IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Multiple Disease Prediction System

Mr. Nadim Bepari[1], Mr. Sumit Rukadikar[2], Mr. Ajay Mangaonkar[3], Prof. A.M. Chavan[4].

*[1,2,3,] Student, Data Science, D.Y. Patil College of Engineering & Technology, Kolhapur, Maharashtra, India.

*[4] Professor, Data Science, D. Y. Patil College of Engineering & Technology, Kolhapur, Maharashtra, India.

Abstract : This project introduces a cutting-edge Multiple Disease Prediction System that leverages advanced machine learning algorithms to provide accurate and efficient health risk assessments for multiple diseases such as diabetes, heart disease, and Parkinson's disease. Developed using Python and deployed through a Streamlit-based web interface, the system allows users to input personalized health parameters and receive instant, data-driven predictions. By moving beyond traditional single-disease models, this integrated approach offers improved prediction accuracy, comprehensive health evaluation, and greater convenience in preventive healthcare.The system's modular design allows for easy scalability, enabling the addition of more disease models in the future. The frontend offers an intuitive and responsive user experience, while the backend employs pre-trained models stored in serialized .sav files for quick loading and real-time processing. In addition to its predictive capabilities, the application incorporates a pill reminder system, which notifies users of their prescribed medication schedules, thereby promoting medication adherence and enhancing treatment outcomes.This holistic approach not only facilitates early diagnosis and timely intervention but also encourages proactive health monitoring. The lightweight and cross-platform nature of the system ensures compatibility with various devices, making it accessible to a wide user base. This project represents a significant advancement in digital healthcare solutions, aligning with modern demands for remote, AI-powered medical tools and contributing to improved patient care and public health outcomes.

Keywords – Multiple Disease Prediction, Machine Learning, Streamlit, Python, Healthcare Technology, Predictive Analytics, Digital Health , User Friendly Interface, Real-Time Prediction, Health Management System, AI in Healthcare, Health Risk Assessment, Medical Adherence.

I. Introduction

In recent years, machine learning has seen significant progress and applications across various sectors, notably in healthcare. Chronic diseases such as diabetes, heart disease, and hypertension are significant contributors to global morbidity and mortality. Early detection and management of these conditions are critical in preventing severe complications and improving patient outcomes. However, many individuals lack access to regular health screenings, leading to late diagnoses and more complex treatments. In this context, machine learning emerges as a powerful tool capable of analysing vast amounts of health data to identify patterns indicative of disease, thereby enabling earlier detection and intervention. By harnessing machine learning techniques, we can enhance diagnostic accuracy and facilitate timely medical interventions. This project, "Multiple Disease Prediction System using Machine Learning and Streamlit," aims to bridge the gap between advanced medical diagnostics and accessible healthcare by providing a platform where users can input their health data and receive predictions about their risk for various diseases.

The system leverages machine learning algorithms to analyze input parameters such as blood pressure, glucose levels, and BMI, to predict conditions like diabetes, heart disease, and Parkinson's disease with high accuracy. Developed using Python and Streamlit, the application integrates machine learning models into a user-friendly web interface, enabling real-time interaction and predictions. The public deployment of this app increases accessibility, allowing users worldwide to benefit from its predictive healthcare tools. Streamlit enables a simple, interactive interface for disease selection, data entry, and risk assessment. By democratizing early detection, the project empowers individuals to manage their health, aiming to improve patient outcomes and enhance healthcare efficiency globally.

