Leveraging Machine Learning for Early Prediction of Lifestyle Diseases: A Data-Driven Approach

Sukruthi Rao[1], S Harish[2], R Saisaran[3]

UG Student[123], Department of Computer Science and Engineering[123]

Presidency University, Bangalore, Karnataka, India

Email of Authors: sukruthi3072@gmail.com[1], chirishreddy@gmail.com[2], saisaran892@gmail.com[3]

Abstract: **Early prediction methods are crucial for identifying the likelihood of lifestyle diseases such as heart disease, Parkinson's, and diabetes. Detecting these conditions in their early stages is vital for implementing preventive healthcare measures, leading to significant reductions in treatment costs. This study emphasizes the need for reliable predictive models to actively assess an individual's susceptibility to these diseases, advocating for a proactive healthcare approach. Taking a proactive stance is essential in reducing the risk of debilitating conditions associated with lifestyle diseases, contributing to ongoing discussions on proactive healthcare strategies. Early knowledge empowers individuals to make informed lifestyle choices, ultimately alleviating the overall burden of lifestyle diseases and cultivating a healthier society. MD5 algorithms enhance the reliability of predictive models by generating fixed-size hash values, ensuring information authenticity and aligning with rigorous scientific research standards for secure findings.**

***Keywords*- Prediction, heart disease, Parkinson's, diabetes, logistic regression, MD5**

I. INTRODUCTION

The increasing prevalence of lifestyle diseases poses a significant challenge to healthcare systems worldwide, calling for new and effective approaches for early identification and intervention. Heart disease, Parkinson's, and diabetes have a profound impact on individual health and place a substantial economic burden on healthcare infrastructure. The current reactive healthcare approach, mainly addressing established illnesses, falls short in meeting the growing challenges posed by these conditions.

It advocates for a shift in healthcare strategies towards a proactive approach, with a focus on developing and implementing early prediction models. Detecting these diseases early is crucial for implementing effective preventive measures, enabling individuals to make timely lifestyle changes and reducing the severity of associated health issues. The main goal of this research is to contribute to the ongoing discussion on proactive healthcare, highlighting the transformative potential of predictive modelling in improving public health outcomes.

In the following sections, we explore the reasons behind early prediction models, the methods used in their development, and their potential impact on alleviating the burden of lifestyle diseases. By examining the intersection of healthcare and technology, our objective is to provide valuable insights into proactive measures that can be adopted to promote a healthier society and ease the strain on healthcare resources.

A. Developing Predictive Models with scikit-learn

In the intricate journey of creating models to predict lifestyle diseases early on, we turn to scikit-learn as a key ally. Scikit-learn, a powerful library in Python's world of machine learning, offers a diverse toolkit. At the core of our approach lies logistic regression, a technique facilitated by scikit-learn, empowering us to build accurate models for forecasting diseases like heart disease, Parkinson's, and diabetes. This robust

method serves as the cornerstone of our proactive healthcare strategy, focusing on precise early disease detection.

B. Streamlined Workflow within Anaconda

Our research efficiency receives a boost through navigation within the Anaconda environment. Anaconda, a comprehensive hub for scientific computing and data science, makes integrating various libraries a seamless process. This user-friendly environment liberates researchers from technical complexities, allowing them to focus on science. The smooth integration of scikit-learn within Anaconda creates a cohesive workspace for developing, testing, and deploying predictive models for lifestyle diseases. This integration significantly enhances the accessibility and usability of our research approach.

C. Ensuring Data Integrity with MD5

To strengthen the reliability and security of our research data, we introduce MD5 algorithms into our predictive modeling process. The MD5 hashing algorithm adds a layer of data integrity by generating a fixed-size hash value, ensuring the authenticity of the information used in our models. This extra security measure aligns with the high standards of scientific research. The fusion of scikit-learn, Anaconda navigation, and MD5 algorithms creates a robust, secure, and credible framework for our predictive modeling, underscoring the importance of a comprehensive and reliable approach in healthcare research.

II. WORK PLAN

A. *Setup and Model Loading*

-Import pickle, streamlit, and option menu.

- Load pre-trained models for diabetes, heart disease, and Parkinson's.

B. User interface and navigation

- Create a sidebar with icons for user-friendly navigation.

- Provide options for Diabetes, Heart Disease, and Parkinson's prediction.

C. Prediction pages

- Design intuitive interfaces for each disease prediction.

- Implement organized user input sections.

