

A Data-Driven Preventive Healthcare System for Multi-Disease Risk Prediction Using Machine Learning

Madhu Tripathi¹,

Prof. Dr. Shivani A. Budhkar²

^{1,2}MCA Department, P E S Modern College of Engineering Pune, India

ABSTRACT

Identifying chronic diseases at an early stage helps doctors take preventive actions before serious problems occur. This study introduces a data-driven prediction system designed to assess the likelihood of diabetes along with cardiovascular conditions using patient clinical data such as blood sugar concentrations, cholesterol measurements, BP, BMI, age, and history of hypertension. Different classification techniques were experimentally compared to determine the most dependable prediction model. Evaluation results indicated that the RF prediction model delivered the most consistent and precise predictive performance among all tested models. The proposed solution was deployed via an interactive browser-accessible healthcare platform where hospital staff can input patient details and receive instant risk scores categorized as low, moderate, or high. Doctors can then review these results, monitor patient records, and generate downloadable reports, making early diagnosis more efficient and organized.

Keywords— Machine Learning, Preventive Healthcare, Disease Prediction, Diabetes Prediction, Heart Prediction, Risk Assessment, Web-Based System

1.INTRODUCTION

Modern healthcare is increasingly focusing on prevention and early diagnosis rather than addressing medical conditions once they become severe. Conditions such as diabetes and heart disease are easier to manage when detected early, before they progress to life-threatening stages. The growth of digital health data and computing tools has supported the creation of develop systems that can automatically screen patients and identify health risks earlier than traditional clinical methods allow. ML-driven analytical systems can interpret substantial volumes of clinical information and identify hidden health-related patterns. These methods assist doctors in making quicker, more informed decisions, especially in busy hospital settings where time is critical. However, traditional patient assessment procedures in many hospitals continue to

depend heavily on manual clinical review, which can slow down the diagnosis process and delay treatment when it is most needed. To address this, this research develops a machine learning-powered preventive healthcare platform that can assess the likelihood of multiple diseases through an accessible web interface. The system allows healthcare professionals to input patient details, generate real-time predictive results, and support physicians in evaluating individual disease risk profiles. By integrating predictive analytics with a user-friendly digital platform, the framework aims to improve the quality and timeliness of early diagnosis and preventive health management. The proposed application is structured as an interactive web tool through which hospital staff can submit patient information and immediately obtain risk predictions grouped into defined categories. Medical professionals can also review past predictions, track patient progress over time, and provide appropriate clinical recommendations. Additional features include automated report generation and basic healthcare analytics capabilities. Together, the system demonstrates how intelligent computational methods can significantly impact frontline clinical workflows.

2.LITERATURE SURVEY

Abdalrada, et al. (2022).

Abdalrada and colleagues developed a machine learning-based framework designed to jointly predict the occurrence of diabetes and cardiovascular conditions using retrospective patient records [1]. Their approach involved classification methods such as LR and RF-based prediction models to identify high-risk individuals. The implemented classifiers demonstrated reliable disease identification capability. The research showed that. However, the work was mainly focused on model benchmarking and did not include the development of an operational web-based platform for real-time clinical screening.

Agrawal, et al. (2023).

Agrawal and co-authors introduced a diabetes forecasting framework based on machine learning on clinical and demographic patient data to differentiate

between diabetic and non-diabetic individuals [2]. Their findings suggested that computational prediction models can assist in recognizing diabetic risk at an earlier stage. While the designed system demonstrated promising classification accuracy, its focus was limited to diabetes alone, leaving the broader challenge of multi-disease risk assessment and healthcare system integration unexplored.

Hennebelle, et al. (2023).

Hennebelle and colleagues introduced HealthEdge, an intelligent healthcare architecture combining IoT devices with edge and cloud computing for predicting Type 2 Diabetes [7]. This distributed approach reduced latency and supported real-time inference across multiple layers. However, the framework required advanced infrastructure and specialized hardware, limiting its applicability in resource-constrained or smaller healthcare facilities. The study was also focused solely on diabetes, without addressing multi-condition prediction.

Cai, et al. (2024).

Cai and co-authors conducted a comprehensive review of AI-based techniques applied for heart disease risk assessment [6]. The review highlighted that intelligent computational approaches frequently outperform conventional statistical assessment methods. While the review confirmed the effectiveness of AI tools for heart-related conditions, it was limited to cardiovascular disease and did not explore multi-disease or combined prediction systems.

