HEALTH GUARD-A Multiple Disease Prediction Model Based on Machine learning

Dr. Indu John
Dept. of Computer Science
Amal Jyothi College of Engineering
Kottayam, India
indujohn@amaljyothi.ac.in

Adithya A
Dept. of Computer Science
Amal Jyothi College of Engineering
Kottayam, India
adithyaa2024@cs.ajce.in

Alwin Rajan
Dept. of Computer Science
Amal Jyothi College of Engineering
Kottayam, India
alwinrajan2024@cs.ajce.in

Amal Biso George
Dept. of Computer Science
Amal Jyothi College of Engineering
Kottayam, India
amalbisogeorge2024@cs.ajce.in

Farhaan M Hussain
Dept. of Computer Science
Amal Jyothi College of Engineering
Kottayam, India
farhaanmhussain2024@cs.ajce.in

***Abstract*—The aim of the paper is to present a new approach to predicting multiple lifestyle diseases such as diabetes and heart disease using machine learning techniques. The proposed approach is based on ensemble learning, genetic algorithm-based recursive feature elimination, and AdaBoost. The data is preprocessed using the Multiple Imputation by Chained Equations (MICE) technique to handle missing data. This technique is used to impute missing values in the dataset by creating multiple imputations and then combining them to create a final dataset. The proposed approach also uses genetic algorithm-based recursive feature elimination to determine the optimal feature subset. This technique uses a genetic algorithm to iteratively remove features from the dataset until the optimal subset is found. The AdaBoost classification model is trained alongside other predictive models for multi-disease prediction. AdaBoost is an ensemble learning technique that combines mul- tiple weak classifiers to create a strong classifier. An extensive comparative study has been conducted to evaluate the effectiveness of the proposed model. The results show that the proposed methodology outperforms existing works in terms of prediction accuracy, precision, and recall. Overall, this study demonstrates the effectiveness of ensemble learning and genetic algorithm-based feature selection in predicting multiple diseases. The proposed approach has the potential to improve disease prediction accuracy and help healthcare professionals make more informed decisions.**

***Index Terms*—AdaBoost, Machine learning**

I. Introduction

In an era marked by increasing rates of chronic diseases such as heart disease and diabetes, the importance of early detection and prevention strategies cannot be overstated. These conditions, which have a significant negative impact on public health worldwide, are frequently characterized by subtle onsets and can go undiagnosed until they are far along, which can result in serious complications and a lower quality of life. Recognizing the urgent need for easily obtainable and effective solutions to tackle these health issues, we introduce ”Health Guard,” an online tool utilizing machine learning (ML) algorithms that forecast an individual’s risk of heart disease and diabetes.

Giving users the capacity to actively monitor their health and take preventative action to lessen the likelihood of developing these fatal diseases is the main goal of Health Guard. Utilizing an array of clinical and demographic data supplied by the user, such as age, sex, physiological parameters, and medical background, the application applies advanced machine learning techniques to produce personalized risk assessments. Through the application of machine learning (ML) algorithms, Health Guard hopes to transform the field of preventive healthcare by providing users with insightful knowledge about their personal risk profiles and assisting them in making well-informed lifestyle decisions and interventions.

Health Guard was first developed using the Support Vector Machine (SVM) algorithm for prediction tasks. To improve its predictive accuracy and resilience, Health Guard underwent iterative refinement and optimization. Afterwards, the implementation of the AdaBoost algorithm was crucial in attaining enhanced performance, resulting in a remarkable accuracy rate of 92%. Not only did AdaBoost’s special ability to group weak learners into a strong classifier improve prediction accuracy, but it also set the foundation for the application’s future scalability and flexibility.

In addition to its predictive powers, Health Guard has new features that encourage ongoing interaction and user health status monitoring. The application stores user-entered data by integrating seamlessly with a centralized database. This allows for long-term tracking of health parameters and makes it easier to compare results over time to evaluate progress. In addition to providing users with real-time insights into their health trajectory, this dynamic feedback loop facilitates personalized interventions tailored to individual needs and goals.