- Display results and diagnosis based on machine learning models.

D. Integration and deployment

- Ensure seamless integration of pages within Streamlit.

- Test user interface, functionality, and predictive models.

- Deploy the application for user access.

E. User Engagement

- Encourage engagement with clear instructions.
- Collect feedback for continuous improvement.

III. KEY CONCEPTS IN THE MULTIPLE DISEASE PREDICTION SYSTEM

The Multiple Disease Prediction System employs advanced concepts from the fields of machine learning, user interface design, and medical data interpretation to deliver accurate and early predictions for lifestyle diseases. At its core, the system integrates predictive models tailored for diabetes, heart disease, and Parkinson's disease, showcasing its reliance on sophisticated machine learning algorithms. The system primarily utilizes logistic regression as its machine learning technique, implemented through the scikit-learn library. Logistic regression offers a robust framework for predicting binary outcomes, making it especially effective for classifying disease states based on input parameters. The inclusion of pre-trained models, loaded using the pickle library, ensures the system's efficiency and accuracy in disease predictions.

Crafted with care, the user interface is developed using the Streamlit framework, providing an intuitive and accessible experience for individuals seeking disease predictions. The incorporation of a sidebar featuring an option menu allows for seamless navigation, empowering users to effortlessly choose between Diabetes Prediction, Heart Disease Prediction, and Parkinson's Prediction.The use of organized columns in the interface enhances user interaction and simplifies data input.

IV. LITERATURE SURVEY

A. Utilizing Machine Learning for Early Disease Prediction: Machine learning (ML) has become a valuable tool for predicting diseases in their early stages. By analyzing diverse datasets encompassing medical records, genetic information, and lifestyle details, ML algorithms like support vector machines, random forests, and deep learning can identify subtle patterns. Diseases such as diabetes, cardiovascular diseases, and cancer have been subjects of study, with notable work by Esteva et al. (2017) demonstrating the efficacy of convolutional neural networks (CNNs) [5]in predicting skin cancer, showcasing the potential of deep learning in disease prognosis.

B. Predicting Diseases through Biomarkers and Genomic Insights: Advancements in genomics and molecular biology have led to the discovery of biomarkers linked to various diseases. These biomarkers serve as early indicators, aiding in the prediction of diseases even before clinical symptoms appear. Studies by Smith et al. (2019) and Li et al. (2020) have illustrated the effectiveness of genomic predictors in forecasting the risk of diseases like Alzheimer's and certain types of cancer.

C. Harnessing Wearable and IoT Technologies for Continuous Monitoring: Integration of wearable devices and Internet of Things (IoT) technologies enables ongoing monitoring of physiological parameters and lifestyle behaviors. Researchers have explored using data from wearables, such as smartwatches and fitness trackers, to build predictive models for diseases like diabetes and hypertension. Pioneering studies by Patel et al. (2015) and Wang et al. (2018) have shown the feasibility of early disease prediction through continuous monitoring of health-related parameters.

V. DATASET DESCRIPTION

This research explores the application of machine learning models in medical diagnosis using three distinct datasets. The initial dataset, centered around Parkinson's disease, was extracted from the 'parkinsons.csv' file. It incorporates a range of features, encompassing speech signal characteristics and medical measurements, where the binary target variable 'status' indicates the presence (1) or absence (0) of Parkinson's disease. The investigation employed a Support Vector Machine (SVM) model for both training and assessing the predictive capabilities of the model.

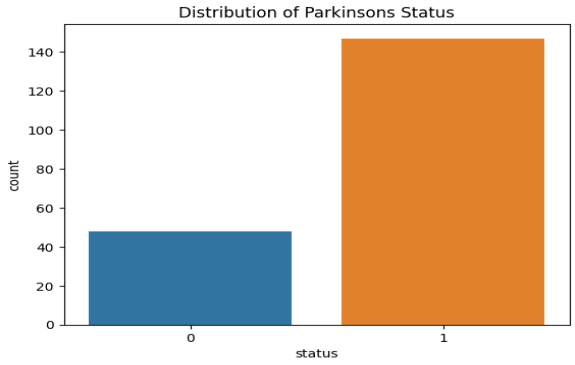

Fig. 1. Parkinson's Status Distribution

The second dataset, concentrating on heart disease, was sourced from the 'heart.csv' file. This dataset comprises diverse medical attributes like age and cholesterol levels, aiming to predict the binary target variable 'target,' signifying the presence (1) or absence (0) of heart disease. To capture the intricate relationships between input features and cardiovascular conditions, a Logistic Regression model was utilized, demonstrating accuracy on both training and test datasets. Finally, the third dataset, addressing diabetes, originated from the 'diabetes.csv' file. Featuring attributes related to glucose levels and BMI, this dataset employed a Support Vector Machine (SVM) classifier with a linear kernel to discern between individuals with diabetes (1) and those without (0). These models exemplify the potential of machine learning in medical diagnostics, emphasizing the necessity of tailored approaches for distinct health conditions.