Ahmad, et al. (2024).

Ahmad and colleagues examined predictive analytics methods used in modern healthcare environments, emphasizing the importance of data-informed decision-making in improving diagnostic accuracy and patient care [3]. The paper provided conceptual insight into the growing role of predictive analytics within healthcare environments. However, it remained theoretical and did not give a practical, deployable solution for multi-disease risk prediction within a clinical setting.

Oikonomou and Khera (2023). Oikonomou and Khera explored the role of machine learning in precision-oriented diabetes care and cardiovascular risk evaluation [10]. Their work highlighted how analytical models can strengthen individualized risk profiling and support data-driven clinical decisions. While the study provided valuable theoretical insights into the clinical relevance of predictive analytics, it did

not develop or implement a practical system suitable for routine patient screening processes.

Puranikath, et al. (2024).

Puranikath and colleagues proposed a machine learning framework able to perform multi-disease prediction [11]. Their study highlighted the feasibility of combining multiple prediction models within a unified healthcare framework. However, the framework lacked real-world operational capabilities commonly found in real-world clinical settings, such as risk-level categorization, patient record management, or interactive data visualization tools.

Islam, et al. (2024).

Islam and co-authors conducted a thorough review of machine learning methods used for chronic disease prediction, with a focus on diabetes and heart disease [8]. The study examined the impact of feature engineering and algorithm selection on predictive accuracy. Although the work covered multiple disease areas, it did not fully address the practical implementation of multi-disease risk prediction in real-world clinical settings.

Al-Nafjan, et al. (2025).

Al-Nafjan and colleagues reviewed emerging artificial intelligence developments in the predictive healthcare domain, demonstrating that AI-based techniques can substantially enhance early diagnosis, continuous patient monitoring, and disease risk profiling [5]. While the review offered forward-looking perspectives on the trajectory of predictive health technologies, it did not address the design or deployment of a ready-to-use system for multi-disease screening in practical healthcare settings.

Mim, et al. (2025).

Mim and co-authors carried out a comparative performance evaluation of multiple predictive algorithms implemented to cardiovascular disease prediction [9]. Their findings indicated that ensemble-based and tree-structured models consistently outperformed classical statistical methods in heart disease classification tasks. The study, however, concentrated solely on algorithmic benchmarking and did not advance toward the development of a comprehensive, deployable healthcare prediction application.

Alhumaidi, et al. (2025).

Alhumaidi and colleagues conducted a broad evaluation of intelligent computational methods in real-world healthcare scenarios, centered on medical

risk forecasting and clinical management [4]. Their analysis highlighted the capacity of data-oriented prediction models to extract actionable insights from large, complex patient datasets and strengthen clinical decision support systems. Despite its thorough coverage, the study did not propose an economical, web-accessible platform capable of simultaneously predicting multiple diseases and presenting interactive risk visualizations.

Research gap:

Existing studies still leave several practical issues unresolved. Despite growing interest in utilizing AI-based models for medical prediction, the following key gaps remain:

1. The majority of existing studies address only a single disease domain—typically diabetes or cardiovascular disorders—rather than providing a unified framework capable of handling multiple disease predictions concurrently.
2. A considerable proportion of published research prioritizes theoretical model evaluation without translating findings into functional, clinically deployable systems.
3. Current predictive models predominantly generate binary classification outcomes (e.g., disease present or absent), lacking nuanced risk stratification or probability-based interpretation of results.
4. Essential operational features such as patient record management, visual analytics dashboards, and report generation capabilities are largely absent from existing healthcare prediction systems.
5. Certain advanced healthcare frameworks are overly dependent on IoT infrastructure, cloud platforms, or edge computing environments, rendering them impractical for implementation in resource-limited clinical settings.

The system proposed in this research seeks to address these shortcomings by delivering a cost-effective, web-deployable healthcare platform that integrates multi-disease risk prediction with patient management capabilities and interactive risk visualization, making it a practical and accessible tool for diverse clinical environments.

