Multiple Disease Prediction Using Machine Learning

Lekhana D S[1], Likhitha N[2], Chaithrashree M C[3], Dr. Prakash[4]
[1,2,3]*UG Student, Department of Computer Science and Engineering, Sir M Visvesvaraya Institute of Technology, Bengaluru, Karnataka, India*
[4]*Professor, Department of Computer Science and Engineering, Sir M Visvesvaraya Institute of Technology, Bengaluru, Karnataka, India*

Abstract*— This study presents a multiple disease prediction system utilizing machine learning techniques to assess the likelihood of diabetes, heart disease, and Parkinson's disease. With the rising prevalence of these chronic conditions, early detection and intervention are crucial for improving health outcomes. The application features a user-friendly web interface developed with Streamlit, allowing users to input relevant health parameters for real-time predictions. Models were trained on established datasets, including the PIMA Diabetes dataset and a heart disease dataset, employing algorithms such as Support Vector Machine (SVM) and Logistic Regression, achieving accuracy scores of approximately 78% and 85%, respectively. Additionally, a model for Parkinson's disease prediction was developed using vocal features. This system enhances accessibility to health information and empowers individuals to take proactive steps in managing their health. The findings highlight the potential of machine learning in healthcare, offering a scalable solution for disease prediction and management.

***Index Terms*— Machine Learning; Support Vector Machine; K-Nearest-Neighbor; Random Forest; Diabetes; Heart disease; Parkinson's disease;**

I. INTRODUCTION

Chronic diseases like diabetes, cardiovascular conditions, and Parkinson's disease are major global health concerns, affecting millions and significantly impacting quality of life. According to the World Health Organization, cardiovascular diseases account for 31% of all deaths globally, while diabetes affects over 422 million people. Traditional diagnostic methods can be invasive, time-consuming, and often inaccessible in low-resource settings. Machine learning (ML) offers a promising alternative by analyzing complex health data to identify individuals at risk with greater speed and accuracy. This study proposes a multiple disease prediction system using ML techniques to assess the likelihood of diabetes, heart disease, and Parkinson's disease. The system features a user-friendly web interface that allows users to input key health parameters and receive real-time risk predictions. By offering accessible and accurate assessments, the application supports early detection, timely intervention, and proactive health management, ultimately contributing to improved public health outcomes.

II. LITERATURE SURVE

Recent advancements in machine learning (ML) have significantly improved early detection of Parkinson's disease (PD), a progressive neurological condition mainly affecting individuals over 50. Multiple data modalities have been investigated for building accurate diagnostic models. Audio-based detection stands out, with Govindu and Palwe (2023) achieving 91.83% accuracy using a Random Forest classifier on MDVP voice data. Their reliance on open-source Python tools enhances accessibility over MATLAB-based methods. Likewise, Bilal et al. used SVM on genetic data, achieving 88.9% accuracy, although voice data appears to offer stronger predictive power.

Movement and gait analysis has also proven valuable for PD diagnosis. Alkhatib et al. proposed a linear model that detected shuffling gait patterns with 95% accuracy, emphasizing the potential of gait as a reliable indicator. Their findings support integrating gait, audio, and sleep data for more robust diagnostic models. In parallel, Ricciardi et al. used spatial-temporal MRI features with algorithms like decision trees, Random Forest, and KNN. Despite promising results, limited sample sizes led them to apply synthetic data augmentation to ensure model consistency and performance.

Traditional ML techniques such as Random Forest and SVM continue to perform well in PD classification, often exceeding 90% accuracy when well-tuned. SVMs are particularly effective with non-linear data using appropriate kernels. KNN and logistic regression are also used but tend to yield

lower accuracy. Deep learning approaches are gaining attention for their superior performance. Wang et al. reported 96.45% accuracy using a custom deep.

Despite promising outcomes, many models rely on limited datasets and single data types, reducing accuracy and generalizability. Deep learning methods face deployment issues due to high computational demands, especially in clinical settings. Future efforts should focus on multimodal data integration, lightweight model design, and standardized protocols for real-world applicability.

III. METHODOLOGY

The methodology for our multiple disease prediction system was structured into four primary stages: data collection and preprocessing, model development and optimization, performance evaluation, and system implementation. Each phase was meticulously designed to ensure accuracy, reproducibility, and user accessibility.

A.Data Collection-and-Preprocessing

The initial phase involved acquiring robust datasets from reliable sources, including the PIMA Indian Diabetes Database for diabetes, the Cleveland Heart Disease dataset for heart conditions, and the Parkinson's Disease dataset from the UCI Machine Learning Repository. These datasets comprised essential features such as demographic data (age, gender), clinical indicators (blood pressure, glucose levels), and disease-specific measurements. Data preprocessing began with handling missing values through median imputation for numerical features and mode imputation for categorical ones. Outliers were identified and treated using the Interquartile Range (IQR) method—values lying beyond 1.5 times the IQR were winsorized to reduce skewness. Feature engineering involved one-hot encoding for categorical variables and standardizing numerical features using z-score normalization. Furthermore, additional derived features such as Body Mass Index (BMI) for diabetes and vocal metrics for Parkinson's disease were incorporated to enhance model performance.