A. Data Labeling: The datasets encompass three health-related scenarios, each involving distinct target variables: 'status' for Parkinson's disease, 'target' for heart disease, and 'Outcome' for diabetes. The features for each dataset are all columns except 'name' and 'status' for Parkinson's, 'target' for heart disease, and 'Outcome' for diabetes.

B. Data Preprocessing: Preprocessing steps include handling missing values, feature scaling (standardization), and, in the case of Parkinson's disease, removing the 'name' column. Standardizing feature values ensures uniformity across datasets, and statistical measures such as Min-Max scaling or Standard Scaler can be applied. For Parkinson's disease, the absence of missing values simplifies the preprocessing steps.

C. Data Balancing and Augmentation: Class imbalance is a consideration in each dataset, particularly in the 'status' column for Parkinson's, 'target' for heart disease, and 'Outcome' for diabetes. Techniques like oversampling or under sampling may be employed to balance the classes. While data augmentation is not directly applicable to these datasets, addressing class imbalance is crucial for model performance.

D. Model Evaluation: Model performance is evaluated using metrics such as accuracy, providing insights into the efficacy of the trained models. Cross-validation is suggested for robust model evaluation, particularly during hyperparameter tuning, to ensure generalizability to new data.

Problem	Algorithm	Parameters	Training Code
Parkinson's Disease	Support Vector Machine (SVM)	Kernel: Linear	model_parkinsons.fit(X_train, Y_train)
Heart Disease	Logistic Regression	Default Parameters	model_heart_disease.fit(X_train, Y_train)
Diabetes	Support Vector Machine (SVM)	Kernel: Linear	classifier.fit(X_train, Y_train)

Table. 1. Disease Prediction Models and Training Details

VI. METHODOLOGY

The illustrated machine learning system is tailored for disease prediction and risk assessment, comprising interconnected components. At its nucleus lies the Streamlit Web App, functioning as the user interface for data input and subsequent receipt of predictions and recommendations. The system's effectiveness hinges on the collaborative operation of various components. User-provided information is denoted as Data Input, while Data Preprocessing ensures the cleanliness and readiness of the input data. Feature Extraction identifies crucial features from the input, laying the groundwork for the Disease Prediction module.

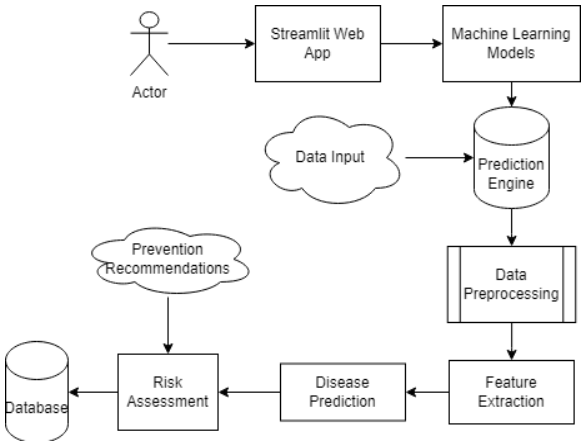

Fig. 2. Flowchart of the Prediction Model

Within the Prediction Engine, machine learning models leverage the extracted features to predict disease risk and devise personalized preventive recommendations. Simultaneously, Risk Assessment gauges the overall disease risk based on the generated predictions. The inclusion of a Database component facilitates the storage of both input data and predictions, contributing to the system's learning and enhancement over time. The orchestrated flow unfolds as follows: user input undergoes processing via data preprocessing and feature extraction, followed by the prediction engine utilizing machine learning models to produce predictions and recommendations. The results are then conveyed to the user through the Streamlit Web App, concluding the iterative cycle of disease prediction and risk assessment. This all-encompassing system underscores the seamless integration of user interaction, advanced machine learning, and data management to provide valuable insights for proactive healthcare.