Looking ahead, Health Guard sees preventive healthcare as a holistic approach that goes beyond conventional disease prediction models to include proactive interventions and

DOI: 10.5281/zenodo.12519878

lifestyle changes. The application hopes to add a comprehensive diet recommendation system that provides individualized nutritional guidance based on user-specific health data by utilizing the power of ML-driven insights. Health Guard seeks to introduce a new phase in preventive healthcare by enabling people to proactively manage their health and well-being by integrating personalized interventions with predictive analytics.

This paper examines the creation, execution, and assessment of the Health Guard online application, clarifying its functionalities, approaches, outcomes, and potential future developments. We seek to highlight the revolutionary potential of ML-driven preventive healthcare solutions in addressing the growing global burden of chronic diseases through a thorough analysis of its capabilities and contributions.

II. Background

Chronic illnesses like diabetes and heart disease present serious obstacles to global public health in the modern world. Heart disease, which includes a variety of disorders that impair the heart's blood vessels and function, continues to be a major global cause of morbidity and death. Similarly, millions of people are afflicted by the crippling effects of diabetes, a condition marked by elevated blood sugar levels, which has epidemic proportions. These illnesses have a severe negative impact on people's health and well-being, but they also place a significant financial strain on healthcare systems and society at large.

Diabetes and heart disease has insidious nature, which emphasizes how crucial early detection and preventative measures are. The incidence and severity of these diseases can be considerably reduced through early risk factor identification and lifestyle modification, which will enhance overall health and quality of life. However, traditional methods of risk assessment often rely on subjective evaluations and may fail to capture subtle nuances in individual risk profiles.

In this context, a new era in disease prediction and preventive healthcare has been heralded in by the development of machine learning (ML) techniques. ML models can examine vast amounts of patient data and spot complex patterns and relationships that might escape conventional diagnostic techniques by utilizing the power of data analytics and computational algorithms. This change in approach has the potential to significantly improve disease prediction models' efficiency and accuracy, enabling more focused interventions and individualized treatment plans.

III. Literature Survey

[1] presents an ensemble learning-based multi-disease prediction system for intelligent decision support. It predicts diabetes and heart disease using AdaBoost and recursive feature elimination based on genetic algorithms. Data preprocessing eliminates redundant information and using the Multiple Imputation by Chained Equations (MICE) method missing data is handled. Comparitive analysis of K-fold cross-validation against benchmark methods shows the system's improved predictive performance. Key contributions include designing a predictive model for two lifestyle diseases and studying prevalent lifestyle diseases, aiding doctors in disease diagnosis and enhancing quality of life.

[2]presents a novel machine learning approach for estimating the risk of diabetes based on community follow-up data. In order to provide basic public health services, they use daily follow-up data gathered by community doctors. This data consists of 252,176 records from people with diabetes in Haizhu District, Guangzhou City, covering the period from 2016 to 2023. The process comprises gathering follow-up data from the community, data preprocessing to remove outliers, deal with missing values, and standardize variables. Then, using algorithms like logistic regression, decision trees, random forests, and support vector machines, a diabetes risk prediction model is created. Ultimately, a diabetes risk scorecard is created to enable quick risk evaluation using indicators of each person's unique life characteristics.

[3]describes a method to predict diabetes using a combination of techniques. Initially, they gather information from the publicly available Pima Indians Diabetes Database. Principal component analysis (PCA) and linear discriminant analysis (LDA) are two multivariate analysis techniques that are used to reduce data dimensionality and identify important predictive features after data preprocessing steps like missing value removal, normalization, and feature selection. Subsequently, a KNN-based classifier is utilized to predict diabetes based on the selected features, leveraging the k-nearest neighbors approach in the feature space. The classifier's performance is then evaluated with metrics such as AUC-ROC, sensitivity, specificity, and accuracy. To ensure accurate and dependable diabetes predictions, this methodology combines feature selection, performance evaluation, and data preprocessing.

[4]addresses the shortcomings of traditional hybrid algorithms and machine learning classifiers by combining the suggested hybrid algorithm with Modified Particle Swarm Optimization (MPSO) and Support Vector Machine (SVM). They evaluated the algorithm's predictive performance using a variety of metrics, including classification accuracy, error, correctness, recall, and F1 score, using datasets from the UCI machine learning repository. The new hybrid algorithm was compared to the conventional Support Vector Machine (SVM), the hybrid Particle Swarm Optimization Support Vector Machine (PSOSVM), and the hybrid Crazy Particle Swarm Optimization Support Vector Machine (CPSOSVM) in the experimental setup. Methodologically, the procedure involved preprocessing the datasets, selecting features using MPSO, and classifying the results using SVM. Notable adjustments were made to improve MPSO's feature selection effectiveness.

