Smart Health Disease Prediction System

Dr. Sachin Goel
Associate Professor, Information Technology
Raj Kumar Goel Institute of Technology
Ghaziabad, India
sachin.viet@gmail.com

Pratibha Tyagi
Information Technology
Raj Kumar Goel Institute of Technology
Ghaziabad, India
pratibhatyagigi@gmail.com

Shalini Singh
Information Technology
Raj Kumar Goel Institute of Technology
Ghaziabad, India
shalinisingh7657@gmail.com

Adesh Rajput
Information Technology
Raj Kumar Goel Institute of Technology
Ghaziabad, India
adeshrajput2002@gmail.com

***Abstract*— This article examines the use of basic machine learning methods to predict the likelihood of diabetes and heart disease. Using archived datasets and powerful algorithms, it aims to demonstrate the potential for early disease detection. The PIMA diabetes database and the ICU heart disease dataset are used to apply logistic regression and develop predictive models.**

Keywords*— Diabetes, Heart Disease, Disease Prediction, Logistic Regression, Machine Learning, Healthcare.

I. INTRODUCTION

Diabetes and heart disease are alarmingly growing health problems worldwide. Traditional diagnostic methods are based on clinical tests, and they can be time-consuming as well as costly. Machine learning can provide an quick as well as cost effective alternative in which, the patient's data is analysed to identify risk factors. This study addresses the need for a simple and accessible method to predict the risk of these diseases using machine learning, particularly valuable for individuals with limited access to regular medical check-ups. Our objectives are to develop basic predictive models, evaluate their performance using standard accuracy metrics, and provide a clear demonstration of machine learning's application in healthcare. This paper contributes a beginner-friendly introduction to disease prediction using machine learning, demonstrating the application of basic algorithms on common healthcare datasets and offering a starting point for further exploration.

A. Types of Diabetes

Type 1 Diabetes: This disease is an autoimmune disease that destroys and eliminates insulin-producing cells in the pancreas. Hence, the body does not maintain sufficient amounts of insulin.

Type 2 Diabetes: In this form, The body either produces insufficient insulin or the cells are unresponsive. no longer respond effectively to it, leading to difficulties in glucose absorption.

Gestational Diabetes: This condition comes on during pregnancy and usually goes away after delivery; However, it can increase the likelihood of moving in the direction of type 2 diabetes in the future.

Prediabetes – The rise in blood sugar level (but lesser than type diabetes) leads to this diabetic condition. This condition also relates the risk of diabetes, heart attack, and stroke in the patients.

B. Kinds of Heart Diseases

Coronary Artery Disease (CAD) – There are multiple possible heart diseases, but CAD is most commonly found one. The rate of blood flow is less than usual.

Heart Attack (Myocardial Infarction) - The blood supply to a specific part of the heart is completely cut off, and makes blood clots.

Heart Failure – blood pumping availability is less than the requirements of the body. It is caused by a variety of factors, including narrowed arteries and elevated blood pressure.

Arrhythmias – The heartbeats become irregular, which can be either too fast, too slow, or erratic.

Stroke – The Stroke in the brain occurs when blood supply is irregular to the brain.

Cardiomyopathy – Cardiomyopathy is a popular heart disease which can make the heart muscles weak, as it make it harder to pump the blood from heart.

Congenital Heart Defects - heart disease problem is present in a human since birth.

II. LITERATURE REVIEW

Several studies have also used logistic regression algorithms for the prediction of diabetes and heart diseases in the patients. Many studies use more complex algorithms, but this study focuses only on basic and relatively simple methods. All of these studies used disease-specific datasets, such as the datasets given by PIMA for Diabetes and the UCI for cardiovascular Disease.

The application and use of ML in the healthcare industry gained increasing importance in recent times. This section provides a summary of previous research related to diabetes and cardiovascular disease prediction, with a particular emphasis on the application of ML algorithms and datasets.

Several studies have been conducted to examineS the application of ML in diabetes prediction. For example, studies utilizing the Pima Diabetes Dataset have demonstrated the effectiveness of several algorithms, such as logistic regression and decision trees, in identifying individuals at risk. Many studies have emphasized the Role of feature selection and preprocessing techniques for improving forecast accuracy. Some scientists have delved further into this area, using more complex models.

