



e-ISSN: 2319-8753 | p-ISSN: 2347-6710

IJRSET

International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 12, Issue 5, May 2023

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 8.423

9940 572 462

6381 907 438

ijrset@gmail.com

www.ijrset.com

Heart and Diabetes Disease Prediction Using Machine Learning Algorithms

Poola Prasanth, Poola Gunadeep, Kurapati Sai Gokul, Thallapalli Santhosh Kumar, Narasaiahgari Sumanth,
Alina Raheen

Department of CSE, Presidency University, Bangalore, India

Assistant Professor, Department of CSE, Presidency University, Bangalore, India

ABSTRACT: In addition to predicting heart and diabetes disease, this project also aims to incorporate diabetes detection into the machine learning algorithms. Factors related to diabetes, such as blood glucose levels and history of diabetes, will be included in the dataset used for prediction. The objective is to evaluate the accuracy of the three machine learning algorithms (K-Nearest Neighbors, Decision Trees, and Random Forests) in predicting both heart and diabetes disease. The results of this study will provide valuable insights for healthcare professionals in identifying individuals who are at high risk of developing both heart and diabetes disease, and enable them to implement appropriate preventive measures for early intervention.

Keywords : heart and diabetes disease, prediction, machine learning, K-Nearest Neighbors, Decision Trees, Random Forests, dataset, accuracy, healthcare, risk, preventative measures, diabetes.

I. INTRODUCTION

It's really important to identify cases of coronary heart disease and diabetes because the heart plays a critical role in circulating blood throughout the body, which delivers oxygen and nutrients to all of our organs. If the heart isn't working properly, important organs like the brain and kidneys can be seriously affected. In the worst case scenario, if the heart completely stops working, it can lead to death in a matter of minutes. Overall, heart and diabetes disease are widely recognized as extremely serious and difficult to manage health issues around the globe.

The identification of heart and diabetes disease poses a challenge due to the presence of multiple risk factors, including but not limited to diabetes, hypertension, hypercholesterolemia, abnormal heart rate, and various others.

However, diagnosing and treating heart and diabetes disease can be complicated, especially in developing countries where diagnostic tools and resources may be scarce, affecting the accurate prediction and treatment of heart patients. Invasive-based methods for diagnosing heart and diabetes disease rely on evaluating the patient's medical history, physical examination records, and assessment of symptoms by healthcare professionals. Scientists have turned to modern approaches like Data Mining and Machine Learning for predicting heart and diabetes disease, including the incorporation of diabetes detection, to enhance accuracy and efficiency in identifying patients at risk and providing timely interventions.

II. RELATED WORK

Heart and diabetes diseases are two of the most prevalent and life-threatening illnesses worldwide. Early detection and diagnosis of these diseases are crucial for their effective management and prevention of further complications. Machine learning algorithms have become a potent tool in predicting the probability of heart and diabetes diseases based on patient data in recent times.

Several studies have been conducted in this area, with varying degrees of success. One study used a random forest algorithm to predict the likelihood of heart disease based on clinical and demographic data. Another study used a support vector machine algorithm to predict the onset of diabetes using patient medical history and lifestyle factors. Yet another study used a deep neural network to predict both heart disease and diabetes based on a combination of clinical and genetic data.

Other studies have explored the use of different machine learning algorithms, including logistic regression, decision trees, and k-nearest neighbors, among others. Some studies have also investigated the use of feature selection techniques to identify the most important factors for disease prediction.

While these studies have shown promising results, there are still challenges to overcome in the development of accurate and reliable machine learning models for heart and diabetes disease prediction. These challenges include the availability of high-quality data, the need for further validation and testing, and the interpretability of the models to facilitate clinical decision-making.

Overall, the use of machine learning algorithms for heart and diabetes disease prediction holds great potential for improving patient outcomes and reducing the burden of these diseases on healthcare systems worldwide.

