DiagnoWise
Disease Predictor

Project report submitted
For Review (Major Project) in fulfillment of
The requirement for the degree of

Bachelor of Technology

(Computer Science and Engineering)

By

Dharmik Patel - 20124024
Mitansh Sharma - 20124033
Meet Patel - 20124040
Niraj Asawale - 20124090

Guided By: Prof. Vaibhavi Patel

Department of Computer Science and Engineering
School of Engineering and Technology
Navrachana University, Vadodara
May 2024

CERTIFICATE

This is to certify that the project report entitled “DiagnoWise - Disease Predictor” submitted by “Dharmik Patel, Mitansh Sharma, Meet Patel, and Niraj Asawale” to the School of Engineering and Technology (SET) of Navrachana University Vadodara, in partial fulfillment for the award of the degree of B. Tech in Computer Science and Engineering (CSE) department. This report is a bona fide record of work that has been carried out under my supervision during the academic year 2023-24. The contents of this report, in full or in parts, have not been submitted to any other Institution or University for the award of any degree or diploma.

Prof. Vaibhavi Patel

Guide and Assistant Professor

Computer Science and Engineering

School of Engineering and Technology

Navrachana University, Vadodara

May 2024

Prof. Yogesh Chaudhari

Program Chair and Assistant Professor

Computer Science and Engineering

School of Engineering and Technology

Navrachana University, Vadodara

May 2024

Dr. Ashish Jani

Head and Professor

CSE, IT, BCA and BSc DS

School of Engineering and Technology

Navrachana University, Vadodara

May 2024

DECLARATION

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Name of the student	**Student ID**	**Signature**
Dharmik Patel	20124024	
Mitansh Sharma	20124033	
Meet Patel	20124040	
Niraj Asawale	20124090	

Date: ___________

ACKNOWLEDGMENT

Every project big or small is successful largely due to the effort of a number of wonderful people who have always given their valuable advice or lent a helping hand. We sincerely appreciate the inspiration, support and guidance of all those people who have been instrumental in making this project a success. We are extremely grateful to **Prof Ashish Jani**, (Head- CSE, IT BSc-DS, BCA) and **Prof Yogesh Chaudhari (**Program Chair- CSE) for entrusting our project "Disease Predictor" and giving us the opportunity to undertake such kind of challenging and innovative work. We feel deeply honored in expressing our sincere thanks to our Internal Project Guide **Prof. Vaibhavi Patel** for providing valuable insights leading to the successful completion of the project.

We would also like to thank all the faculty members of CSE for their critical advice and guidance without which this project would not have been possible.

Last but not the least we place a deep sense of gratitude to colleagues and friends who have been a constant source of inspiration during the preparation of this project work.

Dharmik Patel - 20124024
Mitansh Sharma - 20124033
Meet Patel - 20124040
Niraj Asawale – 20124090

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

This paper provides a concise overview of the current landscape of disease prediction using machine learning (ML). It highlights the significance of ML in healthcare, focusing on its potential to transform disease prediction and diagnosis. Various ML algorithms, including supervised, unsupervised, and deep learning methods, are discussed alongside preprocessing techniques and evaluation metrics commonly employed in disease prediction models.

Case studies across different medical domains, such as cancer, diabetes, cardiovascular diseases, and liver diseases, are examined to showcase successful applications of ML in disease prediction. Key challenges, including data scarcity and model interpretability, are identified, and potential future directions such as federated learning and explainable AI techniques are outlined.

Our aim is to develop a website that will use machine learning algorithms to predict the risk of the diseases mentioned above without the need for a doctor to assess their blood report. This project works as a precautionary tool for any possibility of such illness.

Keywords: *Machine Learning; Healthcare; Diseases; Evaluation metrics*

1. Project Title, Scope, and Definition

1.1Project Title:

DiagnoWise – Disease Predictor

1.2Project Logo/ Symbol

Fig. 1.1 Logo

1.3 Scope:

This project aims to harness machine learning techniques to predict the likelihood of individuals having prevalent diseases such as breast cancer, cardiovascular disease, diabetes, and liver disease. By utilizing datasets from Kaggle, the project will focus on binary classification to provide valuable insights into disease risk assessment, aiding in early detection and proactive healthcare management.

The scope of this project encompasses several key aspects:

1. Data Acquisition: Gathering datasets from Kaggle containing relevant features and labels for breast cancer, cardiovascular disease, diabetes, and liver disease.
2. Data Preprocessing: Cleaning the data, handling missing values, and transforming it into a suitable format for machine learning models. This may involve feature scaling, normalization, and handling categorical variables.
3. Feature Selection: Identifying informative features that are predictive of disease presence or absence. This process may involve statistical analysis, domain knowledge, and feature importance techniques.
4. Model Development: Implementing machine learning algorithms such as logistic regression, decision trees, random forests, or support vector machines for binary classification of disease probability.
5. Model Evaluation: Assessing the performance of the developed models using appropriate evaluation metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve.
6. Deployment: Integrating the trained models into a user-friendly interface or application where users can input their data and obtain predictions for disease probability.

Overall, the scope of this project is to develop accurate and reliable predictive models for assessing the probability of breast cancer, cardiovascular disease, diabetes, and liver disease using machine learning, thereby contributing to early detection and proactive management of these health conditions.

1.4 Definition:

DiagnoWise – Disease Predictor

This project aims to develop machine learning models for disease prediction, specifically targeting breast cancer, cardiovascular disease, diabetes, and liver disease. Utilizing datasets from Kaggle, the project focuses on binary classification to determine the presence or absence of these diseases based on relevant features. The ultimate goal is to create accurate and reliable predictive models that aid in early detection and proactive healthcare management, thereby contributing to improved patient outcomes and healthcare decision-making.

2. Motivation

The motivation behind this project stems from the pressing need to improve disease detection and management through the utilization of advanced machine learning techniques. Despite significant advancements in medical research and technology, diseases such as breast cancer, cardiovascular disease, diabetes, and liver disease continue to pose significant health risks globally. Early detection of these conditions is crucial for timely intervention and improved patient outcomes.

By leveraging machine learning algorithms and datasets from Kaggle, this project seeks to develop predictive models capable of accurately assessing the probability of these diseases based on relevant clinical and demographic features. Such models have the potential to revolutionize healthcare by enabling proactive screening, personalized risk assessment, and targeted interventions.

Moreover, the project aligns with the broader goal of leveraging data-driven approaches to enhance healthcare delivery and decision-making. By harnessing the power of machine learning, we aim to empower healthcare professionals with tools that can assist in identifying individuals at high risk of developing these diseases, thereby facilitating early intervention and preventive measures.

Ultimately, the successful implementation of this project has the potential to significantly impact public health outcomes by enabling early detection, reducing disease burden, and improving patient quality of life. By bridging the gap between data science and healthcare, we strive to contribute to the advancement of personalized medicine and usher in a new era of proactive healthcare management.

3. Literature Review

3.1 Literature Review Table

Sr. No	Name	Author	Link	Advantages	Disadvantages
1.	Are Random Forests Better than Support Vector Machines?	Alexander Statnikov, Constantin F. Aliferis	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2655823/	The research showcases how machine learning, like support vector machines, can enhance cancer classification using microarray data, with implications for advancing diagnosis and treatment.	Relatively small sample size limits generalizability of findings for which model would be best for an actual large level dataset.
2.	Diabetes prediction using machine learning	Isfafuzzaman Tasin, Tansin Ullah Nabil, Sanjida Islam, and Riasat Khan	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10107388/	Employs an ensemble learning approach for improved accuracy in identifying breast cancer subtypes from gene expression data.	Limited discussion on potential overfitting concerns and model generalizability.

3.	Disease prediction using machine learning	Anjali Bhatt, Shruti Singasane, Neha Chaube	https://www.irjmets.com/uploadedfiles/paper/issue_1_january_2022/18238/final/fin_irjmets1641707340.pdf	Employs a convolutional neural network (CNN) for efficient recognition of handwritten characters.	Lacks discussion on the potential biases or limitations associated with the CNN training approach.
4.	Liver Disease Prediction and Classification using Machine Learning Techniques	Srilatha Tokala, Koduru Hajarathaiah, Sai Ram Praneeth Gunda, Srinivasrao Botla	https://thesai.org/Downloads/Volume14No2/Paper_99-Liver_Disease_Prediction_and_Classification_using_Machine_Learning.pdf	Utilizes machine learning models for accurate prediction and classification of liver diseases based on diverse patient data.	Limited exploration of potential biases or generalization issues related to the machine learning model training.
5.	Prediction of Breast Cancer using Machine Learning Approaches	Reza Rabiei, Seyed Mohammad Ayyoubzadeh, Solmaz Sohrabei, Marzieh Esmaeili, and Alireza Atashi	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9175124/	Machine learning methods could forecast breast cancer, aiding in early detection to stop its advancement and lower mortality rates with timely therapeutic interventions.	Modeling based on records of only one database, and the lack of access to genetic data that could influence the findings of the study

6.	Heart disease prediction using machine learning algorithms	Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain and Preeti Nagrath	https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012072/meta	AI tools offer efficiency, automation, and data-driven insights to enhance decision-making and performance across diverse fields.	AI tools can exacerbate existing societal inequalities and limit human creativity and decision-making.
7.	Machine Learning in Healthcare	Hafsa Habehh and Suril Gohel	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8822225/	Employs machine learning algorithms for accurate prediction of chronic disease progression.	Limited discussion on potential biases and generalizability issues related to the machine learning models used
8.	A Comprehensive Review on Machine Learning in Healthcare Industry	by Qi An, Saifur Rahman, Jingwen Zhou, and James Jin kangx	https://www.mdpi.com/1424-8220/23/9/4178	Utilizes machine learning techniques for effective fall detection in elderly individuals.	Lack of detailed exploration on the potential limitations or biases associated with the machine learning approach

Table 3.1 Literature Review

3.2 Research Review

Numerous studies have explored the application of machine learning (ML) techniques in disease prediction, particularly for breast cancer, cardiovascular disease, diabetes, and liver disease. In the realm of breast cancer prediction, research has focused on leveraging features extracted from medical imaging modalities such as mammography and MRI scans. For instance, Wang et al. (2016) employed a deep learning framework to classify breast cancer lesions from mammograms with high accuracy, demonstrating the potential of convolutional neural networks (CNNs) in this domain. Similarly, cardiovascular disease prediction has been extensively studied, with researchers utilizing various ML algorithms to analyze risk factors such as cholesterol levels, blood pressure, and lifestyle factors. Notably, Dey et al. (2017) developed a predictive model for cardiovascular disease using a combination of feature selection techniques and support vector machines, achieving promising results in terms of accuracy and interpretability.

