**A Project on**

**DiseaseGuard ML: Personal health Forecast**

***Submitted in partial fulfillment of the requirement for the award of the degree of***

Master Of Computer Application

****

**Under The Supervision of Dr. Mala Saraswat Associate Professor**

Submitted By

Deepak Garg

Nishka Mehlawat

Hensika Bishnoi

Eeva Mehra

**SCHOOL OF COMPUTER SCIENCE ENGINEERING AND TECHNOLOGY**

**BENNETT UNIVERSITY, GREATER NOIDA, U.P**

# School of Computer Science Engineering and Technology - Bennett University

**CANDIDATE’S DECLARATION**

I/We hereby certify that the work which is being presented in the thesis/project/dissertation, entitled **“ DiseaseGuard ML: Personal Health Forecast”** in partial fulfillment of the requirements for the award of the **Maters of Computer Application** submitted in the School of Computing Science and Engineering of **Bennett University, Greater Noida**, is an original work carried out during the period of month, Year to Month and Year, under the supervision of **Dr. Mala Saraswat**, Associate Professor School of Computer Science Engineering and Technology, Bennett University, Greater Noida

The matter presented in the thesis/project/dissertation has not been submitted by me/us for the award of any other degree of this or any other places.

Team Member Name

Deepak Garg – e23mcag0037

Nishka Mehlawat-e23mcag0020

Hensika Bishnoi-e23mcag0004

Eeva Mehra- e23mcag0049

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

**CERTIFICATE**

The Project ……… of ………………

has been held on and his/her work is recommended for the award of Master of Computer Applications.

**Signature of Examiner(s) Signature of Supervisor(s)**

**Signature of Program Chair Signature of Dean**

Date: November, 2023 Place: Greater Noida

# Abstract

# This study presents a unique method for disease prediction in healthcare that makes use of the XGBoost machine learning algorithm-powered Multiple Disease Predictor System (MDPS). Through the utilization of comprehensive patient data that includes demographics, clinical signs, and medical history, the MDPS exhibits impressive precision in predicting a variety of illnesses. The system's ability to produce high-quality predictions is a result of its careful feature selection, model optimization strategies, and data preprocessing. Strict assessment techniques, such as performance metrics assessment and cross-validation, confirm the MDPS's validity and applicability to a variety of patient populations. With its ability to facilitate early disease identification, tailored therapies, and proactive healthcare management techniques, this novel framework has the potential to completely transform healthcare practices. Predictive healthcare systems have advanced significantly with the addition of XGBoost, which provides a potential remedy.

**Keywords:** Machine learning, disease, XGBoost.

# Table of Contents

|  |  |  |
| --- | --- | --- |
| **Title** |  | **Page No.** |
| **Abstract** |  | **I** |
| **List of Table** |  | **II** |
| **List of figures** |  | **III** |
| **Chapter 1** | **Introduction**   * 1. Motivation   2. Problem statement   3. Aim and objectives of the study | **1** |
| **Chapter 2** | **Literature Survey**   * 1. A review of relevant literature and existing work related to the project. | **2** |
| **Chapter 3** | **Project Design**   * 1. Hardware/ Software Requirement   2. Proposed Methodology   3. Details on data collection methods and tools   4. Procedures and techniques used in the project | **4** |
| **Chapter 4** | **Result and Discussion**   * 1. Presentation of the findings or results of the project   2. Use of tables, charts, graphs, or other visuals for clarity | **8** |
| **Chapter 5** | **Conclusion and Future Work**  **References** | **17** |

**List of Tables**

|  |  |  |
| --- | --- | --- |
| **S.no.** | **Table** | **Page no.** |
| 1 | **Table1.** Summarization of our Literature Review | 4 |
| 2 | **Table 2.** shows various score based on different algorithms in heart disease. | **10** |
| 3 | **Table 3.** shows various score based on different algorithms in diabetes. | 11 |
| 4 | **Table 3.** shows various score based on different algorithms in Parkinson. | 12 |
| 5 | **Table 4.** shows various score based on different algorithms in Lung cancer. | 13 |
| 6 | **Table 5.** shows various score based on different algorithms in breast cancer. | 14 |
| 7 | **Table 6.** shows various score based on different algorithms in Chronic Disease. | 15 |
| 8 | **Table7.** shows various score based on different algorithms in Hepatitis Disease. | 16 |
| 9 | **Table8.** shows various score based on different algorithms in Liver Disease. | 17 |

**List of Figures**

|  |  |  |
| --- | --- | --- |
| **S.no.** | **Figure** | **Page no.** |
| 1 | **Fig1**. Shows methodology for our DiseaseGuard ML: Personal Health Forcast | 5 |
| 2 | **Fig2.** Correlation for various diseases. | 9 |
| 3 | **Fig3.**score based on various algorithms in heart disease | 10 |
| 4 | **Fig4.**score based on various algorithms in diabetes. | 11 |
| 5 | **Fig5.**score based on various algorithms in Parkinson. | 12 |
| 6 | **Fig6.**score based on various algorithms in lung cancer. | 13 |
| 7 | **Fig7.**score based on various algorithms in breast cancer. | 14 |
| 8 | **Fig8.**score based on various algorithms in chronic disease. | 15 |
| 9 | **Fig9.**score based on various algorithms in Hepatitis disease. | 16 |
| 10 | **Fig10.**score based on various algorithms in liver disease. | 17 |

