Health Prognosis System Using Machine Learning

Mohit Aggarwal[1], Manish Yadav[2], Shubham Yadav[3], and Arshad Khan[4]
[1,2,3]*Student, Galgotias College of Engineering and Technology, Greater Noida*
[4] *Assistant Professor, Galgotias College of Engineering and Technology, Greater Noida*

***Abstract*—This research paper explores the development of a Health Prognosis System (HPS) leveraging machine learning (ML) algorithms to predict multiple diseases, such as diabetes, heart disease, and Parkinson's disease. The system integrates logistic regression and support vector machines (SVM) within a user-friendly interface powered by Streamlit. Key objectives include enhancing early detection, improving personalized healthcare, and addressing limitations in existing single-disease prediction models. By employing various datasets, preprocessing techniques, and performance metrics, this paper highlights the efficacy of ML algorithms in providing reliable health prognoses.**

***Keywords*— Machine Learning, Health Prognosis System, Streamlit, Logistic Regression, Support Vector Machines, Multi-Disease Prediction.**

I. INTRODUCTION

A. Overview

The field of healthcare has witnessed significant transformations with the integration of machine learning (ML) technologies, paving the way for enhanced diagnostics and more precise treatment planning. Diseases such as diabetes, heart disease, and Parkinson's disease have a profound impact on global health, often presenting with overlapping risk factors. However, traditional diagnostic systems typically focus on single-disease prediction models, leading to fragmented and inefficient healthcare delivery.

A unified approach to predicting multiple diseases simultaneously can revolutionize preventive healthcare by offering comprehensive insights into a patient's overall health status. This research introduces the Health Prognosis System (HPS), a multi-disease prediction model powered by ML algorithms such as logistic regression and support vector machines (SVM). By analyzing key health parameters like blood pressure, cholesterol levels, and heart rate, the system aims to provide early detection and actionable insights into diabetes, heart disease, and Parkinson's disease.

The proposed system is designed with accessibility in mind, leveraging Streamlit to create an intuitive interface that allows users to input health data and receive real-time predictions. This paper discusses the datasets, preprocessing techniques, and machine learning algorithms utilized to build the HPS, emphasizing its potential to bridge gaps in existing healthcare systems. By demonstrating the effectiveness of multi-disease prediction models, the study highlights the transformative role of ML in improving patient outcomes and promoting personalized healthcare solutions.

B. Problem Statement.

Conventional The current state of healthcare analytics faces significant limitations in addressing the complex interplay between multiple diseases that share overlapping risk factors, such as diabetes, heart disease, and Parkinson's disease. Existing diagnostic systems are predominantly disease-specific, focusing on single-disease prediction through siloed models. This fragmented approach hinders a comprehensive understanding of a patient's overall health, potentially delaying diagnosis and treatment of comorbid conditions.

For example, an individual with symptoms indicative of both heart disease and diabetes might need to consult multiple specialists and undergo separate diagnostic procedures. This not only consumes time and resources but also increases the likelihood of missed or delayed diagnoses, especially for interrelated conditions. Moreover, the lack of integration between these systems creates inefficiencies in healthcare delivery and reduces the overall quality of patient care.

Another challenge lies in accessibility. Many existing diagnostic tools are either confined to clinical settings or require specialized knowledge to operate, limiting

their utility for early intervention, especially in underserved regions. Furthermore, many systems fail to provide actionable insights, leaving users without a clear understanding of their health status or the steps needed to improve it.

The development of a unified, multi-disease prediction model addresses these critical gaps by offering a streamlined framework that evaluates a patient's risk for multiple diseases simultaneously.

Such a system can leverage machine learning algorithms to analyze diverse health parameters and provide early, accurate, and comprehensive prognoses. This integration not only saves time but also empowers healthcare providers and individuals to make informed decisions about preventive care and treatment.

By building a system accessible through an intuitive interface, we aim to democratize advanced healthcare technologies, enabling users to input their health data and receive reliable predictions in real time. The proposed system seeks to overcome the fragmentation and inefficiencies of existing models while promoting a holistic approach to healthcare that emphasizes early
diagnosis, personalized treatment, and improved outcomes.

