iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VII **Month of publication:** July 2026

DOI: https://doi.org/10.22214/ijraset.2026.84334

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VII July 2026- Available at www.ijraset.com

Diabetes Prediction System Using Machine Learning

Sunidhi[1], Mr. Rahul[2], Mr. Parmod Kumar[3], Mr. Upasna Setia[4]
[1]*MTech CSE, Department of Computer Science Engineering, Ganga Institute of Technology and Management, Kablana, Jhajjar*
[2, 3, 4]*Assistant Professor, Department of Computer Science Engineering, Ganga Institute of Technology and Management, Kablana, Jhajjar*

Abstract: ***Healthcare is one of the most significant application domains of Machine Learning, where early disease prediction can help improve patient outcomes and support clinical decision-making. This dissertation presents a Diabetes Prediction and Analysis System Using Machine Learning that predicts the likelihood of a disease based on various patient health parameters and medical records. The system utilizes a healthcare dataset containing attributes such as glucose level, blood pressure, body mass index (BMI), insulin level, age, and other relevant medical factors.***
The collected data is pre-processed through missing value handling, feature normalization, and data partitioning to enhance prediction performance. Multiple Machine Learning algorithms, including Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Decision Tree, and Random Forest, are employed to develop predictive models. The performance of these models is evaluated using metrics such as accuracy, precision, recall, F1-score, and confusion matrix. Comparative analysis is carried out to identify the most suitable algorithm for disease prediction.
The implementation of the proposed system is carried out in MATLAB, utilizing its Machine Learning and data analysis tools for model training, testing, performance evaluation, and result visualization.
Experimental results demonstrate that Machine Learning techniques can effectively predict disease occurrence with high accuracy, thereby assisting healthcare professionals in early diagnosis and treatment planning. The proposed system provides an efficient and reliable approach for disease prediction and analysis, contributing to improved healthcare management and decision support.
Keywords: Disease Prediction, Machine Learning, Healthcare Analytics, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Decision Tree, Random Forest, Medical Diagnosis, Classification, MATLAB.

I. INTRODUCTION

Traditional disease diagnosis and treatment involve doctors and diagnostic labs. This type of disease diagnosis and treatment mainly depend upon doctor`s experience and lab reports. Sometimes inexperienced doctors and intermittent reports prove to be harmful for a patient.

Modern healthcare system involves Artificial Intelligence (AI) and Machine Learning (ML) in order to implement automated system that is able to accurately detect a disease at an early stage and suggest proper treatment. These automated systems help a doctor to predict a disease accurately and prescribe proper medicine.

With the advancement of technology wearable medical equipment such as smart watch, temperature sensors, medical image processors are being used so now smart health monitoring and prediction systems are feasible. AI & ML based disease prediction system make use of a large number of previously available patient data in order to accurately predict a disease at an early stage and help doctor for providing better treatment of the disease. These automated systems are emerging as cheap and efficient option for doctors and patients with respect to prediction and treatment of a disease.

The proposed dissertation, “Diabetes Prediction and Analysis System using Machine Learning” aims at developing a smart healthcare system that is capable of predicting diabetes. The system takes input in the form of patient health parameters such as Pregnancy, BMI, Glucose, Blood pressure, Insulin and so on to predict whether a patient is diabetic or not. There are various Machine Learning techniques which can be employed such as Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Decision Tree. In this work SVM, KNN and Decision tree algorithms are employed for diabetes prediction and their accuracy is compared.

II. ROLL OF MACHINE LEARNING IN HEALTH CARE

Machine Learning is a branch of Artificial Intelligence that enables systems to learn patterns from data and make predictions without explicit programming. In healthcare, machine learning algorithms process large volumes of patient data to identify hidden patterns associated with diseases.

Some commonly used machine learning algorithms in disease prediction include:

- Support Vector Machine (SVM)
- Decision Tree
- Random Forest
- Logistic Regression
- K-Nearest Neighbor (KNN)
- Naïve Bayes

These algorithms are trained using historical datasets and later tested on new patient data to predict diseases accurately. The performance of models is evaluated using parameters such as accuracy, precision, recall, sensitivity, and specificity.

III.SCOPE OF RESEARCH WORK

The proposed system can be used in hospitals, clinics, healthcare centres, and remote healthcare applications. The project aims to assist medical professionals by providing quick preliminary diagnosis results. In future, the system can be extended to include additional diseases, deep learning techniques, cloud-based healthcare monitoring, and real-time patient data analysis using IoT devices.