Table 1: Literature Review

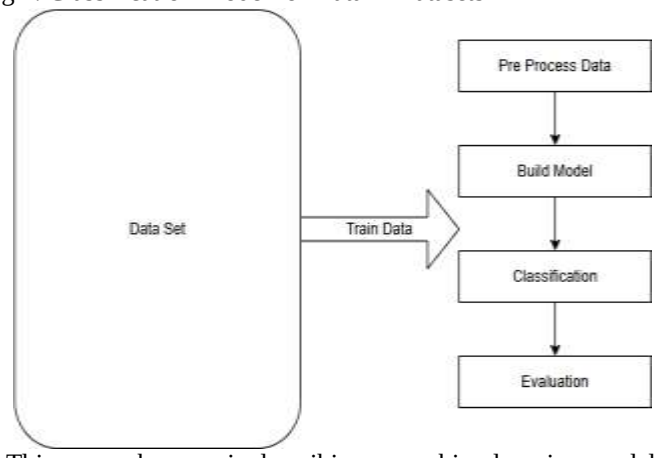
Serial No.	Author Name	Paper Title	Finding	Year
1	Naveed, S.; Geetha, G.; Leninisha, S.	Early Diabetes Discovery From Tongue Images	A non-invasive technique has been suggested to detect diabetes in its early stages without causing any harm.	2022
2	Jahin, S.; Moniruzzaman, M.; Alvee, F.M.; Haque, I.U.; Kalpoma, K.A.	A modern approach to AI assistant for heart disease detection by heart sound through created e-Stethoscope	They developed a broadband wireless acousto-mechanical sensing network and investigated to identify heart disease by analyzing heartbeat sounds.	2022
3	Kibria, H.B.; Nahiduzzaman, M.; Goni, M.O.F.; Ahsan, M.; Haider, J.	An Ensemble Approach for the Prediction of Diabetes Mellitus Using a Soft Voting Classifier with an Explainable AI	By integrating forecasts from six distinct machine learning algorithms, a diabetes detection strategy using soft voting classification is proposed, which yields 90% accuracy for both positive and negative cases.	2022
4	Vincent, P.M.D.R.; Srinivasan, K.; Gutierrez, D.; et al.	IoT-Cloud-Based Smart Healthcare Monitoring System for Heart Disease Prediction via Deep Learning	A smart healthcare system using Bi-LSTM to predict heart disease has been developed, achieving an impressive accuracy rate of 98.86%.	2022
5	Ahmed, A.; Aziz, S.; Abd-alrazaq, A.; et al.	Wearable Devices and Machine Learning for Blood Glucose Level Prediction in Diabetes Mellitus: A Systematic Review	This paper reviews wearable devices that use machine learning to predict blood sugar levels, with an emphasis on the potential for non-invasive monitoring.	2023
6	Shimpei Ogawa, Fuminori Namino, Tomoyo Mori, Ginga Sato, Toshitaka Yamakawa, Shumpei Saito	AI diagnosis of heart sounds differentiated with super StethoScope	The study discussed the development and potential of the Super StethoScope, an AI-integrated device for heart sound analysis, highlighting its ability to overcome limitations of traditional auscultation and its potential for telemedicine applications.	2023
7	Jalpa Shah, Chintan Shah, Premal Patel	Heart Disease Diagnosis Detection Based on Chest Pain Using Machine Learning Stacking Classifier	This study assessed the performance of different machine learning classification algorithms.. including stacking classifiers, for detecting heart disease based on chest pain data, with the goal of identifying effective and computationally efficient methods for risk prediction. This study tested the performance of several machine learning classifiers, including stacking classifiers, for detecting heart disease based on chest pain	2023

