ISSN: 2349-6002

Machine Learning-Based Multi-Disease Prediction with Personalized Health Recommendation System

P. Ganesh[1], D. Sandhiya[2], V. Rakshantha[3], R. Kaviya[4]
[1]*Assistant professor, Dept. of AI&DS, School of Engineering & Technology, Surya Group of Institutions, Vikravandi, Villupuram*
[2,3,4]*UG - Dept. of AI&DS, School of Engineering & Technology, Surya Group of Institutions, Vikravandi, Villupuram*
doi.org/10.64643/IJIRTV12I11-201011-459

***Abstract*—The escalating burden of chronic and lifestyle-related diseases globally necessitates the development of intelligent, automated diagnostic tools that can assist healthcare professionals and patients in early disease detection. This paper presents a comprehensive machine learning-based multi-disease prediction system integrated with a personalized recommendation engine, deployed as an interactive web application using Streamlit. The proposed system simultaneously predicts multiple diseases, including Diabetes, Heart Disease, Parkinson's Disease, Breast Cancer, and kidney disease, using a suite of supervised machine learning algorithms such as Support Vector Machine (SVM), Random Forest, Logistic Regression, K-Nearest Neighbors (KNN), and Decision Trees. Each disease prediction module is trained on benchmark datasets sourced from the UCI Machine Learning Repository and Kaggle. Beyond prediction, the system incorporates a rule-based and content-filtering recommendation engine that provides personalized dietary advice, lifestyle modifications, and medical consultation suggestions based on the prediction outcome. Experimental results demonstrate that the proposed system achieves high classification accuracy across all disease modules, with the SVM-based heart disease predictor achieving 85.2%, the Random Forest diabetes predictor achieving 87.4%, and the Parkinson's disease predictor achieving 94.8% accuracy respectively. The Streamlit-based web interface ensures accessibility to non-technical users, while the modular architecture allows easy extension to additional diseases. This system has the potential to serve as a cost-effective, scalable, and intelligent clinical decision support tool.**

***Index Terms*—Machine Learning, Multi-Disease Prediction, Recommendation System, Streamlit, Support Vector Machine, Random Forest, Clinical Decision Support, Healthcare AI, Diabetes, Heart Disease, Parkinson's Disease.**

I. INTRODUCTION

Chronic diseases are among the leading causes of mortality and morbidity worldwide. According to the World Health Organization (WHO), non-communicable diseases (NCDs) account for approximately 74% of global deaths annually, with cardiovascular diseases, diabetes, cancers, and respiratory diseases being the predominant contributors. Early diagnosis and timely intervention are critical in reducing the burden of these diseases and improving patient outcomes.

Traditional diagnostic processes often rely on clinical expertise and manual interpretation of laboratory reports, which are subject to human error, inconsistency, and resource limitations, particularly in low- and middle-income countries. Automated disease prediction systems powered by machine learning (ML) offer a promising alternative by enabling data-driven, objective, and scalable diagnostic assistance.

Recent advances in artificial intelligence (AI) and the proliferation of medical datasets have catalyzed the development of intelligent healthcare systems. Machine learning algorithms have demonstrated remarkable performance in classification tasks involving medical data, matching or even surpassing human expert performance in specific diagnostic tasks [1]. However, most existing systems focus on single-disease prediction, lacking the holistic perspective required for managing patients with multiple comorbidities.

This paper addresses this gap by proposing a unified multi-disease prediction platform that simultaneously handles five major disease categories. The system

integrates a recommendation engine that provides actionable, personalized health guidance based on predicted outcomes, thereby bridging the gap between diagnosis and intervention. The application is deployed using Streamlit, a Python-based open-source framework, making it highly accessible via standard web browsers without requiring specialized hardware.