II. EXISTING SYSTEM

The existing system for multi-disease prediction uses machine learning algorithms like Support Vector Machines (SVM), Logistic Regression, Naive Bayes, Decision Trees, and Random Forests to predict individual diseases such as diabetes, heart disease, and Parkinson's disease. For diabetes prediction, the SVM classifier achieves 76% accuracy by analyzing factors like blood pressure, cholesterol, and BMI. Logistic Regression is used for heart disease prediction, yielding 79% accuracy, primarily focusing on linear relationships in patient data. For Parkinson's disease, SVM achieves an 85% accuracy rate.

However, the existing system lacks a unified framework to predict multiple diseases simultaneously. Users must interact with separate models for each disease, which can be time-consuming and cumbersome. Additionally, the system does not offer a simple, interactive interface for users, making it less accessible to non-technical individuals. The proposed system seeks to address these limitations by integrating disease predictions into a single, user-friendly platform, improving accessibility and prediction accuracy.

III. PROPOSED SOLUTION

The proposed Health Prognosis System (HPS) is a comprehensive and integrated solution for predicting multiple diseases, designed to overcome the limitations of traditional single-disease models. It utilizes advanced machine learning algorithms and user-centric design principles to provide accurate and timely health predictions.

A. Features

The HPS offers a unified framework for multi-disease prediction, enabling early detection and effective health management. By integrating multiple disease models, the system ensures comprehensive risk assessment for diabetes, heart disease, and Parkinson's disease. An intuitive interface powered by Streamlit ensures accessibility for users with varying levels of technical expertise. Furthermore, the system adheres to stringent data privacy and security standards, ensuring user trust.

B. Technologies Used

The system employs cutting-edge machine learning algorithms tailored for specific diseases. Logistic regression is utilized for binary classification tasks such as heart disease prediction, while support vector machines (SVM) handle non-linear patterns, especially for Parkinson's disease. Random forest algorithms enhance prediction reliability through ensemble learning. Data preprocessing, including cleaning, normalization, and feature selection, ensures robust input for these models. Finally, the Streamlit framework is leveraged to create an interactive and scalable interface, deployed on a cloud-hosted infrastructure for global accessibility.

IV. METHODOLOGY

The methodology for the Health Prognosis System is divided into distinct stages to ensure accurate and reliable disease prediction while maintaining a user-friendly interface. The approach integrates data collection, preprocessing, algorithm selection, model

training, and deployment within an interactive platform.

A. Data Collection

The system relies on publicly available datasets sourced from platforms such as Kaggle and UCI Machine Learning Repository. These datasets include relevant health parameters for diabetes, heart disease, and Parkinson's disease, such as blood pressure, glucose levels, heart rate, and patient demographics. The datasets were chosen for their comprehensive coverage of attributes required for disease prognosis.

a. Data Preprocessing

Effective preprocessing is critical to improve the accuracy of machine learning models. Key preprocessing steps included:

b. Cleaning:

Removal of null values, duplicates, and outliers to ensure data integrity.

c. Normalization:

Scaling numerical data to a standard range, enhancing algorithm performance.

d. Feature Selection:

Utilizing correlation matrices to identify the most influential parameters for each disease, reducing dimensionality and computational complexity.

B Algorithm Selection

The system incorporates a hybrid machine learning approach by leveraging different algorithms tailored to the characteristics of each disease:

a. Logistic Regression: Selected for efficiency in binary classification tasks, particularly for heart disease prediction.

b. Support Vector Machines (SVM): Utilized for its ability to model non-linear relationships, making it suitable for Parkinson's disease prognosis.

c. Random Forest: Deployed for its ensemble learning capability to improve prediction reliability across datasets.

C. Model Training and Evaluation

The datasets were split into training (80%) and testing (20%) subsets to evaluate model performance. During training, hyperparameter optimization techniques, such as grid search, were used to fine-tune the algorithms. Performance metrics, including accuracy, precision, recall, and F1-score, were employed to assess model effectiveness.

a. System Deployment

The trained models were integrated into a web-based application developed using Streamlit, a Python framework for creating interactive data-driven interfaces. The deployment process included:

Developing an intuitive interface for users to input their health data.

Implementing backend support for real-time predictions using the trained machine learning models.

Ensuring security and privacy compliance, such as encrypted data transmission and storage.

