International Journal of Mathematics and Computer Research
ISSN: 2320-7167
Volume 12 Issue 08 August 2024, Page no. –4398-4403
Index Copernicus ICV: 57.55, Impact Factor: 8.316
DOI: 10.47191/ijmcr/v12i8.05

# Comparative Analysis of Supervised Machine Learning Algorithms for Diabetes Prediction

**Okebule Toyin[1], Jackbara-Johnson Timothy[2], Stephen E. Obamiyi[3], Abiola O, B[4], Opani M. Aweh[5], Atachin A. James[6]**
[1,2,3,4,5,6] Department of Computer Science Afe Babalola University Ado-Ekiti, Nigeria

**ARTICLE INFO**

**Published Online:**
**23 August 2024**

Corresponding Author:
**Okebule Toyin**

**ABSTRACT**

Diabetes Mellitus is a chronic disease with far-reaching consequences, necessitating innovative diagnostic approaches. Current methods have limitations, highlighting the need for advanced techniques. This research leveraged machine learning algorithms to predict diabetes using the Pima Indian Diabetes Dataset. Six algorithms were developed and compared: Logistic Regression, Support Vector Machine, K-Nearest Neighbor, Naive Bayes, Random Forest, and Decision Tree. Standard evaluation metrics assessed their performance. The Random Forest Classifier emerged as the top-performing algorithm, achieving an impressive 91% accuracy and demonstrating exceptional reliability. This study showcases machine learning's potential to transform diabetes diagnosis and management. The developed web application provides a user-friendly interface for predicting diabetes, facilitating timely intervention and improved health outcomes. The findings contribute significantly to the development of intelligent disease prediction systems, promoting early detection and enhanced patient care. By harnessing machine learning, this research pioneers a new frontier in diabetes diagnosis, paving the way for improved patient outcomes and enhanced healthcare services.

**KEYWORDS:** Diabetes, K-Nearest Neighbor, Naïve Bayes, Random Forest, Decision Tree, Dataset.

## 1. INTRODUCTION

Diabetes is a metabolic disorder that impairs an individual’s body to process blood glucose, known as blood sugar. This disease is characterized by hyperglycemia resulting from defects in insulin secretion, insulin action, or both [1]. When we eat food, after the digestion process, glucose gets released. Insulin is a blood hormone that moves from blood to cells and instructs cells to consume blood glucose and transform it into energy. When the pancreas cannot produce enough insulin, the cells cannot absorb glucose, and the glucose remains in the blood. Hence the blood glucose/blood sugar increases in the blood at a very unacceptable level. [2]. According to the IDF Diabetes Atlas Ninth edition, in 2019, approximately 463 million adults (20-79 years) were living with diabetes; by 2045 this will rise to 700 million, the proportion of people with type 2 diabetes is increasing in most countries.79% of adults with diabetes were living in low- and middle-income countries**,** 1 in 5 of the people who are above 65 years old have diabetes, 1 in 2 (232 million) people with diabetes were undiagnosed. Diabetes caused 4.2 million deaths. 10% of total global expenditure spent on diabetes. More than 1.1 million children and adolescents are living with type 1 diabetes, more than 20 million live births (1 in 6 live births) are affected by diabetes during pregnancy, and 374 million people are at increased risk of developing type 2 diabetes [3]. Projections indicate that diabetes will be the 7th major illness condition causing deaths in the global population by 2030. The timely identification and the early detection of the onset of diabetes are, therefore, of potential value in the goal of optimizing treatment pathways, providing a better quality of life for people with diabetes, and reducing the number of deaths owing to the condition. Diabetes is a chronic disease that manifests not only in adults but also in young people. It is a multifactorial disease that requires long-term care since it involves major changes in both physical and psychological dimension of each patient. According to the World Health Organization, diabetes will be the 7th leading cause of death in 2030. This is a major problem worldwide today. Machine learning is a component of artificial intelligence that involves learning of patterns within the data and then using the patterns to classify or predict an event related to the problem [4]. It is a subset of artificial intelligence which uses computerized techniques to solve problems based on historical data and information without unnecessarily requiring modification in the core process [5]. Basically, machine learning algorithms are embedded into machines provided so that knowledge and information are extracted and fed into the system for faster and efficient management of processes. Prediction systems refer to algorithms or models designed to provide insights into future occurrences based on data patterns. These systems are designed to make accurate forecasts or estimates about future events, trends, patterns and relationship between variables by analyzing historical and statistical data Prediction systems are concerned with generating accurate and reliable insights on the likelihood that future events, trends and relationships will occur. Predictive analysis is the process of tuning or training the parameters of a model using data algorithm. They are used to find potentially valuable patterns in the data and predict the outcome of the event. The aim of this study is to evaluate and compare different supervised machine learning algorithms and integrate the best fit into a user friendly web application. The study sources data from the Pima Indian Diabetes Dataset (PIDD) acquired from kaggle.com. This project will implement 6 supervised machine learning techniques; Logistic Regression, Support Vector Machine, K-Nearest Neighbor, Naive Bayes, Random Forest and Decision Tree Classifiers for the prediction model. These classifiers are tested and compared using standard evaluation metrics and the best fit is integrated into a user friendly web application.