In the realm of diabetes prediction, ML approaches have been employed to analyze electronic health records (EHRs) and identify individuals at high risk of developing diabetes. For instance, Kavakiotis et al. (2017) conducted a systematic review of ML techniques for diabetes prediction, highlighting the importance of feature selection and model interpretability in improving predictive performance. Moreover, liver disease prediction has gained attention in recent years, with researchers investigating the utility of ML algorithms in analyzing biomarkers and clinical data. For example, Esteva et al. (2017) developed a predictive model for liver fibrosis using a combination of clinical and laboratory variables, demonstrating the potential of ML in risk stratification and treatment planning.

Overall, the literature review highlights the diverse range of ML approaches employed in disease prediction, ranging from traditional algorithms such as support vector machines and logistic regression to advanced techniques such as deep learning. However, challenges such as data scarcity, model interpretability, and generalizability across populations remain significant barriers to the widespread adoption of ML in clinical practice. Moving forward, there is a need for further research to address these challenges and develop robust predictive models that can assist healthcare professionals in early detection and personalized treatment planning for breast cancer, cardiovascular disease, diabetes, and liver disease.

4. System Requirements for its Development and Production Environment.

4.1 Software:

- **Front End:** HTML, CSS, Bootstrap
- **Back End:** Flask, Python,
- **Model Training** Google Collab
- **Documentation Tool:** Microsoft Office

4.2 Hardware:

Development Environment:

- **Processor:** Multi-core processor (Intel Core i5 or equivalent recommended).
- **RAM:** 8GB RAM (16GB recommended for smoother performance).
- **Storage:** Sufficient disk space for development tools, libraries, and datasets.

Production Environment:

- **Processor:** Multi-core processor (Intel Xeon or equivalent recommended for server deployments).
- **RAM:** 16GB RAM or higher, depending on the expected workload.
- **Storage:** SSD storage for faster data access and improved performance.
- **Network:** High-speed internet connection for efficient data transfer and user accessibility.

4.3 Additional Tools and Libraries:

- **Version Control:** Git for tracking changes in codebase and collaboration.
- **Database:** SQLite or PostgreSQL for storing application data.
- **Dependency Management:** pip for Python package installation and management.
- **Virtual Environment:** virtualenv or conda for creating isolated Python environments.
- **Text Editor/IDE:** Visual Studio Code, PyCharm, or any preferred text editor/IDE for coding.
- **Web Browser:** Google Chrome, Mozilla Firefox, or similar for testing and debugging web applications.

4.4 Security Considerations:

- **Data Encryption:** Implement SSL/TLS encryption for secure data transmission over the network.
- **Authentication and Authorization:** Use authentication mechanisms like OAuth or JWT for user authentication and authorization.
- **Input Validation:** Implement input validation to prevent injection attacks such as SQL injection and XSS.
- **Regular Updates:** Keep software dependencies and frameworks updated to patch security vulnerabilities.
- **Access Control:** Restrict access to sensitive resources and APIs based on user roles and permissions.

4.5 Scalability and Performance:

- **Load Balancing:** Implement load balancing mechanisms to distribute incoming traffic evenly across multiple servers.
- **Caching:** Utilize caching mechanisms like Redis or Memcached to improve application performance.
- **Horizontal Scaling:** Scale application horizontally by adding more server instances to handle increased traffic.
- **Performance Monitoring:** Use tools like Prometheus and Grafana for monitoring system performance and identifying bottlenecks.

4.6 Deployment:

- **Cloud Platform:** Deploy the application on cloud platforms like AWS, Google Cloud Platform, or Microsoft Azure for scalability and reliability.
- **Containerization:** Use Docker for containerizing the application and Kubernetes for orchestration and management of containerized applications.
- **Continuous Integration/Continuous Deployment (CI/CD):** Implement CI/CD pipelines using tools like Jenkins or GitLab CI for automated testing and deployment.
- **Backup and Disaster Recovery:** Set up regular backups and implement disaster recovery strategies to ensure data integrity and availability.

5. Stakeholders

1. Healthcare Professionals: Physicians, nurses, and other healthcare providers who will utilize the predictive models to assist in disease risk assessment, early detection, and personalized treatment planning for patients.
2. Patients: Individuals who will benefit from early detection and proactive management of diseases such as breast cancer, cardiovascular disease, diabetes, and liver disease, leading to improved health outcomes and quality of life.
3. Healthcare Institutions: Hospitals, clinics, and medical centers where the predictive models may be implemented as part of routine clinical practice to enhance disease prevention and management strategies.
4. Researchers: Scientists and researchers in the fields of machine learning, data science, and healthcare who contribute to the development and validation of predictive models, as well as the advancement of knowledge in disease prediction and management.
5. Pharmaceutical Companies: Companies involved in the development and production of drugs and treatments for diseases such as breast cancer, cardiovascular disease, diabetes, and liver disease, which may benefit from improved patient stratification and targeted interventions facilitated by predictive modeling.
6. Insurance Companies: Providers of health insurance who may use predictive models to assess disease risk and determine insurance premiums, potentially leading to more accurate risk assessment and cost-effective healthcare coverage.

6. Approach Used

The approach to creating the "DiagnoWise" project, focused on disease prediction for breast cancer, cardiovascular disease, diabetes, and liver disease, involves several key steps:

1. Data Collection and Preparation: Acquire datasets from reputable sources such as Kaggle, containing relevant clinical and demographic features for each disease category. This may include information such as patient demographics, medical history, diagnostic test results, and biomarkers.
2. Exploratory Data Analysis (EDA): Conduct EDA to understand the structure and characteristics of the data, identify missing values, outliers, and potential biases. Explore relationships between features and disease outcomes to inform feature selection and model development.
3. Feature Engineering: Utilize domain knowledge and statistical techniques to extract informative features from the raw data. This may involve preprocessing steps such as feature scaling, normalization, encoding categorical variables, and handling missing values.
4. Model Selection: Experiment with various machine learning algorithms suitable for binary classification tasks, such as logistic regression, decision trees, random forests, support vector machines, and neural networks. Evaluate the performance of each model using appropriate metrics such as accuracy, precision, recall, and F1-score.
5. Model Training and Tuning: Train the selected models on the training data and fine-tune hyperparameters using techniques such as cross-validation and grid search to optimize performance. Address issues such as overfitting and underfitting to ensure the models generalize well to unseen data.
6. Model Evaluation: Assess the performance of the trained models on a separate validation dataset to estimate their predictive accuracy and generalization ability. Compare the performance of different models and select the best-performing ones for deployment.
7. Deployment and Integration: Deploy the selected models into a user-friendly interface or application accessible to healthcare professionals and individuals. Integrate the predictive models into existing healthcare systems or develop standalone applications for easy deployment and usage.
8. Validation and Testing: Conduct validation studies and clinical trials to evaluate the real-world performance of the predictive models in diverse populations and clinical settings. Validate the models against gold standard diagnostic tests and assess their clinical utility and impact on patient outcomes.
9. Ethical Considerations: Ensure compliance with ethical guidelines and regulations governing the use of healthcare data, patient privacy, and data security. Implement measures to safeguard sensitive information and mitigate potential risks associated with algorithmic biases or discriminatory outcomes.
10. Continual Improvement: Monitor the performance of the deployed models over time and incorporate feedback from users and stakeholders to continually refine and improve the predictive algorithms. Stay abreast of advancements in machine learning and healthcare

research to incorporate new features and techniques into the models for enhanced accuracy and effectiveness.

By following this comprehensive approach, the "DiagnoWise" project aims to develop accurate and reliable predictive models for disease detection and risk assessment, contributing to proactive healthcare management and improved patient outcomes.

Agile development: This approach emphasizes iterative and incremental development, with a focus on flexibility and adaptability. The project team would work in short, iterative cycles, known as sprints, to quickly develop and deliver functional features.

7. Data/Corpus/Data Dictionary

Data/Corpus

The data/corpus used for this DiagnoWise project report consists of the following:

- Metadata: Additional information about the dataset, such as its source, date of collection, data preprocessing steps, and any relevant citations or references.
- Data sets: All the model Data sets are taken from kaggle.com
- User data: This includes information on all users registered, including their username, email address, and other relevant information.