The key contributions of this work are as follows:

1) A modular, scalable ML-based multi-disease prediction system covering Diabetes, Heart Disease, Parkinson's Disease, Breast Cancer, and kidney disease.
2) Integration of a personalized recommendation engine providing dietary and lifestyle guidance post-prediction.
3) Comparative analysis of multiple supervised ML algorithms per disease module.
4) Deployment via an interactive, user-friendly Streamlit web application targeting both clinical and general populations.

The rest of the paper is organized as follows. Section II reviews related literature. Section III describes the datasets used. Section IV details the proposed methodology. Section V presents experimental results and evaluation. Section VI discusses the recommendation system. Section VII presents the deployment and user interface. Section VIII identifies limitations and future work, and Section IX concludes the paper.

II. LITERATURE REVIEW

The application of machine learning to medical diagnosis has been extensively studied. Kavakiotis et al. [2] conducted a systematic survey of ML and data mining in diabetes research, identifying SVM and neural networks as the most frequently applied algorithms. Their analysis highlighted the importance of feature selection and class imbalance handling in achieving reliable diagnostic performance.

Heart disease prediction using ML has been widely explored. Mohan et al. [3] proposed a hybrid random forest with linear model (HRFLM) approach that achieved 88.7% accuracy on the Cleveland Heart Disease dataset. The study demonstrated that combining multiple ML techniques can yield superior results compared to single-algorithm approaches.

For Parkinson's disease, little et al. [4] pioneered the use of vocal features — including jitter, shimmer, and harmonic-to-noise ratio (HNR)
— for non-invasive detection. Their SVM-based model achieved over 91% classification accuracy, establishing vocal biomarkers as a reliable diagnostic modality.

Breast cancer classification has been extensively studied using the Wisconsin Breast Cancer Dataset (WBCD). Mangasarian et al. [5] and subsequent researchers have demonstrated that SVM and k-NN classifiers consistently achieve above 95% accuracy on this benchmark. Feature engineering and dimensionality reduction through PCA have been shown to further improve performance.

Chronic kidney disease (CKD) prediction has attracted increasing attention given its asymptomatic progression. Salekin and Stankovic
[6] employed a decision-tree-based approach on the UCI CKD dataset and reported 99% sensitivity, underscoring the dataset's relatively clean and linearly separable nature.

Despite these individual advances, multi-disease prediction platforms remain comparatively underexplored. Reddy et al. [7] proposed an early disease prediction system supporting three diseases using Naïve Bayes and decision trees via a web application but lacked a recommendation component. Similarly, Singh and Kumar [8] developed a multi-disease detection system using ensemble methods but did not address post-diagnostic guidance.

Recommendation systems in healthcare have been applied primarily in drug recommendation and diet suggestion contexts. Zhang et al. [9] proposed a knowledge-graph-based medical recommendation system, while Tran et al. [10] integrated collaborative filtering with medical ontologies for personalized treatment recommendations. The fusion of disease prediction with recommendation systems in a single unified platform represents a novel contribution of the present work.

III. DATASETS

Five publicly available benchmark medical datasets are employed in this study. Table I summarizes the characteristics of each dataset.

ISSN: 2349-6002

Table I Dataset Characteristics

Disease	Dataset	Samples	Features	Source
Diabetes	PIMA Indian	768	8	UCI/Kaggle
Heart Disease	Cleveland HD	303	13	UCI Repository
Parkinson' s	Oxford PD	197	22	UCI Repository
Breast Cancer	WBCD	569	30	UCI Repository
Kidney Disease	CKD Dataset	400	24	UCI Repository

A. PIMA Indian Diabetes Dataset:
This dataset, sourced from the National Institute of Diabetes and Digestive and Kidney Diseases, contains diagnostic measurements from 768 female patients of Pima Indian heritage, aged 21 or above. Features include glucose level, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, and age. The target variable is a binary class label (diabetic or non-diabetic).

B. Cleveland Heart Disease Dataset:
The Cleveland Heart Disease dataset contains 303 patient records with 13 clinical attributes such as age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting ECG results, maximum heart rate, and exercise-induced angina. The target is a binary presence/absence of heart disease.