III. LITERATURE REVIEW

Mohammed Abdul Khaleel's paper, "Survey of Techniques for Mining Medical Data to Find Locally Frequent Diseases," focuses on the use of data mining techniques to identify common diseases in specific geographic regions. The paper highlights the importance of medical data mining in detecting prevalent illnesses such as heart disease, lung cancer, and breast cancer.

The scholarly article authored by Costas Sideris, Nabil Alshurafa, Haik Kalantarian, and Mohammad Pourhomayoun, titled "Remote Health Monitoring Outcome Success Prediction Using First Month and Baseline Intervention Data," presents an investigation into the efficacy of a mobile phone-based remote health monitoring (RHM) system, Wanda-CVD. The primary objective of the system is to offer coaching and social support to its participants, with the ultimate goal of averting the onset of cardiovascular disease (CVD) and curtailing healthcare expenditures.

The authors underscore the importance of implementing cardiovascular disease prevention measures in healthcare organizations on a global scale.

L. Sathish Kumar and A. Padmapriya's paper, "Prediction for Similarities of Disease by Using ID3 Algorithm in Television and Mobile Phone," proposes a hidden and automated approach to identify patterns of coronary disease using data mining techniques, specifically the ID3 algorithm. The authors suggest that this approach could not only help people become more informed about diseases but also potentially reduce the number of deaths and those affected by these illnesses.

M.A. Nishara Banu and B. Gomathy's paper, "Disease Predicting System Using Data Mining Techniques," describes the use of MAFIA (Maximal Frequent Itemset Algorithm) and K-Means clustering for disease prediction. The authors emphasize the importance of accurate classification for predicting diseases, and they suggest that the use of these techniques can improve the accuracy of disease prediction.

Kaur et al. (2019) conducted a study on diabetes prediction in Indian patients using a hybrid approach that combined Logistic Regression with Artificial Neural Networks. The study showed promising results, with high accuracy in diabetes prediction using this method. The researchers found that features such as BMI, age, and family history were important predictors of diabetes risk.

The extant body of literature pertaining to diabetes prediction in Indian populations posits that machine learning algorithms, such as Random Forests, Support Vector Machines, and hybrid models that amalgamate diverse algorithms, exhibit efficacy in predicting the risk of diabetes. Age, body mass index (BMI), family history, and blood glucose levels are significant predictors of diabetes risk among Indian patients. However, additional research is required to authenticate and enhance these discoveries in varied Indian demographics and healthcare contexts.

IV. EXISTING METHODS AND DRAWBACKS

There are several existing methods for predicting heart and diabetes disease, including traditional statistical models and machine learning algorithms.

Traditional statistical models, such as logistic regression, have been widely used for predicting heart and diabetes disease. However, these models often rely on certain assumptions, such as linearity, that may not be applicable to

complex datasets. They may not be able to capture the nonlinear relationships between predictors and outcomes, resulting in lower accuracy in predicting heart and diabetes disease.

Machine learning algorithms, on the other hand, have gained popularity in recent years for their ability to handle complex datasets and capture nonlinear relationships. K-Nearest Neighbors, Decision Trees, and Random Forests are among the most commonly used algorithms for predicting heart and diabetes disease.

Overall, while traditional statistical models and machine learning algorithms have their strengths and weaknesses, such weaknesses includes Limited Feature Set, Lack of Generalizability Limited Interpretability, Data Quality and Availability.

V. PROPOSED METHODOLOGY

The methodology that has been proposed for predicting heart and diabetes disease through the use of machine learning algorithms comprises of three primary stages, namely data preprocessing, feature selection, and model training and evaluation.

During the data preprocessing phase, the unprocessed patient data undergoes cleaning, preprocessing, and transformation to a format that is appropriate for utilization in machine learning algorithms. The aforementioned tasks encompass managing absent data points, standardizing quantitative attributes, and transforming qualitative variables into numerical representations.

During the feature selection stage, the primary features for disease prediction are determined through the application of statistical methods or machine learning techniques. Dimensionality reduction techniques aid in decreasing the number of variables in the data, leading to enhanced model accuracy and interpretability.