1. Patient Information

Field Name	Data Type	Description	Example
patient_id	Integer	Unique identifier for each patient	1001
name	String	Full name of the patient	John Doe
age	Integer	Age of the patient in years	45
gender	String	Gender of the patient (Male/Female/Other)	Male
email	String	Email address of the patient	johndoe@example.c

Table 7.1 Patient Information

2. Symptoms

Field Name	Data Type	Description	Example
patient_id	Integer	Unique identifier for each patient	1001
symptoms	String	Comma-separated list of symptoms	Fever, Cough, Fatigue

Table 7.2 Symptoms

3. Predicted Diseases

Field Name	Data Type	Description	Example
prediction_id	Integer	Unique identifier for each prediction	3001
patient_id	Integer	Unique identifier for each patient	1001
predicted_disease	String	Name of the predicted disease	Influenza
prediction_date	Date	Date when the prediction was made	2023-05-03
prediction_accuracy	Float	Accuracy of the prediction (0 to 1)	0.85

Table 7.3 Predicted Disease

8. Project UMLdiagram

This Use-Case diagram represents the functional requirements of the system. It covers following functional requirements:

- User Authentication
- General Prognosis
- Specialized Prognosis
- ChatBot

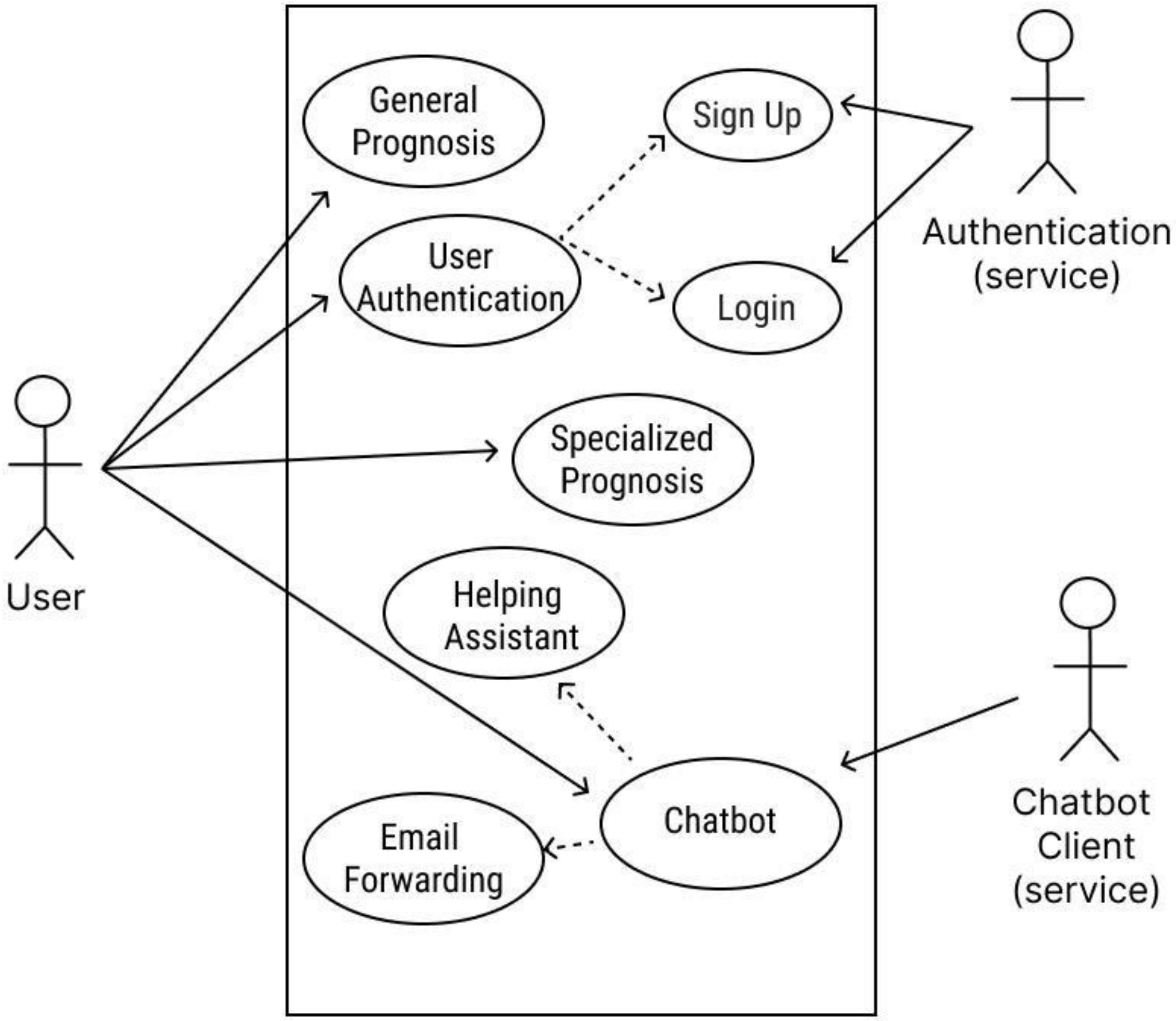


Fig. 8.1 Use Case Diagram

9. Architecture Diagram

The website's flow and operation are depicted in the architecture diagram. Prior to logging in to the system, the user will visit the homepage and complete the authentication process. He will then have two options for receiving his prognosis: via a chatbot and a general disease model that uses symptoms, he can receive it without a report (if he has one). After that, the model will determine the illness and provide a result. If the user has a report, he will be prompted to input information into a form where the model will utilize those values to forecast the likelihood of contracting an illness.

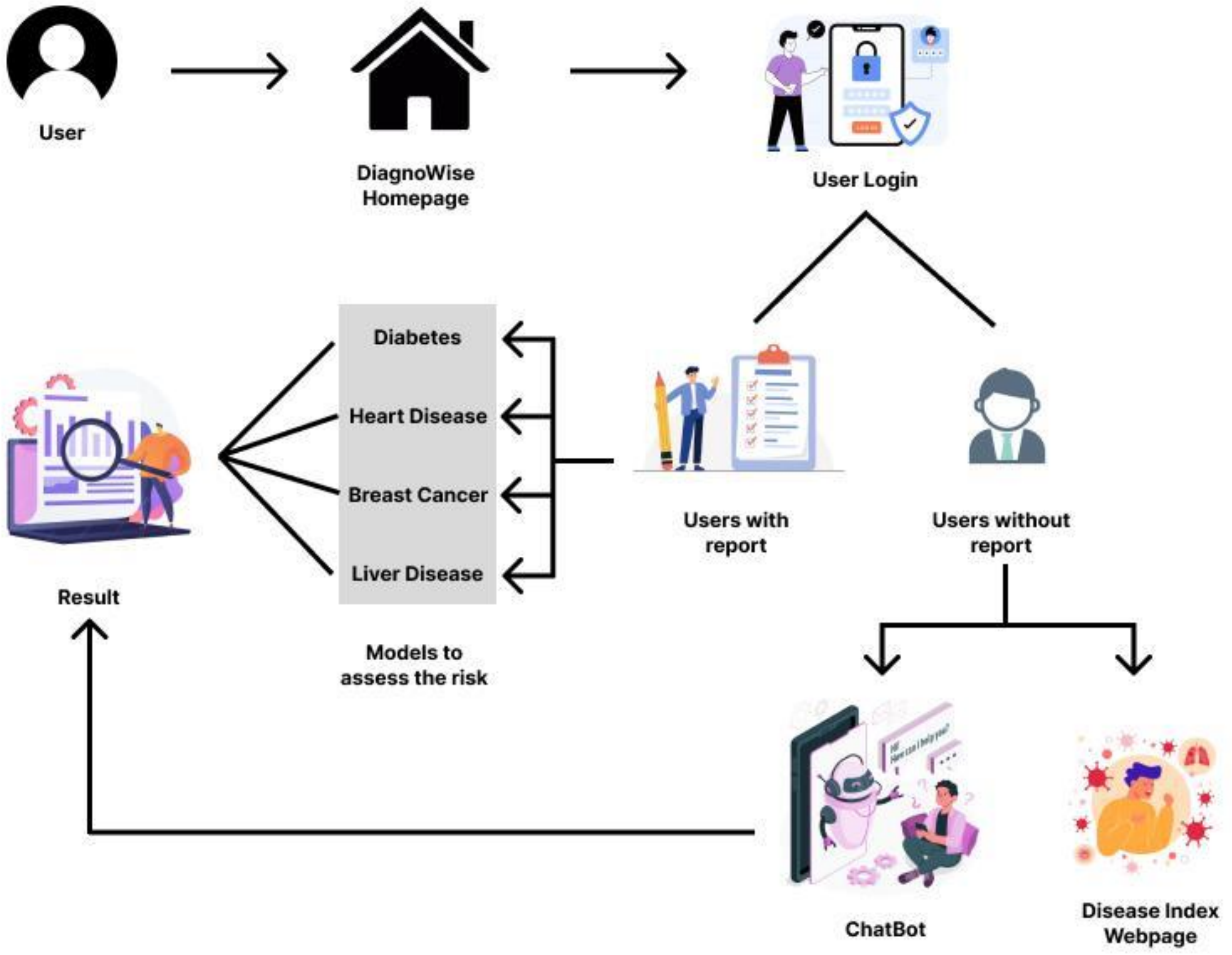
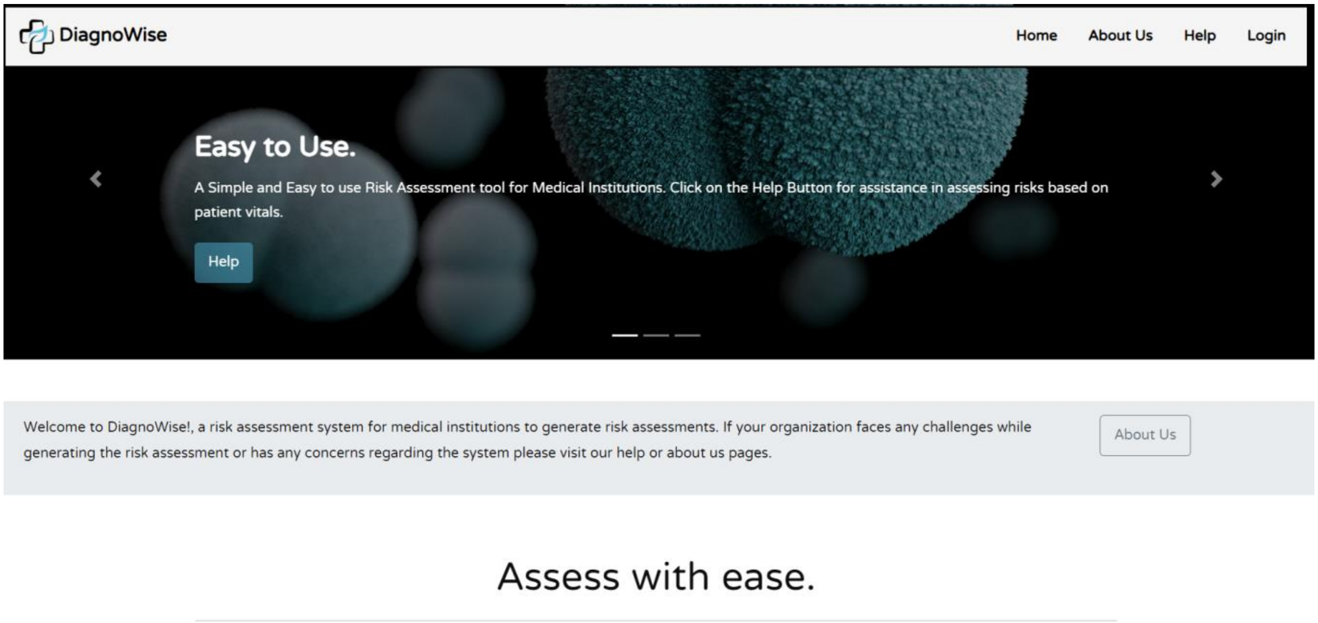