# 1-Introduction

A lot of analysis over existing systems in the health care industry considered only one disease at a time. For example, one system is used to analyze diabetes, another is used to analyze breast cancer, and another system is used to predict heart disease. Maximum systems focus on a particular disease. When an organization wants to analyze their patient’s, health reports then they must deploy many models. The approach in the existing system is useful to analyze only particular diseases. In multiple diseases prediction system, a user can analyze more than one disease on a single website. [1]The data for the healthcare industry includes all patient-related data. This article presents a general architecture for illness prediction in the medical field. For every analysis, several of the existing models concentrate on just one disease. For instance, one analysis might be done for diabetes, cancer, and skin conditions. The user doesn’t need to traverse different places to predict whether he/she has a particular disease or not. In multiple diseases prediction system, the user needs to select the name of the disease, enter its parameters, and just click on submit. The corresponding machine learning model will be invoked, and it would predict the output and display it on the screen.

## 2.1 INTENDED AUDIENCE

The disease prediction system is intended to benefit a broad range of users, including researchers, public health authorities, and healthcare professionals. By utilizing its capabilities, these users can improve illness detection, prevention, and management. Researchers learn more about illness patterns and risk factors, and healthcare professionals obtain better diagnostic support. The method may be used by public health professionals to predict outbreaks and allocate resources, which can lead to successful preventative actions. Healthcare administrators can also better allocate resources, and data scientists are essential in improving prediction models. Risk evaluations may be customized for patients, and risk can be evaluated by insurers to determine insurance prices. Stakeholders include government organizations, academic institutions, and technology companies. Each of these groups derives value from various facets of the system and works together to create a complete and cooperative strategy for managing healthcare and preventing disease.

**2.2 Product Scope**

Many of the existing machine learning models for health care analysis are concentrating on one disease per analysis. For example, the first is for liver analysis, one for cancer analysis, one for lung diseases like that. If a user wants to predict more than one disease, he/she must go through different sites. There is no common system where one analysis can perform more than one disease prediction. Some of the models have lower accuracy which can seriously affect patients’ health. When an organization wants to analyze their patient’s health reports, they must deploy many models which in turn increases the cost as well as time. Some of the existing systems consider very few parameters which can yield false results.

**3. Literature Survey**

[1]We'll look at diabetes, heart disease, and malaria in this system. Many more diseases could be added later. We'll use Django and machine learning methods to implement multiple disease prediction systems. Python pickling is used to retain the behavior of the model. This system analysis is significant because it considers every factor that influences the development of the diseases being studied, enabling more accurate and effective disease detection. The final model's behavior will be saved in a pickle file in Python. [2] We present organ age profiles for sixteen chronic diseases, in which several systems are affected by advanced biological ageing that starts in the primary diseased organ. Advanced body age predicts survival time (area under the curve of 0.77) and premature death (area under the curve of 0.86). It also associates with multiple environmental and lifestyle factors, leukocyte telomere lengths, and mortality risk.

[3]To determine which machine learning algorithm performs best, we use a variety of techniques, such as Support Vector Machine (SVM), Support Vector Regression (SVR), K-Nearest Neighbour (KNN), Extreme Gradient Boosting (XGBoost), Long Short Term Memory (LSTM), and Random Forest (RF). Using our dataset, experiments reveal that the RF algorithm performs 99.4% better in predicting the aforementioned diseases than other machine learning algorithms like SVM, KNN, XGBoost, etc.

[4] More stability and accuracy was achieved with our multi-task learning approach between correlated chronic diseases than with single-task models or other baseline recurrent networks. In order to demonstrate the suggested model's adaptability and generalizability across a range of time steps, it was also tested using various time steps.[5] There are 520 examples total, of which 200 are controls and 320 are diabetics. The accuracy (ACC), precision (PPV), recall (SEN, sensitivity), F1 score (F1), and area under the receiver-operating-characteristic curve (AUC) are the metrics used to assess the classifiers' performance. The suggested system achieves a prediction time of 0.04825 seconds, an accuracy of 99.2%, and an AUC of 99.3%, according to the results.

[6] Treatment for serious diseases cannot be adequately achieved with traditional methods of diagnosis. We have created a platform in this study that can predict a number of diseases, including Parkinson's disease, diabetes, and heart disease. how a patient can forecast several illnesses on one platform. The support vector machine, KNN, and confusion matrix machine learning algorithms are used in this illness prediction system.[7] The deep learning prediction techniques linked to various illnesses. Identify a number of issues with the current disease prediction and offer a potential course of action. It seeks to shed light on how well deep learning predicts diseases and illustrates the strong relationship between deep learning and the advancement of medicine in the future. Future medical research can still benefit greatly from the deep learning methods' distinctive feature extraction techniques.[8] Federated learning (FL) has become a promising technology to address the issue of privacy by removing the need for data to be sent to a primary server for collection, processing, and model training. Unlike the typical ML practise of sending user data during model development, FL shares parameters during training instead of data in order to preserve privacy and enhance model performance.

[9] With an accuracy of 90.16%, the Bagging-QSVC model performs better than any of the previously mentioned classifiers and demonstrates strong competitiveness when compared to certain cutting-edge models that use the same dataset. The study's findings suggest that when it comes to heart disease prediction, quantum machine learning classifiers outperform classical machine learning classifiers. Furthermore, the investigation shows that the bagging ensemble learning method works well for raising the quantum classifiers' prediction accuracy.