			effective and computationally efficient methods for risk prediction.	
8	Yovel Rom, Rachelle Aviv, Gal Yaakov Cohen, Yehudit Eden Friedman, Zack Dvey-Aharon	Diabetes Detection from Diabetic Retinopathy-Absent Images Using Deep Learning Methodology	The study designed a deep learning model capable of identifying diabetes from fundus images of eyes that do not show signs of diabetic retinopathy, achieving an AUC of 0.83 per image and 0.86 per patient, suggesting a novel approach for early diabetes diagnosis.	2023
9	UT Utsha, I Hua Tsai, BI Morshed	A smart health application for real-time cardiac disease detection and diagnosis using machine learning on ECG data	This article presents a health application that uses ML algorithms to spot and assess heart disease in real time using ECG data, highlighting the importance of mobile applications in continuous heart disease monitoring.	2023
10	SP Patro, N Padhy	A secure remote health monitoring for heart disease prediction using machine learning and deep learning techniques in explainable artificial intelligence framework	This paper proposes a trust worthy remote control framework constructed to deliver heart diseases using learning and automatic learning methods. techniques, all within an interpretable artificial intelligence framework, with the goal of achieving a balance between accuracy and transparency in remote cardiac care.	2023
11	Packialatha, A.; Preetha, P.	IoT-Enabled Smart Healthcare System for Heart Disease Prediction Using Deep Learning and Dimensionality Reduction	A sophisticated medical frame is proposed, leveraging the learning models of special networks (CNN) to forecast heart diseases, exhibiting particular precision on data sets procured from Kaggle competitions.	2024
12	Mayya, V.; Kandala, R.N.V.P.S.; Gurupur, V.; et al.	Need for an Artificial Intelligence-based Diabetes Care Management System in India and the United States	To investigate the impact of artificial intelligence (AI) on diabetes care manages in United states and the India, emphasizing the challenges and opportunities related to healthcare systems, electronic health records (EHRs), cultural factors, and data privacy issues.	2024
13	Rabbi, I.R.K.; Zouaghi, H.; Peng, W.	An Intelligent System Approach for Predicting the Risk of Heart Failure	An advanced intelligent system was developed employing a fuzzy inference system (FIS) alongside an artificial neural network (ANN) to predict the probability of heart disease, achieving an accuracy rate of 90.50% with the FIS.	2024
14	Nguyen, T. T., et al.	An Explainable AI Framework for Interpreting Heart Sound Data Collected via Digital Stethoscopes for Early Detection of Cardiac	Developed an explainable AI framework that can analyze heart sound data recorded by digital stethoscopes to detect cardiac anomalies, providing	2024
		Anomalies	insights into the reasoning behind the AI's predictions.	

15	Sharma, A., et al.	A Smart Home-Based System for Detecting Early Signs of Cardiovascular Disease in Elderly Individuals	Proposed a smart home environment equipped with various sensors that can passively collect data on daily activities and physiological parameters. With the help of this proposed environment, the early detection using indicators of cardiovascular disease in elderly individuals.	2025

III. Methodology

Fig 2: Classification Model for Data in Datasets



This research paper is describing a machine learning model which learns on a supervised machine learning approach. This learning will help the model in prediction of diabetes and heart disease risk using the collected datasets. The general methodology of the entire project is divided into several steps: data acquisition and preprocessing, model selection and training, and performance evaluation.

A. Data Acquisition and Preprocessing

Two public available datasets are utilized: the PIMA Diabetes Database and the UCI Heart Disease dataset. The PIMA Diabetes Database has 768 instances and these instances are based on particular 8 features, such as glucose level, BMI, age, and blood pressure. The UCI cardiovascular disease dataset contains 303 instances and these instances are based on particular 13 features likely likely nature of chest pain, blood pressure readings, cholesterol levels, and age.The data underwent pre-processing to enhance its quality and ensure its compatibility with machine learning algorithms, which is a necessary step prior to initiating model training. This procedure includes addressing missing values through mean imputation, wherein the absent numerical values are substituted by the mean of the respective feature. Feature scaling was implemented using the standardscaler from the scikit-learn library to normalize feature ranks. This step is crucial for algorithms that are sensitive to feature scaling, such as logistic regression, as it ensures that no individual feature overshadows the learning process.

B. Model Selection and Training

The model will be developed using two basic classification algorithms: logistic regression and decision trees. Logistic regression, a simple linear model, was chosen for its ease of use in binary classification problems. Logistic regression is used to estimate the probability of an individual falling into a particular disease risk category. Decision trees, a nonlinear model, were chosen for their ability to create tree-like hierarchical structures. Decision trees visually represent decision making based on attribute values.The 80:20 is the ration of dataset which is training sets to testing sets. Python's scikit-learn library is used to train models. Default parameters were used for both logistic regression and decision trees.