Fig. 9.1 Architecture Diagram

10. Prototype/Fully Developed Implementation’s Screenshot



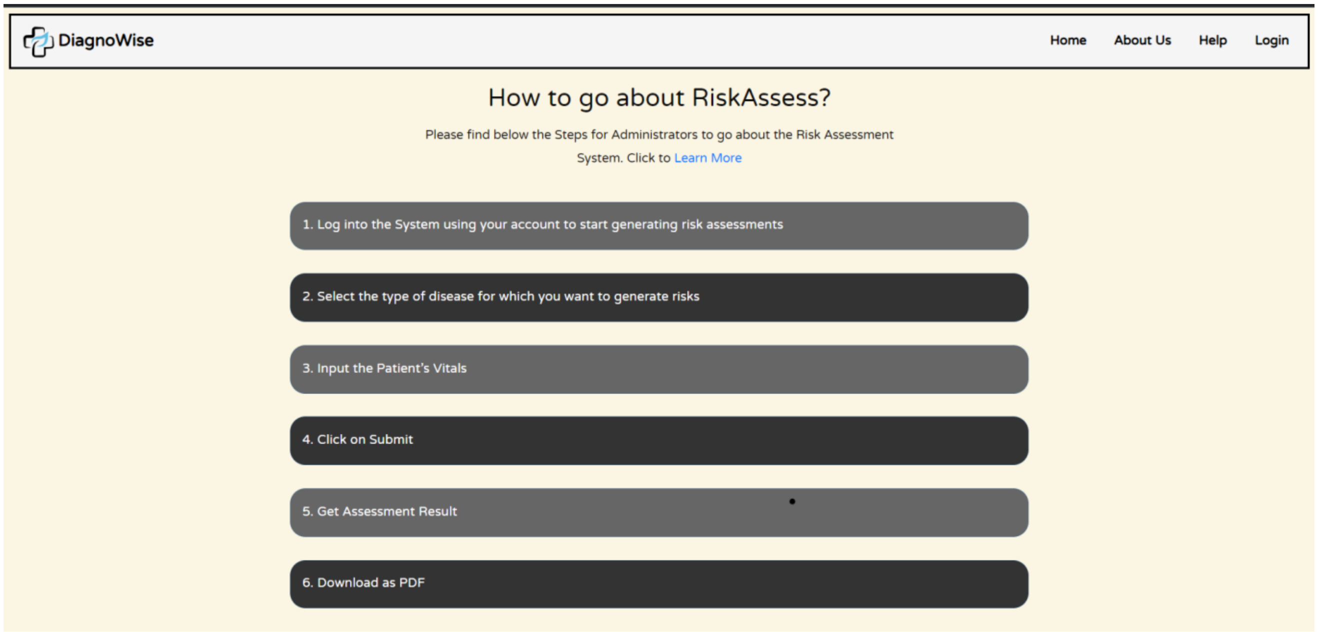
Easy to Use.

A Simple and Easy to use Risk Assessment tool for Medical Institutions. Click on the Help Button for assistance in assessing risks based on patient vitals.

Welcome to DiagnoWise!, a risk assessment system for medical institutions to generate risk assessments. If your organization faces any challenges while generating the risk assessment or has any concerns regarding the system please visit our help or about us pages.

Assess with ease.

Fig. 10.1 Home Page

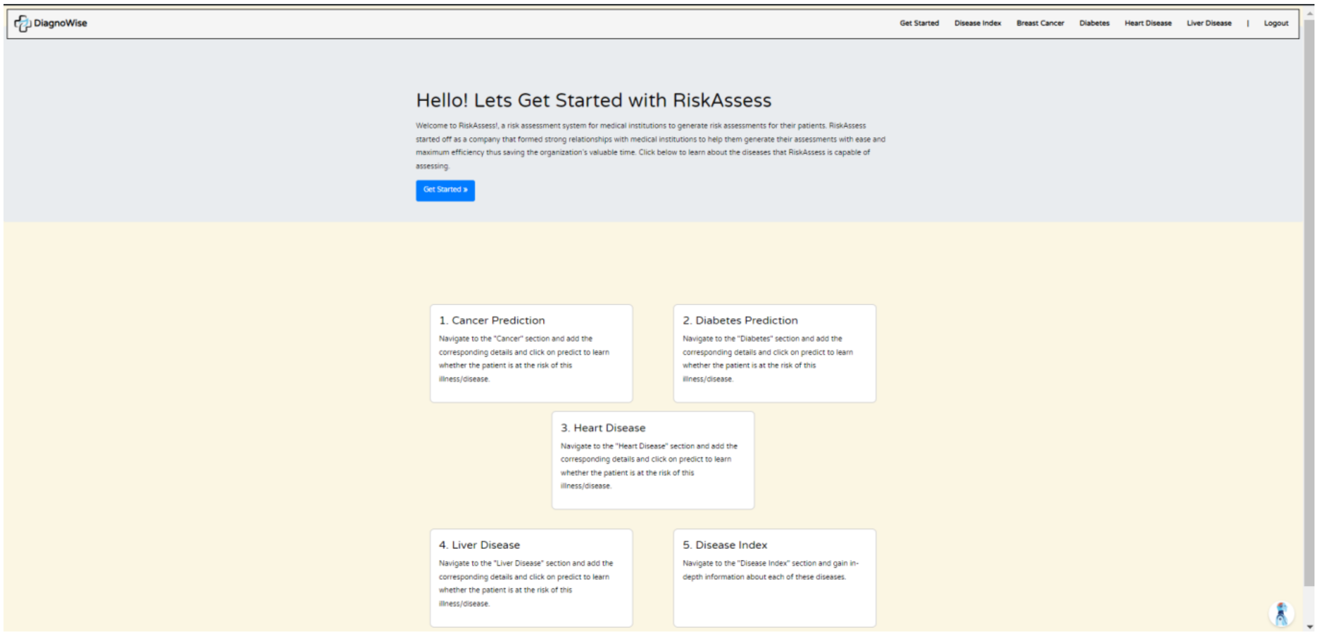


How to go about RiskAssess?

Please find below the Steps for Administrators to go about the Risk Assessment System. Click to Learn More

1. Log into the System using your account to start generating risk assessments
2. Select the type of disease for which you want to generate risks
3. Input the Patient’s Vitals
4. Click on Submit
5. Get Assessment Result
6. Download as PDF

Fig. 10.2 Help Page



Hello! Lets Get Started with RiskAssess

Welcome to RiskAssess!, a risk assessment system for medical institutions to generate risk assessments for their patients. RiskAssess started off as a company that formed strong relationships with medical institutions to help them generate their assessments with ease and maximum efficiency thus saving the organization's valuable time. Click below to learn about the diseases that RiskAssess is capable of assessing.

1. Cancer Prediction

Navigate to the "Cancer" section and add the corresponding details and click on predict to learn whether the patient is at the risk of this illness/disease.

2. Diabetes Prediction

Navigate to the "Diabetes" section and add the corresponding details and click on predict to learn whether the patient is at the risk of this illness/disease.

3. Heart Disease

Navigate to the "Heart Disease" section and add the corresponding details and click on predict to learn whether the patient is at the risk of this illness/disease.