IV. LITERATURE REVIEW

Existing system of disease prediction and analysis is highly dependent on diagnostic laboratory test reports and doctors' expertise. Indian people in remote areas have limited access to experienced doctors and hospitals. Smart AI and ML based diagnostic systems make use of huge amount of patients' data (dataset) to help a doctor in accurate prediction of a disease. These days researchers' have been developing smart diagnostic systems for accurate prediction of diseases such as algorithms such as Heart Disease, Diabetes, Kidney disease, Liver disease and so on. Such smart disease prediction systems have been using various ML algorithms such as Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), K-Nearest Neighbour (KNN), Logistic Regression (LR), and Artificial Neural Networks (ANN). This chapter presents review of important research contributions in the field of smart disease prediction systems.

P. Gandhi and S. Gayathri [1] proposed a diabetes prediction model using machine learning techniques such as SVM, Decision Tree, and Naïve Bayes. The main goal of their research involves improvement of accuracy od diabetes prediction. They performed comparative study

M. Manimekalai et al. [2] have employed Artificial Neural Network algorithm in order to build automated diabetes prediction system. The study have used Pima Indians dataset in order to predict diabetes effectively with 92% accuracy.

M. Samer and S. Naser [3] presented an Artificial Neural Network-based diabetes prediction system. Their proposed model for Neural Network aims at achieving more accuracy in the disease prediction by reducing error during training of the neural network.

K. S. Charan and K. Mahendranath [4] proposed a heart disease prediction model using Random Forest algorithm in their ML model in order to achieve more accurate model for heart disease prediction. Experimentally their Random Forest Model achieved 92% accuracy in prediction of heart disease.

Chaurasia and Pal [5] applied data mining techniques for early prediction of heart disease. They employed three data mining algorithms CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and decision table (DT) extracted from a decision tree or rule-based classifier to develop the prediction models using a large dataset.

Detrano et al. [6] introduced a probability-based automated model which will be able to take decision in the clinical field for coronary artery disease diagnosis. It is a support system for coronary artery disease diagnosis. Their model involved such a method which is based on use of heart disease probabilities. Their model is highly reliable in prediction of heart diseases in patients who have symptoms like pain in chest. Their model is very accurate and reliable in heart disease prediction.

Kora and Kalva [7] proposed automated (ML-Based) platform for prediction of kidney disease which is chronic in nature. They have employed Support Vector Machine Method of Machine Learning for accurate prediction of kidney disease. They also used Ant Colony Optimization method in prediction of kidney disease. They claimed to have very accurate prediction results. Their system is very useful in prediction of chronic kidney disease.

M. Komi et al. [8] proposed diabetes prediction methods using data mining techniques in order to predict diabetes. They used GMM, SVM, Logistic regression, ELM, ANN in order to perform diabetes prediction. They proved that ANN (Artificial Neural Network) model has achieved maximum accuracy in prediction of diabetes disease.

T. Mohanapriya et al. [9] have presented an Artificial Neural Network (ANN) method for prediction of Liver disease. Their model used comprehensive dataset containing huge amount of medical data related to patients age, gender, total amount of bilirubin, direct amount of bilirubin, alkaline phosphatase, and so on. Their model showed highly accurate prediction of liver disease.

Velu et al. [10] presented a Machine Learning based model for prediction of Liver disease. Their model provides a data interface system for patients for prediction of the disease. In this data interface platform patients medical information can be stored and then using various data visualization techniques and data sets of patient information, their disease can be predicted.

Velu et al. [10] presented a Machine Learning based model for prediction of Liver disease. Their model provides a data interface system for patients for prediction of the disease. In this data interface platform patients medical information can be stored and then using various data visualization techniques and data sets of patient information, their disease can be predicted.

Chen et al. [11] proposed a disease prediction model which make use of patients structured as well as non-structured data from hospitals. Their model is based on convolutional Neural Network. They have proved that their model was highly reliable and accurate. Their model can be used accurately in order to predict multi diseases. They have devised CNN-based unimodal for disease prediction. They have claimed that their proposed modal achieves 94.8% accuracy in prediction of a disease.

Jiang et al. [12] reviewed work on automated smart systems for disease prediction. These proposed models are based on AI-ML methods. A comparative study of various such techniques have been reviewed. They have studied various ML algorithms including SVM, modern deep learning as well as natural language processing for unstructured data.