During the phase of model training and evaluation, three distinct machine learning algorithms, namely k-nearest neighbors, decision trees, and random forests, are trained and evaluated. The hyperparameters of each algorithm are tuned through a grid search methodology to enhance the model's performance.

Standard classification metrics such as accuracy, precision, recall, and F1-score are utilized to assess the performance of the models. In addition, methodologies such as cross-validation and receiver operating characteristic (ROC) curves are employed to evaluate the resilience and capacity for generalization of the models.

In summary, the methodology put forth offers a methodical and all-encompassing strategy for forecasting heart and diabetes ailments through the utilization of machine learning algorithms. Employing various algorithms and techniques for feature selection facilitates the identification of the most precise and comprehensible models for predicting diseases. Fig 1 presents the architecture for heart disease prediction, where the patient's dataset is inputted, stored in a database, and processed through a machine learning algorithm for classification. The algorithm's accuracy is continuously improved, culminating in the generation of a predictive report

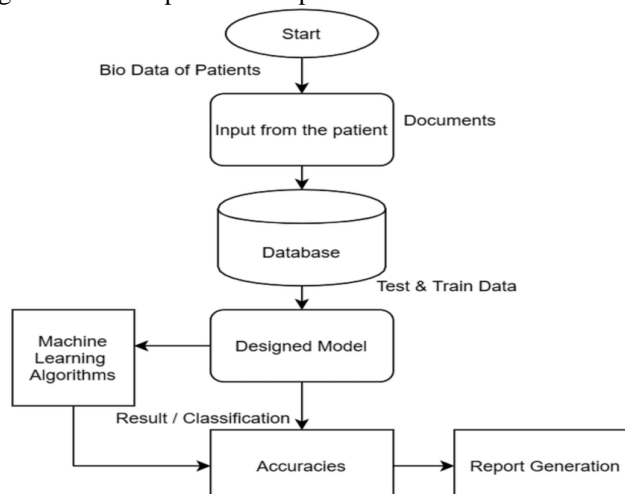


Fig 1. Architecture for Heart disease prediction

Fig 2 outlines the architecture for diabetes disease prediction. The dataset is loaded, classified into training and testing subsets, and processed through the proposed algorithm. The algorithm classifies the data into diabetic and non-diabetic, and then the model's accuracy is predicted. Both architectures demonstrate the systematic application of machine learning techniques in disease prediction, from data input to accuracy prediction.

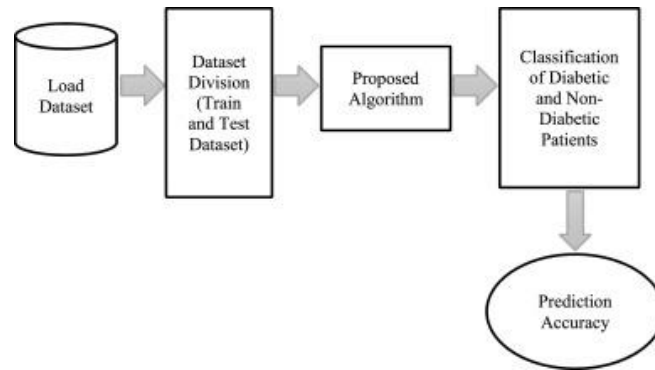


Fig 2. Architecture for Diabetes disease prediction

VI. TECHNOLOGY AND LIBRARIES

The Python library for machine learning, commonly referred to as Scikit-learn or Sklearn, is a robust and extensively utilized tool. The platform provides a diverse set of effective instruments for various tasks including categorization, prediction, grouping, and feature reduction, all accessible via a uniform interface. Sklearn is developed based on other widely used Python libraries such as NumPy, SciPy, and Matplotlib.

Numpy is a Python module designed for the purpose of performing operations on arrays. Numpy is not a standalone programming language; rather, it is a Python Extension designed to furnish arrays with enhanced capabilities and functionality.