3.METHODOLOGY

3.1 Proposed system

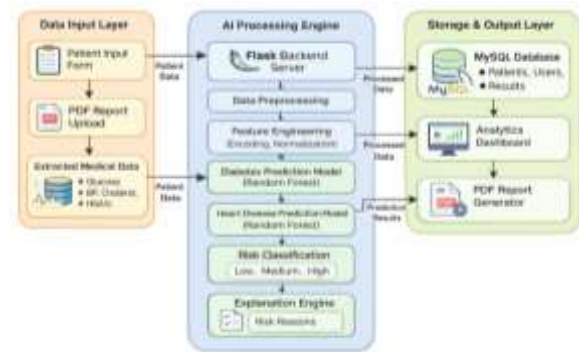


Fig.3.1.1 Architecture of the proposed system

The healthcare prediction framework developed in this study is built to quantify the likelihood of chronic disease onset by applying machine learning techniques to structured patient clinical data. Publicly accessible medical datasets were sourced and utilized to train the underlying predictive models.

Before training the models, the dataset was cleaned by handling missing values, removing incorrect entries, and transforming non-numeric features into machine-readable numbers. This step was important to ensure the algorithms received consistent, usable input. After preprocessing, the dataset was divided into separate subsets for model learning and performance validation.

A set of supervised classification algorithms—comprising Logistic Regression, Decision Tree, K-Nearest Neighbours, and Random Forest—were independently trained and benchmarked against one another through multiple evaluation measures. Upon completing the experimental comparison, the Random Forest algorithm emerged as the most consistently reliable model and was subsequently adopted as the core prediction engine for the system.

Prediction outputs generated by the model are rendered through the web application interface for clinical review. Physicians are able to examine these outputs, document their clinical observations, and prescribe preventive interventions or additional diagnostic procedures as warranted. The system also retains a historical record of patient profiles and prior predictions within a dedicated database, enabling longitudinal monitoring. Downloadable reports containing prediction summaries and supporting visual analytics are available to facilitate documentation and clinical communication. This integrated workflow is

designed to accelerate disease screening and strengthen preventive healthcare decision support.

The finalized predictive models were embedded within a web application constructed using the Flask framework. Through this interface, hospital staff can submit patient data and receive instant disease risk classifications. Physicians retain the ability to scrutinize prediction outcomes, access historical patient data, and formulate evidence-based medical guidance.

3.2 System Workflow

The system workflow is initiated when authorized hospital staff access the web application and enter a patient's relevant medical data through the input interface. The information supplied typically includes demographic and clinical parameters such as age, blood glucose concentration, cholesterol level, BP readings, body mass index, tobacco usage status, and other diagnostic indicators pertinent to disease analysis.

Upon data submission, automated preprocessing routines are triggered to cleanse and standardize the input. This stage addresses incomplete or missing entries, filters out inconsistent values, and maps categorical variables into machine-readable numerical equivalents to ensure compatibility with the trained models.

The conditioned data is then routed to the deployed machine learning models, which evaluate the patient's parameter set and compute a probability estimate for conditions such as diabetes and heart disease. Informed by these probability scores, the system assigns each patient to one of three defined risk tiers: low risk, moderate risk, or high risk.

Prediction results are presented through the application's dashboard, where physicians can review individual assessments, annotate findings with clinical observations, and issue follow-up recommendations. The system preserves all patient interactions and prediction histories within its database, supporting continuity of care and long-term monitoring.

3.3 Model Training and Implementation

The predictive models at the core of this system were built using open-access healthcare datasets obtained from the Kaggle platform. The diabetes dataset encompasses a broad range of patient health indicators, including age, BMI, HbA1c, hypertension status, and cardiac health status. Together, these attributes provide

a comprehensive clinical profile for algorithmic training.

Before model development commenced, a thorough data preparation pipeline was applied to enhance the integrity and predictive utility of the datasets. Data gaps were resolved using appropriate imputation strategies, and categorical fields—such as gender, smoking history, and chest pain classification—were converted to numeric representations via standard encoding techniques. The cleaned datasets were subsequently separated into learning and validation datasets using an 80:20 ratio to support unbiased performance evaluation.

Multiple supervised learning algorithms were implemented and experimentally evaluated to determine the optimal predictive model. The algorithms included in the comparison were Logistic Regression, K-Nearest Neighbours, Decision Tree, Support Vector Machine, and Random Forest. Each algorithm was assessed using a consistent set of performance metrics—accuracy, precision, recall, and F1-score—to enable fair and objective comparison.

Among all tested algorithms, Random Forest produced the most reliable and accurate results for both diseases. Its structure—which combines multiple decision trees—helped it handle the complexity of healthcare data better than single-model approaches. This made it the clear choice for deployment in the final system.