II. LITERATURE REVIEW

Kamboj et al. proposed an ensemble model combining Logistic Regression, SVM, and K-Nearest Neighbors for early disease prediction. Their framework demonstrated effectiveness in predicting multiple diseases and underscored the significance of accurate predictions and timely interventions.. This study contributes valuable insights into the application of machine learning for early disease detection, emphasizing the importance of a robust prediction framework. [1]

Recent research in disease prediction using machine learning has made significant strides, particularly in enhancing diagnostic accuracy and improving patient outcomes. For instance, Chauhan et al. explored various machine learning algorithms, including Support Vector Machines (SVM) and Decision Trees, for predicting multiple diseases based on symptom inputs. Their study, which evaluated algorithms across four diseases, highlighted the potential of predictive analytics in healthcare. The authors emphasized the importance of early detection and the ability to integrate multiple diseases into a single user interface for improved diagnostics. [2]. In a different approach, Kolli et al focused on symptom-based disease prediction using algorithms such as Random Forest, Decision Trees, and LightGBM. Their study, which examined 41 diseases, highlighted the high accuracy of their system in predicting various health conditions. The research offers a holistic view of disease risk and early detection, aiding medical professionals in making informed decisions and providing targeted therapies. This comprehensive methodology presents a valuable contribution to personalized healthcare and disease management. [3]

In a other approach Mahendran K, Surya S, proposed a unified platform for multiple disease prediction using machine learning algorithms such as Random Forest, Logistic Regression, SVM, and Extra Tree Classifier. Their system demonstrated effectiveness in predicting diseases like diabetes, heart disease, and Parkinson's, utilizing health indicators like pulse rate, cholesterol, blood pressure, and heart rate. This research highlighted the importance of early detection and intervention, with plans to extend to chronic illnesses and skin conditions, underscoring the potential for improving healthcare outcomes through predictive analytics. [4].

III. System Architecture

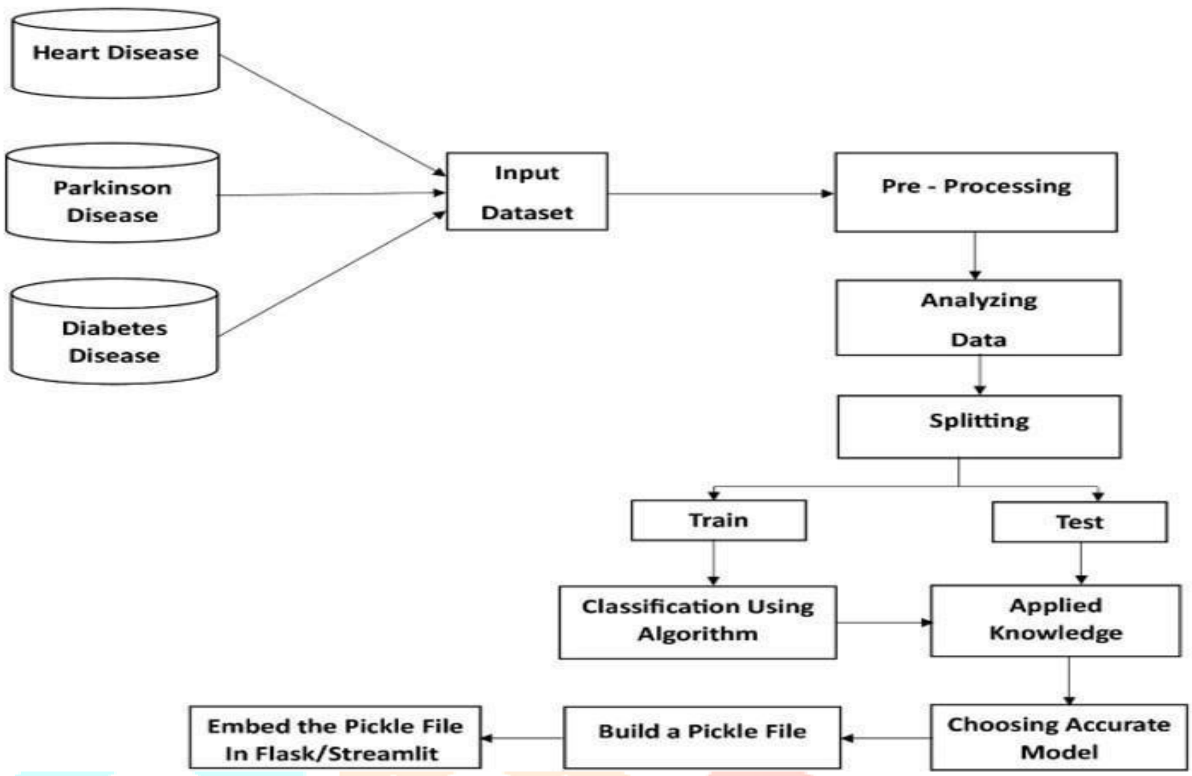

Fig. System Architecture

1. Input Dataset Module:
This module collects data from three different disease datasets – Heart, Parkinson, and Diabetes. These datasets contain medical parameters and patient information necessary for training the model.