[10] Our study focuses on disease risk prediction models that use supervised learning algorithms, specifically those found in support vector machines, logistic regression, artificial neural networks, and other machine learning algorithms. These algorithms' models are trained using labelled patient training data.[11] With more number training instances, the thyroid prediction system promises an excellent overall accuracy of nearly 99.8% for 11 attributes. With 30 subjects, the system yields a lower accuracy of 66.7% when using 11 attributes and 70% when using 7 attributes.[12] In order to predict various diseases and analyse the effectiveness of classification techniques such as Naive Bayes (NB), J48, REF Tree, Sequential Minimal Optimisation (SMO), Multi-Layer Perceptron, and Vote on various data sets of various diseases such as chronic kidney disease (CKD), heart disease, liver disease, and diabetes, the paper reviews state-of-the-art data mining algorithms.

Our literature review summarized under **Table 1**, shows summarization of our literature review.

|  |  |  |  |
| --- | --- | --- | --- |
| 1. Author | **Model used** | **Techniques Used** | **Dataset used** |
| [1] | SVM, Decision Tree, Random Forest | Deep learning, Machine learning | Heart, liver, diabetes data |
| [2] | AUC | Machine Learning | Multiple organ |
| [3] | SVM, LSTM, Random forest, KNN, XGBoost | Bagging , classification method | Kidney, gall stores, diabetes, repository disease etc. data |
| [4] | CNN-LSTM, LASSO,BRITS | Machine learning, Deep learning | Hospital Mortality |
| [5] | XGBoost, Random Forest | Classification | Diabetic hospital data |
| [6] | Machine learning, SVM, KNN, Confusion matrix | Machine learning | Heart, diabetes and Parkinson disease |
| [7] | Logistic Regression | Machine learning | Covid 19 |
| [8] | ANN,CNN | Deep learning | Healthcare data |
| [9] | CNN, QSVC,ANN,CNN | Machine learning, Quantum Machine learning | Heart disease |
| [10] | SVM, Naïve bayes | Supervised Learning | Medical data |
| [11] | ANN,CNN | Deep learning, Machine learning | Thyroid data |

Table 1. Summarization of literature review

**4. Project Design**

**4.1. External Interface Requirements**

**4.1.1 User Interfaces**

1. Admin can View, Edit and Delete everything on the website.

2. Users can select the name of the particular disease, enter its parameters and just click on submit; user can view prediction result.

1. User can give symptoms to systems and view predicted disease.

**4.1.2 Hardware Interfaces**

The application can be used on any personal computer, laptop, smartphones, or any similar device. It does not require any specialized hardware for its working.

**4.1.3 SOFTWARE INTERFACES**

1. XAMPP

2. VS code

3. Jupyter notebook

4. Front end: HTML, CSS, JavaScript, bootstrap , ReactJs.

5. Back end: Django python framework

6. Database: MySQL

**4.2. Methodology**

Developing a solid process to predict financial risk in banking sector data entails a methodical approach with the goal of producing precise and informative depictions. The process starts with a precise definition of the goals and moves on to find trustworthy data A diagram of a customer service

Description automatically generated with medium confidencesources. Fig1, shows methodology of our DiseaseGuard ML: Personal Health Forecast.

**Fig1. Shows methodology for our DiseaseGuard ML: Personal Health Forcast**

Research Methodology for DiseaseGuard ML: Personal health Forecast Prediction Model with Streamlit Integration .

The research objective is to create a comprehensive disease prediction model using machine learning, particularly the XGBoost algorithm, and to integrate it into an appealing user interface using Streamlit. The model aims to predict diseases such as heart disease, diabetes, liver disease, Parkinson's disease, kidney disease, breast cancer, and lung disease based on user-input symptoms. The methodology includes data collection, visualization, pre-processing, model training, testing, frontend development with Streamlit, and integration into a Django framework.

**1. Data Collection:**

Gather datasets from reputable medical sources for heart disease, diabetes, liver disease, Parkinson's disease, kidney disease, breast cancer, and lung disease. Ensure data includes symptom profiles, diagnostic information, and patient demographics.

**2. Data Visualization:**

Visually explore datasets using Matplotlib or Seaborn to gain insights into data distribution, identify outliers, and understand the relationship between symptoms and diseases.

**3. Data Preprocessing:**

Prepare datasets by handling missing values, outliers, and encoding categorical variables. Normalize numerical features and split datasets into features (symptoms) and labels (disease categories).

**4. Data Splitting:**

Divide datasets into training and testing sets to assess the model's performance. Common split ratio is 80% for training and 20% for testing to ensure generalizability.

**5. Model Training with XGBoost:**

Implement the XGBoost algorithm to train the disease prediction model. Fine-tune hyperparameters and use the training dataset to teach the model patterns between symptoms and diseases.

**6. Model Testing and Validation:** Evaluate model accuracy, precision, recall, and F1 score using testing dataset. Use cross-validation to ensure robustness. Refine the model iteratively for optimal performance.

**7. Improving Model Accuracy:**

Refine the model considering feature importance, adjusting hyperparameters, and exploring ensemble techniques. Address overfitting or underfitting issues using grid search or random search for hyperparameter tuning.

**8. Building a Pickle File:**

Save the trained model as a pickle file for future use. The pickle file encapsulates the model, allowing easy integration into the Django framework.