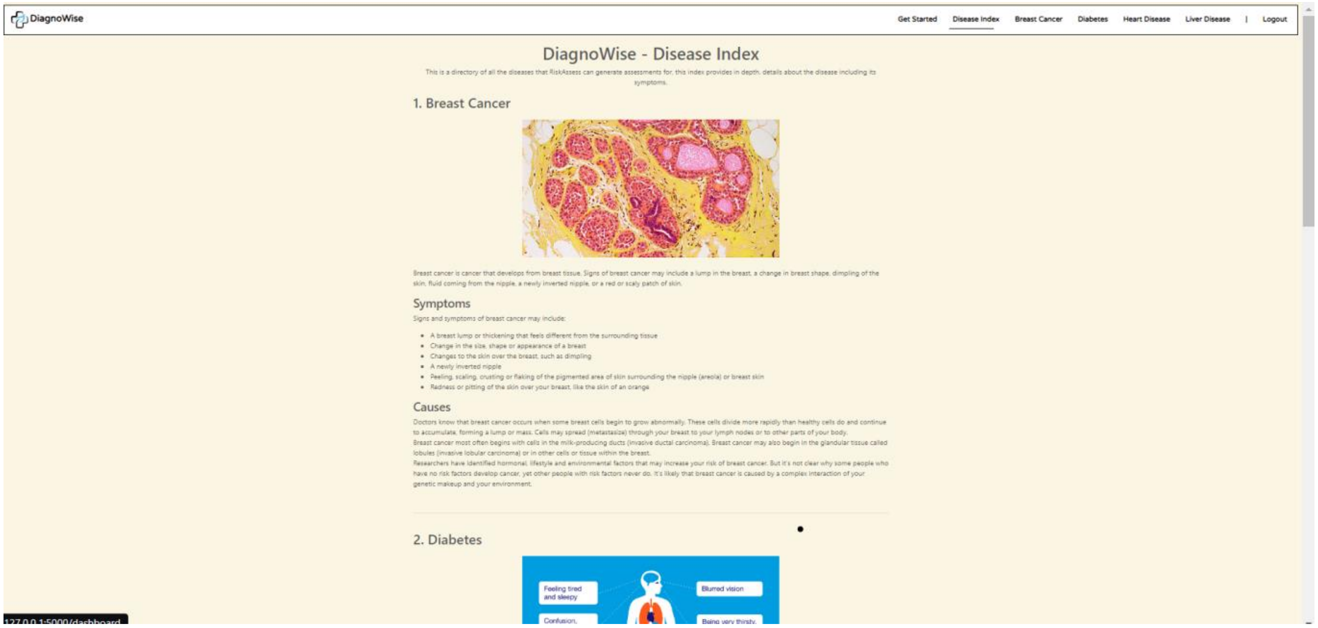
4. Liver Disease

Navigate to the "Liver Disease" section and add the corresponding details and click on predict to learn whether the patient is at the risk of this illness/disease.

5. Disease Index

Navigate to the "Disease Index" section and gain in-depth information about each of these diseases.

Fig. 10.3 Get Started



DiagnoWise - Disease Index

This is a directory of all the diseases that RiskAssess can generate assessments for, this index provides in depth, details about the disease including its symptoms.

1. Breast Cancer

Breast cancer is cancer that develops from breast tissue. Signs of breast cancer may include a lump in the breast, a change in breast shape, dimpling of the skin, fluid coming from the nipple, a newly inverted nipple, or a red or scaly patch of skin.

Symptoms

Signs and symptoms of breast cancer may include:

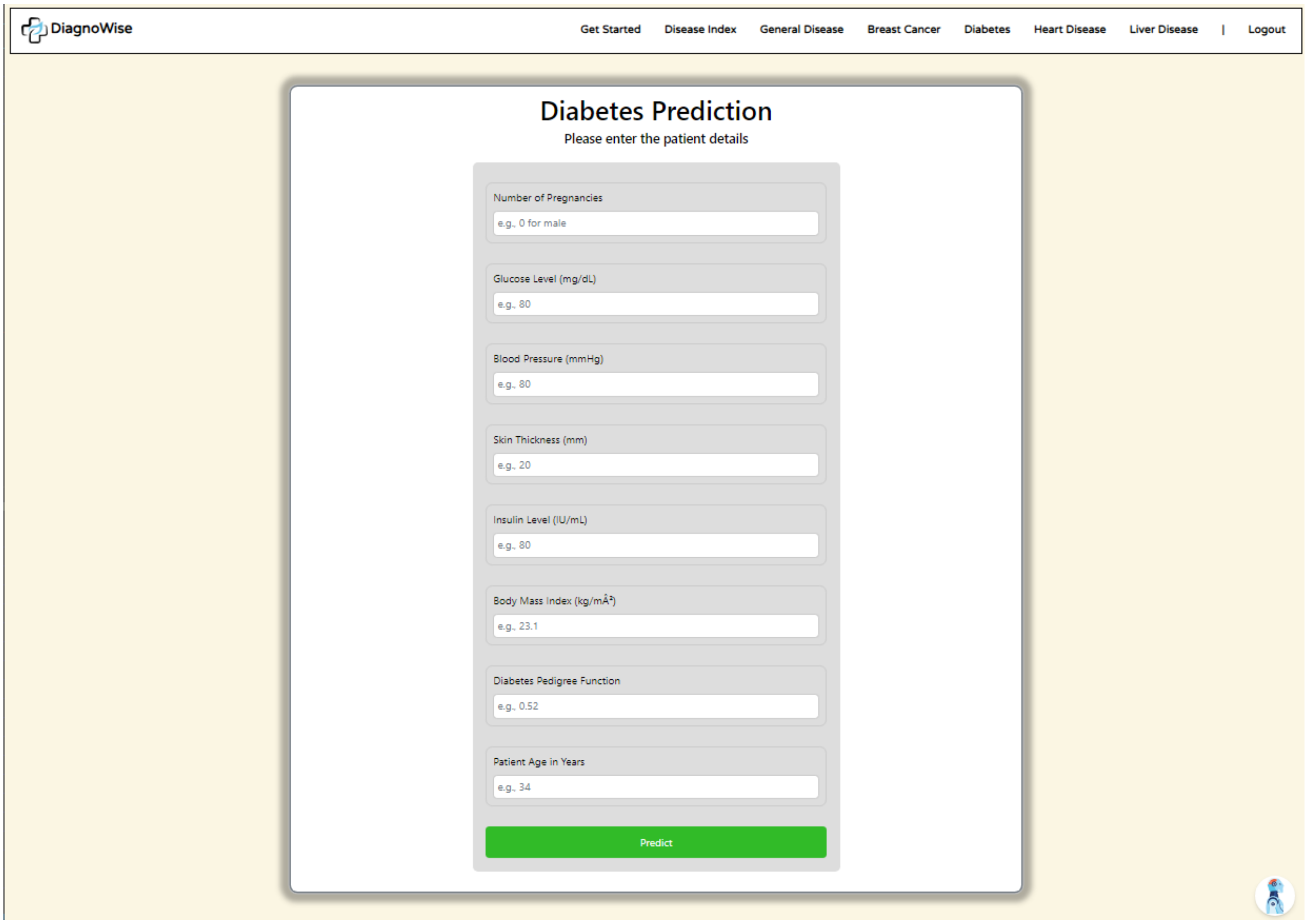
- A breast lump or thickening that feels different from the surrounding tissue
- Change in the size, shape or appearance of a breast
- Changes to the skin over the breast, such as dimpling
- A newly inverted nipple
- Peeling, scaling, crusting or flaking of the pigmented area of skin surrounding the nipple (areola) or breast skin
- Redness or pitting of the skin over your breast, like the skin of an orange

Causes

Doctors know that breast cancer occurs when some breast cells begin to grow abnormally. These cells divide more rapidly than healthy cells do and continue to accumulate, forming a lump or mass. Cells may spread (metastasize) through your breast to your lymph nodes or to other parts of your body.
Breast cancer most often begins with cells in the milk-producing ducts (invasive ductal carcinoma). Breast cancer may also begin in the glandular tissue called lobules (invasive lobular carcinoma) or in other cells or tissue within the breast.
Researchers have identified hormonal, lifestyle and environmental factors that may increase your risk of breast cancer. But it's not clear why some people who have no risk factors develop cancer, yet other people with risk factors never do. It's likely that breast cancer is caused by a complex interaction of your genetic makeup and your environment.