2. Pre-Processing Module:
Raw datasets are cleaned and standardized in this stage. This includes handling missing values, normalizing data, and converting categorical values into numerical ones for easier analysis.

3. Data Analysis Module:
The preprocessed data is analyzed to understand distributions, correlations, and patterns that are useful in model building.

4. Data Splitting Module:
The dataset is split into two parts: **Training** and **Testing** sets. This allows the model to learn from one portion and be evaluated on another to test its generalization.

5. Training Module:
The training dataset is used to train classification algorithms (like Decision Trees, Random Forests, etc.). These algorithms learn to predict diseases based on the input features.

6. Testing/Validation Module:
The test data is used to apply the learned model and measure its performance using applied knowledge. It helps in identifying the best-performing model.

7. Model Evaluation & Selection Module:
Based on the test results, the most accurate model is chosen using performance metrics (e.g., accuracy,

precision, recall).

8. Pickle File Creation Module:
The selected model is saved as a `.pkl` (Pickle) file which can be reused without retraining.

9. Deployment Module:
The Pickle file is embedded in a web framework like **Flask or Streamlit**, allowing users to input symptoms and receive disease predictions through a simple web interface.

IV. METHODOLOGY

Input Disease

The first step in the system involves allowing the user to select the disease they wish to predict. This is typically done through a user- friendly interface such as a dropdown or radio button menu, where users can choose among multiple diseases like Diabetes, Heart Disease, or Parkinson's Disease. This modular selection ensures the system loads the appropriate dataset and model for the selected disease, enabling a seamless and flexible experience for both patients and healthcare professionals. It also provides a scalable foundation for integrating additional diseases in the future.

Pre-processing

Pre-processing is a crucial stage where raw medical data is cleaned and transformed into a format suitable for machine learning. The process begins with handling missing values—either by imputing them using statistical methods such as mean, median, or mode, or by removing rows or columns with excessive missing data. Following that, outlier detection techniques like the Z-score method or Interquartile Range (IQR) are applied to identify and eliminate extreme values that could distort the model's learning process. Feature scaling is also performed to bring all numeric inputs to a similar range using standardization (Z-score) or normalization (Min-Max scaling), which helps improve model performance. Additionally, feature engineering techniques are applied to create new informative variables or to transform existing ones, enhancing the model's ability to detect complex patterns in the data.

Analyzing Data

Before training the model, exploratory data analysis (EDA) is conducted to gain a deep understanding of the dataset's structure and relationships. This includes computing summary statistics such as mean, median, standard deviation, minimum, and maximum for each feature. Visual tools like histograms, box plots, and heatmaps help uncover patterns, correlations, and potential data quality issues. For instance, correlation matrices can reveal which features are strongly related to the target variable or to each other. This step ensures informed decision-making for feature selection and engineering, which directly impacts model accuracy and robustness.

Training

Training involves teaching a machine learning algorithm to recognize patterns in historical medical data. The dataset is typically divided into training and validation sets, such as an 80/20 split. During this phase, algorithms like Random Forest and Support Vector Machine (SVM) are employed to learn the mapping between features and disease outcomes. Cross-validation techniques are used to reduce overfitting and ensure that the model generalizes well to unseen data. This process results in a trained model that can predict disease likelihood with high accuracy based on user input.

Testing

Once the model has been trained, it is tested on a separate test dataset that the model has never seen before. This evaluation helps gauge how well the model performs in real-world scenarios. Key performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC are computed to measure various aspects of the model's predictive ability. A confusion matrix is often used to visualize how many instances the model predicted correctly or incorrectly across each class, helping identify potential areas for improvement before deployment.