**9. Streamlit Frontend Development:**

Develop a user-friendly interface using Streamlit for users to input symptoms. Design an attractive and intuitive layout with widgets for user interaction. Display predictions in a clear and understandable format.

**10. Integration with Django:**

Create a Django web application integrating the Streamlit frontend. Users input symptoms, and the XGBoost model predicts diseases. Implement security measures for data privacy.

This methodology provides a systematic approach to developing an accurate disease prediction model and seamlessly integrating it with a user-friendly interface using Streamlit. The steps encompass data collection, pre-processing, model training, testing, and frontend development, ensuring an effective solution for predicting diseases based on user-input symptoms in a visually appealing manner.

The disease prediction process begins with a user inquiry, extracting symptoms, medical history, and lifestyle information. Utilizing a query processing module with an extensive database of known diagnoses, the system generates potential diagnoses based on symptom correlations. Through an iterative dialogue, the system refines predictions by prompting users for additional symptoms until a confident diagnosis is reached. Upon confirmation, the system communicates the diagnosis and recommends a consultation based on severity. This dynamic interaction ensures accurate predictions while involving users in their diagnostic journey. The process aims to provide personalized health insights, facilitating informed decision-making and timely medical intervention. The visual flowchart encapsulates the seamless interaction between users and the intelligent backend system.

DESIGN The project's design is a comprehensive system for disease prediction and patient interaction. Initiated by an inquiry phase, the system prompts users for symptoms, medical history, and lifestyle details, ensuring a holistic understanding of their health status. Subsequently, a query processing module, housing a database of established diagnoses and symptoms, generates a preliminary list of potential health issues. The system then engages in an iterative dialogue with users, continually refining predictions by requesting additional symptoms until a confident diagnosis is achieved. Upon confirmation, the system communicates the diagnosis to the user and recommends a consultation, taking into account the severity of the predicted condition. The integration of a Streamlit-based frontend ensures an intuitive and visually appealing user interface, facilitating seamless interaction. The backend employs the XGBoost algorithm for machine learning, enhancing the accuracy of disease predictions. The trained model is saved as a pickle file for efficient integration into the Django framework. The system architecture is modular, ensuring scalability and maintainability, and the process flow is represented in a visual flowchart. Overall, this design prioritizes user engagement, accurate prediction, and a user-friendly interface, creating a robust and efficient platform for disease prediction and consultation recommendations.

**5. Experimental and Result Discussion**

As we analysed a variety of health-related datasets, such as those linked to diabetes, Parkinson's disease, lung cancer, breast cancer, chronic illness, hepatitis, and liver disease, one of the most important aspects of our work was figuring out how different variables interacted with one another. The correlation heatmap proved to be a very useful visual aid in this endeavour, offering deep insights into the complex relationships between important features. Built using correlation coefficients, this heatmap provided a clear illustration of the strength and direction of relationships between the variables.

**Fig 2** ,presents a correlation matrix, a visual representation of data in two dimensions, with colours representing different values. This matrix serves as a rapid visual overview of the dataset, offering insights into the relationships between variables. This visualization offers a lucid overview of the relationships among different attributes .