C. Performance Evaluation

Accuracy evaluation is an important aspect in calculating the performance of trained models The precision means that the case reports are precisely classified according to the number of total samples. It serves as a proper measure of the model in precise classification of the individuals. Performance is assessed for determining the degree that the model can conform to the unknown or intangible data.. The results were recorded and compared to calculate the algorithm's efficiency and accuracy in predicting diabetes and heart disease risk.

IV. Results and Discussion

The findings derived from the analysis performed utilizing logistic regression and decision trees. These machine learning techniques will subsequently be employed to forecast the likelihood of diabetes and cardiovascular disease.

A. Results

The accuracy of the models was evaluated to determine their performance. The representation of the proportion of correctly predictions is performed using it. The following table summarizes the accuracy achieved by each model on each dataset:

Table 2: Accuracy and Results

Dataset	Model	Accuracy (%)
PIMA Diabetes	Logistic Regression	76.30
PIMA Diabetes	Decision Trees	72.10
UCI Heart Disease	Logistic Regression	82.00
UCI Heart Disease	Decision Trees	78.50

As shown in the table, the logistic regression and decision tree algorithms perform well on both data sets.

B. Discussion

The observed differences in performance between Logistic Regression and Decision Trees are inherited characteristics of these algorithms and the nature of the datasets. Logistic Regression, being a linear model, effectively handles linear relationships between features and the target variables. In contrast, Decision Trees, a non-linear relationships, may have chances of overfitting.

The higher accuracy achieved by Logistic Regression on the UCI Heart Disease dataset declares that the features in this dataset have linear correlations with high heart disease risk. The relatively lower accuracy of Decision Trees on the PIMA Diabetes Database indicates that the decision boundaries for diabetes prediction are more complex, requiring more tree structured algorithms.

The precision illustrates the practicality of employing fundamental machine learning models for predicting disease risk. The obtained and concluded precisions are acceptable for initial risk assessment, but are not sufficient for clinical decision making. In addition, the simplicity and accuracy of the models used for the evaluation metrics limit the scope of this study.

As described in Methods, the knowledge-based representations in the trained models allowed them to identify important patterns and connections in the data. However, the lack of readable terms or coefficients makes the results difficult to interpret.

In conclusion, this study emphasizing the importance and promise of basic automatic learning techniques to assess the risk of diabetes and heart disease. However, an additional survey is necessary to improve their prediction accuracy. and power of the model, refine its interpretation, and validate the results in real-world clinical settings.

V. Knowledge Representation

This research and documentation provides the information needed to extract parameters and structures from machine learning, logistic regression, and decision tree models. These models have the ability to identify patterns and create relationships between input data, effectively representing the knowledge needed to predict diabetes and cardiovascular disease risk.

A. Logistic Regression

Logistic regression represents knowledge through its coefficients. These coefficients are learned early during the training process and reflect the effect of each input feature on the predicted probability of disease risk. This model derives knowledge from a linear combination of input features, where each feature is multiplied by a corresponding coefficient. Therefore, the knowledge is represented in the form of a linear equation, in which the coefficients are the learned knowledge.

B. Decision Trees

Decision trees represent knowledge through their hierarchical structure. A decision based attribute is used and represented by each node in the Decision Tree. The branches symbolize the possible outcomes of that choice. The tree structure provides rules for classifying individuals according to the values of their attributes. The acquired knowledge is represented through some rules that lead to the classification. For example, “If glucose levels are higher than X and BMI is higher than Y, then it predicts diabetes.” The depth of the tree and the attributes used at each node reflect the importance of each attribute in the acquired knowledge.

C. Implicit Representation

Both logistic regression and decision trees use implicit representations of knowledge, where learned patterns are incorporated into the model parameters or structure. This approach allows for automating models and extracting relevant information from data. User-defined rules are not explicitly used. This is in contrast to symbolic representations of knowledge, where it is represented through formal logic or semantic networks.

D. Feature Importance

These models show the importance of each feature. Logistic regression coefficients and decision trees allow us to see which features have the greatest impact on the prediction.