## 2. LITERATURE REVIEW

This section to examine other research works and gives a detailed explanation of literature associated with diabetes prediction as well as machine learning and related concepts breaking them down and analyzing different opportunities for our system within their scope.

The authors in [3] proposed a comparative analysis of machine learning algorithms to build a predictive model for diabetes disease where classification in supervised learning machine learning technique was used to predict diabetes mellitus. The training set consisted of some percentage of data that were initially gathered, and these data were used to train the system, using various algorithms which are Decision tree, linear regression, Support vector machine and KNN. However, the correct prediction accuracy of current machine learning algorithms in this study was low. In [6], the prediction of type 2 diabetes mellitus using hematological factors based on machine learning approaches were proposed: a cohort study analysis using MASHAD study dataset to implement Decision Tree and Logistic Regression algorithms. Although, the MASHAD dataset was limited to patients in a particular geographical area in Iran (north-eastern Iran) who conform to a particular habit and lifestyle making it impossible to generalize the result to other countries or even the whole of Iran. The authors in [7] implemented existential risk prediction models for diabetes mellitus adopting an exploratory descriptive approach. The study explored the existential predictive models and subsequently describes them analytically. However, no feature selection or calculation was done to ascertain feature importance. In [8] multiple disease prediction using machine learning were developed. The Cleveland dataset is utilized as input. The dataset was fed into the machine learning classifiers such as SVM, Nave Bayes and Decision tree. A drawback in this study was the insufficient data available for use in prediction. The authors in [9] utilized the development and comparison of three data Models for predicting diabetes mellitus using risk factors in a Nigerian population. This comparative cross-sectional study was conducted among 733 participants aged 18 and older. Participants were recruited from the endocrinology and general outpatient clinics of a large tertiary hospital in Lagos State, Nigeria between September 1, 2019 and December 31, 2019. A minimum sample size of 560 was calculated using the sample size formula for a diagnostic test. Although, the self-reported nature of some of the variables may have been affected by recall or social desirability bias and the cross-sectional nature of the study does not allow for causal inferences. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning was utilized in [10]. This study used dataset acquired from the National Health and Nutrition Examination Survey (NHANES) and various machine learning algorithms to predict variables that are a major cause of diabetes. The authors used ensemble models by combining the performance of weaker models. Their work was further divided into laboratory and non-laboratory dataset. Laboratory results were any feature variables in the dataset that were obtained from blood or urine tests. The exact number of variables used in non-laboratory data is not reported, so it cannot be concluded and their approach might not be useful in general situations. In [11] Implementation of predictive models for diabetes mellitus using machine learning techniques was developed. A predictive model was built to identify Canadian patients at risk of diabetes based on demographic data. Logistic Regression and Gradient Boosting Machine techniques were used in this system. However, the authors did not mention their performance inaccuracy or specificity, thus their performance cannot be generalized.

