDGE

This research focuses on deploying the Random Forest and Histogram Gradient Boosting Classifier, which is the machine learning algorithms with the highest accuracy among the models developed in this research. In addition, the model was deployed in real-time using Django, a Python framework that allows machine learning models to interact with web applications for seamless application. Furthermore, the user-friendly web application that was created enabled users to make predictions with the system. This study made an important contribution by successfully implementing machine learning technique with the highest accuracy for Diabetes Prediction System, which focuses on predicting in real-time once users input the variables from the dataset needed to predict diabetes.

VI. FUTURE RESEARCH DIRECTION

- Firstly, institutions should consider using either Random Forest or Histogram Gradient Boosting Classifier for tasks requiring high accuracy and robust generalization.
- This study was able to contribute to the system's adaptability over time.
- Logistic Regression is recommended for scenarios where interpretability of results is crucial.
- Future research could explore ensemble techniques that combine the strengths of multiple models to enhance predictive accuracy and mitigate limitations observed in individual models, such as the variability in SVM's performance across different classes.
- More advanced machine learning and deep learning algorithms should also be included.
- Additionally, integrating validated machine learning models into healthcare systems could enable early detection of an individual who is at risk of developing diabetes illness, facilitating targeted solution as well improving overall health outcomes.
- Future work should concentrate on frequent checking and upgrading the model with newest data.

VII. CONCLUSION

These results indicate that the Random Forest and Histogram Gradient Boosting Classifier models displayed superior predictive capabilities, achieving high accuracy metrics across all evaluated classes. These models effectively classified diabetes outcomes without misclassifications, highlighting their suitability for applications requiring high precision and recall. Logistic Regression also showed strong performance, delivering interpretable results with balanced precision-recall metrics and an overall accuracy of 96%. The Support Vector Machine (SVM) demonstrated strong predictive power, particularly in identifying instances of diabetes, although it exhibited some variability in performance across different classes.

REFERENCES

[1]. Jaiswal, V., Negi, A., & Pal, T. (2021). A review on current advances in machine learning based diabetes prediction. Primary Care Diabetes, 15(3), 435-443.
[2]. Rastogi, R., & Bansal, M. (2023). Diabetes prediction model using data mining techniques. Measurement: Sensors, 25, 100605.
[3]. Pradhan, N., Rani, G., Dhaka, V. S., & SPoonia, R. C. (2020). Diabetes prediction using artificial neural network. In Deep Learning Techniques for Biomedical and Health Informatics (pp. 327-339). Academic Press.
[4]. Manikandababu, C. S., IndhuLekha, S., Jeniefer, J., & Theodora, T. A. (2022). Prediction of Diabetes using Machine Learning. In 2022 International Conference on Edge Computing and Applications (ICECAA) (pp. 1121-1127). IEEE.
[5]. Larabi-Marie-Sainte L Aburahman, R Almohaini, T Sab current techniques for doaetes prediction review & case study Applied Science A (21), 4604, 2019.
[6]. Mujumdar &Vaidehi, et al., 2021
[7]. Ikegami, H. Y. Hiromine, S. Noso, Insulin-dependent diabetes mellitus in older adults: current status and

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1760

future prospects, Geriatric. Gerontol. Int. 22 (8) (2022) 549–553. [9] Y. Liu, Q. Wang, K. Wu, Z. Sun, Z. Tan.

[8]. Gollapalli, M. A. Alansari, H. Alkhorasani, M. Alsubaii, R. Sakloua, R. Alzahrani, W. Albaker, A novel stacking ensemble for detecting three types of diabetes mellitus using a Saudi Arabian dataset: pre-diabetes, T1DM, and T2DM, Computer. Biol. Med. 147 (2022), 105757.

[9]. Oputa, R. N., & Oputa, P. U. (2024). Chronic Complications of Diabetes Mellitus: West Africa Journal of Medicine, 41 (8), 904-908.

[10]. Dinic et al. “Type 1 Diabetes Mellitus: retrospect and prospect”. Bulletin of the National Research Center, 2024.

[11]. Singh S. (2024). Deciphering the complex interplay of risk factors I types 2 diabetes mellitus in adults: A review. International Journal of Environmental Research and Publication.

[12]. Khan, S.M.A. (2023). Waterfall Model used I Software Reference: Software Requirements Engineering Waterfall Model [Technical Report SRE-008]. National University of Computer and Emerging Science. D01: 10.13140/RG.2.2.29580.69764.

[13]. Analytics Vidaya. (2024), Machine Learning Lifecycle Explained. Analytics Vidhya Blog. Retrieved from: https;//www.analyticsvidhya.com/log/2021/05/machine learning-life-cycle-explained/.