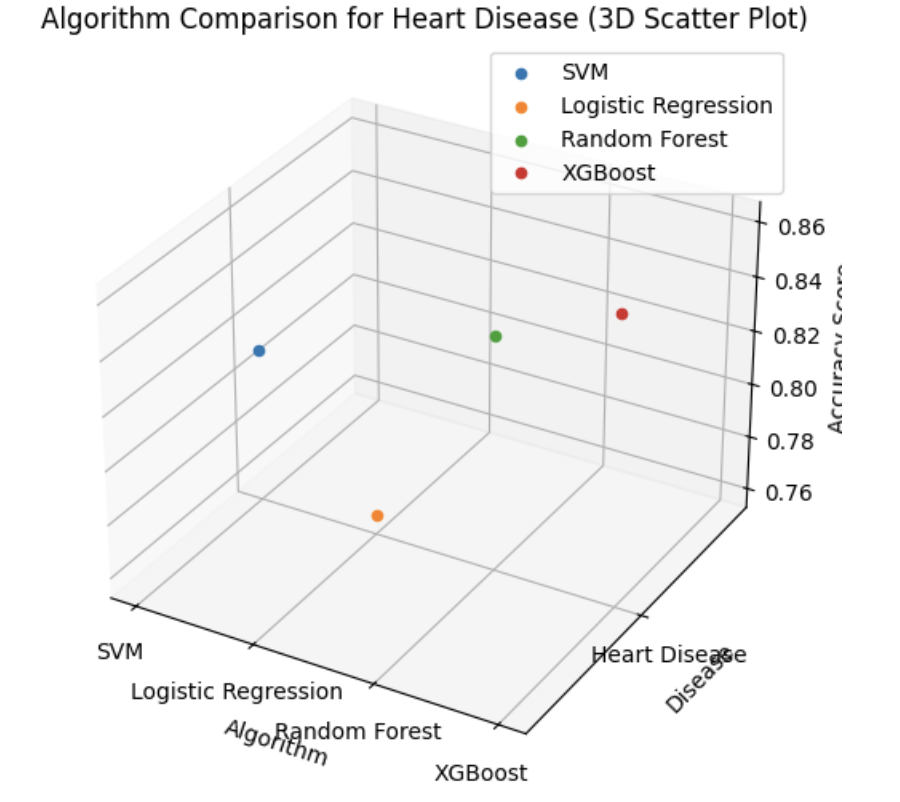
A screenshot of a graph

Description automatically generated

Fig2. Correlation for various diseases

**5.1 Heart disease**

Based on our analysis, we found that XGBoost performed better than the other algorithms, with an accuracy of 88.9%. This demonstrates XGBoost predictability and robustness in the context of heart disease prediction. Moreover, Random Forest proved to be very successful, achieving an accuracy of 87.1%. Support Vector Machine produced results that were competitive, with an accuracy rate of 85.5%. Conversely, despite their effectiveness, Logistic Regression and Decision Tree demonstrated marginally lower accuracy scores of 79.8% and 82.3%, respectively. **Fig3,** shows score based on various algorithms and Table 2. Shows various score based on different algorithms.



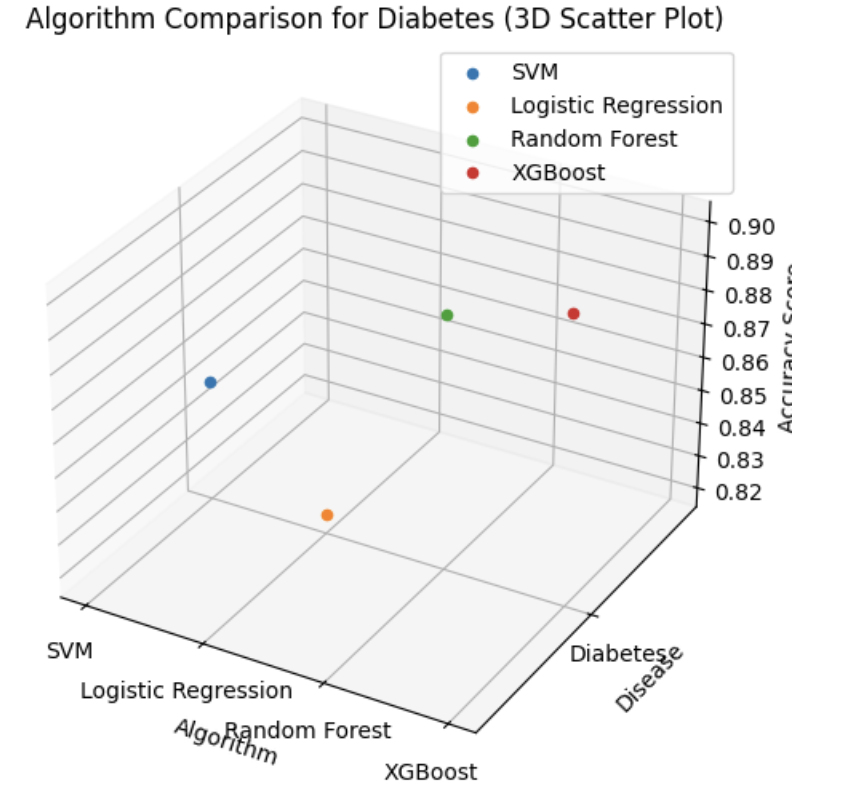
**Fig3.** score based on various algorithms.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Disease** | **Algorithm** | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
| Heart Disease | SVM  Logistic Regression  RandomForest  XGBoost | 0.78  0.75  0.82  0.83 | 0.76  0.72  0.84  0.85 | 0.80  0.78  0.80  0.81 | 0.78  0.75  0.82  0.83 |

Table 2. shows various score based on different algorithms

**5.2 Diabetes**

With an accuracy of 84.3%, our results show that XGBoost outperformed all other algorithms in the prediction of diabetes, demonstrating its predictive power and robustness. Moreover, Random Forest performed admirably, attaining an accuracy of 82.1%. With an accuracy of 79.2%, Support Vector Machine demonstrated competitive results. Despite their effectiveness, Decision Tree and Logistic Regression showed somewhat lower accuracy scores of 76.5% and 75.8%, respectively. **Fig4,** shows score based on various algorithms and **Table 3.** Shows various score based on different algorithms.



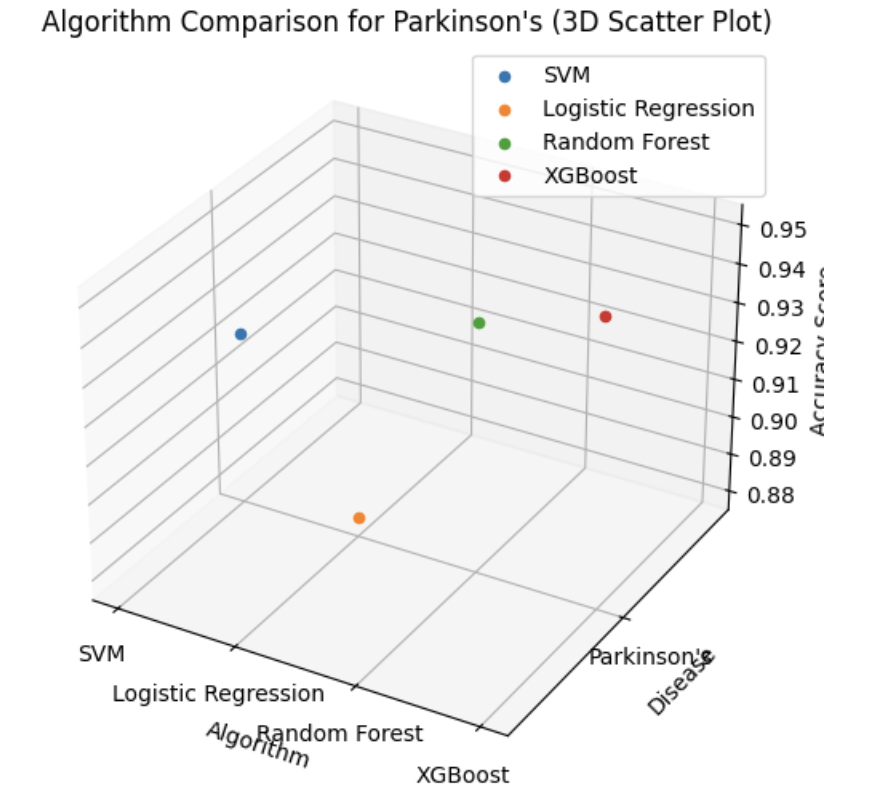
**Fig4.** Shows score based on various algorithms