## 3. METHODOLOGY

This research project adopted a systematic approach to create an expert system for predicting diabetes. The Anaconda platform was leveraged, with Jupiter Notebook serving as the primary development interface. The dataset used was the Pima Indian Diabetes collection in CSV format, and Python was selected as the programming language due to its versatility and extensive machine learning libraries. The model was saved and loaded using the Joblib library, while the user interface was built in the VsCode environment using the Streamlit framework. Several models were developed, evaluated, and compared to select the most effective one, which was then integrated into the web application. This methodology facilitated the creation of a reliable and accurate expert system for diabetes prediction. Figure 1 shows the System Architecture of the study.

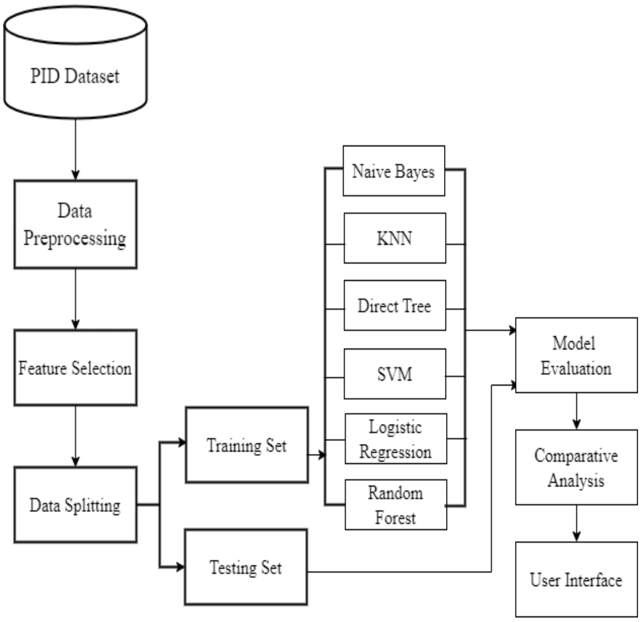


**Fig 1: System Architecture Diagram**

### 3.1 Data Collection and Description

The dataset was downloaded from Kaggle repository. It was downloaded as a csv file and accessed using “pd.read_csv()”. Figure 1 shows the attributes of the datasets.

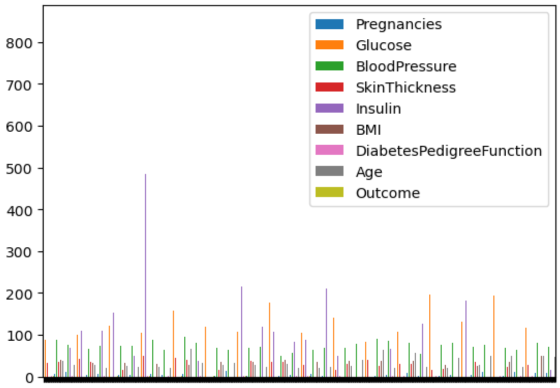


**Fig 2: Bar Chart of Attributes**

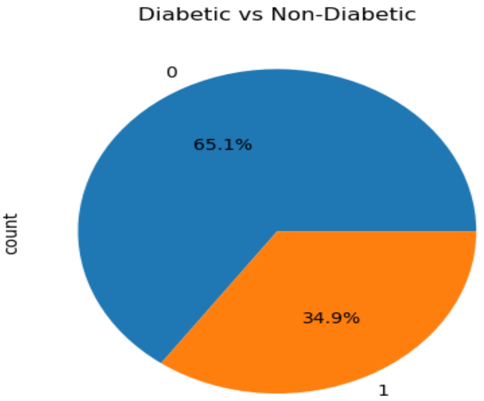


**Fig 3: Pie Chart of Outcome**

Figure 3 shows the outcome of the patients in terms of diabetics and non-diabetics. The results indicate that 65.1% of the patients have diabetes, while 34.9% are non-diabetic.