The trained models were serialized using the Joblib library and integrated into the Flask-based web application to enable real-time inference. To support result interpretation, statistical plots were generated with Matplotlib library, and patient-specific PDF reports—containing prediction summaries, confidence scores, and visual analytics—were produced using the ReportLab library.

The complete implementation yields a fully functional healthcare prediction platform capable of storing patient records, generating disease risk predictions on demand, and delivering actionable analytical outputs to support clinical professionals in preventive healthcare management.

4. TOOLS AND TECHNOLOGY USED

The proposed healthcare prediction platform was constructed by integrating machine learning methodologies with contemporary web development technologies. The frontend user interface was engineered using HTML, CSS, and JavaScript,

delivering an intuitive, responsive experience for both healthcare staff and medical professionals.

The server-side logic of the application was built on Python in conjunction with the Flask micro-framework, which handles request routing, user authentication, session management, and communication with the embedded prediction models. Machine learning model development and training were done by Scikit-learn library, which provided implementations of all evaluated classification algorithms and supported the preprocessing and evaluation pipeline.

Once trained, the finalized models were persisted to disk using the Joblib serialization library, enabling efficient loading within the live web application for real-time risk inference. A MySQL relational database was integrated into the backend to manage persistent storage of patient information, user credentials, prediction records, and generated reports. The database architecture also supports role-based access controls, whereby system administrators oversee platform operations, clinical staff submit patient data, and physicians review and act upon prediction outputs.

The Matplotlib library was employed to generate graphical representations of prediction results and statistical trends, while the ReportLab toolkit was used to produce downloadable PDF documents comprising patient summaries, prediction confidence values, and embedded visualizations. Collectively, these technologies form a cohesive, production-ready healthcare intelligence system.

5. Result and Discussion

The developed web application was able to generate disease risk predictions successfully in real time. It brings together trained prediction models and an easy-to-use interface, allowing hospital staff and doctors to work with patient data quickly and with greater confidence.

Two publicly available datasets were employed throughout the experimental phase. The diabetes dataset comprised approximately 100,000 patient records encompassing attributes such as age, BMI, blood glucose level, HbA1c value, hypertension history, and smoking behavior. The heart disease dataset included 920 records characterized by clinical variables including cholesterol concentration, chest pain classification, blood pressure, and related indicators.

Prior to model training, preprocessing operations—including missing value handling, feature normalization, and categorical attribute encoding—were applied to both datasets to improve data quality and ensure consistent model performance. The preprocessed data was then used to train and evaluate five classification algorithms: Logistic Regression, K-Nearest Neighbours, Decision Tree, Support Vector Machine, and Random Forest.

For diabetes prediction, the Random Forest model delivered the highest accuracy, approaching 97%, consistently outperforming all other algorithms under evaluation. Logistic Regression also produced competitive results, while K-Nearest Neighbours, Decision Tree, and SVM exhibited comparatively weaker predictive capability on this dataset.

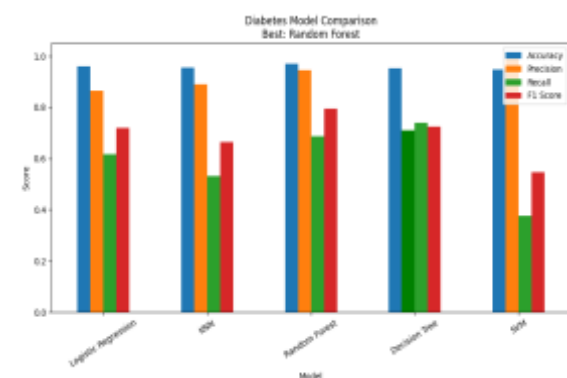


Fig.5.1 Best Diabetes Model

In the heart disease prediction task, Random Forest again achieved the strongest overall performance, recording an accuracy exceeding 85%. Logistic Regression performed competitively, while SVM attained moderate accuracy relative to the ensemble-based approach. Across both prediction tasks, the Random Forest algorithm demonstrated the greatest stability and reliability, attributed to its ensemble structure that aggregates multiple decision trees to reduce variance and improve generalization.