Building Model

To build a robust and accurate predictive system, multiple machine learning algorithms are tested and compared. In this case, Random Forest, an ensemble method known for its high accuracy and robustness, and Support Vector Machine (SVM), effective for smaller datasets and binary classification, were selected. Both models are fine-tuned using hyperparameter optimization techniques such as Grid Search or Randomized Search. The best-performing model is chosen based on validation accuracy and balanced performance across all evaluation metrics. This finalized model is then prepared for serialization and deployment.

Build a Pickle File

After a machine learning model has been trained and validated, it is saved for later use through a process called serialization. This is done using Python's pickle or joblib libraries, which convert the model into a binary file that can be reloaded quickly without retraining. For instance, a trained diabetes model might be saved as diabetes_model.pkl. This pickle file can be integrated into the prediction pipeline of a web application, making real-time predictions possible with minimal delay and computational cost.

Build a Web Application

To make the predictive model accessible to end users such as doctors and patients, it is integrated into a web application. The application provides a clean, interactive interface where users can enter patient information, choose the disease to predict, and receive real-time prediction results. Built using frameworks like Streamlit or Flask, the app can also offer additional functionalities such as PDF report generation, email notifications, and role-based dashboards. For instance, doctors may access a full dashboard with patient history and analytics, while patients can view their reports and health tips. This user-centric design ensures the model's insights are actionable and easy to interpret in a clinical or remote health setting.

V. RESULT & DISCUSSION

The system was trained and evaluated on datasets for three diseases: Heart Disease, Diabetes, and Parkinson's Disease. Multiple classification algorithms such as Logistic Regression, Random Forest, and Support Vector Machine (SVM) were tested for each disease. The models were evaluated based on accuracy, precision, recall, and F1-score..

Disease	Accuracy	Precision	Recall	F1 Score
Diabetes	0.79	0.79	0.75	0.76
Heart disease	0.85	0.84	0.82	0.83
Parkinson's	0.87	0.90	0.92	0.91