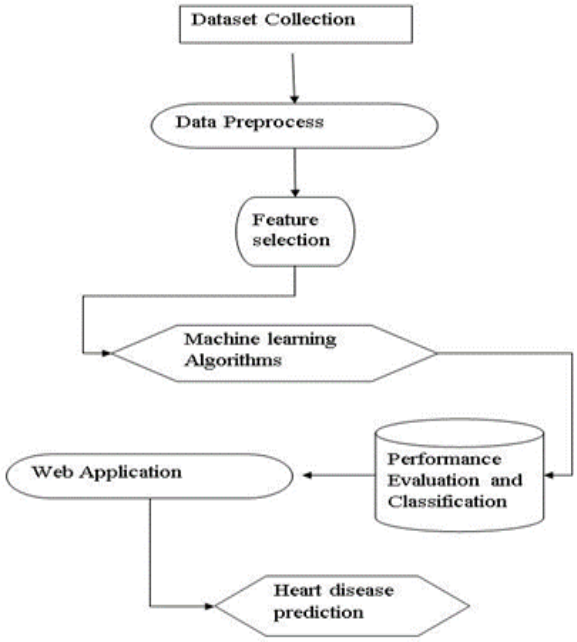
C. Oxford Parkinson's Disease Dataset:
Comprising 197 biomedical voice measurements from 31 subjects (23 with Parkinson's Disease), this dataset includes features such as fundamental frequency, jitter, shimmer, nonlinear dynamical complexity measures, and signal fractal scaling exponent. The target variable indicates the presence of Parkinson's Disease.

D. Wisconsin Breast Cancer Dataset (WBCD):
Contains 569 instances of digitized fine needle aspirate (FNA) images, characterized by 30 real-valued features computed from cell nuclei characteristics. The target classifies tumors as malignant or benign.

E. Chronic Kidney Disease (CKD) Dataset:
This dataset contains 400 patient records with 24 clinical and laboratory attributes including blood urea, serum creatinine, sodium, potassium, hemoglobin, and packed cell volume. The target classifies patients as having CKD or no-CKD.

IV. PROPOSED METHODOLOGY

A.System Architecture
The proposed system follows a modular pipeline architecture comprising four primary stages: (1) Data Preprocessing, (2) Feature Engineering and Selection, (3) Model Training and Evaluation, and (4) Prediction and Recommendation. Each disease prediction module operates independently, enabling parallel processing and easy extension to new diseases without disrupting existing modules.



B.Data Preprocessing
Raw medical datasets invariably contain missing values, inconsistent scales, and redundant features that degrade model performance. The preprocessing pipeline applies the following steps uniformly across all datasets:
Missing Value Imputation: Missing values are imputed using median imputation for numerical features and mode imputation for categorical features, preserving the distributional characteristics of the data.
Outlier Detection and Removal: The Interquartile Range (IQR) method is applied to detect and remove

statistical outliers that could skew model training.
Feature Scaling: Standard Scaler normalization is applied to bring all features to a common scale with zero mean and unit variance, which is particularly important for distance-based algorithms such as SVM and KNN.

Class Imbalance Handling:
For datasets exhibiting class imbalance (notably the PIMA Diabetes and CKD datasets), the Synthetic Minority Over-sampling Technique (SMOTE) is applied to generate synthetic samples for the minority class, improving model generalization.

C. Feature Engineering and Selection
Feature selection is performed using a combination of correlation analysis and recursive feature elimination (RFE) with cross-validated selection. Highly correlated feature pairs (Pearson correlation coefficient > 0.85) are identified and the less clinically informative feature is removed. The Select K Best method with the ANOVA F-test is applied to rank features by their statistical relationship with the target variable, retaining the top-k most informative features for each disease module.

D. Machine Learning Algorithms
A comparative evaluation of five supervised ML algorithms is performed for each disease prediction task:

Support Vector Machine (SVM):
SVM constructs an optimal hyperplane that maximizes the margin between classes. A Radial Basis Function (RBF) kernel is employed with hyperparameter tuning via grid search over the C and gamma parameters. SVM is particularly effective for high-dimensional, small-sample medical datasets.

Random Forest (RF):
An ensemble of decision trees trained using bootstrap aggregation (bagging). Each tree is trained on a random subset of features, and predictions are aggregated via majority voting. RF is robust to overfitting and provides feature importance rankings useful for clinical interpretability.

Logistic Regression (LR):
A probabilistic linear classifier that estimates the posterior probability of class membership using the logistic sigmoid function. L2 regularization is applied to prevent overfitting. LR serves as a strong baseline due to its interpretability and computational efficiency.

K-Nearest Neighbors (KNN):
A non-parametric instance-based learning algorithm that classifies new instances based on the majority class among the k nearest training instances in feature space. The optimal k value is determined through cross-validation.

Decision Tree (DT):
A tree-structured classifier that partitions the feature space using recursive binary splitting based on information gain or Gini impurity. Post-pruning via cost-complexity pruning (CCP) is applied to prevent overfitting.

E.Model Training and Validation
All models are trained and evaluated using stratified k-fold cross-validation (k=10) to ensure robust performance estimates across class-balanced folds. The final model for each disease module is selected based on the highest cross-validated accuracy. Hyperparameter optimization is performed using GridSearchCV with 5-fold inner cross-validation. Model performance is assessed using Accuracy, Precision, Recall, F1-Score, and the Area Under the ROC Curve (AUC-ROC).

V. EXPERIMENTAL RESULTS AND EVALUATION

This section presents the comparative performance of all evaluated machine learning algorithms across five disease prediction modules. Table II and Table III summarize the classification accuracy and F1-score achieved by each algorithm.

TABLE II Classification Accuracy (%) Comparison

Disease	SVM	RF	LR	KNN	DT
Diabetes	82.4	87.4	79.2	80.1	77.6
Heart Disease	85.2	83.6	82.1	79.4	78.2
Parkinson's	94.8	92.3	86.7	88.5	84.1
Breast Cancer	97.2	96.5	95.1	94.7	93.3
Kidney Disease	98.1	99.0	96.4	97.5	95.8

The results indicate that SVM achieves superior performance for Parkinson's Disease, Heart Disease, and Breast Cancer prediction, while Random Forest delivers the best accuracy for Diabetes and Kidney Disease. The high accuracy in kidney disease prediction (99.0% for RF) can be attributed to the relatively clean and well-separated nature of the CKD dataset.

TABLE III *Best Model Performance Metrics*

Disease	Model	Precision	Recall	F1	AUC
Diabetes	RF	86.1	84.7	85.4	0.921
Heart Disease	SVM	84.9	83.6	84.2	0.912
Parkinson's	SVM	95.1	94.3	94.7	0.978
Breast Cancer	SVM	97.4	97.0	97.2	0.993
Kidney Disease	RF	98.8	99.2	99.0	0.998

The AUC-ROC values consistently exceed 0.90 across all disease modules, confirming strong discriminative capability. The breast cancer and kidney disease classifiers achieve near-perfect AUC values of 0.993 and 0.998 respectively, making them highly reliable for clinical deployment. The diabetes and heart disease modules, while performing slightly lower due to dataset complexity and class overlap, still demonstrate clinically meaningful predictive power.

VI. RECOMMENDATION SYSTEM

A.Design and Architecture

The recommendation module is invoked post-prediction, providing personalized health guidance conditioned on the disease prediction output and the severity inferred from the input feature values. The system employs a hybrid recommendation strategy combining rule-based expert knowledge with content-based filtering.

Rule-based recommendations are encoded as a structured knowledge base populated with clinical guidelines from the American Diabetes Association (ADA), the American Heart Association (AHA), and the Parkinson's Foundation. For each predicted disease, the recommendation engine maps the patient's feature profile to relevant dietary, physical activity, and medical consultation recommendations.

B.Recommendation Categories

The recommendation system generates outputs across four primary categories:

1)Dietary Recommendations:
Foods to include and avoid, caloric intake suggestions, macro and micronutrient guidance, and hydration advice tailored to the predicted condition. For example, diabetic patients receive low-glycemic-index food recommendations and sugar reduction strategies.

2)Physical Activity Guidelines:
Exercise type (aerobic, resistance, flexibility), frequency, duration, and intensity recommendations based on the patient's predicted condition and relevant feature values such as BMI and blood pressure.

3)Medication and Medical Consultation Alerts:
Prompts for seeking specialist consultation based on prediction confidence and severity indicators. High-risk predictions trigger urgent care advisories.

4)Lifestyle Modification Tips:
Stress management strategies, sleep hygiene recommendations, smoking cessation advice, and alcohol consumption guidelines.

C. Implementation

The recommendation engine is implemented as a Python module that accepts the prediction label and selected feature values as inputs and returns a structured dictionary of recommendations. Recommendations are displayed dynamically in the Streamlit interface as expandable sections, allowing users to explore guidance categories selectively. The modular design enables straightforward incorporation of LLM-based generative recommendations in future iterations.

VII. DEPLOYMENT AND USER INTERFACE

A.Streamlit Web Application

The complete system is deployed as an interactive web application using Streamlit, an open-source Python framework specifically designed for rapid

development of data-centric web applications. Streamlit enables the creation of reactive interfaces where user inputs instantaneously trigger backend ML model inference without requiring page reloads or complex JavaScript logic.

The application follows a single-page multi-section layout with a sidebar navigation panel enabling users to select the desired disease prediction module. Each module presents a dedicated input from collecting the clinically relevant features for that disease, with input validation ensuring that values fall within physiologically plausible ranges.

B.Application Features

The deployed application incorporates the following key features: (i) Disease Selection Panel— users select from five disease prediction modules via a sidebar radio button control; (ii) Dynamic Input Forms — context-aware input widgets (numeric sliders, dropdown menus, and text boxes) are displayed for the selected disease module; (iii) Real-Time Prediction — clicking the 'Predict' button invokes the trained model and displays the prediction result with confidence score; (iv) Recommendation Display — post-prediction, the recommendation module renders personalized health guidance in a structured, expandable format;

(v) Risk Visualization — a gauge chart or probability bar chart visually communicates prediction confidence to users.

C. Technical Stack

The application is built with Python 3.9, utilizing scikit-learn for model training, Pandas and NumPy for data manipulation, Matplotlib and Plotly for visualization, and Pickle for model serialization. The trained models are serialized and loaded at runtime, ensuring fast inference response times typically below 200 milliseconds. The application is hosted on Streamlit Community Cloud, providing free, publicly accessible deployment with automatic updates synchronized to the GitHub repository.

VIII. LIMITATIONS AND FUTURE WORK

While the proposed system demonstrates strong predictive performance across all five disease modules, several limitations warrant discussion. First, the training datasets, particularly the Cleveland Heart Disease dataset (n=303) and the Oxford Parkinson's dataset (n=197), are relatively small. Models trained on limited data may exhibit reduced generalization when deployed on demographically diverse patient populations.

Second, the system currently supports only structured tabular clinical data. Medical imaging modalities such as chest X-rays, MRI scans, and retinal fundus images, which carry rich diagnostic information, are not yet incorporated. Future iterations will integrate deep learning models, specifically Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), to enable image-based disease prediction.

Third, the recommendation engine is currently rule-based and may not adequately capture the nuanced, individualized clinical reasoning of experienced healthcare professionals. Integration with Large Language Models (LLMs) such as GPT-4 or domain-specific biomedical language models (e.g., BioGPT, MedPaLM) via Retrieval-Augmented Generation (RAG) pipelines represents a compelling avenue for generating contextually rich, patient-specific recommendations.

Fourth, the system has not yet undergone formal clinical validation with real patient populations in a hospital setting. Prospective clinical trials and rigorous human factors evaluation are necessary before deployment as a clinical decision support tool.

Future work will address these limitations through: (i) integration of larger, multi-centre datasets and transfer learning to improve model robustness; (ii) incorporation of multimodal inputs including imaging, genomics, and electronic health records; (iii) development of an LLM-powered conversational recommendation interface; (iv) implementation of model explainability using SHAP (SHapley Additive exPlanations) values to enhance clinical trust and transparency; and (v) formal usability and clinical validation studies.

IX. CONCLUSION

This paper has presented a comprehensive machine learning-based multi-disease prediction system with an integrated personalized recommendation engine, deployed as an accessible Streamlit web application.

The proposed system addresses a critical need in modern healthcare — the availability of intelligent, scalable, and user-friendly tools that can assist in the early detection of multiple chronic diseases simultaneously.

Five disease prediction modules — Diabetes, Heart Disease, Parkinson's Disease, Breast Cancer, and chronic kidney disease — were implemented using a comparative evaluation of Support Vector Machine, Random Forest, Logistic Regression, K-Nearest Neighbors, and Decision Tree classifiers. Experimental results demonstrate strong predictive performance, with classification accuracies ranging from 77.6% to 99.0% across disease modules and algorithms. SVM and Random Forest consistently emerged as the top-performing algorithms, achieving the best balance between precision, recall, and AUC-ROC.

The integrated recommendation engine augments raw prediction outputs with actionable, personalized health guidance across dietary, physical activity, medical consultation, and lifestyle modification dimensions, positioning the system as a holistic clinical decision support tool rather than a mere classifier.

The Streamlit deployment ensures broad accessibility without specialized infrastructure requirements, making the system suitable for deployment in resource-constrained healthcare environments. The modular architecture facilitates seamless extension to additional diseases, input modalities, and recommendation strategies.

In conclusion, the proposed system represents a meaningful contribution toward the democratization of AI-powered healthcare diagnostics, with the potential to meaningfully improve early disease detection rates, patient health outcomes, and healthcare resource utilization globally.

REFERENCES

[1] M. Topol, “High-performance medicine: The convergence of human and artificial intelligence,” *Nature Medicine*, vol. 25, no. 1, pp. 44–56, Jan. 2019.

[2] Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, “Machine learning and data mining methods in diabetes research,” *Computational and Structural Biotechnology Journal*, vol. 15, pp. 104–116, 2017.

[3] S. Mohan, C. Thirumalai, and G. Srivastava, “Effective heart disease prediction using hybrid machine learning techniques,” *IEEE Access*, vol. 7, pp. 81542–81554, 2019.

[4] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. Costello, and I. M. Moroz, “Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection,” *BioMedical Engineering OnLine*, vol. 6, no. 1, p. 23, 2007.

[5] O. L. Mangasarian, W. N. Street, and W. H. Wolberg, “Breast cancer diagnosis and prognosis via linear programming,” *Operations Research*, vol. 43, no. 4, pp. 570–577, 1995.

[6] Salekin and J. Stankovic, “Detection of chronic kidney disease and selecting important predictive attributes,” in *Proc. IEEE Int. Conf. Healthcare Informatics (ICHI)*, 2016, pp. 262–270.

[7] T. Reddy, M. P. K. Reddy, K. Lakshmanna, D. S. Rajput, R. Kaluri, and G. Srivastava, “Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis,” *Evolutionary Intelligence*, vol. 13, no. 2, pp. 185–196, 2020.

[8] Singh and R. Kumar, “heart disease prediction using machine learning algorithms,” in *Proc. Int. Conf. Electrical and Electronics Engineering (ICE3)*, 2020, pp. 452–457.

[9] Zhang, N. J. Yuan, D. Lian, X. Xie, and W. Y. Ma, “Collaborative knowledge base embedding for recommender systems,” in *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 353–362.