B. J. Sunday et al. [13] have presented a model for health care which make use of medical data taken from various hospitals. Therefore, huge number of patients data was used to make the model for disease prediction. They have shown performance of various machine learning models such as SVM, Random Forest, KNN and Neural Network in the development of their model of disease prediction.

L. P. Sharn et al. [14] presented comparison of various machine learning based techniques for prediction of various diseases such as heart disease, Liver disease, diabetes. They have proposed a disease prediction model which used Decision Tree classifier, Random forest classifier, and Naïve Bayes classifier. They have also presented the comparative study of the results of the above algorithms used.

K. A. Parane et al. [15] introduced a Cloud based healthcare system that consists of a computing device and number of sensors mounted on patient's body. They have presented such a health Care system where patient wears health monitoring devices that are IOT enabled. These devices send patient's data to cloud for processing by doctors at hospitals.

B. Godi et al. [16] have presented IOT based automated model for disease prediction. Patients can send their health parameters to main server using their nodes that is computer systems and seek medical advice from their model of disease prediction, They claimed to give accurate disease prediction. Patients from remote area will mainly get benefit from this chine learning (ML) techniques to design an advanced automation system. Using this system it will connect, monitor and decisions making for proper diagnosis.

Tomar and Agarwal [17] reviewed various data mining and machine learning approaches used in healthcare systems and highlighted their advantages in disease diagnosis applications.

S. A. Tanim et al. in [18] have proposed machine learning based type-2 diabetes prediction system. They have used three types of data sets separately to comparatively study the performance of their proposed automated prediction system for diabetes. They achieved test accuracy above 90% in all the three cases. They proved that performance of Deepnetx2 is much more superior than existing model.

V. THE PROPOSED DIABETES PREDICTION SYSTEM

A. Introduction

This chapter explains how the proposed Diabetes Prediction and Analysis System using Machine Learning was developed, implemented and tested. The system has been developed and implemented using MATLAB programming and by using various machine learning algorithms such as Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Decision Tree. The primary goal of the proposed model is to determine whether an individual is likely to have diabetes by analyzing various medical parameters such as blood glucose level, BMI, age of the individual, number of pregnancies, insulin level etc.

The implementation process involves several stages such as collecting the dataset, preprocessing, feature scaling, model training, testing, prediction, and performance evaluation. The obtained results are then analyzed using confusion matrices and graphical comparison plots. Using these matrices and plots the effectiveness of the models is evaluated which helps in comparing their performance and accuracy of the different models. The evaluation methods help in measuring the effectiveness, accuracy and performance of each algorithm making it easier to compare and identify the most suitable model for diabetes prediction.

B. Dataset Discription

The diabetes dataset used in the proposed work is collected from the PIMA Indian Diabetes Dataset. The PIMA Indian Diabetes Dataset is a widely used dataset for diabetes prediction research. The dataset contains several medical attributes that are useful for predicting diabetes condition among patients.

TABLE I
The dataset consists of the following attributes:

Attribute	Description
Pregnancies	Number of pregnancies
Glucose	Plasma glucose concentration
Blood Pressure	Diastolic blood pressure
Skin Thickness	Thickness of the triceps skin fold thickness
Insulin	Serum insulin level
BMI	Body Mass Index
DPF	Diabetes Pedigree Function
Age	Age of the patient
Outcome	Diabetic or non-diabetic

The outcome variable contains:

- 1 → Diabetes Detected, indicates that the patient is diabetic
- 0 → No Diabetes, indicates that the patient is non-diabetic

This dataset serves as a foundation for training and testing the various machine learning models which have been used in the proposed diabetes prediction system.

C. Proposed work

The proposed work is implemented using MATLAB and consists of the following stages:

Dataset Loading

1. Data Preprocessing
2. Dataset Splitting
3. Feature Scaling
4. Machine Learning Model Training
5. Prediction and Testing
6. Performance Evaluation
7. New Patient Prediction

D. MATLAB Implementation

1) Dataset Loading: The diabetes dataset is loaded into MATLAB using the readtable() function. Which allows the data to be stored in structured tabular form for future analysis.

```
data = readtable('diabetes.csv');
```

The first few dataset samples are displayed using:

```
disp(head(data));
```

2) Feature and Label Separation: After loading the dataset input features and output labels are separated as follows:

```
X = data{:,1:end-1}; Y = data{:,end};
```

where:

- X represents input features
- Y represents output class labels

The input features contain the medical attributes which are used for diabetes prediction. While the output class labels represents the diabetes status of each patient. This separation is needed for better evaluation and training of different machine learning models.