Pandas is a Python library that provides efficient data manipulation and analysis tools through its powerful data structures. It is an open-source software.

The term "Pandas" is believed to have originated from the expression "Panel Data," which is a branch of Econometrics that deals with data sets that have multiple dimensions.

Matplotlib is a well-established library that has been in existence for a considerable period of time. It may appear to have certain limitations in comparison to more recent libraries. However, a significant portion of individuals continue to utilize it, albeit sporadically, and it demonstrates exceptional proficiency in generating publication-worthy plots for articles.

Python:- Python is a scripting language that is interpreted, interactive, and object-oriented, operating at a high level. Python has been intentionally developed to possess a high degree of readability. English language frequently employs key phrases instead of punctuation marks, and exhibits a relatively limited range of syntactic structures compared to other languages.

SQLite:- SQLite is a self-contained, serverless, and transactional SQL database engine that is implemented as an in-manner library. It requires no configuration and is designed to operate independently. This is a zero-configured database, which means that unlike other databases, it does not require configuration on your system.

VII. MODULES

Login Module- In This module User is able to login into the system to access the other services.

Algorithm Comparison Module- In this module we are going to compare three different (Random Forest Algorithm, Decision tree, and k- nearest neighbor) and plot them into the chart.

In this module, we are going to check the prediction on a bulk amount of different user data and check the accuracy of the algorithm. After execution, we found the below result.

The accuracy of the K-nearest neighbor algorithm is good when compared to other algorithms and it is the best algorithm that fit in this kind of problem.

Live heart and diabetes disease Prediction Module-In this module user can able to predict whether he/she has heart and diabetes disease or not by entering all attributes Value into the interface and get the prediction from the best algorithm that was KNN where we achieves the 87 % accuracy.