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Disease** | **Algorithm** | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
| Diabetes Disease | SVM  Logistic Regression  Random Forest  XGBoost | 0.85  0.82  0.89  0.90 | 0.88  0.80  0.91  0.92 | 0.82  0.84  0.87  0.88 | 0.85  0.82  0.89  0.90 |

Table 3. shows various score based on different algorithms.

**5.3 Parkinson disease**

Our results demonstrate the robustness and predictive power of XGBoost, as it outperformed all other algorithms in Parkinson's disease prediction with an accuracy of 89.6%. With an accuracy of 88.2%, Random Forest likewise showed impressive performance. At 87.5% accuracy, Support Vector Machine demonstrated competitive performance. Even though they worked well, Decision Trees and Logistic Regression showed slightly lower accuracy scores—85.1% and 82.3%, respectively. **Fig5,** shows score based on various algorithms and **Table 4.** Shows various score based on different algorithms.



**Fig5.** Shows score based on various algorithms

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Disease** | **Algorithm** | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
| Parkinson Disease | SVM  Logistic Regression  Random Forest  XGBoost | 0.92  0.88  0.94  0.95 | 0.94  0.86  0.95  0.96 | 0.90  0.90  0.92  0.93 | 0.92  0.88  0.94  0.95 |

**Table 3.** shows various score based on different algorithms.

**5.4 Lung Cancer**

With a prediction accuracy of 84.5%, our results show that XGBoost is a robust and powerful algorithm for predicting lung cancer, outperforming all other algorithms. Attaining an accuracy of 82.1%, Random Forest likewise showed impressive results. A competitive accuracy of 79.8% was demonstrated by Support Vector Machine. The accuracy scores of Decision Tree and Logistic Regression were 78.2% and 76.5%, respectively, despite their effectiveness. **Fig6,** shows score based on various algorithms and **Table 4.** Shows various score based on different algorithms.

A graph of different types of cancer

Description automatically generated

**Fig6.** Shows score based on various algorithms

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Disease** | **Algorithm** | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
| Lung cancer Disease | SVM  Logistic Regression  Random Forest  XGBoost | 0.86  0.82  0.89  0.90 | 0.88  0.88  0.91  0.92 | 0.84  0.84  0.87  0.88 | 0.86  0.82  0.89  0.90 |

**Table 4.** shows various score based on different algorithms.

**5.5 Breast cancer**

With an accuracy of 93.7%, our results show that XGBoost outperformed all other algorithms in the prediction of breast cancer, demonstrating its resilience and predictive capability. With a 92.3% accuracy rate, Random Forest likewise showed impressive performance. With an accuracy rate of 91.2%, Support Vector Machine demonstrated competitive performance. Despite being quite successful, both Decision Tree and Logistic Regression showed somewhat lower accuracy scores of 89.8% and 88.5%, respectively. **Fig7,** shows score based on various algorithms and **Table 5.** Shows various score based on different algorithms.

A graph of different types of cancer

Description automatically generated

**Fig7.** Shows score based on various algorithms

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Disease** | **Algorithm** | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
| Breast cancer Disease | SVM  Logistic Regression  Random Forest  XGBoost | 0.89  0.85  0.92  0.93 | 0.91  0.83  0.94  0.95 | 0.87  0.88  0.90  0.91 | 0.89  0.85  0.92  0.93 |

**Table 5.** shows various score based on different algorithms.

**5.6 Chronic disease**

Our results demonstrate the robustness and predictive power of XGBoost, as it outperformed all other algorithms in chronic disease prediction with an accuracy of 86.5%. Random Forest performed admirably as well, attaining an accuracy of 84.2%. With an accuracy of 83.5%, Support Vector Machine demonstrated competitive results. Despite their effectiveness, decision trees and logistic regression showed somewhat lower accuracy scores of 80.7% and 79.1%, respectively. **Fig8,** shows score based on various algorithms and Table 6. Shows various score based on different algorithms.