[5]presents a novel non-linear kernel that addresses limitations in existing kernels to improve the accuracy of diabetes prediction. It does this by combining the radial basis function (RBF) and RBF city block kernels. The PIMA dataset, which includes demographic and diagnostic data for female Pima Indians, is used by to thoroughly assess the suggested kernel. Careful preprocessing and data normalization methods are

used to make sure the dataset is suitable for SVM classification. Area under the curve (AUC), recall, accuracy, precision, F1-score, and polynomial and sigmoid kernels are among the performance metrics used in the comparative analysis against these well-known kernels. The proposed kernel outperforms other machine learning techniques, as evidenced by the experimental results, which show an amazing 85.5% accuracy in predicting type II diabetes on the PIMA dataset.

[6]involves obtaining the Pima Indians Diabetes Database dataset from Kaggle, then preprocessing it to eliminate any missing or unnecessary data and normalizing it for consistent feature scaling. To find pertinent features, feature selection is done with the Chi-Square test. Several machine learning models, such as Naive Bayes, K-Nearest Neighbor (KNN), support vector machine, random forest, and decision tree, are trained with the preprocessed dataset. A confusion matrix visualization is used in conjunction with performance metrics like accuracy, precision, recall, and F-score. A comparative analysis is carried out in order to find the best model. A Naive Bayes application for diabetes prediction is developed, and Particle Swarm Optimization (PSO) is used to maximize the algorithm's performance.

[7]examines several machine learning algorithms using the Pima Indian Diabetes dataset from the UCI Machine Learning Repository, including Naive Bayes, Support Vector Machine, Logistic Regression, Adaboost, Random Forest, K Nearest Neighbor, Decision Tree, and Neural Network. The Pearson's correlation approach is used to select features, and a variety of metrics are used to carefully evaluate each algorithm's performance. Remarkable accuracies of 77.6% and 76% are demonstrated by the Support Vector Machine and Logistic Regression models, respectively. Furthermore, an astounding accuracy of 88.6% is attained by a Neural Network model that has two hidden layers. The study emphasizes the accuracy of the two-layered neural network model and the effectiveness of the Logistic Regression and Support Vector Machine models. The adoption of these algorithms for early diabetes detection is strongly advocated.

[8]compares a suggested algorithm's effectiveness using the Cleveland heart disease dataset from the UCI Machine Learning Repository. The steps in its methodology are feature selection, classification, and preprocessing. While feature selection employs the EGA algorithm, preprocessing makes use of K-means clustering. Fuzzy weights are integrated into the FWSVM algorithm for classification, which assesses a variety of inputs including patient's age, sex, angina, and smoking. Accuracy, sensitivity, specificity, and F1-score are just a few of the metrics that are used in a thorough evaluation to show how well the suggested algorithm performs in comparison to other data mining algorithms such as Support Vector Machine, Neural Network, Logistic Regression, Decision Tree, and Naive Bayes classifiers, especially in terms of accuracy and efficiency. Using the EGA algorithm, specifically, optimizes feature selection and chooses more pertinent features based on the heart disease dataset, thereby increasing accuracy.

IV. Methodology

1) Dataset Acquisition: Finding appropriate datasets for the predictive models' training and testing was the first stage in creating the "Health Guard" web application. The Pima Diabetes Dataset and the Cleveland Heart Disease Dataset were obtained from Kaggle after extensive investigation. The wide range of information required to forecast heart disease and diabetes risk factors was provided by these datasets.

2) Model Development: Initially, the Support Vector Machine (SVM) algorithm was used to build the diabetes and heart disease predictive models. SVM is an effective supervised learning algorithm that excels at handling classification tasks. Using characteristics like age, sex, physiological measurements, and medical history, the SVM model was trained on the obtained datasets to predict a patient's risk of developing diabetes and heart disease.