### 3.2 Data Pre-processing

Data preprocessing is defined as the transformation of unprocessed data to make it appropriate for analysis. The dataset used for this study is the Pima Indian Diabetes Dataset, originally developed by the National Institute of Diabetes and Digestive and Kidney Diseases. The goal of data preprocessing is to clean, transform, and prepare the data in a format suitable for modeling. In this study, the dataset was preprocessed and prepared in a format suitable for modeling.

### 3.3 Feature Selection

Feature selection is the process of selecting a subset of relevant features (variables, attributes, or predictors) from a larger set of features in a dataset. The goal of feature selection is to reduce the dimensionality of the dataset, remove irrelevant or redundant features, and retain only the most informative features for a given analysis or modeling task.

### 3.4 Data Splitting

Data Splitting is the process of partitioning data into training and testing sets to build models and access performance. The training set is used to train the model to learn and understand relationships between data attributes. The testing set is used to evaluate the model’s performance using unseen data to access the understanding. In this study, the data is split using 70% - 30% for training and testing respectively. The data was split using “train_test_split” module in sklearn.model selection library.

### 3.6 Description of the model used:

In this study, six different supervised machine learning algorithms were used for comparative - Naive Bayes Classifier, Decision Tree Classifier, Support Vector Machine, K-Nearest Neighbor Classifier, Logistic Regression and Random Forest Classifier. The descriptions and explanations for each classifier are as follows:

**(a) Naive Bayes Classifier**

Naive Bayes is a classification technique based on the Bayes Theorem with an assumption of independence among predictors. The algorithm creates Bayesian networks which are generated based on the probability of occurrence of outcome based on probabilities imposed on the input variables. It assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

It is represented as follows:

$$\mathrm{P}(\frac{\mathrm{c}}{\mathrm{x}}) = \frac{P(\frac{x}{c}) * P(c)}{P(x)} \qquad (1)$$

Where:

$\mathrm{P}(\frac{\mathrm{c}}{\mathrm{x}}) = \mathit{Posterior\ Probability}$

$\boldsymbol{P(\frac{x}{c}) = Class\ Prior\ Probability}$

$\boldsymbol{P(x) = Predictor\ Prior\ Probability}$

**(b) Decision Tree Classifier**

Direct Tree is a classification algorithm used when the response variable is categorical. This is a logically based algorithm represented as a logical tree having a range of conditions and conclusions as nodes and branches that connects the conditions with conclusions. In figure 4, the nodes represent variable groups for classification and the branches represent the values that the attribute provide.

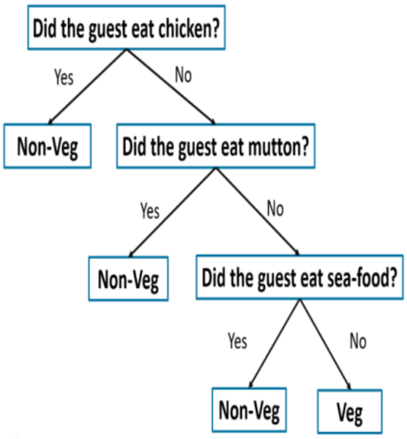

Did the guest eat chicken?
Yes
No
Non-Veg
Did the guest eat mutton?
Yes
No
Non-Veg
Did the guest eat sea-food?
Yes
No
Non-Veg
Veg

**Fig 4: Decision Tree Sample**

### (c) Support Vector Machine

This classifier aims at forming a hyper plane that can separate the classes as much as possible by adjusting the distance between the data points and the hyper plane. It performs a non-linear classification using a “kernel trick”, implicitly mapping their inputs into high dimensional feature spaces.

It is represented as follows:

$$\frac{1}{n}\sum_{i=1}^{n} max([0,1 - y(w.x_i - b)] + \mu || w ||^2) \qquad (2)$$

### (d) K-Nearest Neighbor Classifier

K-Nearest Neighbor is applied to find out the class to which new unlabeled objects belong. For this, a “k” is decided (where k is number of neighbors to be considered) which is generally odd and the distance between the data points that are nearest to the objects is calculated using equations like Euclidean’s distance.