B. Model Development and Optimization

Three primary machine learning algorithms were employed for disease prediction: Support Vector Machine (SVM) for diabetes, Logistic Regression for heart disease, and Random Forest for Parkinson's disease.The selection of these models was based on their strengths—SVM's suitability for high-dimensional data, Logistic Regression's interpretability, and Random Forest's robustness through ensemble learning. The Random Forest model included 200 decision trees with a maximum depth of 15 to prevent overfitting. Hyperparameter tuning was performed using GridSearchCV with 5-fold cross-validation. For the SVM model, kernel types (linear, polynomial, RBF) and regularization parameter C (0.1, 1, 10) were evaluated. Logistic Regression parameters included L1 and L2 regularization with similar C values. The Random Forest model's tuning involved varying the number of estimators (100, 200, 300) and maximum tree depths (10, 15, 20). Root Mean Squared Error (RMSE) was used as the primary metric for optimization, while R^2 was monitored to maintain explanatory strength.

C.Performance-Evaluation

The predictive performance of each model was rigorously evaluated using several standard metrics: accuracy, precision, recall, F1 score, and ROC-AUC. Accuracy provided a general measure of correctness, while precision and recall assessed the model's capability to identify true positives without excessive false alarms. The F1 score, being the harmonic mean of precision and recall, balanced both metrics. ROC-AUC served as an effective measure of class separability. To ensure unbiased evaluation, each dataset was split into 80% training and 20% testing data using stratified sampling, preserving the proportion of positive and negative cases across both sets. All performance metrics were computed on the test set to reflect the models' generalizability to unseen data.

D.System-Implementation

The trained models were deployed through a user-friendly web application developed using Streamlit. This interface enabled users to input relevant health parameters—such as age, BMI, and glucose levels—and receive real-time disease risk predictions. Dropdown menus facilitated categorical selections like gender, and validation checks ensured that inputs were within medically appropriate ranges. The

system also provided visual feedback indicating the predicted risk level for each disease. The entire solution was encapsulated in a web-based interface, requiring no additional installations, thereby improving accessibility.

IV. RESULTS AND DISCUSSIONS

The predictive models developed for diabetes, heart disease, and Parkinson's disease were evaluated using standard performance metrics, including accuracy, precision, recall, and F1 score. The Support Vector Machine (SVM) model for diabetes prediction achieved an accuracy of approximately 78%, with a precision of 0.75 and a recall of 0.80, leading to an F1 score of 0.77. These results indicate a reasonably reliable model for early detection of diabetes, particularly in identifying at-risk individuals. The Logistic Regression model used for heart disease prediction performed the best among the three, achieving an accuracy of 85%, with a precision of 0.82, recall of 0.88, and an F1 score of 0.85. This high level of performance highlights the model's effectiveness in correctly classifying individuals at risk of heart disease. For Parkinson's disease prediction, the Random Forest model achieved an accuracy of 82%, precision of 0.80, recall of 0.84, and an F1 score of 0.82. Feature importance analysis further confirmed that vocal characteristics were among the most influential predictors, aligning with existing literature on Parkinson's diagnosis.

When comparing the performance of all three models, the Logistic Regression model for heart disease emerged as the most accurate, likely due to the strong linear relationships between its features and the target variable. The SVM model for diabetes showed comparatively lower performance, possibly due to the complexity and variance in the dataset, which may have required more advanced feature selection or additional data preprocessing. The Random Forest model for Parkinson's disease demonstrated its strength in capturing complex, non-linear relationships, which is particularly beneficial given the intricate patterns present in vocal and neurological features. These results affirm the importance of selecting algorithms that align with the data characteristics of each disease domain.

The outcomes of this study carry meaningful implications for healthcare delivery, particularly in preventive medicine. By enabling early risk assessment, the developed machine learning models can support timely medical consultations and interventions.

The integration of these models into a user-friendly web application further enhances accessibility, allowing individuals to input their health parameters and receive immediate, interpretable feedback regarding their disease risk. This ease of use can promote health awareness and encourage proactive health management among users, even those without technical expertise.

V. ACKNOWLEDGMENT

The authors thank Dr .Prakash A for his expert guidance and valuable insights throughout this research project. Thanks are extended to the Department of Computer Science and Engineering at Sir M Visvesvaraya Institute of Technology for providing the necessary resources and infrastructure. The authors also acknowledge the healthcare institutions whose data were used in estimations for this study.

REFERENCES

[1] R. Shukla and R. Sawant, "Multiple Disease Prediction System Using Machine Learning," 2023 International Conference on Electrical, Electronics, Communication and Computers (ELEXCOM), Roorkee, India, 2023, pp. 1-6.

[2] A. Mangal and V. Jain, "Performance analysis of machine learning models for prediction of diabetes," 2022 2nd International Conference on Innovative Sustainable Computational Technologies (CISCT), Dehradun, India, 2022, pp. 1-4.

[3] S. S, A. S, G. V. V. Rao, P. V, K. Mohanraj and R. Azhagumurugan, "Parkinson's Disease Prediction Using Machine Learning Algorithm," 2022 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS), Chennai, India, 2022, pp. 1-5.

[4] M. Rahman, S. Islam, S. B. Sarowar and M. T. Zaman, "Multiple Disease Prediction using Machine Learning and Deep Learning with the Implementation of Web Technology," 2023 IEEE International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings), Mount Pleasant, MI, USA, 2023, pp. 1-7.