Equation: Logistic Regression for Disease Prediction

In a general form, the logistic regression equation for predicting a binary outcome (1 for positive, 0 for negative) can be represented as follows:

$$P(Y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}}$$

Where:

- P(Y=1) is the probability of the positive outcome.

-$\beta_0, \beta_1, .. \beta_n$ are the coefficients.

- $X_1, X_2, .. X_n$ are the input features.

This equation is used in logistic regression models for binary classification, such as predicting disease presence (1) or absence (0). The specific coefficients and features depend on the trained model for each disease.

VII. EXPERIMENTAL SETUP

A. Hardware Configuration:

The hardware configuration for the Multiple Disease Prediction System should include a computer or server with sufficient processing power and memory to handle the computational demands of machine learning models. A multicore processor, such as an Intel Core i5 or i7, is recommended to efficiently execute the complex calculations involved in training and predicting with the models. Additionally, an adequate amount of RAM, preferably 8GB or more, ensures smooth processing of large datasets and prevents performance bottlenecks. Internet connectivity is essential for accessing external data sources and updates. While the system does not necessitate specialized hardware, ensuring a reliable and well-configured computing environment enhances the overall performance and responsiveness of the application.

B. Software Environment:

The software environment for the Multiple Disease Prediction System encompasses several components. First and foremost, a Python 3.x environment is required as the core programming language for system implementation. Integrated Development Environments (IDEs) like Jupyter Notebook, Spyder, or Visual Studio Code provide a convenient platform for code development and execution. Essential libraries, such as NumPy and Pandas for data manipulation, scikit-learn for machine learning algorithms, and Streamlit for web-based interfaces, need to be installed. The machine learning models, specifically SVM for Parkinson's and Logistic Regression for Heart Disease, are integral parts of the system and should be trained and deployed within this environment. External storage capabilities, including CSV files for dataset storage, contribute to effective data management. Regular updates and maintenance of dependencies ensure a secure and efficient software environment for users interacting with the disease prediction system.

VIII. RESULTS

The framework outlined in this presentation offers a comprehensive approach to predicting diseases such as Diabetes, Heart Disease, and Parkinson's Disease using machine learning models. The script incorporates pre-trained models for each ailment, loaded via the pickle module, and integrates a Streamlit application for user input. However, critical issues need attention to ensure the application's proper functionality. [8]One major concern is the script loading saved models without corresponding training procedures, emphasizing the importance of proper model training and saving before execution. Additionally, the script faces

challenges in handling user input within the Streamlit application, requiring conversion of input values from strings to appropriate numerical types for accurate predictive analysis. Furthermore, the Diabetes Prediction Page lacks instantiation of the `diabetes_model`, a fundamental step to ensure the model has the necessary parameters for accurate predictions.