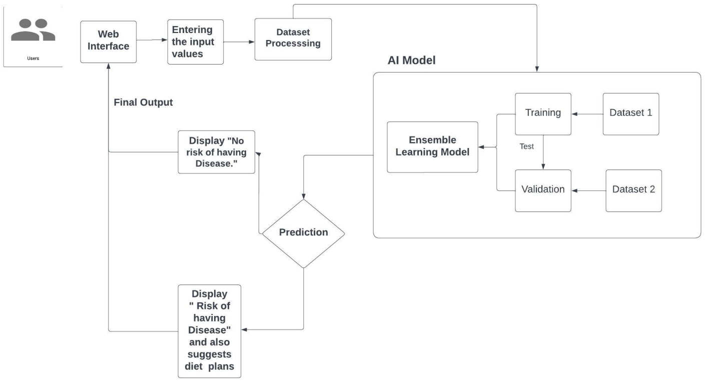


Fig. 1. Architecture Diagram

3) Algorithm Selection and Optimization: Although the SVM model produced preliminary predictions, further research indicated that predictive accuracy could be improved. It was discovered through iterative testing and analysis that the AdaBoost algorithm outperformed SVM in terms of performance. Adaptive Boosting, or AdaBoost, is an ensemble learning technique that builds a strong classifier by combining several weak learners. In "Health Guard", AdaBoost was used to train a series of weak classifiers iteratively, with each new classifier concentrating on cases that the preceding ones had misclassified. AdaBoost maximizes generalization and predictive accuracy by efficiently focusing on difficult-to-classify data points by assigning greater weight to misclassified instances. Because of this feature, AdaBoost is especially well-suited for datasets containing complex patterns and class imbalances, such as those utilized to predict diabetes and heart disease.

4) Website Development: After the predictive models were created, the "Health Guard" web application's user interface and functionality were developed during the following stage of development. This involved designing and implementing a user-friendly and intuitive website using a combination of HTML, CSS, Python, and Bootstrap. The website allows users enter clinical and demographic information, including age,

sex, and pertinent health parameters, to get personalized risk assessments for diabetes and heart disease. Additionally, a strong user authentication mechanism was put in place to guarantee the security and privacy of user data. Because of this, users had to enter unique credentials (password and username) in order to use the platform. Additionally, for effective user data storage and retrieval, the Health Guard web application integrated with the Firebase real-time database. User-provided data were safely stored in the Firebase database, allowing for persistence and long-term monitoring of user-specific health metrics.

5) Integration and Deployment: The last stage was integrating the trained predictive models into the Health Guard platform after the website was developed successfully. With the help of Flask, this integration made it possible to predict the risk of diabetes and heart disease in real time using information entered by the user. Users can access the integrated web application on various platforms and devices thanks to its deployment on a web server. The application's functionality and performance were maintained through ongoing updates and monitoring. The "Health Guard" web application accurately and securely assessed users' risk of diabetes and heart disease by using this thorough methodology, enabling them to make decisions about their health and well-being.

V. Results

In order to properly handle missing data, the Multiple Imputation by Chained Equations (MICE) technique was employed during the data preprocessing phase. MICE ensured dataset integrity for further analysis by combining and imputed missing values through multiple imputations. Because missing data can introduce biases and impair the accuracy of predictive models, this preprocessing step was essential to preserving the quality and reliability of the dataset. Additionally, MICE took into account the intrinsic complexity of lifestyle disease datasets, where missing values are frequently the result of a variety of issues like patient non-compliance or incomplete medical records.

A key factor in improving the efficacy and efficiency of disease prediction models was feature selection. Genetic algorithm-based recursive feature elimination (GA-RFE) repeatedly eliminated features from the dataset until the ideal subset was found. This strategy minimized overfitting, decreased model complexity, and increased prediction accuracy by giving priority to the most pertinent features. Only the most informative features were kept in the final model due to the iterative nature of feature selection, which enabled a thorough exploration of the feature space. Furthermore, the application of genetic algorithms offered a strong and flexible framework for feature selection that could manage datasets with many dimensions and intricate feature relationships.