3) Missing Value Handler: In the dataset missing values are replaced with zero values by using the following MATLAB function:

```
X = fill missing (X,'constant',0);
```

Proper handling of missing values helps in maintaining consistency within the database and also helps in improving the performance of the model.

4) Data Splitting: After preprocessing the dataset is divided into training and testing subsets using the Hold-Out validation approach.

```
cv = cv partition(size(X,1),'HoldOut',0.2);
```

The dataset distribution is:

- 80% Training Data
- 20% Testing Data

Typically, 80% of the data is used for training the machine learning models while remaining 20% is reserved for testing. This division of dataset allows the model to learn from one portion of the data and the models are tested and evaluated on basis of test dataset hence providing fair assessment of their capability of predicting diseases.

Feature normalization is performed using Z-score normalization to improve classification performance. Feature scaling is an important preprocessing step in machine learning that ensures all input features are on a similar scale. In this work, **Zscore normalization (Standardization)** is applied to improve the performance and convergence of machine learning algorithms. The method transforms the original feature values into standardized values with a mean of 0 and a standard deviation of 1.

The normalization equation is:

$$z = \frac{x-\mu}{\sigma} \quad (3)$$

For example consider:

$$x = 1.2,\ \mu = 0.0,\ \sigma = 1.0 \quad (4)$$

Substituting these values into the normalization equation:

$$z = \frac{x-\mu}{\sigma} \approx 1.2 \quad (5)$$

Thus, the normalized value is:

$$z \approx 1.2 \quad (6)$$

The cumulative probability corresponding to the standardized value can be obtained from the standard normal distribution table:

$$\phi(z) \approx 88.5\% \quad (7)$$

This indicates that approximately 88.5% of the observations in a standard normal distribution lie below a Z-score of 1.2.

a) The machine learning models considered in this study are:

b) Support Vector Machine (SVM) – A supervised learning algorithm that finds the optimal decision boundary to separate different classes with maximum margin.

c) Random Forest (RF) – An ensemble learning method that combines multiple decision trees to improve classification accuracy and reduce overfitting.

d) Decision Tree (DT) – A tree-structured classifier that makes decisions based on feature values and generates interpretable classification rules.

e) K-Nearest Neighbour (KNN) – An instance-based learning algorithm that classifies a new sample based on the majority class of its nearest neighbours.

These algorithms are trained and tested using the prepared dataset, and their performances are compared using evaluation metrics such as accuracy, precision, recall, F1-score, and confusion matrix. The comparative analysis helps in selecting the most suitable model for accurate disease prediction.

- Diabetes Prediction using SVM Model:

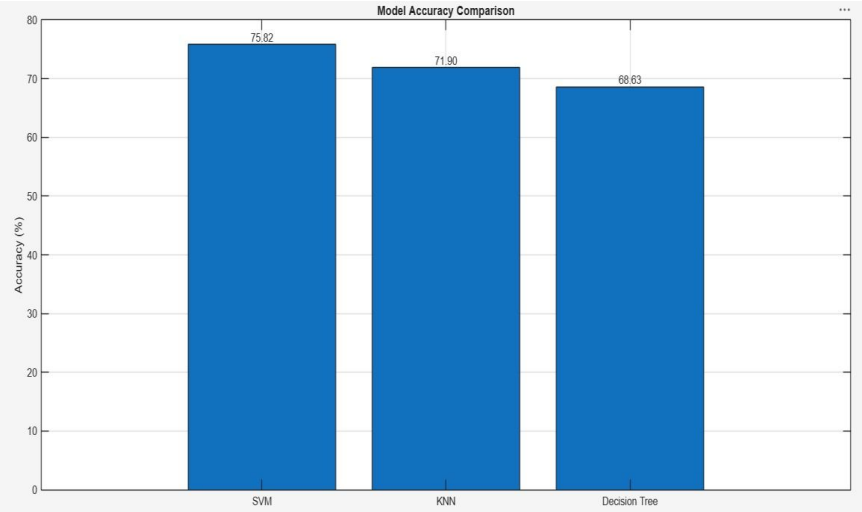
```
C:\Users\hp\Desktop\Matlab_Project\Disease_Predict4.m
1   clc;
2   clear;
3   close all;
4
5   %% Step 1: Load Dataset
6   data = readtable('diabetes.csv');
7   disp(head(data));
8
9   %% Step 2: Separate Features and Labels
10  X = data{:,1:end-1};
```