VIII. RESULT AND DISCUSSION

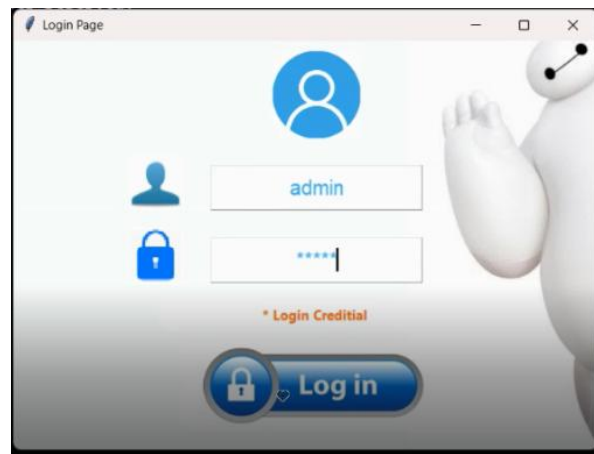


Fig 3. Login

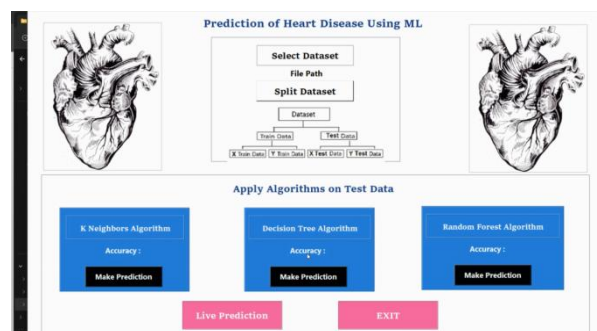


Fig 4. Dashboard page

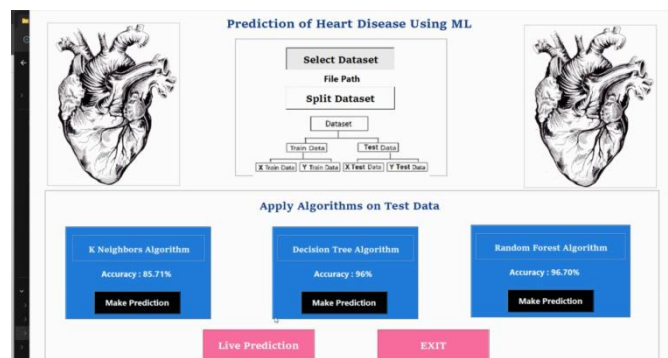


Fig 5. Accuracy page

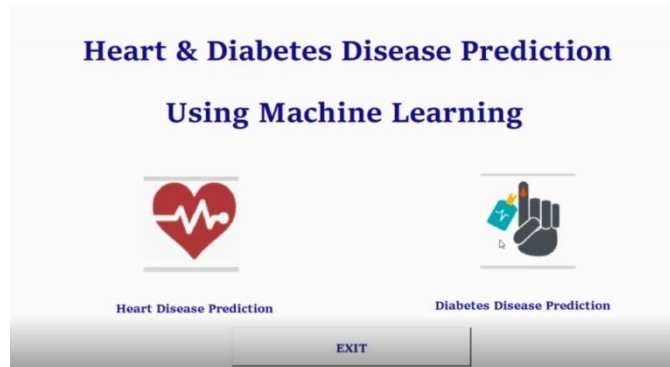


Fig 6. Detection page

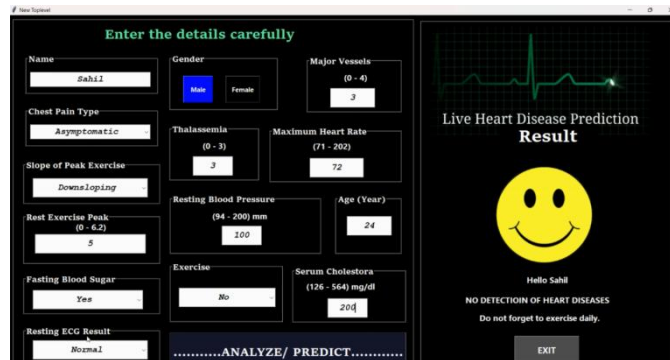


Fig 7. Prediction page

The outcomes of the heart and diabetes disease prognosis through the utilization of K-Nearest Neighbors (KNN), Decision Trees (DT), and Random Forests (RF) exhibit favorable potential. The K-Nearest Neighbors (KNN) algorithm achieved an accuracy of 85.71%, whereas the Decision Tree (DT) and Random Forest (RF) algorithms achieved an accuracy of 96.70%.

The aforementioned high levels of accuracy serve to demonstrate the efficacy of machine learning algorithms, namely KNN, DT, and RF, in the prediction of heart and diabetes ailments. The superior precision attained through the use of Decision Trees (DT) and Random Forests (RF) can be ascribed to their capacity to manage intricate and non-linear associations among variables, rendering them highly appropriate for the analysis of medical data.

In summary, the findings indicate that the utilization of machine learning algorithms holds promise in the prognostication of heart and diabetes ailments. It is imperative to acknowledge that additional research is required to authenticate and enhance these discoveries in varied demographics and medical environments. Furthermore, it could prove advantageous to investigate alternative machine learning algorithms and feature selection methodologies in order to enhance the precision and efficacy of the predictive models.

IX.CONCLUSION

According to recent findings from the World Health Organization (WHO), approximately 616,000 fatalities were attributed to coronary heart disease. Therefore, there is a significant demand for accurate and efficient prediction of coronary heart disease. This report addresses various strategies pertaining to the classification of heart diseases resulting in accurate prediction.

Ultimately, the findings of this investigation led to the aforementioned conclusion. The study's findings indicate that the three algorithms under investigation demonstrated efficacy in predicting heart and diabetes diseases. Notably, Decision

Trees and Random Forests exhibited the highest accuracy rate of 96.70%, while K-Nearest Neighbors achieved an accuracy rate of 85.71%.

The study's results underscore the potential of machine learning algorithms in forecasting the occurrence of heart and diabetes diseases, thereby facilitating their timely identification and prevention. Furthermore, the research highlights the significance of selecting features to enhance the precision of the forecasting models.