D. Validation and Testing

The system underwent rigorous validation using cross-validation techniques to minimize overfitting and ensure robust performance across varied datasets. Usability testing was conducted to refine the interface and ensure a seamless user experience.

V. IMPLEMENTATION

The implementation of the Health Prognosis System (HPS) involved a detailed, step-by-step process, incorporating data handling, model development, interface design, and deployment. Below, the implementation stages are elaborated to demonstrate the development workflow.

A. Data Handling and Preparation

Data handling formed the backbone of the system, ensuring the input data was suitable for machine learning tasks. The datasets for diabetes, heart disease, and Parkinson's disease were acquired from open-access repositories like Kaggle and UCI Machine Learning Repository. Each dataset underwent the following preprocessing steps:

a. Data Cleaning:

Missing values were imputed using mean or median values based on the distribution of data.

Outliers were identified using statistical methods such as the Z-score and removed to enhance data quality.

b. Feature Scaling and Normalization:
Continuous attributes like blood glucose levels and cholesterol were scaled using Min-Max normalization to a range of 0 to 1, ensuring uniformity for algorithms sensitive to feature magnitude.

c. Feature Engineering:
Correlation matrices were used to identify the most significant predictors for each disease.
Unnecessary features were eliminated to reduce computational overhead and prevent overfitting.

B. Model Development
The machine learning models were developed using Python's Scikit-learn library, chosen for its flexibility and robustness. The implementation process for each model is detailed below:
a. Logistic Regression:
Used for heart disease prediction due to its simplicity and effectiveness in binary classification problems.
Hyperparameters such as regularization strength (C) were optimized using grid search.

C. Support Vector Machines (SVM):
Implemented for Parkinson's disease, leveraging its ability to handle non-linear relationships via the Radial Basis Function (RBF) kernel.
Hyperparameters like kernel type, regularization parameter, and gamma were fine-tuned using cross-validation.
a. Random Forest:
Integrated to combine predictions for enhanced accuracy and reliability, especially for diabetes prognosis.
The number of estimators and maximum depth were adjusted to achieve a balance between performance and computational efficiency.
The models were trained on 80% of the data, with the remaining 20% used for evaluation. Stratified sampling ensured balanced representation of classes in both training and testing sets.

D. User Interface Development
The user interface was designed using Streamlit, a Python-based library that facilitates the rapid development of web applications. The implementation steps included:
a. Input Mechanism:
Forms and sliders were created for users to enter key health parameters such as blood pressure, glucose levels, and heart rate.
Validations were applied to ensure users input realistic and clinically relevant data ranges.
b. Output Display:
Predictions were presented in a clear and actionable manner, with disease probabilities displayed as percentages alongside confidence intervals.
Color-coded indicators (e.g., green for low risk, red for high risk) enhanced interpretability.

E. Backend Integration
The backend connected the interface to the trained machine learning models. Key steps included:
a. Model Serialization:
Models were serialized using Python's joblib library for efficient loading and inference during runtime.
b. Prediction Logic:
User inputs were processed to match the format of the training data.
Predictions were generated in real time by invoking the appropriate model based on the disease being analyzed.

F. Deployment and Hosting
The finalized system was deployed on a cloud platform using Streamlit Cloud for accessibility. Key considerations during deployment included:
a. Scalability:
The system was tested to handle multiple simultaneous user requests without degradation in performance.
b. Security:
HTTPS was enforced to secure data transmission.
User inputs were not stored or logged, maintaining privacy and compliance with data protection standards.

G. Validation and Testing
The implementation was validated through rigorous testing phases:
a. Functional Testing:
Ensured the system performed as intended, delivering accurate predictions and responsive user interactions.
b. Performance Testing:
Measured response times and system stability under varying loads.
c. User Testing:
Conducted usability tests with a sample audience to gather feedback on the interface and prediction clarity.
The meticulous implementation ensured the Health Prognosis System met its objectives of providing a

reliable, secure, and user-friendly platform for multi-disease prediction.
Stratified sampling ensured balanced representation of classes in both training and testing sets.
providing a reliable, secure, and user-friendly platform for multi-disease prediction.