The Euclidean’s distance between 2 points, P and Q is defined by the following equation:

$$d(P, Q) = \sum_{i=1}^{n}(P_i - Q_i)^2 \qquad (3)$$

### (e) Logistic Regression

Logistic Regression model is used where the dependent variable is categorical. The model is used to estimate the probability of the response variable based on one or more predictor variables. The aim of logistic regression is to best fit which is responsible for describing the relationship between target and predictor variable.

It is represented using the following equation:

$$\boldsymbol{f(x)} = \frac{\mathbf{L}}{\mathbf{1 + e^{-k(x + x_0)}}} \qquad \mathbf{(4)}$$

Where:

$X_0$ = *The value of the sigmoid's midpoint*

$L$ = *The curve's maximum point*

$K$ = *The logistic regression growth rate or steepness of the curve*

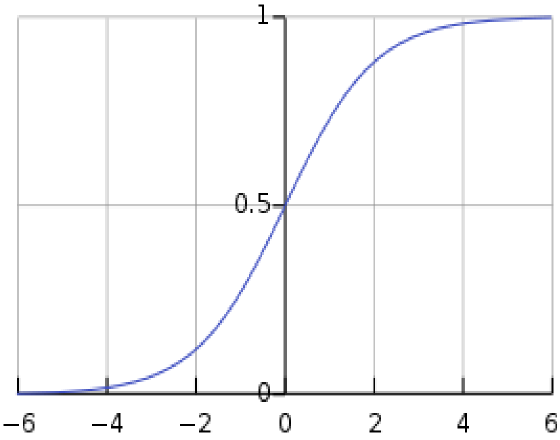


**Fig 5: Logistic Regression Curve**

### (f) Random Forest Classifier

Random Forest Classifier is an ensemble learning method that operates by constructing a multitude of decision trees at training time and outputs the class that is the mode of the classes or classification or regression of the individual trees. It uses a bootstrap randomized re-sampling method to extract multiple versions of the sample sets from the original training sets.

The Random Forest finds the best split using the Gini - Index Cost Function which is given by:

$$\text{Gini} = \sum_{k=1}^{n} P_k * (1 - P_k) \qquad (5)$$

Where:

$\boldsymbol{k}$ = ***Each Class***

$P$ = *Proportion of training instances*

## 4. RESULTS AND DISCUSSION

The system testing was carried out after the system was developed using the training set. The testing was aimed at validating the ability of the system to predict the diabetic status of patients using their vital signs and using the best algorithms to enhance the performance of the prediction system and also to ensure that all the components being coded and put together worked effectively. In this study, six algorithms were built and trained using the diabetes prediction dataset. They are compared using the evaluation metrics; Accuracy, Precision, F1 score and Recall.

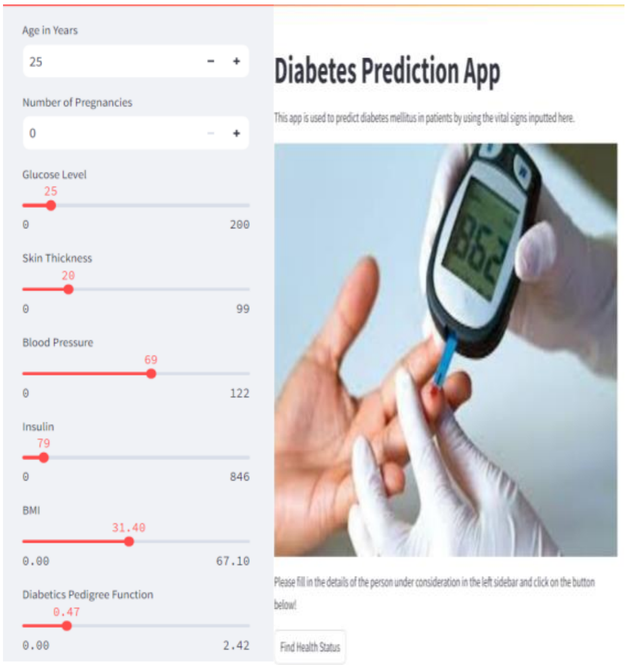
Table 1 shows the comparative analysis of the machine learning models employed in this study.

