OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

Machine Learning Based Predictive Analysis of Heart Failure And Type 2 Diabetes

Prof. Shiva Phulari[1], Mahek Bhartiya[2], Pavan Gadave[3], Aditya Dhas[4] , Ayush Gawai[5]

Department of Computer Engineering PDEA's College of Engineering, Savitribai Phule Pune University ,Pune, India[2,3,4,5]

Associate Professor Department of Computer Engineering PDEA's College of Engineering, Savitribai Phule Pune University[1] Pune,

shivaphulari6666@gmail.com[1] mahekbhartiya123@gmail.com[2], pavangadave2003@gmail.com[3], dhasaditya84@gmail.com[4], ayushgawai3107@gmail.com[5]

Abstract: ***Heart disease and type 2 diabetes represent two of the most pressing global health challenges, with significant impacts on mortality rates and healthcare systems worldwide. The early and accurate prediction of these conditions is critical for effective prevention, timely intervention, and efficient management. Traditional diagnostic approaches, while valuable, often lack the precision and scalability required to address the growing prevalence of these diseases. To overcome these limitations, this study proposes a machine learning-based predictive system tailored for heart disease and type 2 diabetes, leveraging diverse and extensive datasets.The proposed framework integrates advanced machine learning techniques, including feature selection, dimensionality reduction, and cross-validation, to enhance model performance and reliability. A custom ensemble algorithm combining hard and soft voting classifiers was developed to handle different cardiovascular and diabetes-related datasets. This system achieves robust and high-precision results, with prediction accuracies consistently in the range of 95%.***

Our framework incorporates essential preprocessing steps such as handling missing data, balancing datasets using SMOTE, and optimizing features to mitigate biases and improve predictive accuracy. The system is designed to be scalable and adaptable, ensuring compatibility with diverse real-world healthcare scenarios. By providing personalized risk assessments, it assists healthcare professionals and individuals in early detection and informed decision-making.

Keywords: Heart failure, Cardiovascular diseases, Type-II Diabetese, Machine Learning.

JEL Classification Number: I10, C88, L86.

I. Introduction

I. INTRODUCTION

In 2019, Nearly 18 million people died of cardiovascular diseases, accounting for 32 percent of all world deaths. Many cardiovascular illnesses can be avoided by addressing risk factors such as unhealthy diet, tobacco use, physical inactivity and harmful consumption of alcohol. It is important to detect heart disease early so that management begins with counselling and medication. Cardiovascular diseases are the most life- threatening diseases. They have gotten quite popular over time and have now reached the healthcare systems of several countries. We aim this project for effective prediction of cardiovascular disease that is heart disease and diabetes using a machine learning framework. This will help doctors as well as patients for early prediction of disease so that they get proper treatment. Other researchers and scientists have approached it with different techniques and methods. Cardiovascular illnesses are the most deadly syndromes in the world, with the greatest fatality rate. These diseases have recently grown exceedingly widespread, putting a strain on countries' healthcare systems. For effective and high-precision prediction of heart disease and diabetes, a machine learning-based system is proposed. Cardiovascular Disease Prediction Using Machine Learning The diagnostic is quite accurate and can be used in the real world to detect cardiovascular illnesses early. Using various datasets and approaches, researchers have suggested multiple algorithms for the prediction of cardiovascular illnesses throughout the last decade. Heart disease, Cleveland, Framingham, and Cardiovascular Disease are some of the most common datasets used for prediction. These datasets are made up of many attributes that are used to forecast these diseases. Family history, age etc are considered under non-modifiable factors because this cannot be changed. Whereas, Smoking habits, unhealthy lifestyle (having fast food, not maintaining quality standard), blood pressure, and cholesterol can be considered under modifiable risk factors these can be changed and controlled by taking certain precautions and medication We also have scientific methods like medicines and surgery. The use of computer tools, namely machine learning algorithms, to diagnose the disease before it becomes infected is one of the most prevalent preventative strategies. In India, more than 74 million people are affected by diabetes which is the highest

of any country in the world, and the worldwide count of diabetes is 422 million. By 2045 this number is likely to increase to 124 million in India. Females have a higher chance of having diabetes than males. Deaths by diabetes are increasing each year. In 2019, over 1.5 million deaths are caused by diabetes. Diabetes can cause heart attacks and can affect blood vessels, skin, eyes, and feet. According to research, it is found that many diabetes patients are from urban areas. The changing lifestyle of people consuming food is affecting their. health as more sugary and fewer nutrient products are consumed in urban areas. Diabetes can be caused by abnormally high sugar in the blood, absence or insufficient production of insulin, obesity, and family history of diabetes. Although diabetes is not fully curable if it is detected in the early stage it can be treated and controlled. This project will help people to check whether they are affected by diabetes or not by providing some information. To predict this disease manually several parameters are collected like glucose level, insulin level, triceps skinfold thickness, BMI, and age, which may take more time to analyze and make the decision but these machine learning techniques will help in predicting in less time than usual. Human error can be reduced in this proposed method of predicting diabetes. Various machine learning algorithms are used in this research, namely random forest classifier, decision tree, K-nearest neighbor, support vector machine, LightGBM, cat boost, and stacking algorithm.