Command Window

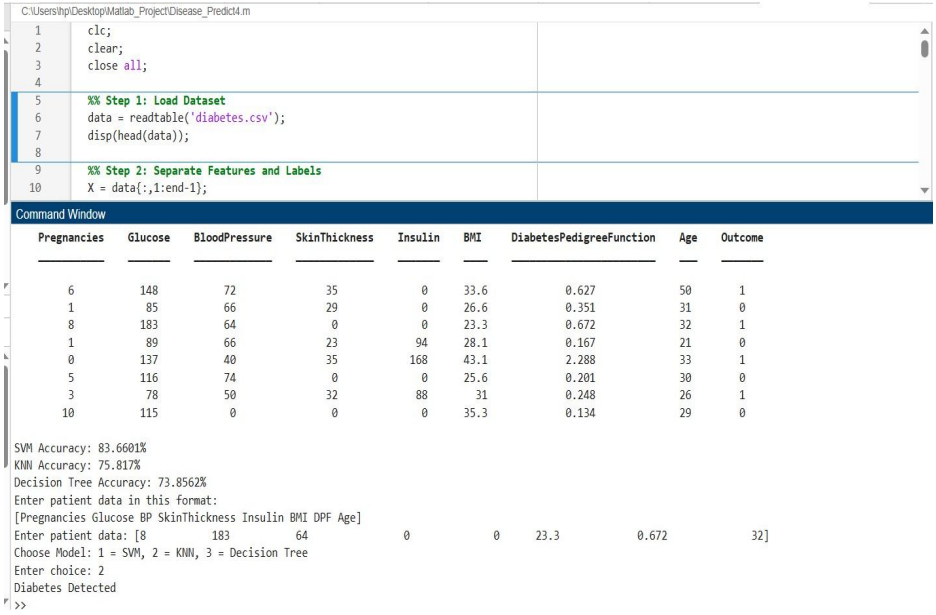
Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0

```
SVM Accuracy: 75.817%
KNN Accuracy: 71.8954%
Decision Tree Accuracy: 68.6275%
Enter patient data in this format:
[Pregnancies Glucose BP SkinThickness Insulin BMI DPF Age]
Enter patient data: [6          148          72          35          0      33.6          0.627          50 ]
Choose Model: 1 = SVM, 2 = KNN, 3 = Decision Tree
Enter choice: 1
Diabetes Detected
>>
```

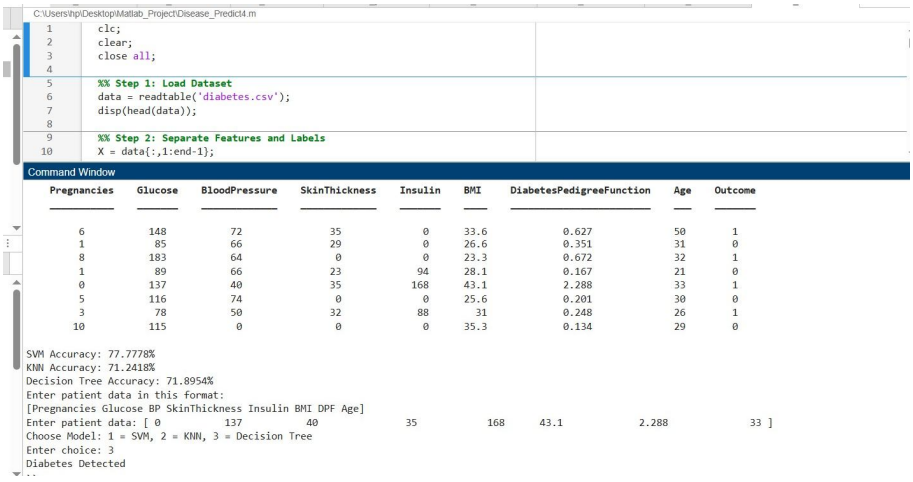
- Models' Accuracy Comparison for Prediction of Diabetes



- Diabetes Prediction using KNN Model



- Diabetes Prediction using Decision Tree Model



VI. ADVANTAGES OF PROPOSED SYSTEM

The proposed system offers the following advantages:

- Early diabetes prediction
- Improved prediction accuracy
- Reduced diagnosis time
- User-friendly implementation in MATLAB
- Comparative analysis of multiple machine learning models
- Cost-effective healthcare support system

VII. CONCLUSION

This chapter presented the implementation and results of the proposed AI-based diabetes disease prediction system using machine learning algorithms in MATLAB. The system was trained and tested using the diabetes dataset and evaluated using accuracy analysis and confusion matrices. Experimental results showed that the Support Vector Machine classifier achieved better prediction accuracy compared to KNN and Decision Tree algorithms. The proposed system demonstrates the effectiveness of machine learning techniques for intelligent healthcare applications and disease prediction systems.

REFERENCES

[1] P. Gandhi and S. Gayathri, “Diabetes Prediction Using Machine Learning Techniques,” International Journal of Research Publication and Reviews, vol. 3, no. 2, pp. 77-82, 2022.

[2] M. Manimekalai, S. Divya, “Prediction of Diabetes using Artificial Neural Network Classification Technique”, in Journal of Emerging Technology and Innovative Research, vol. 6, issue 6, 2019, pp. 9-13.

[3] N. Samer and S. Naser, “Diabetes Prediction Using Artificial Neural Network”, International Journal of Advanced Science and Technology, vol.124 , pp. 1-10, 2018.

[4] K. S. Charan and K. S. Mahendranath, “Heart Disease Prediction Using Random Forest Algorithm,” International Research Journal of Engineering and Technology, vol. 9, no. 3, pp. 1191-1194, 2022.

[5] V. Chaurasia and S. Pal, ““Early Prediction of Heart Diseases Using Data Mining Techniques,” Caribbean Journal of Science and Technology, vol. 1, pp. 208-217, 2013.

[6] R. Detrano et al., “International Application of a New Probability Algorithm Algorithm for the Diagnosis of Coronary Artery Disease ,” American Journal of Cardiology, vol. 64 no. 5, pp. 304–310, 1989.

[7] S. Reshma, S. Shaji, A. Janisha “Chronic Kidney Disease Prediction Using Machine Learning,” International Journal of Engineering Research and Technology, vol. 9, no. 7, pp. 137-140, 2020.

[8] M. Komi, Jun Li, Yongxin Zhai and Xianguo Zhang, "Application of data mining methods in diabetes prediction," 2017 2nd International Conference on Image, Vision and Computing (ICIVC), Chengdu, 2017, pp. 1006-1010, doi: 10.1109/ICIVC.2017.7984706..

[9] T. Mohanapriya, M. Priyadarshini, H. Sirajudeen and N. Angayarkanni “Liver Disease Prediction Using Classification Techniques,” International Journal of Computer Science and Information Technologies, vol. 15, no. 3, pp. 19–22, 2024.

[10] Velu, S.R., Ravi, V. & Tabianan, K. Data mining in predicting liver patients using classification model.Disease Diagnosis”, Health Technol **12**, 1211–1235 (2022). https://doi.org/10.1007/s12553-022-00713-3

[11] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, “Disease Prediction by Machine Learning over Big Data from Healthcare Communities,” IEEE Access, vol. 5, pp. 8869– 8879, 2017.

[12] J. Jiang et al., “Artificial Intelligence in Healthcare: Past, Present and Future”, Stroke and Vascular Neurology, vol. 2, no. 4, pp. 230–243, 2017.

[13] B. J. Sunday et al., “A Comparative Study of Machine Learning Algorithms for Predictive Health care Systems,” International Journal of Engineering and Information Systems, vol. 9, no. 5, pp. 82-877, 2025.

