A Machine Learning Framework for Multi-Disease Prediction from Clinical Data

Om Tapkire[1], Mayur Bari[2], Mohit Narkhede[3], Pooja Chaudhari[4]
[1,2,3]*UG Students, Department of Artificial Intelligence and Data Science, SSGBCOET Bhusawal, Maharastra, India*
[4]*Assistant Professor, Department of Computer Science Engineering and Data Science, SSGBCOET Bhusawal, Maharastra, India*

***Abstract*—Early and accurate diagnosis of chronic diseases is a critical challenge in modern healthcare. This paper presents a Machine Learning Framework for Multi-Disease Prediction from Clinical Data a unified, web-based system capable of simultaneously predicting Diabetes, Heart Disease, and Parkinson's Disease using supervised machine learning models. The proposed system leverages pre-trained classification algorithms including Support Vector Machines, Random Forest, and Logistic Regression trained on established medical datasets. Deployed through an interactive Streamlit interface, the system enables healthcare professionals to input patient-specific clinical parameters and receive immediate binary predictions. The framework follows a modular architecture with distinct components for model loading, input validation, feature extraction, and result visualization. Experimental evaluation demonstrates clinically relevant accuracy across all three disease modules: approximately 78–82% for diabetes, 85–88% for heart disease, and 90–93% for Parkinson's detection. The system prioritizes local data processing to ensure patient privacy and is designed for extensibility, supporting future integration of additional disease modules. This work demonstrates the feasibility of integrated multi-disease prediction as a practical clinical decision support tool.**

***Index Terms*—Machine Learning, Disease Prediction, Diabetes, Heart Disease, Parkinson's Disease, Streamlit, Clinical Decision Support, Random Forest, Support Vector Machine, Healthcare Analytics**

I. INTRODUCTION

Chronic diseases such as diabetes, cardiovascular disease, and Parkinson's disease represent a growing global health burden. The World Health Organization estimates that non-communicable diseases account for over 70% of global deaths annually. Early and accurate diagnosis of these conditions is pivotal to improving patient outcomes, reducing healthcare expenditure, and enhancing quality of life.

Traditional diagnostic pathways are reactive typically triggered only after clinical symptoms manifest. This reactive paradigm creates missed opportunities for early intervention that could dramatically alter disease trajectories. Moreover, diagnostic tools tend to be disease-specific, requiring separate clinical workflows for each condition, which increases fragmentation and inefficiency in healthcare delivery.

Machine learning offers a transformative approach by identifying subtle patterns in clinical data that may not be immediately apparent to human practitioners. By training predictive models on curated medical datasets, it becomes possible to assess disease risk from routine clinical parameters before overt symptoms appear.

This paper presents a unified, web-based Multi-Disease Prediction System that simultaneously addresses three major chronic conditions Diabetes, Heart Disease, and Parkinson's Disease through a single cohesive interface. The system leverages pre-trained supervised classification models and is deployed via Streamlit, a Python-based rapid web application framework, enabling accessible and intuitive risk assessment for healthcare professionals and patients alike.

The key contributions of this work are: (i) an integrated multi-disease prediction platform eliminating fragmentation found in single-disease tools; (ii) a modular and extensible software architecture enabling future addition of disease modules; (iii) a locally-processed, privacy-preserving

deployment that requires no external data transmission; and (iv) a user-friendly web interface requiring no data science expertise to operate.

II. RELATED WORK

Machine learning has emerged as a powerful tool for clinical prediction tasks. Kourou et al. [1] provided a comprehensive review of ML applications in cancer prognosis, demonstrating the effectiveness of algorithms such as Support Vector Machines (SVM) and Artificial Neural Networks. Fatima and Pasha [2] surveyed healthcare prediction systems and highlighted the dominance of ensemble and probabilistic methods.

A. Diabetes Prediction

Zou et al. [3] demonstrated that models trained on the Pima Indians Diabetes Dataset achieve accuracy rates of 75–85% using features such as glucose levels, BMI, and family history. Sisodia and Sisodia [4] compared multiple classification algorithms for diabetes prediction and found ensemble methods particularly effective, with Random Forest outperforming Naive Bayes and Decision Trees on standard benchmark datasets.

B. Heart Disease Prediction

Reddy et al. [5] showed that models incorporating clinical parameters such as blood pressure, cholesterol, and electrocardiographic readings could predict heart disease with accuracy exceeding 85%. Latha and Jeeva [6] emphasized the importance of feature selection in improving cardiovascular risk classification, demonstrating that ensemble techniques improve both sensitivity and specificity.

C. Parkinson's Disease Detection

Little et al. [7] pioneered the use of acoustic voice measurements for Parkinson's detection, showing that features such as fundamental frequency variations (Jitter, Shimmer) and nonlinear dynamical features (RPDE, DFA) could effectively differentiate Parkinson's patients from healthy controls. Subsequent studies have confirmed that SVM classifiers achieve over 90% accuracy on such voice datasets.

D. Multi-Disease and Integrated Systems

Shah et al. [8] demonstrated that unified multi-disease prediction platforms improve clinical workflow efficiency and user adoption compared to isolated single-disease tools. Tiwari et al. [9] and Gopisetti and Kummera [10] have specifically implemented Streamlit-based multi-disease prediction systems, validating the practical viability of this approach. The present work extends these efforts with a more rigorous modular architecture, explicit feature validation, and a comprehensive system design that supports extensibility.

III. SYSTEM ARCHITECTURE

The proposed system follows a layered, modular architecture comprising four primary components: the Presentation Layer (Streamlit UI), the Application Logic Layer, the Model Management Layer, and the Data/Model Storage Layer. Figure 1 provides a conceptual overview of this architecture.

A. Presentation Layer

The Streamlit framework serves as the frontend. The interface presents a sidebar navigation menu with hospital-themed icons enabling users to switch between three disease prediction modules. Each module renders a structured form with labeled input fields corresponding to the clinical parameters required by the respective model. Results are displayed inline as color-coded success or warning banners immediately upon prediction.

B. Application Logic Layer

This layer handles input collection, data validation, feature vector construction, and coordination between the UI and the model management layer. Upon form submission, inputs are validated for completeness and type consistency (numeric floating-point values). Validated inputs are assembled into ordered feature vectors matching each model's expected input schema before being forwarded to the prediction engine.

C. Model Management Layer

The model management layer is responsible for loading pre-serialized machine learning models from disk using Python's pickle module. A dedicated ML Model Loader component dynamically selects and loads the appropriate model based on the user's navigation choice. This decoupling ensures that model updates can be performed independently of interface changes.

D. Data / Model Storage Layer

Pre-trained models are stored as serialized .sav files within a local Saved_Models directory. This local persistence strategy ensures that the system can operate entirely offline after initial setup, with no external API dependency. Three model files are maintained: diabetes_model.sav, heart_disease_model.sav, and parkinsons_model.sav.

IV. METHODOLOGY

A. Dataset Description

Three publicly available, well-established medical datasets are used for training the prediction models:

- Diabetes: The Pima Indians Diabetes Dataset (768 samples, 8 features) includes pregnancies, plasma glucose concentration, diastolic blood pressure, triceps skinfold thickness, serum insulin, BMI, diabetes pedigree function, and age.
- Heart Disease: The Cleveland Heart Disease Dataset (303 samples, 13 features) includes age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting ECG, maximum heart rate, exercise-induced angina, ST depression, ST slope, number of major vessels, and thalassemia status.
- Parkinson's Disease: The UCI Parkinson's Voice Dataset (195 samples, 22 acoustic features) includes fundamental frequency measures (Fo, Fhi, Flo), jitter variants (%), shimmer variants (dB), noise-to-harmonics ratio (NHR), harmonics-to-noise ratio (HNR), and nonlinear dynamical complexity measures (RPDE, DFA, spread1, spread2, D2, PPE).

B. Preprocessing

Preprocessing steps applied during model training include: (i) missing value imputation using median substitution; (ii) feature scaling via standardization (zero mean, unit variance) for algorithms sensitive to magnitude; (iii) class imbalance handling through stratified sampling; and (iv) train-test splitting with an 80:20 ratio and k-fold cross-validation (k=5) to produce generalization-robust estimates.

At inference time, the system performs lightweight preprocessing: type conversion of string inputs to float, ordered feature vector assembly, and input validation with user feedback for out-of-range or missing values.

C. Classification Algorithms

Algorithm selection was guided by literature benchmarks for each disease domain:

- Diabetes: Logistic Regression and Support Vector Machine with RBF kernel, selected for their interpretability and strong performance on tabular medical data.
- Heart Disease: Random Forest classifier (100 estimators) selected for its robustness to outliers and built-in feature importance ranking.
- Parkinson's: Support Vector Machine with RBF kernel, consistent with the seminal work of Little et al. [7] and subsequent confirmatory studies.

D. Model Serialization and Deployment

Trained models are serialized using Python's pickle library and stored as .sav binary files. At application startup, all three models are loaded into memory simultaneously to eliminate per-prediction loading latency. The system uses the os module to construct platform-independent file paths, ensuring cross-OS compatibility (Windows, macOS, Linux).

V. IMPLEMENTATION

A. Technology Stack

The system is implemented entirely in Python 3.8+. Key libraries include: Streamlit (v1.x) for the web interface; scikit-learn for model training and inference; pickle for model persistence; NumPy for numerical operations; Pandas for data preprocessing; and streamlit-option-menu for sidebar navigation.

Hardware requirements are minimal: an Intel Core i3 processor (i5 recommended), 4 GB RAM (8 GB recommended), and 200 MB disk space. The application runs locally after initial dependency installation, with no continuous internet connection required.

B. Feature Input Modules

Each disease module presents clinically organized input fields: the Diabetes module collects 8 parameters in a 3-column layout; the heart disease module collects 13 parameters across a 3-column form; and the Parkinson's module collects 22 acoustic features in a 5-column layout to maximize screen real estate efficiency.

C. Prediction Pipeline

Upon clicking the prediction button, the system: (1) retrieves all field values; (2) converts inputs to float, catching TypeError or ValueError for invalid entries; (3) assembles inputs into a Python list in model-expected order; (4) calls model.predict([input_list]) to obtain a binary output (0 or 1); (5) maps the output to a human-readable diagnosis string; and (6) renders the result using st.success() for positive outcomes and a corresponding banner for negative outcomes.

VI. RESULTS AND EVALUATION

Table I presents the classification accuracy of each disease prediction module evaluated on held-out test sets using stratified 5-fold cross-validation.

Table I Prediction Accuracy by Disease Module

Disease Module	Algorithm	Accuracy (%)	Dataset	Features
Diabetes	SVM / Logistic Reg.	78–82%	Pima Indians Diabetes	8
Heart Disease	Random Forest	85–88%	Cleveland Heart Disease	13
Parkinson's	SVM	90–93%	UCI Parkinson's Voice	22

The Parkinson's prediction module achieves the highest accuracy (90–93%), consistent with literature findings that voice-based acoustic features provide high discriminative power. The Heart Disease module (85–88%) benefits from the rich clinical feature set of the Cleveland dataset and the ensemble nature of Random Forest. The Diabetes module (78–82%) reflects the inherent class imbalance and feature overlap in the Pima dataset.

Table II provides a comparative analysis of the proposed system against traditional single-disease tools and conventional diagnostic approaches.

Table II Comparative System Analysis

Feature	Proposed System	Traditional Tools	Single-Disease ML
Multi-disease support	Yes (3 diseases)	No	No
Web-based UI	Yes (Streamlit)	Varies	Limited
Local data processing	Yes	No	Varies
No specialist required	Yes	No	Partial
Extensible architecture	Yes	No	Partial
Real-time prediction	Yes	No	Yes

The proposed system demonstrates clear advantages in multi-disease coverage, accessibility, and privacy preservation compared to existing tools.

VII. CONCLUSION

This paper has presented a unified Machine Learning Framework for Multi-Disease Prediction from Clinical Data. The system integrates three supervised classification models for Diabetes, Heart Disease, and Parkinson's Disease prediction within a single, accessible, web-based interface. The modular architecture supports extensibility, local data processing ensures patient privacy, and the Streamlit-based UI eliminates the need for specialized technical expertise.

Experimental results confirm clinically relevant accuracy across all three modules. The system successfully addresses key limitations of existing approaches, including disease fragmentation, accessibility barriers, and usability challenges in clinical environments.

Future work will focus on: (i) expanding disease coverage to include chronic kidney disease, liver disease, and respiratory conditions; (ii) cloud deployment for multi-user clinical environments; (iii) integration of explainable AI techniques (SHAP/LIME) to provide feature-level interpretability; (iv) persistent patient data storage for longitudinal health tracking; and (v) development of a mobile application for broader accessibility.

REFERENCES

[1] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8–17, 2015.

[2] M. Fatima and M. Pasha, "Survey of machine learning algorithms for disease diagnostic," *Journal of Intelligent Learning Systems and Applications*, vol. 9, no. 1, pp. 1–16, 2017.

[3] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting diabetes mellitus with machine learning techniques," *Frontiers in Genetics*, vol. 9, Art. no. 515, 2018.

[4] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia Computer Science*, vol. 132, pp. 1578–1585, 2018.

[5] G. 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, pp. 185–196, 2020.

[6] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics in Medicine Unlocked*, vol. 16, Art. no. 100203, 2019.

[7] M. A. Little, P. E. McSharry, E. J. Hunter, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 4, pp. 1015–1022, 2009.

[8] D. Shah, S. Isah, and F. Zulkernine, "Predicting complications in pregnancy with machine learning algorithms," in *Proc. IEEE HEALTHCOM*, 2020.

[9] S. Tiwari *et al*., "Multiple disease prediction using machine learning," *International Journal of Innovative Science and Research Technology*, vol. 8, 2023.

[10] S. Gopisetti and S. Kummera, "Multiple disease prediction system using machine learning and Streamlit," in *Proc. 9th Int. Conf. Inventive Systems and Control (ICISC)*, 2025.

[11] G. Battineni, N. Chintalapudi, and F. Amenta, "Machine learning in medicine: Performance calculation of dementia prediction by support vector machines," *Informatics in Medicine Unlocked*, vol. 16, Art. no. 100200, 2019.

[12] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, "Development of disease prediction model based on ensemble learning approach for diabetes and heart disease," *IEEE Access*, vol. 7, pp. 144777–144789, 2019.