Diabetes can be caused by abnormally high sugar in the blood, absence or insufficient production of insulin, obesity, and family history of diabetes. Although diabetes is not fully curable if it is detected in the early stage it can be treated and controlled. This project will help people to check whether they are affected by diabetes or not by providing some information. To predict this disease manually several parameters are collected like glucose level, insulin level, triceps skinfold thickness, BMI, and age, which may take more time to analyze and make the decision but these machine learning techniques will help in predicting in less time than usual. Human error can be reduced in this proposed method of predicting diabetes. Various machine learning algorithms are used in this research, namely random forest classifier, decision tree, K-nearest neighbor, support vector machine, LightGBM, cat boost, and stacking algorithm

II.LITERATURE REVIEW

2.1 Machine Learning Techniques for Cardiovascular Disease Prediction: The Cleveland heart disease dataset has been widely used for predicting cardiovascular diseases. Hybrid models, combining decision trees and random forests, achieved an accuracy of **88.7%**. Sequential feature selection combined with random forest classifiers further enhanced accuracy to **99%** when applied to datasets like Cleveland and Hungary. These results highlight the significance of using optimized feature selection and advanced hybrid models in cardiovascular disease prediction.

2.2 Frameworks and Algorithms for Disease Prediction: Integrated frameworks focusing on preprocessing and feature engineering have shown exceptional performance. For instance, a framework combining Logistic Regression (LR) and K-Nearest Neighbour (KNN) achieved accuracies of **99.1%, 98.0%, and 95.5%** on the Framingham, Heart Disease, and Cleveland datasets, respectively. Machine learning algorithms like Naïve Bayes and Random Forest have also demonstrated significant promise, with accuracies of **96%** and **90.16%**, respectively. These studies emphasize the importance of combining preprocessing techniques and algorithmic optimization for disease prediction.

2.3 Techniques for Diabetes Prediction: Preprocessing methods like outlier rejection and dimensionality reduction using PCA have improved the quality of diabetes prediction datasets. For example, the Random Forest algorithm achieved an accuracy of **83%**, compared to **81.4%** with Support Vector Machines (SVM). Deep Neural Networks (DNN) have further enhanced prediction capabilities, outperforming other machine learning techniques with superior performance metrics, including accuracy, precision, specificity, and F1-score. These advancements underscore the potential of advanced preprocessing and deep learning techniques in diabetes prediction.

2.4 Challenges in Disease Prediction Models: While significant progress has been made in improving the accuracy of disease prediction models, several challenges persist. Issues such as data imbalance, missing values, and the lack of diverse datasets limit the generalizability of models. Additionally, most research focuses heavily on improving accuracy without adequately addressing other crucial factors like dataset balancing, outlier detection, and robust evaluation metrics. Addressing these challenges is critical for creating reliable and scalable prediction systems.

2.5 Future Directions in Machine Learning for Disease Prediction: To advance the field further, future research should emphasize developing models that integrate multiple datasets and diverse feature sets to ensure robustness. Combining traditional machine learning techniques with emerging technologies like ensemble learning and deep learning can yield better results. The use of culturally sensitive algorithms, particularly in underrepresented regions, and focusing on real-world implementation challenges will be pivotal for improving healthcare outcomes globally.

III.The Model

3.1 Conceptual Framework

The conceptual framework for our system revolves around leveraging the potential of machine learning to enhance the early prediction of heart disease and diabetes. At its core, the framework is designed to process large volumes of medical data, identify complex patterns, and deliver accurate predictions that aid healthcare professionals in early diagnosis and treatment planning. The primary focus is on creating a predictive model that combines diverse datasets, advanced algorithms, and robust preprocessing techniques to ensure reliability and accuracy across varying patient demographics.

Fig.3.1.1 Heart Disease Prediction Page

The framework incorporates a multi-phase process that begins with data collection and preprocessing. Medical datasets are collected from reputable sources such as Cleveland, Framingham, and the PIMA dataset, representing a broad spectrum of attributes including demographic, lifestyle, and clinical factors. Preprocessing includes handling missing data, balancing the datasets using SMOTE, and optimizing features through dimensionality reduction techniques like PCA. These steps are critical for eliminating biases and enhancing the quality of input data.