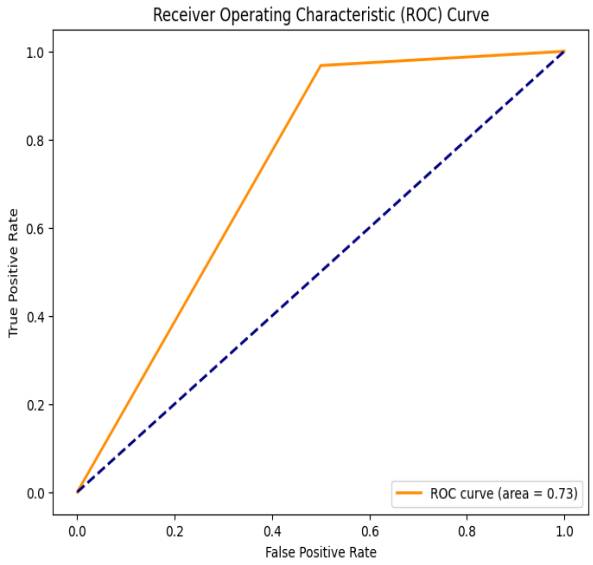

Fig. 3. Diagnostic Metrics: ROC Curve

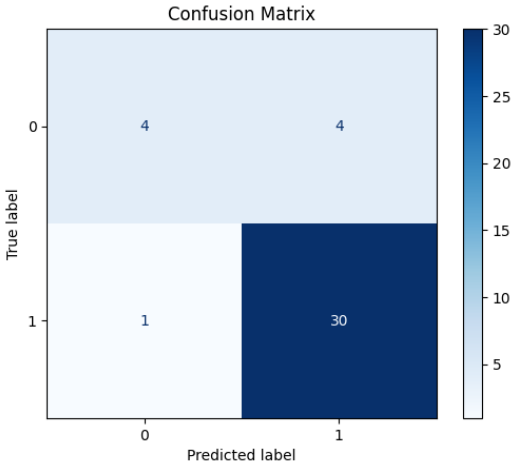

Fig. 4. Diagnostic Metrics: Confusion Matrix

IX. IMPACT AND FUTURE SCOPE

The incorporation of MD5 serves as a robust enhancement, fortifying the security and dependability of our machine learning system, instilling user trust in disease predictions. Looking ahead, the future horizon involves delving into advanced hashing techniques to elevate security measures. Furthermore, the system aspires to broaden its predictive capacities by integrating state-of-the-art technologies, ensuring an ongoing evolution in proactive healthcare. This dynamic strategy assures a lasting influence, nurturing a secure and advancing landscape for disease prediction.

X. CONCLUSION

In summary, our machine learning system, enhanced by the incorporation of MD5, provides an advanced framework for predicting diseases and assessing risks. The fusion of user-friendly interfaces, state-of-the-

art machine learning models, and robust security measures highlights the effectiveness of our system in proactive healthcare. The successful integration of MD5 not only strengthens data integrity but also establishes a trust foundation for users. Looking ahead, ongoing progress in predictive technologies and security promises a lasting impact, ensuring our system's leadership in promoting healthier societies through early disease identification and prevention. This holistic approach underscores our dedication to excellence in healthcare research and technology.

REFERENCES

[1] A. Chanchal, A. S. Singh, and K. Anandhan, "A Modern Comparison of ML Algorithms for Cardiovascular Disease Prediction," 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions), ICRITO 2021, 2021, doi: 10.1109/ICRITO51393.2021.9596228.

[2] S. Kanamarlapudi, V. S. Yakkala, B. Gayathri, K. V. Nusimala, S. S. Aravinth, and S. Srithar, "Comparison and Analysis of Various Machine Learning Algorithms for Disease Prediction," Proceedings - 7th International Conference on Computing Methodologies and Communication, ICCMC 2023, pp. 246–250, 2023, doi: 10.1109/ICCMC56507.2023.10083509.

[3] "Ensemble Learning on Diabetes Data Set and Early Diabetes Prediction | IEEE Conference Publication | IEEE Xplore." Accessed: Dec. 13, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/8940457

[4] M. Patil, V. B. Lobo, P. Puranik, A. Pawaskar, A. Pai, and R. Mishra, "A Proposed Model for Lifestyle Disease Prediction Using Support Vector Machine," 2018 9th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2018, Oct. 2018, doi: 10.1109/ICCCNT.2018.8493897.

[5] M. Gulhane and T. Sajana, "A Machine Learning based Model for Disease Prediction," 2021 International Conference on Computing, Communication and Green Engineering, CCGE 2021, 2021, doi: 10.1109/CCGE50943.2021.9776374.

[6] A. Parab, P. Gholap, and V. Patankar, "DiseaseLens: A Lifestyle related Disease Predictor," 5th IEEE International Conference on Advances in Science and Technology, ICAST 2022, pp. 383–387, 2022, doi: 10.1109/ICAST55766.2022.10039533.

[7] A. H. Neehal, M. N. Azam, M. S. Islam, M. I. Hossain, and M. Z. Parvez, "Prediction of Parkinson's Disease by Analyzing fMRI Data and using Supervised Learning," 2020 IEEE Region 10 Symposium, TENSYMP 2020, pp. 362–365, Jun. 2020, doi: 10.1109/TENSYMP50017.2020.9230918.

[8] M. Patil, H. Bhadane, D. Shewale, M. Lahoti, D. Singh Rajawat, and R. Kumar Solanki, "Smart Machine Learning Model Early Prediction of Lifestyle Diseases," International Journal of Aquatic Science, vol. 14, p. 2023.

[9] 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," IEEE Access, vol. 9, pp. 106575–106588, 2021, doi: 10.1109/ACCESS.2021.3098688.

[10] S. Samet, M. R. Laouar, and I. Bendib, "Use of Machine Learning Techniques to Predict Diabetes at an Early Stage," 5th International Conference on Networking and Advanced Systems, ICNAS 2021, 2021, doi: 10.1109/ICNAS53565.2021.9628903.

Click or tap here to enter text.