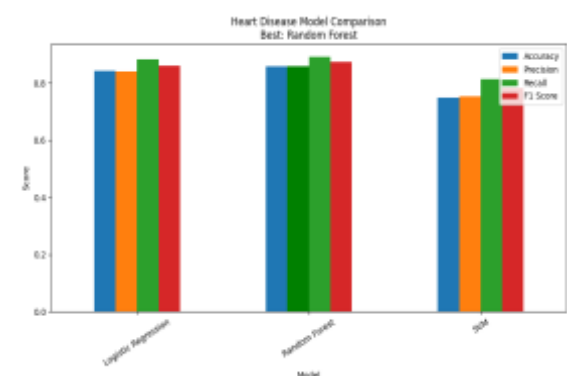


Fig.5.2 Best Heart Model

Based on these findings, Random Forest was confirmed as the final model for integration into the web application. The deployed system allows hospital staff to input patient clinical details and receive immediate risk assessments. Physicians can subsequently examine prediction outputs, consult historical patient records, and issue tailored medical recommendations through the system dashboard.

The experimental findings validate that machine learning-based healthcare platforms hold substantial promise for early disease identification, proactive health planning, and informed clinical decision support in real-world settings.

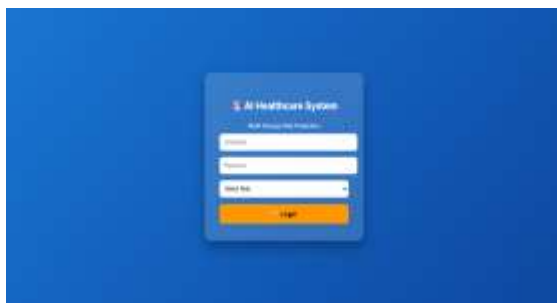


Fig.5.3 Login page



Fig.5.4 Home page



Fig.5.5 System Analytics

ID	First Name	Last Name	Age	Gender	Weight	Height	Blood Pressure	Heart Rate	Diabetes	Cholesterol	Smoking	Overall Risk
1001	John	Doe	45	Male	180.0	175.0	120/80	75	Yes	200	Yes	High
1002	Jane	Doe	35	Female	160.0	160.0	110/70	65	No	150	No	Medium
1003	John	Doe	55	Male	190.0	185.0	130/90	85	Yes	250	Yes	Very High
1004	Jane	Doe	25	Female	150.0	150.0	100/60	55	No	120	No	Low
1005	John	Doe	65	Male	200.0	195.0	140/100	95	Yes	300	Yes	Critical
1006	Jane	Doe	40	Female	170.0	170.0	115/75	70	No	180	No	Medium-High
1007	John	Doe	30	Male	165.0	165.0	105/65	60	No	130	No	Low-Medium
1008	Jane	Doe	50	Female	175.0	175.0	125/85	80	Yes	220	Yes	High