VI. RESULT AND ANALYSIS

The results of the multi-disease prediction system using machine learning techniques such as Support Vector Machines (SVM), Logistic Regression, and other algorithms have shown promising potential in predicting various diseases, including diabetes, heart disease, and Parkinson's disease. The system's performance was evaluated using several performance metrics like accuracy, precision, recall, and F1-score. Below is a detailed analysis of the results obtained for each disease prediction model.

A. Diabetes Prediction:

The prediction of diabetes was performed using the Support Vector Machine (SVM) classifier. The dataset included several key features such as blood pressure, cholesterol levels, BMI, age, and family history of diabetes. After preprocessing the data, including normalization and handling missing values, the SVM model was trained and tested using a dataset split into 70% for training and 30% for testing.

a. Accuracy:

The SVM model achieved an accuracy of 76% for diabetes prediction, meaning that 76% of the predictions made by the model were correct. This result is a strong indication that the SVM model can correctly identify individuals at risk of diabetes based on the given features.

b. Precision and Recall:

The precision for diabetes prediction was around 74%, which signifies the proportion of true positive predictions out of all positive predictions made by the model. The recall was 78%, which reflects how well the model identifies actual positive instances of diabetes.

c. Analysis:

While the 76% accuracy is promising, the model could be further improved with more extensive data, fine-tuning of hyperparameters, and the inclusion of additional features such as lifestyle factors, diet, and activity levels. Also, the SVM algorithm performed well due to its ability to handle non-linear relationships between input features, but it still struggles with more complex cases that may require additional data.

B. Heart Disease Prediction:

For heart disease prediction, the Logistic Regression model was employed. Logistic Regression is particularly suitable for binary classification problems like predicting the presence or absence of heart disease. The input features for heart disease prediction included age, sex, chest pain type, resting blood pressure, cholesterol levels, and fasting blood sugar.

a. Accuracy: Logistic Regression achieved an accuracy of 79%, which is higher than the diabetes prediction model. This shows that the Logistic Regression model can effectively classify heart disease patients based on the selected features.
b. Precision and Recall: The precision was 80%, meaning that the model correctly identified heart disease patients 80% of the time out of all predicted positive cases. The recall was 77%, showing that the model identified 77% of actual heart disease cases.
c. Analysis: The Logistic Regression model showed effective performance for heart disease prediction. The higher accuracy could be attributed to the linear relationships between the predictors and the outcome. However, like the diabetes prediction model, further improvements could include using more advanced techniques or incorporating a broader set of patient data to refine predictions.

C. Parkinson's Disease Prediction:

Parkinson's disease was predicted using the SVM classifier, with key features such as motor function, speech patterns, age, and gender. Parkinson's disease prediction is inherently complex due to the subtle nature of its early symptoms, making this an interesting test case for machine learning models.

a. Accuracy: The SVM model achieved an impressive accuracy of 89% for Parkinson's disease prediction. This high level of accuracy suggests that the model is highly capable of distinguishing between individuals with and without Parkinson's disease.
b. Precision and Recall: The precision for Parkinson's disease prediction was 88%, while the recall stood at 91%. These results indicate that the model was able to identify nearly 91%

of all actual Parkinson's disease cases and made very few false positive predictions.

c. Analysis: The SVM model demonstrated strong performance in predicting Parkinson's disease, likely due to its ability to handle the complex, non-linear relationships in the dataset. The higher accuracy and recall reflect that SVM is particularly effective in situations with subtle differences between the classes, which is characteristic of Parkinson's disease.

Comparative Analysis of All Models:

a. Diabetes: SVM performed reasonably well with a 76% accuracy, but there is still room for improvement. Future work could involve experimenting with other models such as Random Forest or Neural Networks to see if they yield higher performance.
b. Heart Disease: Logistic Regression showed a solid 79% accuracy, demonstrating that linear models can perform well with structured medical data.
c. Parkinson's Disease: The SVM model outperformed the other models with an 89% accuracy, suggesting that it is the most effective algorithm for this specific task among those tested.

Overall, the proposed multi-disease prediction system has proven effective in predicting diseases using machine learning algorithms, with particular success in predicting Parkinson's disease. While the accuracy of each model is commendable, there is potential for improving the system's performance by integrating additional data, fine-tuning model parameters, and exploring more complex models. The accuracy and reliability of predictions could also be enhanced by implementing an ensemble approach or combining different algorithms for each disease.