**Table 1: Comparative Analysis of Machine Learning Model**

| Model | Accuracy | Precision | F1 Score | Recall |
|---|---|---|---|---|
| Naive Bayes | 0.79 | 0.67 | 0.64 | 0.62 |
| Decision Tree | 0.84 | 0.84 | 0.76 | 0.69 |
| Support Vector Machine | 0.77 | 0.53 | 0.61 | 0.73 |
| K-Nearest Neighbour | 0.81 | 0.71 | 0.73 | 0.74 |
| Logistic Regression | 0.77 | 0.69 | 0.67 | 0.65 |
| Random Forest | 0.91 | 0.86 | 0.87 | 0.89 |

The performance of the classification model used is measured using evaluation metrics. This entails training the model on the diabetes prediction dataset and using the model to predict the probability of diabetes in patients based on stated inputs. The predictions are compared with the actual values to check the validity of the predictions made. Based on the given evaluation metrics, the Random Forest Classifier is the most suited for the diabetes prediction system with the highest evaluation metrics recorded.

In Fig. 6, after inputting the vital signs, the application provides a health status indication, which in this case, indicates that the patient is healthy.



Age in Years
25
Number of Pregnancies
0
Glucose Level
25
0 200
Skin Thickness
20
0 99
Blood Pressure
69
0 122
Insulin
79
0 846
BMI
31.40
0.00 67.10
Diabetics Pedigree Function
0.47
0.00 2.42

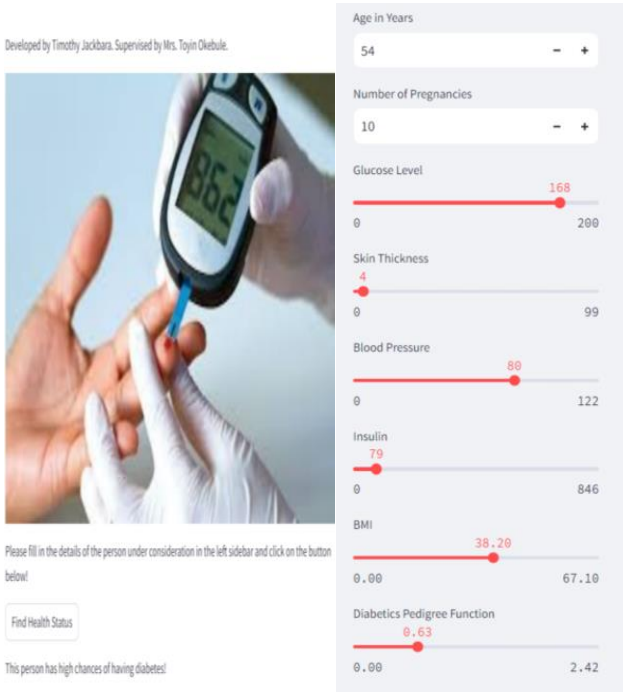
Diabetes Prediction App

This app is used to predict diabetes mellitus in patients by using the vital signs inputted here.

Please fill in the details of the person under consideration in the left sidebar and click on the button below!

**Figure 6: Diabetics Patient Status A**

In Fig. 7, after inputting the vital signs, the application provides a health status indication, which in this case, shows that the person is at risk of having diabetes.



Developed by Timothy Jackbara. Supervised by Mrs. Toyin Okebule.

Please fill in the details of the person under consideration in the left sidebar and click on the button below!

This person has high chances of having diabetes!