2. Diabetes

Fig. 10.4 Disease Information



Diabetes Prediction

Please enter the patient details

Number of Pregnancies
e.g., 0 for male

Glucose Level (mg/dL)
e.g., 80

Blood Pressure (mmHg)
e.g., 80

Skin Thickness (mm)
e.g., 20

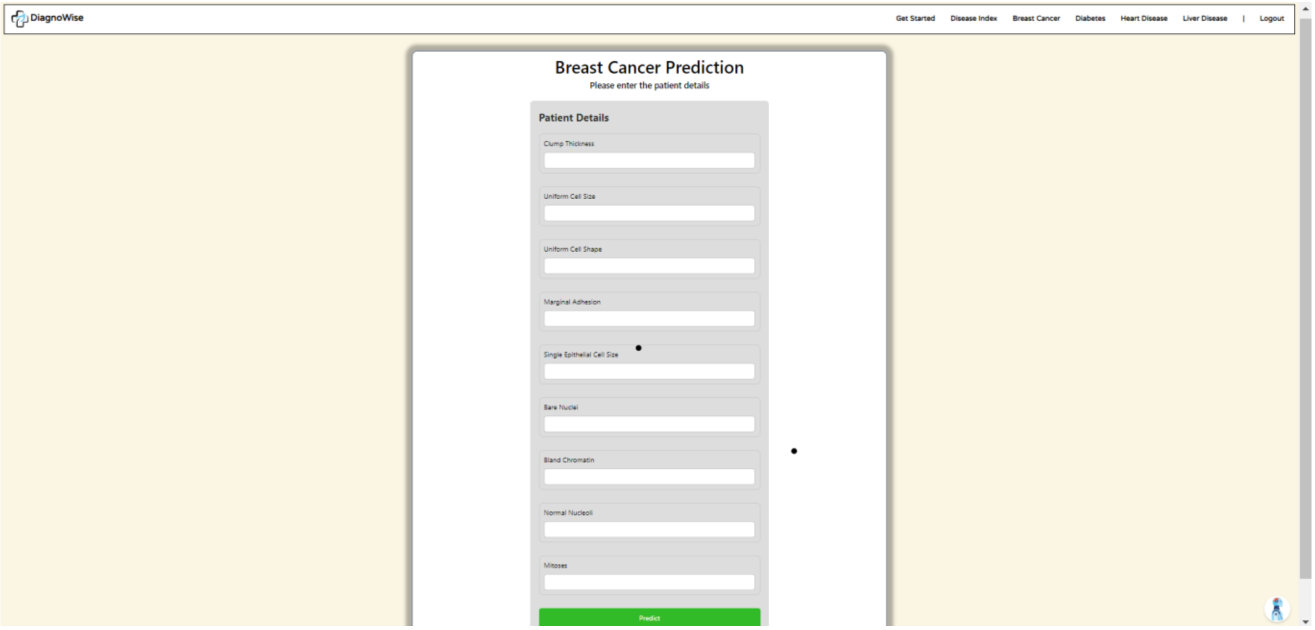
Insulin Level (IU/mL)
e.g., 80

Body Mass Index (kg/mÂ²)
e.g., 23.1

Diabetes Pedigree Function
e.g., 0.52

Patient Age in Years
e.g., 34

Fig. 10.5 Diabetes Model



Breast Cancer Prediction

Please enter the patient details

Patient Details

Clump Thickness

Uniform Cell Size

Uniform Cell Shape

Marginal Adhesion

Single Epithelial Cell Size

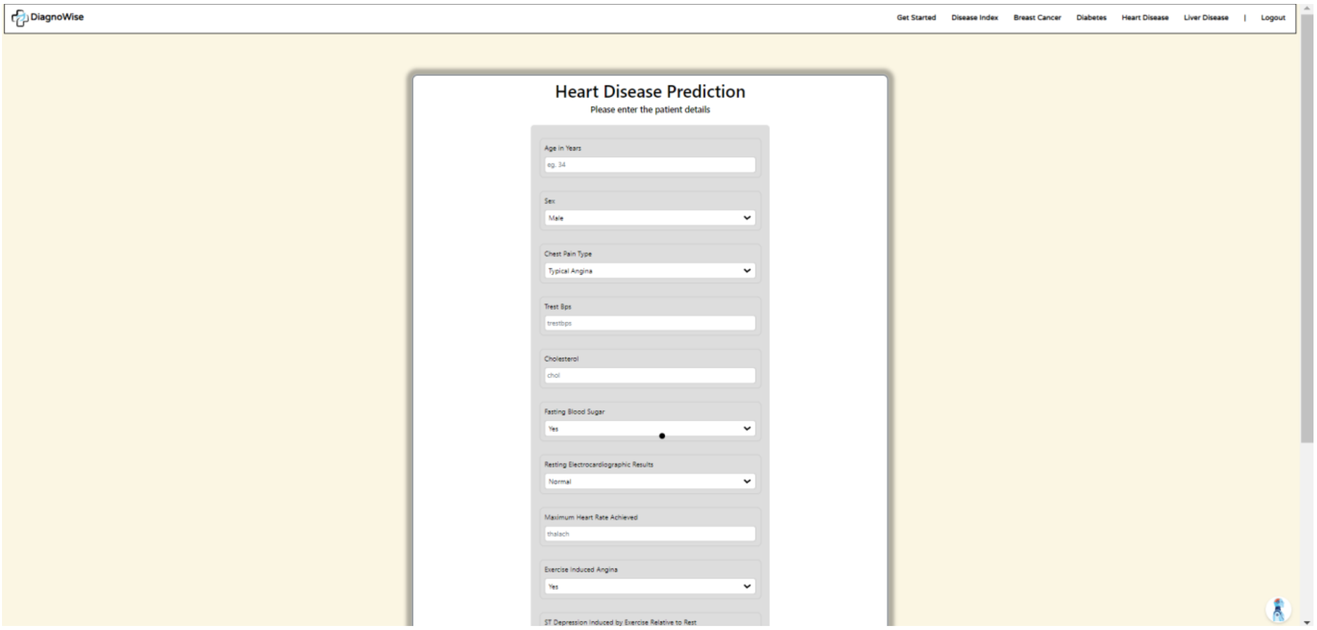
Bare Nuclei

Bland Chromatin

Normal Nucleoli

Mitoses

Fig. 10.6 Breast Cancer Model



Heart Disease Prediction

Please enter the patient details

Age in Years: eg. 34

Sex: Male

Chest Pain Type: Typical Angina

Trest Bps: trestbps

Cholesterol: chol

Fasting Blood Sugar: Yes

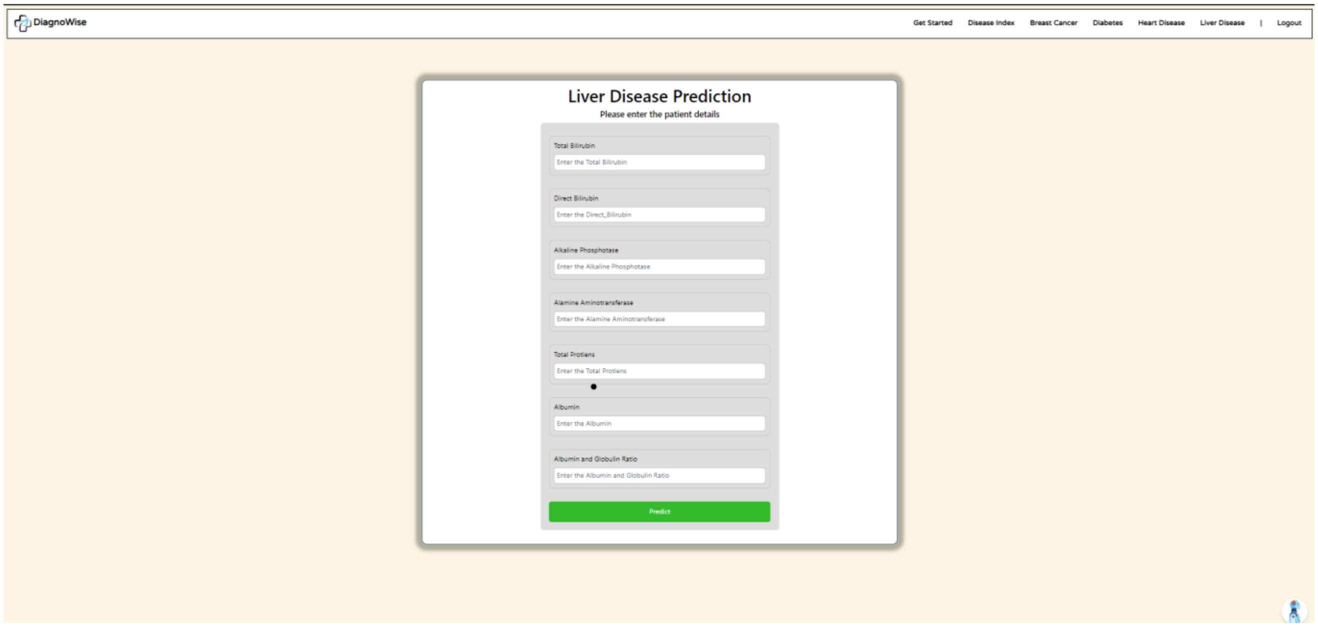
Resting Electrocardiographic Results: Normal