Based on underlying knowledge representation techniques, machine learning models can effectively predict diabetes and cardiovascular disease risk based on provided patient data.

VI. Conclusion

This research finds the application of critical automatic learning techniques, such as logistics regression and allied techniques. decision trees, to forecast the probability of diabetes and cardiovascular diseases by evaluating the Pima Diabetes Database and the UCI Dataset.

The results of the experiments show that logistic regression and decision trees provide good accuracy in predicting these outcomes. These results highlight the capabilities of machine learning in early assessment of disease risk, especially where there is limited access to comprehensive medical evaluation.

This study presents a basic framework for beginners interested in implementing machine learning algorithms in the healthcare sector. By focusing on public datasets and simple algorithms, the goal is to clarify the process and encourage the exploration of more advanced techniques. The knowledge representations contained in the trained models are effectively documented. Also, any

relevant trends and connections in the data allowing accurate predictions about patient characteristics.

However, it is also important to consider the limitations of this study.Use of basic algorithms and reliance on accuracy as the sole evaluation measure does not fully capture the complexity of disease prediction.

Ultimately, this research paper starts a fundamental point for the development of more robust and reliable machine learning systems. These systems can predict diseases and aid to the overall aim of improving early detection and preventive care in healthcare.

REFERENCES

[1] S. Naveed, G. Geetha, and S. Leninisha, *"Early Diabetes Discovery From Tongue Images,"* 2022.
[2] S. Jahin, M. Moniruzzaman, F. M. Alvee, I. U. Haque, and K. A. Kalpoma, "A modern approach to AI assistant for heart disease detection by heart sound through created e-Stethoscope," 2022.
[3] H. B. Kibria, M. Nahiduzzaman, M. O. F. Goni, M. Ahsan, and J. Haider, "An Ensemble Approach for the Prediction of Diabetes Mellitus Using a Soft Voting Classifier with an Explainable AI," 2022.
[4] P. M. D. R. Vincent, K. Srinivasan, D. Gutierrez, et al., "IoT-Cloud-Based Smart Healthcare Monitoring System for Heart Disease Prediction via Deep Learning," 2022.
[5] A. Ahmed, S. Aziz, A. Abd-alrazaq, et al., "Wearable Devices and Machine Learning for Blood Glucose Level Prediction in Diabetes Mellitus: A Systematic Review," 2023.
[6] S. Ogawa, F. Namino, T. Mori, G. Sato, T. Yamakawa, and S. Saito, "AI diagnosis of heart sounds differentiated with super StethoScope," 2023.
[7] J. Shah, C. Shah, and P. Patel, "Heart Disease Diagnosis Detection Based on Chest Pain Using Machine Learning Stacking Classifier," 2023.
[8] Y. Rom, R. Aviv, G. Y. Cohen, Y. E. Friedman, and Z. Dvey-Aharon, "Diabetes Detection from Diabetic Retinopathy-Absent Images Using Deep Learning Methodology," 2023.
[9] U. T. Utsha, I. H. Tsai, and B. I. Morshed, "A smart health application for real-time cardiac disease detection and diagnosis using machine learning on ECG data," 2023.
[10] S. P. Patro and N. Padhy, "A secure remote health monitoring for heart disease prediction using machine learning and deep learning techniques in explainable artificial intelligence framework," 2023.
[11] A. Packialatha and P. Preetha, "IoT-Enabled Smart Healthcare System for Heart Disease Prediction Using Deep Learning and Dimensionality Reduction," 2024.
[12] V. Mayya, R. N. V. P. S. Kandala, V. Gurupur, *et al.*, "Need for an Artificial Intelligence-based Diabetes Care Management System in India and the United States," 2024.
[13] I. R. K. Rabbi, H. Zouaghi, and W. Peng, "An Intelligent System Approach for Predicting the Risk of Heart Failure," 2024.
[14] T. T. Nguyen, *et al.*, "An Explainable AI Framework for Interpreting Heart Sound Data Collected via Digital Stethoscopes for Early Detection of Cardiac Anomalies," 2024.
[15] A. Sharma, *et al.*, "A Smart Home-Based System for Detecting Early Signs of Cardiovascular Disease in Elderly Individuals," 2025.