Age in Years
54
Number of Pregnancies
10
Glucose Level
168
0 200
Skin Thickness
4
0 99
Blood Pressure
80
0 122
Insulin
79
0 846
BMI
38.20
0.00 67.10
Diabetics Pedigree Function
0.63
0.00 2.42

**Figure 7: Diabetics Patient status B**

### 3.7 Web Application

Fig 7. Shows the landing page of the web application. This is the first page the user see when they visit the web address. By the left, it shows the input section for the user to input his/her vital signs. To aid user friendliness, the home page includes a sentence explaining the page functionality, a set of instructions on how to operate the application and a sidebar with numerical limits to help patients fill in accurate data.

## 3 CONCLUSION

In this project, the system created is a diabetes prediction system which comprises of a Random Forest Classifier model which was trained using the Pima Indian diabetes dataset. The model was tested on the remaining dataset and produced an accuracy of 0.90909 (91%), a precision of 0.85714 (86%), an F1 score of 0.87272 (87%) and a recall of 0.88888 (88%). The model was then integrated into a user friendly web application. Although, some improvements are still possible, such as the use of unsupervised machine learning methods and deep learning algorithms for future study.

## REFERENCES

1. Butt, U. M., *Letchmunan*, S., Ali, M., Hassan, F. H., Baqir, A., & Sherazi, H. H. R. (2021). Machine learning based diabetes classification and prediction for healthcare applications. *Journal of healthcare engineering*, *2021*.

2. Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *Ict Express*, *7*(4), 432-439.
3. Victoria, O. O., and Adeyemo, A. F. (2021). A Comparative Analysis of Machine-Learning Algorithms to Build A Predictive Model for Diabetes.
4. Alpaydın, E. (2014). *Introduction to machine learning*. Cambridge, MA: MIT Press. *8*(18), 2100275.
5. Sandhu, T. H. (2018). Machine learning and natural language processing—A review. *International Journal of Advanced Research in Computer Science, 9*(2), 582–584.
6. Mansoori, A., Sahranavard, T., Hosseini, Z. S., Soflaei, S. S., Emrani, N., Nazar, E., and Mobarhan, M. G. (2023). Prediction of type 2 diabetes mellitus using hematological factors based on machine learning approaches: a cohort study analysis. *Scientific Reports*, *13*(1), 663.
7. Ahmed, U., Issa, G. F., Khan, M. A., Aftab, S., Khan, M. F., Said, R. A., and Ahmad, M. (2022). Prediction of diabetes empowered with fused machine learning. *IEEE Access*, *10*, 8529-8538.
8. Arumugam, K., Naved, M., Shinde, P. P., Leiva-Chauca, O., Huaman-Osorio, A., & Gonzales-Yanac, T. (2023). Multiple disease prediction using Machine learning algorithms. *Materials Today: Proceedings*, *80*, 3682-3685.
9. Odukoya, O., Nwaneri, S., Odeniyi, I., Akodu, B., Oluwole, E., Olorunfemi, G., and Osuntoki, A. (2022). Development and Comparison of Three Data Models for Predicting Diabetes Mellitus Using Risk Factors in a Nigerian Population. *Healthcare informatics research*, *28*(1), 58-67.
10. Dinh, A.; Miertschin, S.; Young, A.; Mohanty, S. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Med. Inform. Decis. Mak.* 2019, *19*.
11. La, H.; Huang, H.; Keshavjee, K.; Guergachi, A.; Gao, X. Predictive models for diabetes mellitus using machine learning techniques. *BMC Endocr. Disord.* 2019, *19*, 1–9.
12. Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR). [Internet]*, *9*(1), 381-386.
13. Rani, K. J. (2020). Diabetes prediction using machine learning. *International Journal of Scientific Research in Computer Science Engineering and Information Technology*, *6*, 294-305.
14. Soni, M., & Varma, S. (2020). Diabetes prediction using machine learning techniques. *International Journal of Engineering Research & Technology (Ijert) Volume*, *9*.
15. Abaker, A. A., and Saeed, F. A. (2021). A comparative analysis of machine learning algorithms to build a predictive model for detecting diabetes complications. *Informatica*, *45*(1).