The machine learning layer of the framework employs both traditional algorithms and advanced techniques, such as ensemble learning, to build a predictive model. This custom ensemble combines hard and soft voting classifiers to ensure that the predictions are not only accurate but also robust across diverse scenarios. The integration of feature selection methods enhances the interpretability of the model, enabling it to focus on the most significant factors contributing to heart disease and diabetes.

Finally, the framework emphasizes scalability and adaptability, ensuring it can handle increasing data volumes and accommodate future enhancements. By designing a modular and flexible architecture, the framework is well-suited for integration into real-world healthcare systems, making it a valuable tool for early detection and intervention.

3.2 Functional Architecture

The functional architecture is built to support end-to-end data handling, model training, and prediction workflows. It consists of the following layers:

1. **Data Layer**: Includes data preprocessing (e.g., handling missing values, outlier detection, and SMOTE for data balancing).
2. **Feature Engineering Layer**: Applies dimensionality reduction techniques (e.g., PCA) and feature selection methods to enhance performance.
3. **Model Layer:** Implements multiple machine learning algorithms, such as Logistic Regression, Random Forest, and Neural Networks, in a hybrid or ensemble structure.
4. **Output Layer**: Provides risk scores, classifications, and diagnostic insights with interpretable results.

IV.ACKNOWLEDGMENTS

We would like to express our heartfelt gratitude to PDEA's College of Engineering, affiliated with Savitribai Phule Pune University, for providing the necessary facilities and environment for this research. We sincerely thankful to Prof. S.V. Phulari, Associate Professor, of CE department for is constant guidance, support and encouragement throughout this study.

We extent our special thanks to Dr. M. P. Borawake, Head of the Department of Computer Engineering, and all the faculty member for their invaluable support and motivation that helped drive this research forward.

V.FUTURE SCOPE

This study opens several avenues for future research and development:

1. Real-Time Risk Assessment: Enhance model inference speed to support real-time prediction during routine clinical visits or remote monitoring.
2. Advanced AI Models: Integrate deep learning and ensemble methods to improve predictive accuracy for early detection and disease progression.
3. Multimodel Data Fusion: Combine EHR, ECG, wearable sensor data, imaging, and lab results to create more comprehensive prediction models. Portability: Developing compact and cost-effective HSI system for widespread clinical use.
4. Portability and Deployment: Develop lightweight, deployable models (e.g., via mobile apps or edge devices) for use in primary care or rural settings.
5. Large-scale Validation: Conduct prospective, multicenter trials with diverse populations to validate model generalizability. Cloud Support: Implementing Cloud-based platforms for remote data analysis.
6. Comorbidity Analysis: Extend predictive models to account for comorbid conditions like hypertension, kidney disease, and obesity

IMPACT AND BENEFITS

The comparative analysis of Machine Learning Predictive Analysis For Heart Failure and Type II Diabetes offers significant advantages in clinical practice:

1. Early Diagnosis: ML models enable early identification of high-risk individuals by analyzing patterns in clinical, laboratory, and lifestyle data—supporting preventive interventions before severe complications arise. Non-Invasive Precision: These techniques reduce the need for invasive diagnostics by accurately visualizing and analyzing mucosal changes.
2. Non-Invasive Risk Prediction: Predictive analytics use existing non-invasive data (e.g., EHR, wearable data, blood tests) to assess patient risk, reducing the need for costly or invasive procedures. Cost-Effectiveness: Early detection reduces long-term treatment costs by preventing ulcer-related complications such as bleeding and perforation.

3. Advancement in cardio-Metabolic Care:This research fosters innovation in disease prediction and monitoring, encouraging widespread adoption of AI in cardiology and endocrinology.
4. Cost Effectiveness: Predictive analytics use existing non-invasive data (e.g., EHR, wearable data, blood tests) to assess patient risk, reducing the need for costly or invasive procedures.

VI.CONCLUSION

This project aimed to develop a robust and efficient system for predicting heart disease and diabetes using machine learning techniques. The literature survey revealed significant advancements in the domain, showcasing the utility of various algorithms and models for disease prediction. By analyzing and integrating these approaches, we proposed a custom ensemble-based system to ensure high accuracy, reliability, and usability. The system's implementation focused on creating a user-friendly interface that provides personalized risk assessments for heart disease and diabetes. With features like secure login, onboarding guidance, and dedicated prediction pages, the platform ensures a seamless user experience. The integration of scalable, culturally sensitive, and real-world-ready algorithms further enhances its applicability in diverse healthcare settings.

The results emphasize the importance of combining advanced technologies such as deep learning, ensemble learning, and data preprocessing techniques for achieving high predictive accuracy. Moreover, the system highlights the significance of considering real-world implementation challenges, including data imbalance, feature optimization, and cultural sensitivity, to ensure meaningful healthcare outcomes.