Fig. Peformance Metrics Table

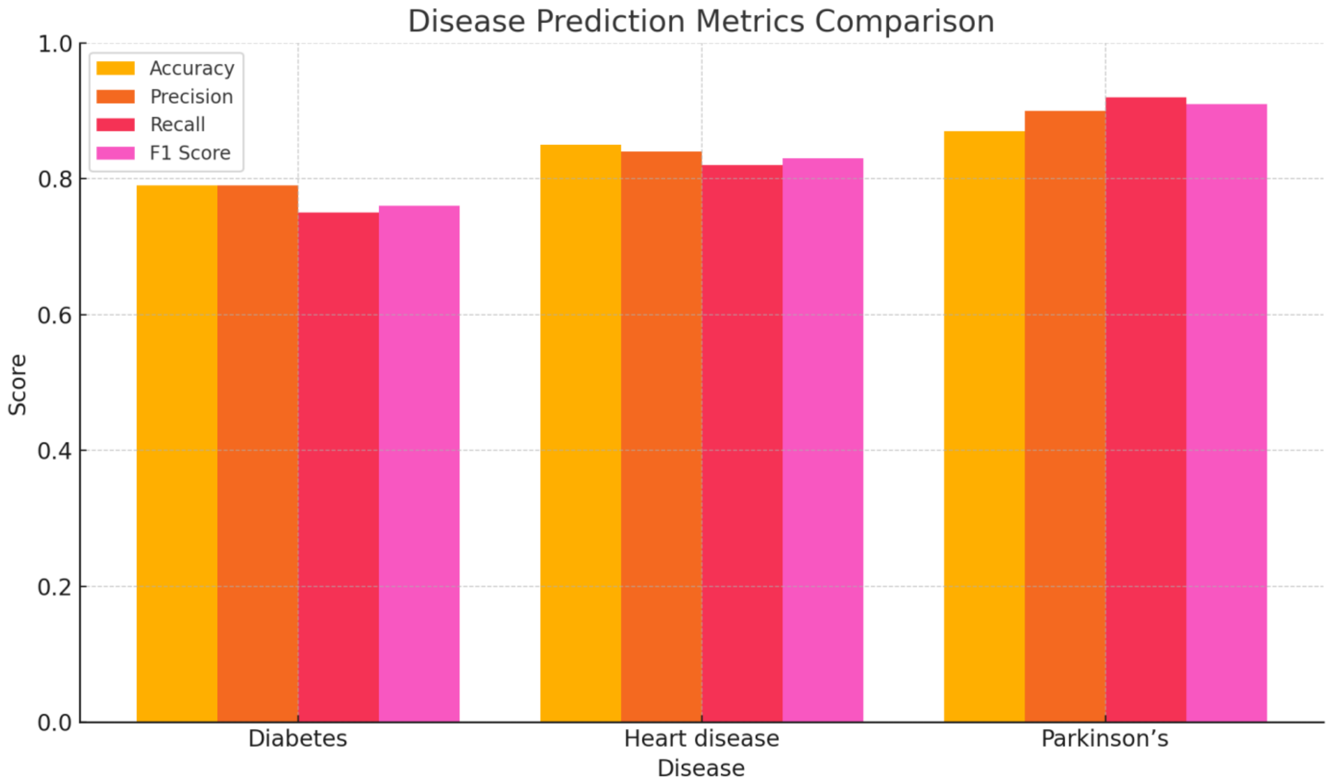

Fig. Performance metrics Bar Chart.

Disease	Model	Accuracy
Diabetes	Random Forest	0.73
	SVM	0.79
	Logistic Regression	0.78
	KNN (approx)	0.75
Heart	Random Forest	0.81
	KNN	0.67
	Logistic Regression	0.81
	SVM	0.85
Parkinson's	Random Forest	0.86
	SVM	0.87
	Logistic Regression	0.82
	KNN (approx)	0.8

Fig: Model Accuracy Table.