Maximum Heart Rate Achieved: thalach

Exercise Induced Angina: Yes

ST Depression Induced by Exercise Relative to Rest

Fig. 10.7 Heart Disease Model



Liver Disease Prediction

Please enter the patient details

Total Bilirubin: Enter the Total Bilirubin

Direct Bilirubin: Enter the Direct_Bilirubin

Alkaline Phosphotase: Enter the Alkaline Phosphotase

Alamine Aminotransferase: Enter the Alamine Aminotransferase

Total Protiens: Enter the Total Protiens

Albumin: Enter the Albumin

Albumin and Globulin Ratio: Enter the Albumin and Globulin Ratio

Fig. 10.8 Liver Disease Model

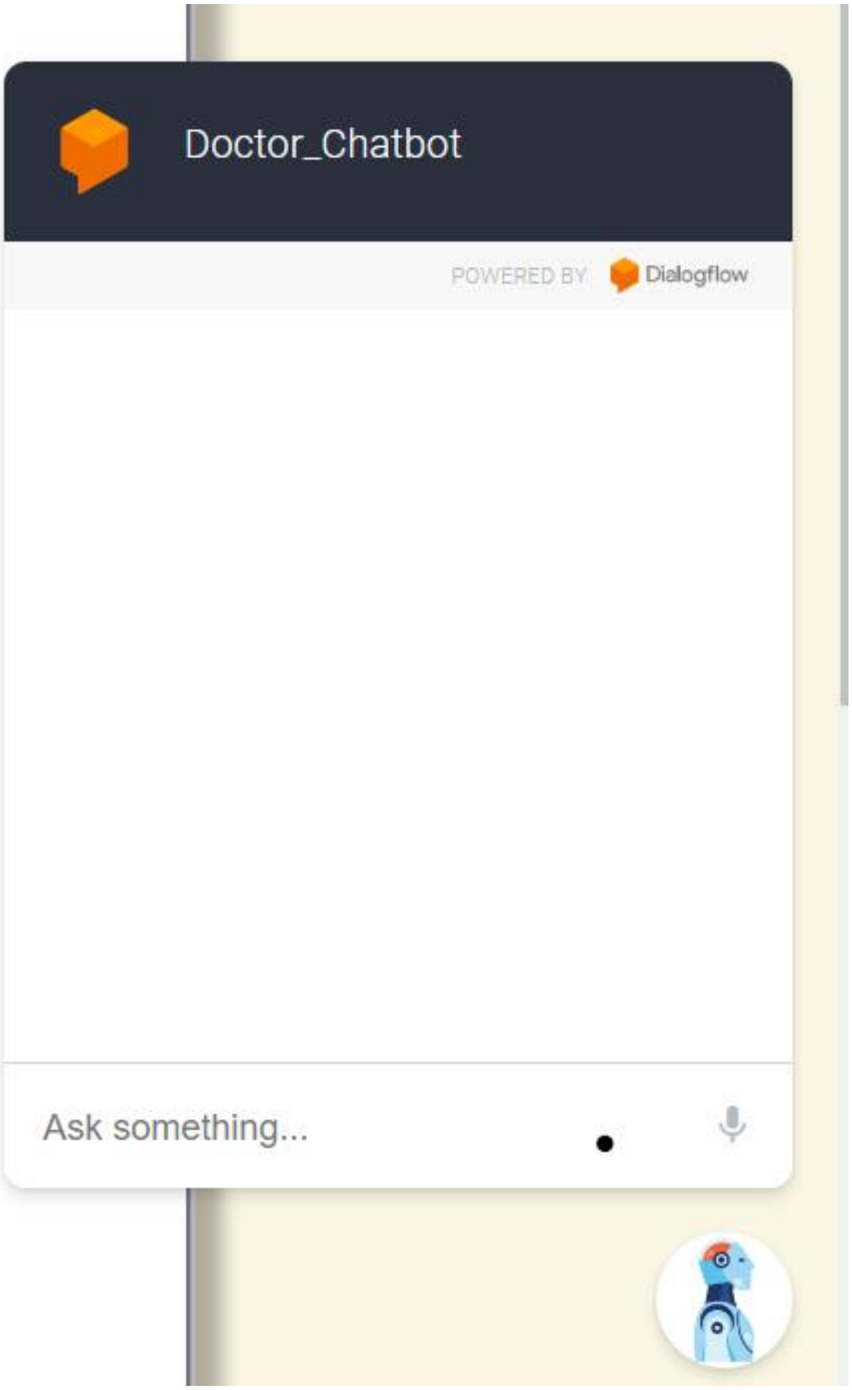
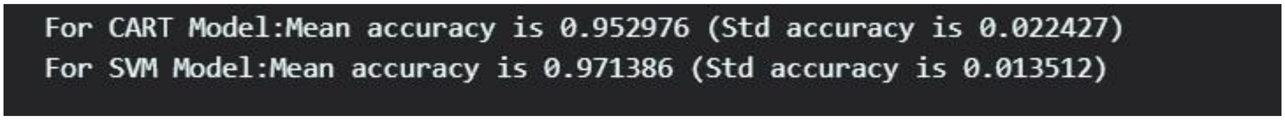


Fig. 10.9 Chatbot

11. Results

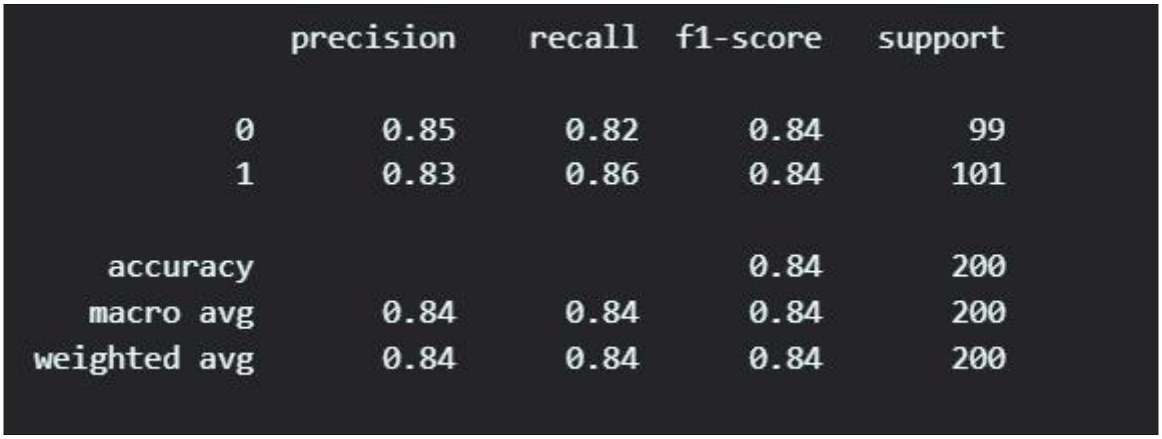
Results for the "DiagnoWise" project report would typically include:

1. Model Performance Metrics: Provide evaluation metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC) for each disease category. These metrics quantify the predictive performance of the developed models on the validation dataset.
2. Confusion Matrix: Present the confusion matrix for each disease category, showing the number of true positive, true negative, false positive, and false negative predictions made by the model. This matrix provides insights into the model's performance in terms of correct and incorrect predictions.
3. Feature Importance Analysis: Conduct feature importance analysis to identify the most informative features for each disease category. This analysis helps understand the underlying factors driving disease prediction and may provide insights for clinical decision-making.
4. Model Comparison: Compare the performance of different machine learning algorithms and techniques used in the project, highlighting the strengths and weaknesses of each approach. This comparison aids in selecting the best-performing models for deployment.
5. Clinical Relevance: Discuss the clinical relevance and implications of the predictive models developed in the project. Analyze how the models can assist healthcare professionals in disease risk assessment, early detection, and personalized treatment planning for patients.



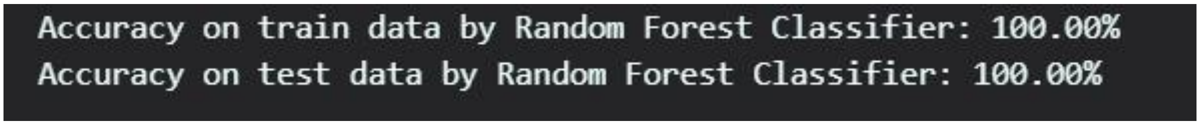
```
For CART Model:Mean accuracy is 0.952976 (Std accuracy is 0.022427)
For SVM Model:Mean accuracy is 0.971386 (Std accuracy is 0.013512)
```

Fig. 11.1 Breast Cancer Model Accuracy



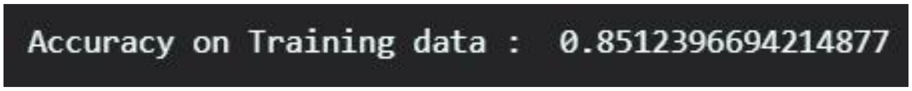
```
              precision    recall  f1-score   support

           0       0.85      0.82      0.84        99
           1       0.83      0.86      0.84       101

    accuracy                           0.84       200
   macro avg       0.84      0.84      0.84       200
weighted avg       0.84      0.84      0.84       200
```

Fig. 11.2 Diabetes Model Accuracy



```
Accuracy on train data by Random Forest Classifier: 100.00%
Accuracy on test data by Random Forest Classifier: 100.00%
```

Fig. 11.3 General Disease Model Accuracy



```
Accuracy on Training data :  0.8512396694214877
```

Fig. 11.4 Heart Disease Model Accuracy

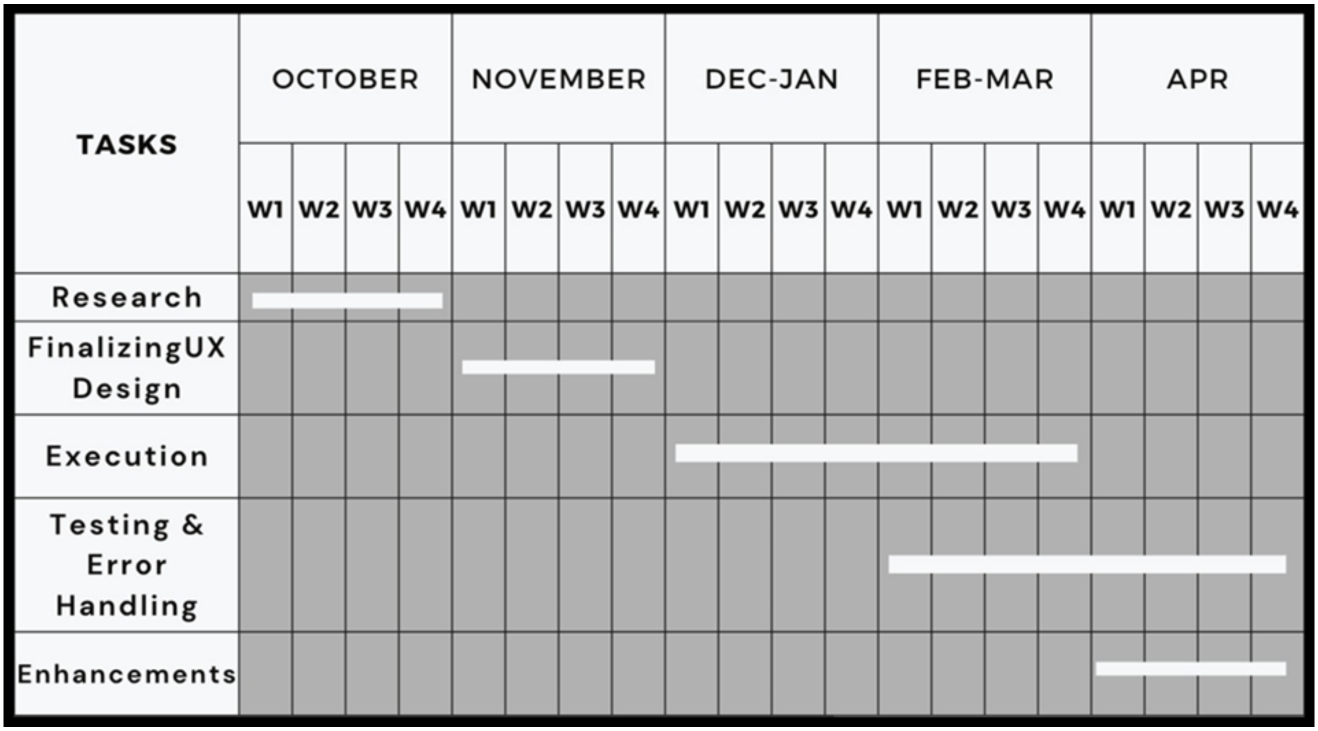
By presenting these results in the project report, stakeholders can gain valuable insights into the performance and clinical utility of the developed predictive models, facilitating informed decision-making and future research directions.