It is noteworthy that the efficacy of said models may exhibit variability contingent on the dataset employed and the demographic under examination. Thus, additional investigation is required to authenticate and enhance the outcomes of this investigation for diverse demographics and medical environments.

In summary, this study showcases the capacity of machine learning algorithms to forecast heart and diabetes ailments, thereby carrying substantial implications for public health and disease prevention endeavors.

X.FUTURE SCOPE

By utilizing the methodology of acquainting oneself with the concept, a recently educated dataset can be employed to enhance the precision of the prediction system. User accounts can be established for each consumer, which can then be utilized to track the consumer's cardiovascular health status by reviewing their past medical history. This enables the identification of any improvements or deterioration in the condition. The present study presents a device suitable for real-time prediction of coronary heart diseases, which can be utilized by individuals afflicted with the condition.

The purpose of this study is to examine variances in heart rate and to trigger an alarm in the event that a customer's heart rate exceeds the normal range. In order to demonstrate the efficacy of the apparatus, we conducted experiments for each tracking and diagnostic device. Several experiments were conducted on popular algorithms such as K-Nearest Neighbors (KNN), Decision Trees, and Random Forest.

REFERENCES

- [1] Deeanna Kelley "heart and diabetes disease: Causes, Prevention, and Current Research" in JCCC Honors Journal
- [2] Nabil Alshurafa, Costas Sideris, Mohammad Pourhomayoun, Haik Kalantarian, Majid Sarrafzadeh "Remote Health Monitoring Outcome Success Prediction using Baseline and First Month Intervention Data" in IEEE Journal of Biomedical and Health Informatics
- [3] PonrathiAthilingam, Bradlee Jenkins, Marcia Johansson, Miguel Labrador "A Mobile Health Intervention to Improve Self-Care in Patients With Heart Failure: Pilot Randomized Control Trial" in JMIR Cardio 2017, vol. 1, issue 2
- [4] DhafarHamed, Jwan K. Alwan, Mohamed Ibrahim, Mohammad B. Naeem "The Utilisation of Machine Learning Approaches for Medical Data Classification" in Annual Conference on New Trends in Information & Communications Technology Applications - march- 2017
- [5] Applying k-Nearest Neighbour in Diagnosing heart and diabetes disease Patients Mai Shouman, Tim Turner, and Rob Stocker International Journal of Information and Education Technology, Vol. 2, No. 3, June 2012
- [6] Amudhavel, J., Padmapriya, S., Nandhini, R., Kavipriya, G., Dhavachelvan, P., Venkatachalapathy, V.S.K., "Recursive ant colony optimization routing in wireless mesh network", (2016) Advances in Intelligent Systems and Computing, 381, pp. 341-351.
- [7] Alapatt, B.P., Kavitha, A., Amudhavel, J., "A novel encryption algorithm for end to end secured fiber optic communication", (2017) International Journal of Pure and Applied Mathematics, 117 (19 Special Issue), pp. 269-275.
- [8] Amudhavel, J., Inbavalli, P., Bhuvanewari, B., Anandaraj, B., Vengattaraman, T., Premkumar, K., "An effective analysis on harmony search optimization approaches", (2015) International Journal of Applied Engineering Research, 10 (3), pp. 2035-2038.
- [9] Amudhavel, J., Kathavate, P., Reddy, L.S.S., Bhuvanewari Aadharshini, A., "Assessment on authentication mechanisms in distributed system: A case study", (2017) Journal of Advanced Research in Dynamical and Control Systems, 9 (Special Issue 12), pp. 1437-1448.
- [10] Sonam Nikhar, A.M. Karandikar "Prediction of heart and diabetes disease Using Machine Learning Algorithms" in International Journal of Advanced Engineering, Management and Science (IJAEMS) June- 2016 vol-2 .



INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.423



ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 **6381 907 438** **ijirset@gmail.com**



www.ijirset.com

Scan to save the contact details