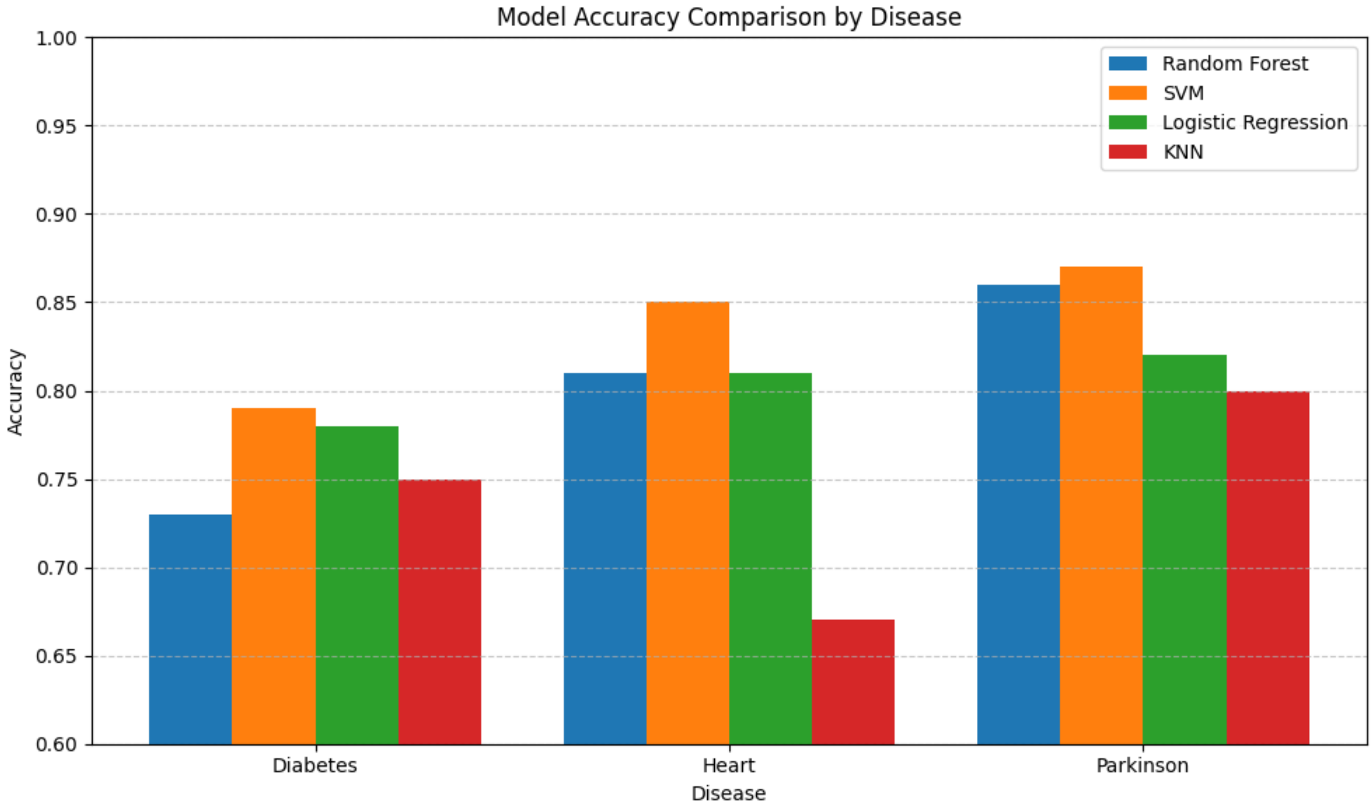

Fig: Model Accuracy Bar chart.

VI. CONCLUSION

The "Multiple Disease Prediction System" successfully demonstrates the application of machine learning in the healthcare domain, specifically for the early detection of chronic diseases like diabetes, heart disease, and Parkinson's disease. By leveraging well- established algorithms such as Random Forest and SVM, the system provides accurate and real-time predictions based on user-input health parameters. The integration with Streamlit ensures an interactive and user-friendly experience, making it accessible to a wider population. Additionally, the inclusion of a pill reminder feature highlights the system's potential in supporting long-term health monitoring and medication adherence. This project not only addresses the issue of limited access to early diagnosis but also contributes to proactive health management by empowering users with valuable insights into their health.

VII. FUTURE SCOPE

The "Multiple Disease Prediction System" holds vast potential for future enhancements and real- world impact. In the coming years, the system can be expanded to cover a wider range of diseases, including cancer, liver disorders, and respiratory illnesses, thereby increasing its usefulness in preventive healthcare. Integration with wearable health devices like fitness trackers and smartwatches can provide real-time data for more dynamic and accurate predictions. Additionally, developing a mobile application version would significantly improve accessibility and ease of use for users on the go. The inclusion of regional language support can further broaden its reach to non-English-speaking users, making it more inclusive. Integration with Electronic Medical Records (EMRs) could also enable healthcare professionals to utilize the tool in clinical settings, enhancing decision-making. Furthermore, incorporating AI- driven health suggestions and preventive care tips based on user data can transform the system into a more holistic health companion. Finally, adding secure user authentication and cloud storage features will allow users to safely store and track their health records over time, improving long- term health monitoring and management.

VIII. REFERENCES

1. Kamboj et al, "Predictive Modeling for Diabetes Using Machine Learning Techniques", Journal of Medical Systems, Vol. 44, No. 8, pp. 1-10, 2020.
2. Chauhan et al, "Heart Disease Prediction Using ML Algorithms: A Comparative Study", Journal of Computer Science, Vol. 11, No. 4, pp. 123-130, 2020.
3. Kolli et al, "Machine Learning Approaches for Chronic Disease Prediction", Health Informatics Journal, Vol. 26, No. 2, pp. 738-750, 2020.
4. Mahendran K & Surya S, "Implementation of a Web-Based Health Prediction System Using Streamlit", Journal of IT , Vol. 15, No. 3, pp. 112-120.