VII. DISCUSSION

The results highlight the potential of machine learning algorithms in transforming healthcare diagnostics, making early disease detection more accessible and efficient. The system offers several advantages, such as the ability to simultaneously predict multiple diseases, which can save time and improve overall patient care. Additionally, the user-friendly interface powered by the Streamlit library ensures that even non-expert users can easily interact with the system.

However, the accuracy of predictions is dependent on the quality of the data. For more reliable results, it is essential to ensure that the data is comprehensive and accurate. The performance of the models could also be further improved by employing advanced techniques like feature engineering, deep learning, and incorporating additional clinical data, such as genetic information, lifestyle habits, and family medical history.

In conclusion, the system demonstrates the effectiveness of machine learning in predicting multiple diseases and provides a foundation for developing a more advanced, real-time diagnostic tool for healthcare providers. With continuous enhancements and data enrichment, this system can play a crucial role in improving early detection, patient management, and overall healthcare outcomes.

VIII. CONCLUSION

In conclusion, the multi-disease prediction system using machine learning algorithms such as Support Vector Machine (SVM) and Logistic Regression has shown significant promise in the early detection of diseases such as diabetes, heart disease, and Parkinson's disease. By leveraging various input features, the models were able to predict the likelihood of these diseases with reasonably high accuracy. Among the models tested, SVM performed particularly well for Parkinson's disease, achieving an impressive 89% accuracy, while Logistic Regression demonstrated strong results for heart disease prediction with a 79% accuracy.

Despite the promising results, there remains room for improvement. The accuracy of the diabetes prediction model, for example, was slightly lower (76%), indicating the need for further refinement, possibly through the use of more advanced algorithms, hyperparameter tuning, or by incorporating additional data. The performance of these models could also be enhanced by utilizing more diverse and comprehensive datasets, as well as exploring ensemble methods to combine the strengths of multiple algorithms.

The system's ability to predict multiple diseases simultaneously opens new possibilities for healthcare, potentially reducing diagnostic time and improving patient outcomes. By implementing a

user-friendly interface with tools like Streamlit, the system is also accessible to healthcare professionals who may not have extensive experience with machine learning, making it a valuable tool in clinical settings.

Overall, this system demonstrates the vast potential of artificial intelligence and machine learning in revolutionizing healthcare, providing early disease detection, and assisting medical professionals in making informed decisions. With continuous improvements and data integration, the system can be further optimized to deliver more accurate and timely predictions, ultimately leading to better prevention, management, and treatment of diseases.

REFERENCE

[1] Katarya, R., & Srinivas, P. (2020). Predicting heart disease at early stages using machine learning: A survey. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 302–305.

[2] Poudel, R. P., Lamichhane, S., Kumar, A., et al. (2018). Predicting the risk of type 2 diabetes mellitus using data mining techniques. *Journal of Diabetes Research*, 2018, 1686023.

[3] Ismaeel, S., Miri, A., & Chourishi, D. (2015). Using the extreme learning machine (ELM) technique for heart disease diagnosis. In *2015 IEEE Canada International Humanitarian Technology Conference (IHTC 2015)*, 1–3.

[4] Gavhane, A., Kokkula, G., Pandya, I., & Devadkar, K. (2018). Prediction of heart disease using machine learning. In *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 1275–1278.

[5] Deo, R. C. (2015). Machine learning in medicine. *Circulation*, 132(20), 1920–1930. .

[6] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth and Brooks.

[7] Parashar, A., Gupta, A., & Gupta, A. (2014). Machine learning techniques for diabetes prediction. *International Journal of Emerging Technology and Advanced Engineering*, 4(3), 672–675.

[8] Paniagua, J. A., Molina-Antonio, J. D., Lopez-Martinez, F., et al. (2019). Heart disease prediction using random forests. *Journal of Medical Systems*, 43(10), 329.

[9] Subramanian, S., & Raj, P. (2020). A comparative study of machine learning models for heart disease prediction. *International Journal of Advanced Research in Computer Science and Software Engineering*, 10(3), 12–16.

[10] Sushmitha, P., & Geetha, G. (2019). Early detection of diabetes using machine learning. *Journal of Engineering Research and Applications*, 9(2), 15–18.