As part of an ensemble learning strategy, the AdaBoost classification model was trained alongside other predictive models. In addition, an ensemble of multi-disease prediction predictive models is trained, with AdaBoost included as one of the classifiers because of its capacity to fuse weak classifiers into a strong one. We assessed the performance of the suggested model by a thorough comparative analysis with other models, taking prediction accuracy, precision, and recall into account. AdaBoost is chosen as the ultimate predictive model after a careful evaluation since it shows the highest level of accuracy out of all the models examined, it successfully captured a variety of patterns and relationships in the data by combining the strengths of several weak classifiers to produce a strong classifier. AdaBoost showed facilitated model diversity and robustness, mitigating the risk of overfitting and improving generalization performance. Moreover, further optimization and customization based on particular application requirements were made possible by the ease of integration of additional models or algorithms made possible by the flexibility of ensemble learning.

A thorough comparative analysis (Table 1) of the suggested methodology confirmed its effectiveness in forecasting a variety of lifestyle diseases. Using several datasets and experimental setups, the study evaluated a number of performance metrics, such as prediction accuracy, precision, and recall. The outcomes clearly showed how much better the suggested strategy was than the current ones, indicating that it could have practical uses in hospital environments. Through sensitivity analyses and cross-validation experiments, the methodology's robustness and scalability were further confirmed, guaranteeing consistent and dependable results under a variety of conditions.

TABLE I
Comparative Analysis

Sl.no	ML models	Training Percentage	Heart Disease	Diabetes Disease
			Accuracy	Accuracy
1	SVM	80%	78.02%	86.7%
2	KNN	80%	79.32%	74.58%
3	Decision Tree	80%	85.47%	94.47%
4	AdaBoost	80%	92.56%	83.71%

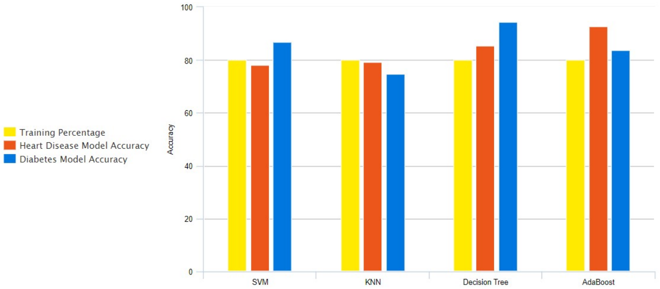
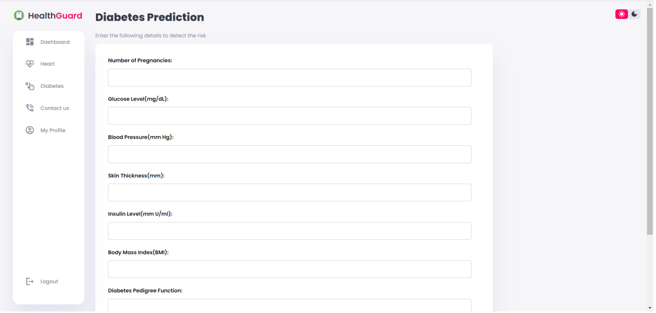


Fig. 2. Comparative Analysis

The suggested approach also takes into account dietary and nutritional guidelines specific to each person's health situation. Personalized diet recommendations are given to people whose eating habits indicate a higher risk of lifestyle diseases in order to encourage healthier eating and reduce risk factors for disease. With the help of these suggestions, people should be able to take charge of their health and make wise lifestyle decisions. By integrating diet and nutrition recommendations into the predictive model, the proposed approach offers a comprehensive solution for disease prevention and management, emphasizing the importance of dietary interventions in promoting long-term health outcomes.

Ultimately, the results of this study highlight how well ensemble learning and feature selection based on genetic algorithms can predict a variety of lifestyle diseases. This approach has the potential to redefine disease prediction and management strategies by integrating personalized interventions with advanced machine learning techniques paving the way for a more proactive and personalized approach to healthcare delivery.