In conclusion, As the prevalence of heart disease continues to escalate, the imperative to develop robust forecasting systems becomes increasingly evident. Our approach involved crafting tailored ensemble models for both datasets, delivering optimal accuracy. For the combined dataset, our bespoke ensemble model, comprising Logistic Regression, Decision Tree, Random Forest, Gaussian NB, and XGBoost classifiers, achieved an impressive 97% accuracy using a soft voting classifier. Meanwhile, for the Framingham dataset, custom ensemble models, incorporating decision tree, random forest, and XGBoost classifiers with hard voting, yielded a commendable 95% accuracy. Given the alarming rise in diabetes cases, early detection becomes paramount for effective treatment. Our proposed method boasts over 90% accuracy in diabetes prediction, facilitated by the utilization of diverse ensemble algorithms across two distinct datasets.

VII.REFERENCE

1. M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai and R. S. Suraj, "Heart Disease Prediction using Hybrid machine Learning Model," 2021 6th Interna- tional Conference on Inventive Computation Technologies (ICICT), 2021, pp. 1329-1333, doi: 10.1109/ICICT50816.2021.9358597.
2. N. Ahmad, Shafiullah, A. Algethami, H. Fatima and S. M. H. Akhter, "Comparative study of Optimum Medical Diagnosis of Human Heart Disease using Machine Learning Technique with and without Sequential Feature Selec- tion," in IEEE Access, doi: 10.1109/ACCESS.2022.3153047.
3. Riyaz L., Butt M.A., Zaman M., Ayob O. (2022) Heart Disease Prediction Using Machine Learning Techniques: A Quantitative Review. In: Khanna A., Gupta D., Bhattacharyya S., Hassanien A.E., Anand S., Jaiswal A. (eds) In- ternational Conference on Innovative Computing and Communications. Ad- vances in Intelligent Systems and Computing, vol 1394. Springer, Singapore.
4. D. P. Yadav, P. Saini and P. Mittal, "Feature Optimization Based Heart Dis- ease Prediction using Machine Learning," 2021 5th International Conference on Information Systems and Computer Networks (ISCON), 2021, pp. 1-5,doi: 10.1109/ISCON52037.2021.9702410.
5. A. Rahim, Y. Rasheed, F. Azam, M. W. Anwar, M. A. Rahim and A. W. Muzaffar, "An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases," in IEEE Access, vol. 9, pp. 106575-106588, 2021, doi:
6. 10.1109/ACCESS.2021.3098688.
7. Garg, Apurv Sharma, Bhartendu Khan, Rizwan. (2021). Heart disease pre- diction using machine learning techniques. IOP Conference Series: Materials Science and Engineering. 1022. 012046. 10.1088/1757-899X/1022/1/012046.
8. Rajdhan, Apurb Agarwal, Avi Sai, Milan Ghuli, Poonam. (2020). Heart Dis- ease Prediction using Machine Learning. International Journal of Engineering Research and. V9. 10.17577/IJERTV9IS040614
9. Combine heart disease dataset: https://www.kaggle.com/fedesoriano/heart-failure-prediction
10. Framingham Dataset: https://framinghamheartstudy.org/
11. Mitushi Soni, Dr. Sunita Varma, "Diabetes Prediction using Machine Learn- ing Techniques", International Journal of Engineering Research Technology (IJERT), IJERTV9IS090496,Vol. 9 Issue 09, September-2020, ISSN: 2278-0181
12. Aishwarya Mujumdar, V Vaidehi, "Diabetes Prediction using Machine Learn- ing Algorithms", INTERNATIONAL CONFERENCE ON RECENT TRENDS IN ADVANCED COMPUTING 2019, ICRTAC 2019
13. Jingyu Xue, Fanchao Min, Fengying Ma, "Research on Diabetes Prediction Method Based on Machine Learning" Publication: Journal of Physics: Confer- ence Series, Volume 1684, Issue 1, article id. 012062 (2020). Pub Date: Novem- ber 2020 ,DOI: 10.1088/1742-6596/1684/1/012062

14. Quan Zou, Kaiyang Qu, Yamei Luo, Ying Ju, "Predicting Diabetes Melli- tus With Machine Learning Techniques" Front. Genet., 06 November 2018 — https://doi.org/10.3389/fgene.2018.00515
15. Md. Ashraful Alam, Dola Das, Eklas Husain, Mahmuddal Hasan "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers", March 2022, DOI:10.48550/arXiv.2203.04921
16. S. Ananya, J. Aravinth, R. Karthika, " Diabetes Prediction us- ing Machine Learning Algorithms with Feature Selection and Dimen- sionality Reduction" Published in: 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), DOI: 10.1109/ICACCS51430.2021.9441935
17. Tawfik Beghriche, Mohamed Djerioui, Youcef Brik, Bilal Attallah, and Samir Brahim Belhaouari, "An Efficient Prediction System for Diabetes Dis- ease Based on Deep