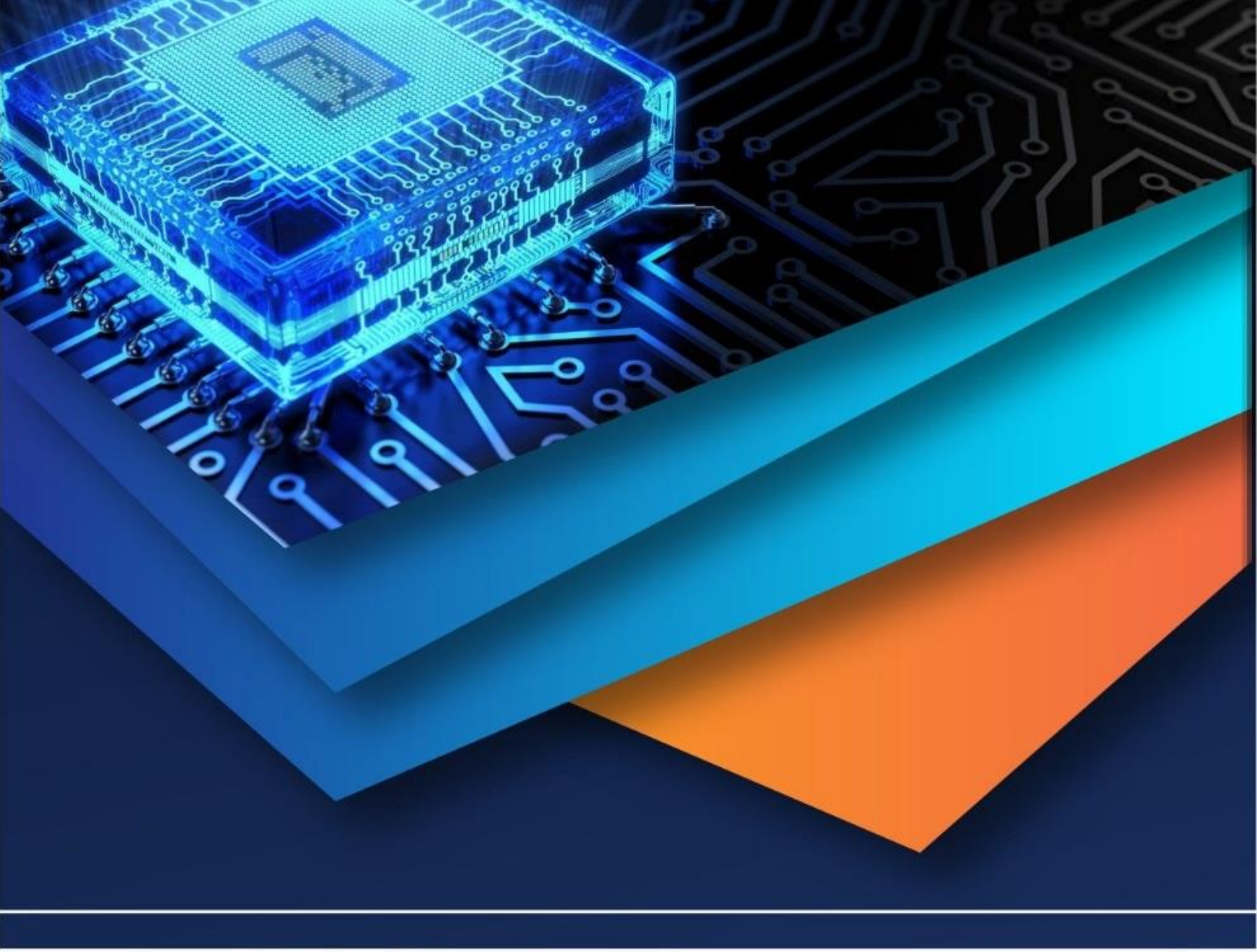
[14] L. P. Sharan et al., “Disease Prediction Using Machine Learning Algorithms,” International Journal of Advance Research in Science Communication and Technology , vol. 3, no. 1, pp. 5-12, 2023.

[15] K. A. Parane, N. C. Patil, S. R. Poojara and T. S. Kamble, "Cloud based Intelligent Healthcare Monitoring System," 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), Ghaziabad, India, 2014, pp. 697-701, doi: 10.1109/ICICICT.2014.6781365.

[16] B. Godi, S. Viswanadham, A. S. Muttipati, O. P. Samantray and S. R. Gadiraju, "EHealthcare Monitoring System using IoT with Machine Learning Approaches," 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA), Gunupur, India, 2020, pp. 1-5, doi: 10.1109/ICCSEA49143.2020.9132937..

[17] D. Tomar and S. Agarwal, “A Survey on Data Mining Approaches for Healthcare,” International Journal of Bio-Technology, vol. 5, pp. 241-266, 2013.

[18] S. A. Tanim et al, “Explainable deep learning for diabetes diagnosis with DeepNetX2,” Biomedical Signal Processing and Control, Springer Nature, vol. 99, 106902, ISSN 17468094, 2025.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY