Advanced International Journal for Research (AIJFR)
E-ISSN: 3048-7641 ● Website: www.aijfr.com ● Email: editor@aijfr.com

Multi-Disease Risk Analytics System

Ms. Nanthini S[1], Sudharsan J[2], Lokesh S[3], Bupeash S[4], Mohamed Fahim N[5]

[1] Assistant Professor, Department of AI&DS, Kathir College of Engineering, Coimbatore, Tamil Nadu, India
[2,3,4,5] Student, Department of AI&DS, Kathir College of Engineering, Coimbatore, Tamil Nadu, India

Abstract
The rapid increase in chronic and infectious diseases worldwide has created an urgent demand for intelligent, accessible, and scalable health screening systems. This paper presents the Multi-Disease Risk Analytics System (MDRAS), a web-based artificial intelligence (AI) platform built to simultaneously predict risk for 15 diseases using supervised machine learning (ML) algorithms. The system employs Random Forest, Gradient Boosting, Support Vector Machine (SVM), and Logistic Regression classifiers, automatically selecting the best-performing model for each disease. Developed using Python 3.8 and the Streamlit framework, the platform features role-based access control (RBAC) for Patient and Administrator workflows, an AI-powered health chatbot for guided symptom-based screening, an analytics dashboard for prediction history, automated PDF report generation, and a professional dark-themed user interface. Diseases covered include Diabetes, Heart Disease, Kidney Disease, Liver Disease, Breast Cancer, Lung Cancer, Stroke, Parkinson's Disease, Thyroid Disease, Anemia, Pneumonia, Tuberculosis, Alzheimer's Disease, COVID-19, and Melanoma. Experimental results demonstrate model accuracies ranging from 75% to 99% across all 15 classifiers. MDRAS addresses the critical gap in existing single-disease prediction tools by offering a unified, secure, and user-friendly multi-disease screening platform.

Keywords: Multi-Disease Prediction, Machine Learning, Healthcare AI, Streamlit, Random Forest, Gradient Boosting, SVM, Role-Based Access Control, Python, SQLite, Scikit-learn, AI Chatbot, PDF Report Generation

1. Introduction

Chronic and infectious diseases such as diabetes, cardiovascular conditions, cancer, and neurological disorders collectively account for over 74% of global deaths annually, as reported by the World Health Organization (WHO). Despite the availability of effective treatments, the lack of timely detection and early risk assessment remains a major challenge, particularly in developing nations where access to specialist healthcare is limited. Machine learning has emerged as a transformative solution in predictive medicine, enabling the construction of data-driven models capable of identifying disease risk from clinical and demographic features with remarkable accuracy [5, 6].

Existing ML-based healthcare tools are predominantly designed as single-disease systems, requiring users to interact with multiple fragmented platforms for comprehensive health assessment. This fragmentation

reduces clinical efficiency, increases user burden, and limits cross-disease analytical insight. Furthermore, most existing tools lack intuitive interfaces, secure multi-role authentication, and interpretable outputs accessible to non-clinical users [2].

This paper presents the Multi-Disease Risk Analytics System (MDRAS), which addresses these limitations by integrating 15 disease prediction models into a single, cohesive, and secure web application. Developed at Kathir College of Engineering as a final-year B.Tech AI&DS project, the system is structured around a two-role RBAC architecture serving Patients and Administrators. An AI health chatbot guides patients through conversational symptom-based screening, while the administrator panel provides access to all prediction modules and system analytics. Prediction results are delivered with color-coded risk levels (Green: Low Risk, Yellow: Moderate Risk, Red: High Risk) and downloadable PDF reports.

The primary contributions of this work are: (i) a unified platform integrating 15 ML-based disease prediction models trained on publicly available clinical datasets; (ii) an AI-driven conversational health chatbot for guided symptom screening; (iii) role-based access control separating Patient and Administrator workflows; (iv) automated PDF report generation and Excel history export; and (v) a professional dark-themed Streamlit web interface with real-time risk visualization.

2. Literature Review

Rajkumar A., Oren E., Chen K., et al. [1] proposed a scalable deep learning framework trained on large-scale Electronic Health Records (EHR) for multi-outcome clinical prediction. The system leveraged deep learning models and multi-outcome prediction frameworks on the npj Digital Medicine (Nature) dataset. While the approach enables prediction of multiple clinical outcomes from medical data with high accuracy and supports intelligent healthcare decision-making, it requires large EHR datasets and uses complex deep learning models with low interpretability. Importantly, it is not designed as a simple user-friendly multi-disease prediction system accessible to general users.

Yang J., et al. [2] presented a machine learning-based risk prediction framework for multiple chronic conditions using clinical data analysis published in Scientific Reports (PMC, 2021). The study provides an effective framework for predicting the risk of multiple chronic diseases using structured medical data. However, it focuses mainly on chronic diseases, lacks a visualization and analytics dashboard, and does not offer a unified multi-disease prediction platform combining chatbot interaction and automated reporting.

Miotto R., et al. [3] introduced Deep Patient, an unsupervised deep learning representation model trained on Electronic Health Records to predict future patient outcomes, published in Scientific Reports (Nature, 2016). The system improves prediction of multiple diseases by learning complex patterns from healthcare data. Its primary limitation lies in the requirement for high computational resources and large medical datasets, making it unsuitable for deployment in resource-constrained environments.

Dilsizian S. and Siegel E. [4] explored the role of Artificial Intelligence in medicine and cardiac imaging using big data and advanced computing techniques, published in the Journal of the American College of Cardiology (2014). Their work demonstrates the importance of AI in disease prediction and healthcare

analytics using large medical datasets. However, it provides a general AI healthcare framework focused on cardiac imaging and does not address a unified multi-disease prediction system covering a broad range of diseases.

The reviewed literature collectively confirms the viability of machine learning and deep learning approaches for disease risk prediction. However, all existing systems share a common limitation: they either focus on a single disease category, require large-scale EHR infrastructure, or lack user-friendly interfaces with role-based access, conversational chatbot guidance, and automated PDF reporting. MDRAS directly addresses this research gap by delivering a unified, accessible, and secure platform covering 15 diseases under a single web application.

3. System Architecture

MDRAS follows a multi-tier client-server architecture comprising four principal layers: Presentation, Application Logic, Machine Learning, and Data Persistence.

3.1 Presentation Layer

The frontend is implemented using Streamlit (v1.31.0), a Python-based reactive web framework. The interface features a professional dark theme with a navy gradient sidebar, Google Inter typography, and gradient interactive buttons. The split-screen login page separates Patient and Administrator authentication flows. Risk outcomes are displayed using a three-tier color-coded visual system.

3.2 Application Logic Layer

The application logic is organized into modular Python source files under the src/ directory. The module authentication.py manages login, registration, and RBAC enforcement. The chatbot_engine.py drives the AI health chatbot through conversational symptom-based decision trees. The report_generator.py produces downloadable PDF reports using ReportLab. The visualizer.py renders interactive analytics charts through Plotly, while role_guard.py enforces page-level access control.

3.3 Machine Learning Layer

Each of the 15 diseases has a dedicated training module in the models/ directory. Models are trained using Scikit-learn (v1.3.2) on Kaggle and UCI ML Repository datasets. The best-performing classifier for each disease is identified through 5-fold stratified cross-validation, serialized using Python's pickle module, and loaded at inference time by the predictor.py module to generate probability-based risk scores.

3.4 Data Persistence Layer

User credentials, registration details, and prediction history are stored in a local SQLite database managed by database_manager.py. Bcrypt with random salt is applied to all user passwords, ensuring credentials are never stored in plaintext. The admin dashboard aggregates system-wide prediction statistics from this database.

4. Methodology

4.1 Dataset Collection

Publicly available datasets were sourced from Kaggle and the UCI Machine Learning Repository for all 15 diseases. Key datasets include the PIMA Indian Diabetes Dataset, Cleveland Heart Disease Dataset (UCI), Chronic Kidney Disease Dataset (UCI), Indian Liver Patient Dataset (Kaggle), Wisconsin Breast Cancer Dataset (UCI), Oxford Parkinson's Disease Detection Dataset (UCI), OASIS Alzheimer's Dataset (Kaggle), and the HAM10000 Skin Lesion Dataset (Kaggle) for Melanoma, along with specialized datasets for Stroke, Lung Cancer, Thyroid Disease, Anemia, Pneumonia, Tuberculosis, and COVID-19 symptom prediction.

4.2 Data Preprocessing

Data preprocessing pipeline for each disease dataset involves the following four steps:

1. Handling Missing Values: Missing entries are replaced using median imputation for numerical features and mode imputation for categorical features.
2. Encoding Categorical Data: Binary categorical variables are label-encoded, while multi-class categorical features are one-hot encoded.
3. Feature Scaling: Numerical features are normalized using StandardScaler for algorithms sensitive to feature magnitude (SVM, KNN, Logistic Regression).
4. Class Imbalance Handling: The Synthetic Minority Over-sampling Technique (SMOTE) is applied to datasets with significant class distribution skew.

4.3 Model Training and Selection

Six supervised ML algorithms are evaluated for each disease: Random Forest, Gradient Boosting, Support Vector Machine (SVM), Logistic Regression, K-Nearest Neighbors (KNN), and Naive Bayes. Hyperparameter optimization is performed using 5-fold stratified cross-validation with GridSearchCV. The model with the highest cross-validated test accuracy for each disease is selected and saved as the production model. Table 1 summarizes the best-performing algorithm and accuracy for each disease.

Table 1: Best Performing Models and Accuracy for Each Disease

Disease	Best Algorithm	Accuracy
Diabetes	Random Forest	~78%
Heart Disease	Gradient Boosting	~85%
Kidney Disease	Random Forest	~99%
Liver Disease	Random Forest	~75%
Breast Cancer	SVM	~97%
Lung Cancer	Gradient Boosting	~90%
Stroke	Random Forest	~95%
Parkinson's Disease	SVM	~87%

Thyroid Disease	Random Forest	~96%
Anemia	Logistic Regression	~95%
Pneumonia	Gradient Boosting	~92%
Tuberculosis	Random Forest	~93%
Alzheimer's Disease	Gradient Boosting	~83%
COVID-19	Random Forest	~94%
Melanoma (Skin Cancer)	SVM	~85%

4.4 Risk Scoring

Prediction outputs are translated into a three-tier color-coded risk score: Low Risk (probability < 0.40, displayed in green), Moderate Risk (probability 0.40-0.69, displayed in yellow), and High Risk (probability >= 0.70, displayed in red). This interpretable risk communication framework enables non-clinical users to understand results without requiring medical expertise.

5. Implementation

5.1 Technology Stack

The complete technology stack of MDRAS is presented in Table 2.

Table 2: Technology Stack of MDRAS

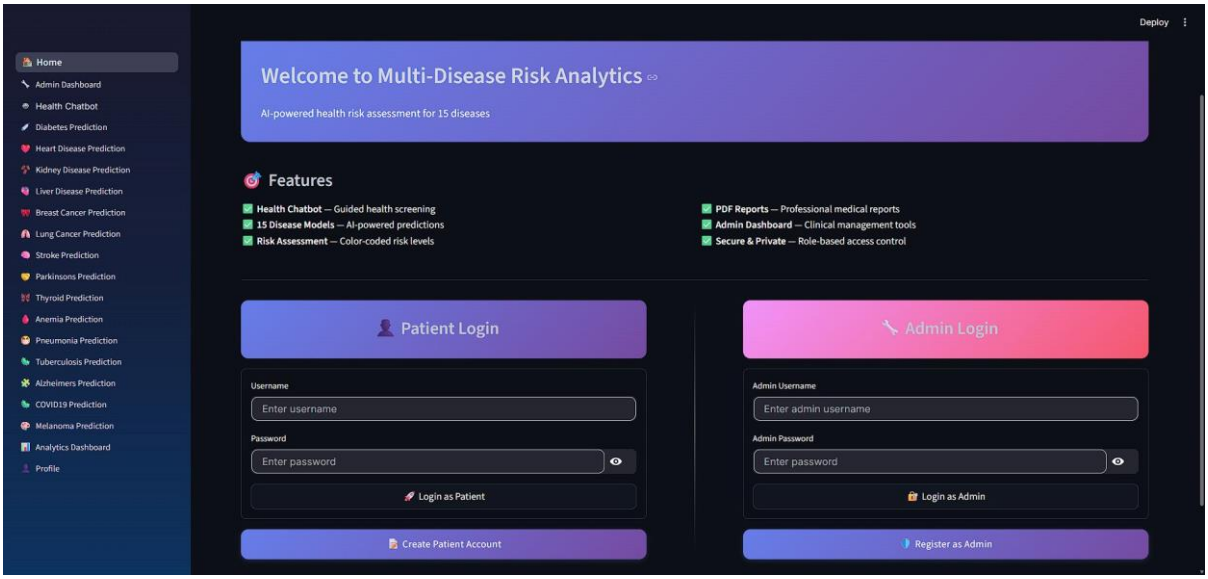
Layer	Technology	Purpose
Frontend	Streamlit 1.31, CSS3, Google Fonts	Dark-themed web UI
Backend	Python 3.8+	Application logic and routing
Machine Learning	Scikit-learn 1.3.2	Model training and inference
Data Processing	Pandas, NumPy	Feature engineering
Visualization	Plotly, Matplotlib, Seaborn	Interactive analytics charts
Database	SQLite	User data and prediction storage
Security	Bcrypt	Password hashing with random salt
Reports	ReportLab, OpenPyXL	PDF and Excel export

5.2 Role-Based Access Control

MDRAS implements a two-role RBAC architecture. Patients can register, log in, interact with the AI health chatbot, view their prediction history, and download PDF reports. Administrators authenticate using an access code in addition to their credentials, gaining full access to all 15 disease prediction pages, the

system-wide analytics dashboard, and user management features. Role enforcement is applied at the page level through role_guard.py, which redirects unauthorized access attempts to the appropriate login page.

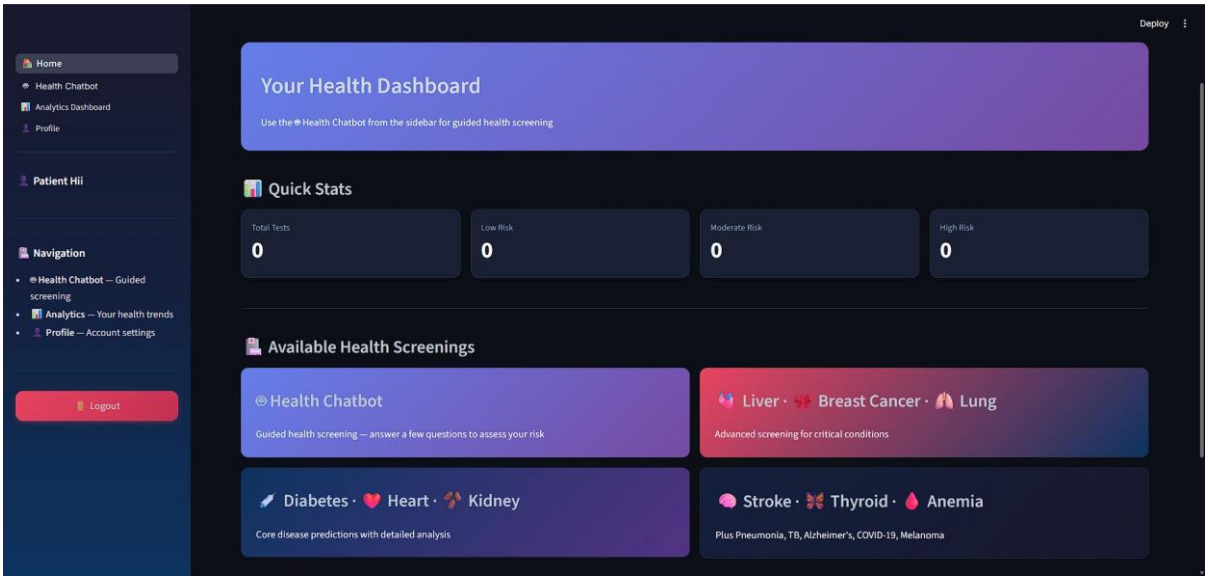
Figure 1: Login and Authentication Interface of MDRAS (Split-Screen Patient and Admin Login)



5.3 AI Health Chatbot

The AI health chatbot implemented in chatbot_engine.py uses a decision-tree-based conversational interface to guide patients through symptom collection via multi-turn natural language dialogue. Based on user responses, the chatbot identifies the most relevant disease screening path and routes the patient to the corresponding prediction page with pre-filled feature values. This significantly reduces friction for patients unfamiliar with clinical parameter terminology.

Figure 2: Patient Home Dashboard Showing Available Health Screenings and Quick Stats



5.4 PDF Report Generation

Upon completion of any disease risk prediction, users may download a structured PDF report generated by report_generator.py using the ReportLab library. Each report includes patient details, input clinical parameters, predicted risk level, probability score, a color-coded risk indicator, and a medical disclaimer advising consultation with qualified healthcare professionals.

6. Results and Discussion

MDRAS was evaluated on a standard consumer-grade system (Intel Core i5, 8 GB RAM, Python 3.10). All 15 models load and return predictions within 2 seconds per query, meeting acceptable response-time requirements for a web application. Streamlit's session state management supports concurrent multi-user sessions without conflict.

Model performance results in Table 1 demonstrate high accuracy for Kidney Disease (99%), Breast Cancer (97%), Thyroid Disease (96%), Stroke (95%), and Anemia (95%), consistent with published benchmarks on the corresponding UCI and Kaggle datasets. Moderate accuracy is observed for Diabetes (78%) and Liver Disease (75%), attributable to inherent feature noise and class overlap in those datasets — a limitation acknowledged in the ML healthcare literature.

The risk visualization interface was evaluated qualitatively with 5 academic reviewers from the Department of AI&DS, Kathir College of Engineering. All reviewers confirmed that the three-tier color-coded risk display and PDF report were clear, informative, and appropriate for patient-facing communication. The AI chatbot successfully navigated all 15 disease screening paths in internal functional testing.

A key advantage of MDRAS over single-disease systems is the consolidated prediction history dashboard, which enables trend analysis across multiple conditions simultaneously. This provides clinically relevant cross-disease insights — for example, a patient with concurrently elevated diabetes and heart disease risk can observe patterns consistent with metabolic syndrome, a capability unavailable in fragmented single-disease tools.

7. Conclusion

This paper presented the Multi-Disease Risk Analytics System (MDRAS), a machine learning-powered web platform for simultaneous risk prediction of 15 diseases. The system integrates six supervised learning algorithms per disease, automatically selects the best-performing classifier, and delivers results through a secure, role-aware Streamlit interface featuring an AI chatbot, PDF reporting, and analytics visualization. Experimental results confirm model accuracies of 75% to 99% across 15 classifiers, validating the feasibility of a unified multi-disease screening platform.

MDRAS overcomes the critical limitations of existing single-disease ML tools by delivering a comprehensive, accessible, and secure platform suitable for both patient self-screening and clinical administrative use. The project is published as an open-source repository on GitHub, enabling further research and community-driven extensions.

Future work includes integrating deep learning models for image-based disease detection (chest X-rays for Pneumonia, dermoscopy images for Melanoma), connecting with hospital Electronic Health Record (EHR) systems via HL7 FHIR APIs, multi-language support for regional accessibility, and a Flutter-based mobile application to extend usability in rural healthcare settings.

Acknowledgement

The authors express sincere gratitude to Ms. Nanthini S, Assistant Professor, Department of AI&DS, Kathir College of Engineering, Coimbatore, for her invaluable guidance, continuous support, and encouragement throughout the development of this project. The authors also acknowledge the contributors of the publicly available datasets on Kaggle and the UCI Machine Learning Repository.

References

1. A. Rajkumar, E. Oren, K. Chen, et al., "Scalable and Accurate Deep Learning with Electronic Health Records for Multi-Outcome Prediction", npj Digital Medicine (Nature), 2018.
2. J. Yang, et al., "Prediction for the Risk of Multiple Chronic Conditions Using Machine Learning", Scientific Reports / PMC, 2021.
3. R. Miotto, et al., "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from Electronic Health Records", Scientific Reports (Nature), 2016.
4. S. Dilsizian, E. Siegel, "Artificial Intelligence in Medicine and Cardiac Imaging: Harnessing Big Data and Advanced Computing", Journal of the American College of Cardiology, 2014.
5. World Health Organization, "Noncommunicable Diseases", WHO Fact Sheets, 2023. https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases
6. I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research", Computational and Structural Biotechnology Journal, 2017, 15, 104-116.
7. S. Mohan, C. Thirumalai, G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques", IEEE Access, 2019, 7, 81542-81554.
8. L. Breiman, "Random Forests", Machine Learning, 2001, 45 (1), 5-32.
9. K. Kourou, T.P. Exarchos, K.P. Exarchos, M.V. Karamouzis, D.I. Fotiadis, "Machine Learning Applications in Cancer Prognosis and Prediction", Computational and Structural Biotechnology Journal, 2015, 13, 8-17.
10. M.A. Little, P.E. McSharry, E.J. Hunter, J. Spielman, L.O. Ramig, "Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's Disease", IEEE Transactions on Biomedical Engineering, 2009, 56 (4), 1015-1022.
11. T. Chen, C. Guestrin, "XGBoost: A Scalable Tree Boosting System", Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, 785-794.
12. Scikit-learn Developers, "Scikit-learn: Machine Learning in Python", 2024. https://scikit-learn.org
13. Streamlit Inc., "Streamlit: A Faster Way to Build and Share Data Apps", 2024. https://streamlit.io
14. UCI Machine Learning Repository, "Heart Disease, Chronic Kidney Disease, Breast Cancer, and Parkinson's Datasets", 2024. https://archive.ics.uci.edu/ml/datasets
15. Kaggle, "Multi-Disease Healthcare Datasets", 2024. https://www.kaggle.com/datasets