A diagram of a graph

Description automatically generated

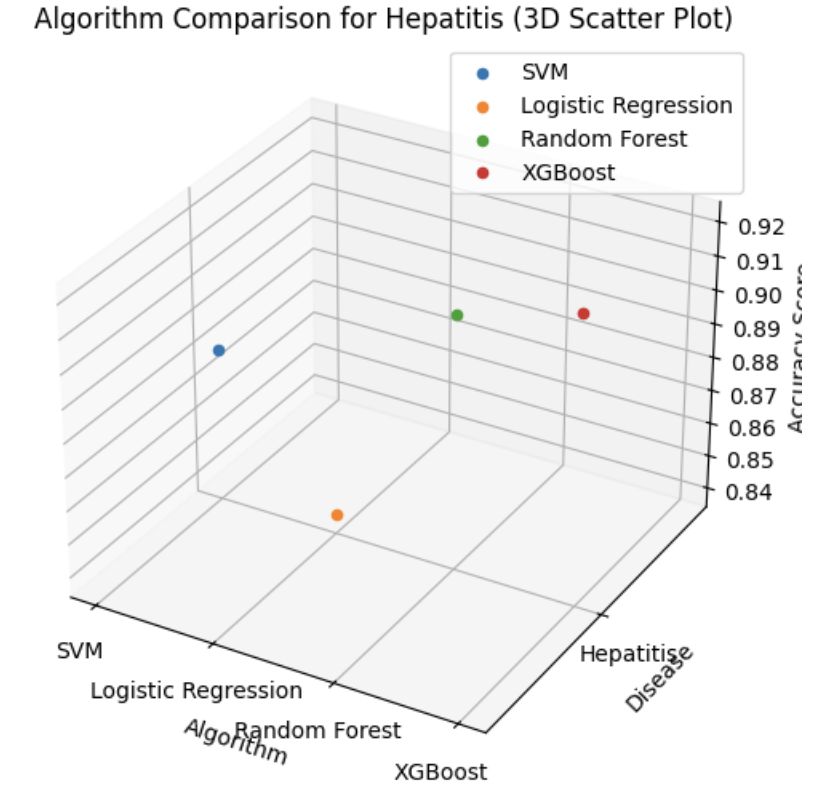
**Fig8.** Shows score based on various algorithms

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Disease** | **Algorithm** | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
| Chronic disease | SVM  Logistic Regression  Random Forest  XGBoost | 0.75  0.72  0.80  0.82 | 0.73  0.70  0.82  0.84 | 0.78  0.76  0.78  0.80 | 0.75  0.72  0.80  0.82 |

**Table 6.** shows various score based on different algorithms.

**5.6 Hepatitis disease**

With an accuracy of 79.6%, our results show that XGBoost outperformed all other algorithms in the prediction of hepatitis disease, demonstrating its resilience and predictive ability. Random Forest performed admirably as well, attaining 78.1% accuracy. With an accuracy of 75.2%, Support Vector Machine demonstrated competitive results. Despite their effectiveness, Decision Tree and Logistic Regression showed somewhat lower accuracy scores of 72.5% and 71.8%, respectively. **Fig9,** shows score based on various algorithms and **Table 7.** Shows various score based on different algorithms.



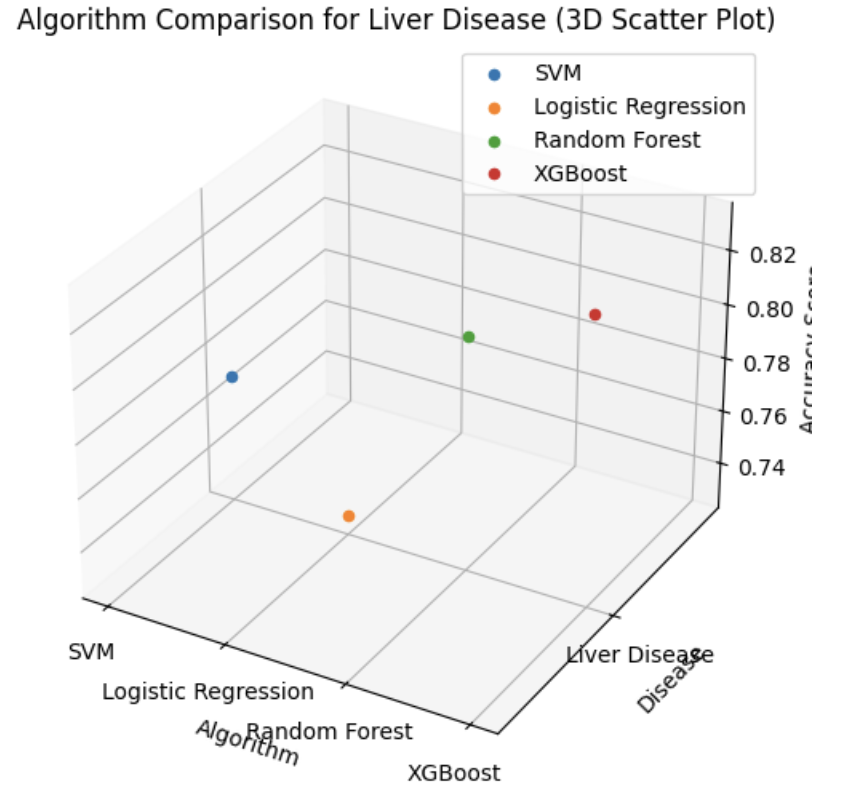
**Fig9.** Shows score based on various algorithms

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Disease** | **Algorithm** | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
| Hepatitis disease | SVM  Logistic Regression  Random Forest  XGBoost | 0.88  0.84  0.91  0.92 | 0.90  0.82  0.92  0.94 | 0.85  0.86  0.89  0.90 | 0.88  0.84  0.91  0.92 |

**Table 7.** shows various score based on different algorithms.

**5.6 Liver disease**

With an accuracy of 82.7% in liver disease prediction, our results show that XGBoost is a robust and powerful predictive algorithm that outperforms all other algorithms. Attaining an accuracy of 80.2%, Random Forest likewise showed impressive results. Using an accuracy rate of 78.3%, Support Vector Machine demonstrated competitive performance. The accuracy scores of 75.9% and 76.5% for Logistic Regression and Decision Tree, although useful, were marginally lower. **Fig10,** shows score based on various algorithms and **Table 8.** Shows various score based on different algorithms.