Fig.5.6 Patient Records

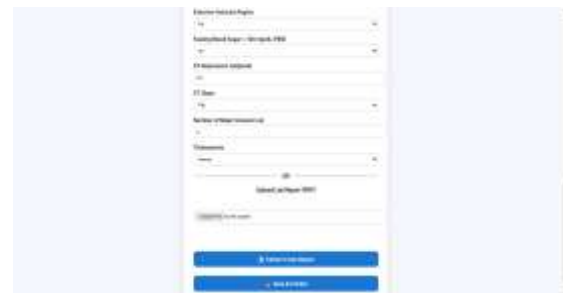


Fig.5.7 Add Patient

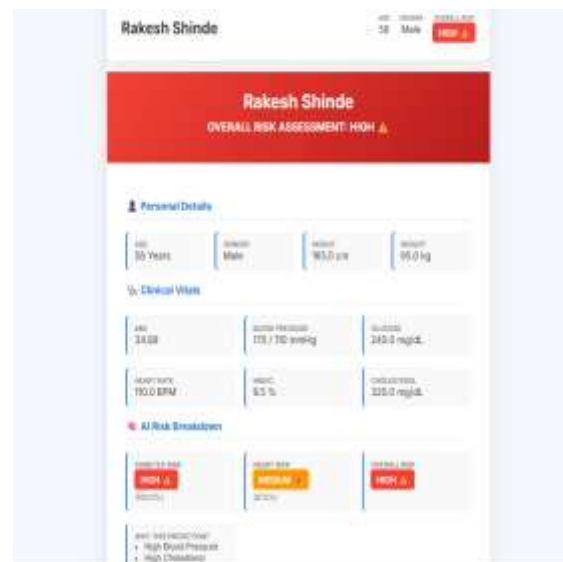


Fig.5.8 Patient Profile

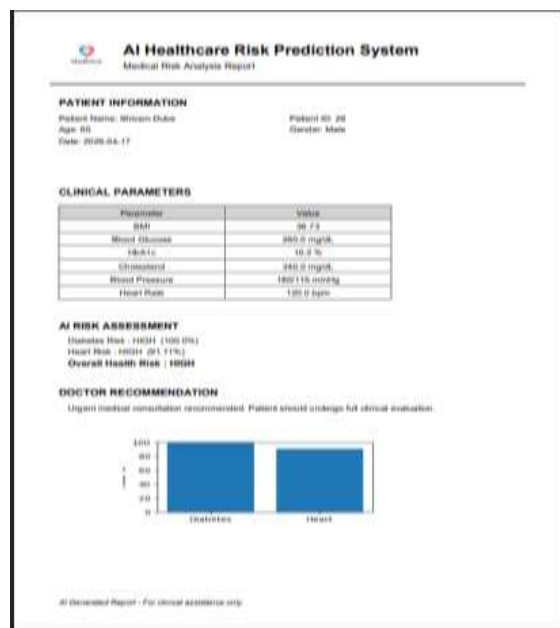


Fig.5.9 Patient Report

6. Conclusion

The presented work developed an intelligent healthcare prediction platform designed for early multi-disease risk assessment. A systematic comparison of multiple classification-based algorithms was conducted to identify the best performing predictive approach for each disease domain.

The experimental evaluation confirmed that the tree-based ensemble model consistently achieved superior predictive accuracy and delivered dependable results across both disease datasets. The trained models were successfully embedded within a web-based healthcare application, enabling real-time disease risk generation in a clinically accessible format.

The system helps doctors and hospital staff perform faster disease screening and make better decisions about early treatment and prevention. Additional features like patient record storage, report generation, and visual charts make the tool more useful in day-to-day hospital operations.

Prospective extensions of this work may encompass the incorporation of prediction modules for additional disease categories, expansion of the training datasets to enhance model generalizability, experimentation with deep learning architectures for improved accuracy, and the integration of advanced monitoring dashboards and visualization tools to better serve real-world clinical application requirements.

7. Bibliography

1. S. Abdalrada, J. Abawajy, T. Al-Quraishi, and S. M. S. Islam, "Machine learning models for prediction of co-occurrence of diabetes and cardiovascular diseases: A retrospective cohort study," 2022.
2. A. J. Agrawal, L. H. Patil, and S. S. Aote, "Diabetes Prediction Using Medical Data and Disease Influence Measures using Machine Learning," 2023.
3. A. Ahmad, H. Gupta, and S. Singh, "Predictive Analytics in Healthcare," SRMCEM, Lucknow, India.
4. N. H. Alhumaidi, D. Dermawan, H. F. Kamaruzaman, and N. Alotaiq, "The Use of Machine Learning for Analysing Real-World Data in Disease Prediction and Management: Systematic Review."
5. A. Al-Naffjan, A. Aljuhani, A. Alshebel, A. Alharbi, and A. Alshehri, "Artificial Intelligence in Predictive Healthcare: A Systematic Review."
6. Y. Cai *et al.*, "Artificial intelligence in the risk prediction models of cardiovascular disease and development of an independent validation screening tool: A systematic review."
7. A. Hennebelle, H. Materwala, and L. Ismail, "HealthEdge: A Machine Learning-Based Smart Healthcare Framework for Prediction of Type 2 Diabetes in an Integrated IoT, Edge, and Cloud Computing System."
8. R. Islam, A. Sultana, and M. R. Islam, "A comprehensive review for chronic disease prediction using machine learning algorithms."
9. T. J. H. Mim, M. M. Hassan, B. K. Paul, and M. S. H. Mollah, "Machine Learning Approaches for Cardiovascular Disease Prediction: A Comparative Study," 2025.
10. E. K. Oikonomou and R. Khera, "Machine learning in precision diabetes care and cardiovascular risk prediction."
11. J. Puranikath, F. Farheen, A. Patil, and A. K. M., "Analysis and Prediction of Multiple Disease using Machine Learning," Ballari Institute of Technology and Management.