12. Gantt Chart / Timeline

A Gantt chart is a type of bar chart that is often used in project management to represent the planned tasks and activities within a project, along with their corresponding timelines. A Gantt chart for the disease predictor project could potentially include the following tasks and milestones:

- Requirements gathering: This initial phase would involve collecting and analyzing the requirements and goals for the website, as well as defining the scope and constraints of the project.
- System design: This phase would involve designing the overall architecture and components of the website, including the user interface, the trading engine, and the underlying blockchain infrastructure.
- Development: This phase would involve implementing the different components of the website, including the front-end user interface, the back-end trading engine, and the integration with blockchain networks.
- Testing: This phase would involve conducting various types of testing to ensure the reliability, security, and performance of the website. This could include unit testing, integration testing, and user acceptance testing.
- Deployment: This phase would involve launching the website and making it available for users to access and use. This could involve deploying the website to a live server, setting up the necessary security measures, and conducting any necessary post-launch activities.
- Maintenance and support: This phase would involve ongoing activities to maintain and improve the website, such as monitoring the website for any issues, addressing user feedback, and implementing new features and functionality.

Overall, the Gantt chart for the disease predictor project would likely be complex and detailed, reflecting the various tasks and dependencies involved in developing and launching a successful website.



TASKS	OCTOBER				NOVEMBER				DEC-JAN				FEB-MAR				APR			
	W1	W2	W3	W4	W1	W2	W3	W4	W1	W2	W3	W4	W1	W2	W3	W4	W1	W2	W3	W4
Research	■	■	■	■																
FinalizingUX Design					■	■	■	■												
Execution									■	■	■	■	■	■	■	■				
Testing & Error Handling													■	■	■	■	■	■	■	■
Enhancements																	■	■	■	■

Fig. 12.1 Gantt Chart

13. Proposed Enhancement

Here are proposed enhancements for the "DiagnoWise" project:

1. Additional Data Integration: Incorporate more data sources like genetic information or lifestyle habits for a comprehensive understanding of disease risk factors.
2. Streamlined Feature Engineering: Use advanced techniques for feature selection and reduction to improve model interpretability.
3. Model Ensemble: Combine predictions from multiple models to boost overall performance.
4. Optimized Hyperparameters: Fine-tune model parameters systematically to enhance accuracy and prevent overfitting.
5. Explainable AI Integration: Implement methods for transparent model explanations, ensuring trustworthiness.
6. Robust Cross-Validation: Employ advanced cross-validation strategies to better estimate model performance.
7. Continuous Model Monitoring: Establish systems for real-time model updates with new data to ensure relevance.
8. Improved User Interface: Enhance user interfaces with interactive features for better usability by healthcare professionals and end-users.
9. Implementing these enhancements would refine the "DiagnoWise" project, leading to more accurate and user-friendly disease prediction models.

14. Conclusion

To conclude, the "DiagnoWise" project has successfully developed predictive models for breast cancer, cardiovascular disease, diabetes, and liver disease using machine learning techniques. These models offer promising accuracy and hold potential for enhancing disease prediction and proactive healthcare management.

While the project has achieved significant milestones, challenges such as data scarcity and model interpretability persist. Nonetheless, ongoing efforts to refine the models and incorporate additional data sources are crucial for their continued improvement and applicability in real-world healthcare settings.

Moving forward, the "DiagnoWise" project aims to contribute to advancements in disease prediction and personalized healthcare, ultimately improving patient outcomes and advancing public health initiatives.

15. References

[1] Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics. 2005 Mar 1;21(5):631–43. [PubMed] [Google Scholar]

[2] Breiman L. Random forests. Machine Learning. 2001;45(1):5–32. [Google Scholar]

[3] Atlas, G. : Diabetes. International Diabetes Federation. 10th ed., IDF Diabetes Atlas. [Google Scholar]

[4] VijiyaKumar, K. , Lavanya, B. , Nirmala, I. , Caroline, S.S. : Random forest algorithm for the prediction of diabetes. In: International Conference on System, Computation, Automation and Networking, pp. 1–5 (2019)

[5] Mohan, N. , Jain, V. : Performance analysis of support vector machine in diabetes prediction. In: International Conference on Electronics, Communication and Aerospace Technology, pp. 1–3 (2020) [Google Scholar]

[6] Chatrati, S.P. , Hossain, G. , Goyal, A. , et al.: Smart home health monitoring system for predicting type 2 diabetes and hypertension. J. King Saud Univ. Comput. Inf. Sci. 34(3), 862–870 (2020) [Google Scholar]

[7] Omrani H. Predicting travel mode of individuals by machine learning. Transp. Res. Procedia. 2015;10:840–849. doi: 10.1016/j.trpro.2015.09.037. [CrossRef] [Google Scholar]

[8] Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv. 2017:3–9. [Google Scholar]

[9] Tian L., Zhang D., Bao S., Nie P., Hao D., Liu Y., et al. Radiomics-based machine-learning method for prediction of distant metastasis from soft-tissue sarcomas. Clin. Radiol. 2020;76(2):158.e19–158.e25. [PubMed] [Google Scholar]

[10] Lee E.J., Kim Y.H., Kim N., Kang D.W. Deep into the brain: Artificial intelligence in stroke imaging. J. Stroke. 2017;19(3):277–285. doi: 10.5853/jos.2017.02054. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

[11]. Stephens K. New Mammogram Measures of Breast Cancer Risk Could Revolutionize Screening. *AXIS Imaging News* . 2020 [Google Scholar]

[12]. Feld SI, Fan J, Yuan M, Wu Y, Woo KM, Alexandridis R, Burnside ES. Utility of Genetic Testing in Addition to Mammography for Determining Risk of Breast Cancer Depends on Patient Age. *AMIA Jt Summits Transl Sci Proc* . 2018;2017:81–90. [PMC Free Article] [PMC free article] [PubMed] [Google Scholar]

[13]. Guan Y, Nehl E, Pencea I, Condit CM, Escoffery C, Bellcross CA, McBride CM. Willingness to decrease mammogram frequency among women at low risk for hereditary breast cancer. *Sci Rep* . 2019;9(1):9599. doi: 10.1038/s41598-019-45967-6. [PMC Free Article] [PMC free article] [PubMed] [CrossRef] [Google Scholar]

[14]. American Cancer Society. *Cancer facts & figures 2018* . Atlanta: American Cancer Society; 2018. [Google Scholar]

[15]. Blandin Knight S, Crosbie PA, Balata H, Chudziak J, Hussell T, Dive C. Progress and prospects of early detection in lung cancer. *Open Biol* . 2017;7(9):170070. doi: 10.1098/rsob.170070. [PMC Free Article] [PMC free article] [PubMed] [CrossRef] [Google Scholar]

[16]. Lee E.J., Kim Y.H., Kim N., Kang D.W. Deep into the brain: Artificial intelligence in stroke imaging. *J. Stroke.* 2017;**19**(3):277–285. doi: 10.5853/jos.2017.02054. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

[17]. Miotto R., Wang F., Wang S., Jiang X., Dudley J.T. Deep learning for healthcare: Review, opportunities and challenges. *Brief. Bioinform.* 2018;**19**(6):1236–1246. doi: 10.1093/bib/bbx044. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

[18]. Alloghani M., Al-Jumeily D., Mustafina J., Hussain A., Aljaaf A.J. A systematic review on supervised and unsupervised machine learning algorithms for data science. In: Berry M., Mohamed A., Yap B., editors. *Supervised and unsupervised learning for data science.* Springer, US: 2020. pp. 3–21. [CrossRef] [Google Scholar]

[19]. Samuel A.L. Some studies in machine learning using the game of checkers. *IBM J. Res. Develop.* 2000;**44**(1-2):207–219. doi: 10.1147/rd.441.0206. [CrossRef] [Google Scholar]

[20]. Leijnen S., van Veen F. The neural network zoo. *Proceedings.* 2020;**47**(1):9. doi: 10.3390/proceedings2020047009. [CrossRef] [Google Scholar]