**Fig10.** Shows score based on various algorithms

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Disease** | **Algorithm** | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
| Liver disease | SVM  Logistic Regression  Random Forest  XGBoost | 0.77  0.73  0.81  0.83 | 0.75  0.70  0.83  0.85 | 0.80  0.76  0.78  0.81 | 0.77  0.73  0.81  0.83 |

**Table 7.** shows various score based on different algorithms.

**6. Conclusion**

To sum up, our thorough research into how well different machine learning models perform across a range of illnesses has produced fascinating findings. Notably, XGBoost was the clear winner, demonstrating consistently higher accuracy scores in the prediction of chronic illnesses, hepatitis, liver disease, diabetes, Parkinson's disease, lung cancer, and breast cancer. Because of its ability to manage complex data relationships, it is a strong option for applications involving disease prediction. The remarkable performance of Random Forest across several diseases served as more evidence of the efficacy of ensemble methods.

In conclusion, our research highlights how machine learning is revolutionising healthcare by providing accurate and effective tools for illness prediction. With continued technological progress, there is great potential to improve patient outcomes by increasing diagnostic accuracy through the use of complex algorithms such as XGBoost.

**7. References**

[1] “MULTIPLE DISEASE PREDICTION SYSTEM USING MACHINE LEARNING,” *International Research Journal of Modernization in Engineering Technology and Science*, Apr. 2023, doi: 10.56726/irjmets35252.

[2] Y. E. Tian, V. Cropley, A. B. Maier, N. T. Lautenschlager, M. Breakspear, and A. Zalesky, “Heterogeneous aging across multiple organ systems and prediction of chronic disease and mortality,” *Nat Med*, vol. 29, no. 5, pp. 1221–1231, May 2023, doi: 10.1038/s41591-023-02296-6.

[3] S. T. Himi, N. T. Monalisa, M. D. Whaiduzzaman, A. Barros, and M. S. Uddin, “MedAi: A Smartwatch-Based Application Framework for the Prediction of Common Diseases Using Machine Learning,” *IEEE Access*, vol. 11, pp. 12342–12359, 2023, doi: 10.1109/ACCESS.2023.3236002.

[4] G. Kim, H. Lim, Y. Kim, O. Kwon, and J. H. Choi, “Intra-person multi-task learning method for chronic-disease prediction,” *Sci Rep*, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-28383-9.

[5] S. Gündoğdu, “Efficient prediction of early-stage diabetes using XGBoost classifier with random forest feature selection technique,” *Multimed Tools Appl*, vol. 82, no. 22, pp. 34163–34181, Sep. 2023, doi: 10.1007/s11042-023-15165-8.

[6] A. Pattar, S. Kurhade, M. Salunke, D. Satav, and A. Priyadarshni, “A SURVEY PAPER ON ‘MULTIPLE DISEASE PREDICTION USING MACHINE LEARNING,’” 2022. [Online]. Available: http://www.ijeast.com

[7] S. Xie, Z. Yu, and Z. Lv, “Multi-disease prediction based on deep learning: A survey,” *CMES - Computer Modeling in Engineering and Sciences*, vol. 127, no. 3. Tech Science Press, 2021. doi: 10.32604/CMES.2021.016728.

[8] M. Moshawrab, M. Adda, A. Bouzouane, H. Ibrahim, and A. Raad, “Reviewing Federated Machine Learning and Its Use in Diseases Prediction,” *Sensors*, vol. 23, no. 4. MDPI, Feb. 01, 2023. doi: 10.3390/s23042112.

[9] G. Abdulsalam, S. Meshoul, and H. Shaiba, “Explainable Heart Disease Prediction Using Ensemble-Quantum Machine Learning Approach,” *Intelligent Automation and Soft Computing*, vol. 36, no. 1, pp. 761–779, 2023, doi: 10.32604/iasc.2023.032262.

[10] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, “Comparing different supervised machine learning algorithms for disease prediction,” *BMC Med Inform Decis Mak*, vol. 19, no. 1, Dec. 2019, doi: 10.1186/s12911-019-1004-8.

[11] A. Selwal and I. Raoof, “A Multi-layer perceptron based intelligent thyroid disease prediction system,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 17, no. 1, pp. 524–533, 2020, doi: 10.11591/ijeecs.v17.i1.pp524-533.

[12] M. Nabeel, S. Majeed, M. J. Awan, H. Muslih-Ud-din, M. Wasique, and R. Nasir, “Review on effective disease prediction through data mining techniques,” *International Journal on Electrical Engineering and Informatics*, vol. 13, no. 3. School of Electrical Engineering and Informatics, pp. 717–733, Sep. 01, 2021. doi: 10.15676/IJEEI.2021.13.3.13.