Diabetes Prediction

Enter the following details to detect the risk

Number of Pregnancies:

Glucose Level(mg/dL):

Blood Pressure(mm Hg):

Skin Thickness(mm):

Insulin Level(mm U/ml):

Body Mass Index(BMI):

Diabetes Pedigree Function:

Fig. 3. User Interface of Health Guard

VI. Conclusion

In conclusion, the ”Health Guard” online application is a noteworthy development in the field of preventive healthcare, providing users with an easy-to-use interface to evaluate their risk of diabetes and heart disease. Health Guard enables people to take proactive measures to improve their health outcomes by combining machine learning algorithms with a stylish and user-friendly interface.

The Firebase database integration allows for the effective storage and retrieval of data for customized risk assessments, while the user authentication feature guarantees the security and privacy of user data. Health Guard uses Flask to integrate its models and provides real-time, accurate risk predictions for diabetes and heart disease.

For people who want to make well-informed decisions about their health and wellbeing, Health Guard is a useful tool because of its remarkable accuracy rate and dedication to user-centric design. Looking ahead, Health Guard’s ongoing development and improvement show promise for boosting public health outcomes and enabling people to live healthier lives.

References

[1] “An optimal multi-disease prediction framework using hybrid machine learning” Aditya Gupta,Amritpal Singh,2021,Dept. of Computer Science and Engineering, Dr. B R Ambedkar National Institute of Technology, Jalandhar, India

[2] Diabetes risk prediction model based on community follow-up data using machine learning. Liangjun Jiang, Zhenhua Xia, Ronghui Zhu, Haimei Gong, Jing Wang, Juan Li, Lei Wangl, 2023, College of Information and Communication Engineering, State Key Lab of Marine Resource Utilisation in South China Sea, Hainan University, Haikou, China , Electronics & Information School of Yangtze University, Jingzhou, China Shenzhen Nanshan Medical Group HQ, Shenzhen, China , E-link Wisdom Co., Ltd, Shenzhen, China , Haizhu District Community Health Development Guidance Center, Guangzhou, China.

[3] Predicting diabetes with multivariate analysis an innovative KNN-based classifier approach. BVVS Prasad, Naiwrita Borah, Hitendra Kumar Lautre, 2023,Department of CSE (School of Engineering), Anurag University, Hyderabad, Telangana, India ,Department of Computer Science Engineering, Jain (Deemed to be university) Bangalore, India , Department of ECE, Saveetha school of Engineering, Sriperumbudur, Thandalam, Tamil Nadu 602105, India , Department of chemistry, BYOS SCIENTIFIC LAB, Mowa Raipur, Chhattisgarh 492007, India ,Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur Dist, AP, India

[4] A Hybrid Machine Learning algorithm for Heart and Liver Disease Prediction Using Modified Particle Swarm Optimization with Support Vector Machine,Mandakini Priyadarshani Behera, Archana Sarangi, Debahuti Mishra, Shubhendu, 2023,ITER, Faculty of Engineering & Technology Siksha ’O’ Anusandhan Deemed to be University, Bhubaneswar, Odisha, India

[5] Improving SVM performance for type II diabetes prediction with an improved non-linear kernel: Insights from the PIMA dataset, Md.Shamim Reza , Umme Hafsha, Ruhul Amin ,Rubia Yasmin, Sabba Ruhi, 2023, Department of Statistics, Deep Statistical Learning & Research Lab, Pabna University of Science & Technology, Pabna 6600, Bangladesh , Department of Statistics, Pabna University of Science & Technology, Pabna 6600, Bangladesh

[6] Diabetes prediction using supervised machine learning. Muhammad Exell Febriana, Fransiskus Xaverius Ferdinana, Gustian Paul Sendani, Kristien Margi Suryanigrum, Rezki Yunanda, 2023, Computer Science Department,School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia.

[7] A comparison of machine learning algorithms for diabetes prediction. Jobeda Jamal Khanam, Simon Y. Foo, Department of Electrical and Computer Engineering, FAMU-FSU College of Engineering, Tallahassee, FL 32310, USA

[8] Earlier identification of heart disease using enhanced genetic algorithm and fuzzy weight based support vector machine algorithm. G. Sugendran, S. Sujatha ,2023,Department of Computer Science, Dr.G.R. Damodaran College of Science, Bharathiar University, India ,Department of Computer Science, Dr.G.R. Damodaran College of Science, Bharathiar University, Coimbatore, 641016, India