International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 03 March 2025
Page No: 448-455
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0062

Multi-Disease Detection with Doctor Recommendation System

Prof. Aparna Kulkarni[1], Mr. Ritesh Jalkote[2], Mr. Yash Khedkar[3], Mr. Omkar Bhor[4], Mr. Harsh Sutar[5]
[1,2,3,4,5]Department of Artificial Intelligence & Data Science Engineering, Dr. D. Y. Patil Institute of Technology Pimpri, Pune, India.
Emails: *Aparna.kulkarni@dypvp.edu.in[1], jalkoteritesh2@gmail.com[2], yashkhedkar2522@gmail.com[3], omkarbhor360@gmail.com[4], harshsutar304@gmail.com[5]*

Abstract
AI and machine learning, from self-driving cars to health care, have turned out to be an indispensable tool for several industries. In medicine, the big availability of patient records now opens new possibilities for applying techniques of machine learning for disease detection and diagnosis at earlier stages. The main objective of this work is to provide an advanced prediction system capable of detecting multiple diseases. This will overcome one of the major disadvantages of most systems, which usually target single diseases and may or may not do so very accurately. Our system will target the following major five diseases: Heart Disease, Liver Disease, Diabetes, Lung Cancer, and Parkinson's Disease-specifically, although it has the flexibility for future gains on other conditions. The following project incorporates different disease-specific parameters, wherein a user can enter their health data to get accurate predictions of the presence of a disease. This project would have a greater influence on enabling people to monitor and take precautions for better maintenance of health, thereby helping to increase the life span of an individual. In using machine learning in this respect, such a system may support individual well-being by providing the most accurate disease predictions possible, saving lives if that be the case.

Keywords: *Exploratory data analysis (EDA); k-Nearest Neighbours algorithm (KNN); Logistic regression; Streamlit Cloud; Support vector machine (SVM).*

1. Introduction
The overall machine learning-based solution developed to analyse the range of medical conditions and predict the possibility of the illness in a patient is called the Multiple Disease Prediction System. Using machine learning, it's assumed that we will be able to enhance diagnostic precision and therapeutic interventions accordingly. This system makes use of predictive modelling to estimate the probabilities of various diseases, thus improving health outcomes. From analysing huge volumes of medical data and applying sophisticated machine learning algorithms, we provide timely, accurate predictions that help healthcare providers and support improvements in health outcomes globally.

1.1 Description
Traditional medical systems can monitor only one disease at any one time. For instance, different analyses, such as diabetes, diabetic retinopathy, or heart disease, require different systems. Most of the existing platforms are specialized in single-disease analysis; during comprehensive patient health data analysis, multiple tools have to be used. Traditional systems are capable of finding only specific diseases one by one; it can also require users to operate each system separately. In multi-disease prediction, though, a user can detect the potential presence of several illnesses in one place. Our system allows users to select any disease they would like, enter parameters related to it, and view immediate predictions. This makes the process more integrated in the detection of diseases and hence easier and efficient.

1.2 Problem Statement
Most of the current applications of machine learning in healthcare relate to single-disease analysis, where conditions such as liver disease, cancer, or lung disease are treated independently. This kind of fragmented setup contributes to difficulties when users want to make predictions across multiple diseases; they usually have to navigate between

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 03 March 2025
Page No: 448-455
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0062

various platforms. Furthermore, some models have treaty accuracy, which can be harmful to patients in certain cases. The different models make it expensive and time-consuming for healthcare organizations to deploy several models to get full-fledged analysis. Systems based on minimal parameters carry the risk of producing results that are less accurate. [1-5]

1.3 Proposed System

We propose a system aimed at solving multi-disease prediction in an integrated manner. This, therefore, implies that this system will allow the prediction of several diseases on one platform. The users will have easy access to results regarding different conditions with the advantage of a model that integrates a great set of parameters for better accuracy and reliability. This negates the need for several platforms and hence streamlines the prediction process, enhancing healthcare literally. We will perform the multi-disease analysis using machine learning algorithms and further developing the interface by using Streamlit, which provides options to users for the selection of diseases, input of parameters required, and display of predictive results. This research advances healthcare analytics—one of the more holistic, transformative approaches toward disease prediction and bettering patient care across the globe.

2. Background

This is the area where most recent healthcare advancement has been influenced by machine learning techniques. The high volume of health data intertwined with increased computing power has converted machine learning into an applicable method for the prediction and diagnosis of various diseases. Some of the major advantages of multi-disease predictions with the use of machine learning are manifold. First, health professionals can target single individuals with a high risk for multiple conditions by providing timely and effective interventions and preventive policies. Secondly, it assists in the appropriate resource allocation in health services by prioritizing people at high risk so that timely interventions may be facilitated. Moreover, the algorithms developed for machine learning may reveal patterns of diseases and risk factors, which can contribute to formulating specific programs of public health. However, some of the challenges persist in machine learning-based multi-disease prediction, such as health data availability and quality, patient privacy and data security issues, and improvement in interpretability of the predictive model. Integrating machine learning algorithms into the existing healthcare system also poses challenges in terms of ensuring regulatory frameworks are adequate, ethical standards are satisfied, and compatibility with the established workflows of healthcare. In a nutshell, machine learning has immense potential to bring transformational changes in health care through multi-disease prediction. These algorithms can enable health care providers to identify people at risk much earlier, improve diagnosis, and optimize the efficacy of treatment. However, data quality, privacy, interpretability, and regulatory compliance are the major challenges that need to be addressed for successful implementation of machine learning-based predictive models in health care settings.

3. Tools Used

- **Kaggle:** It provides access to a large number of varied datasets, which are essential for the training and testing of many machine learning models.
- **Google Colab:** is a web-based environment for data analysis and machine learning. And Colab supports both collaborative development and model training.
- **Anaconda:** A distribution platform to smoothen package management and deployment, hence giving a very efficient environment for Python data science workflows.
- **Visual Studio Code:** VSCode: An open-source code editor. With extensions and tools for data science, it extends the ability for coders to create code, debug, and organize code.
- **Streamlit Cloud:** A platform where users can deploy, manage, and share applications directly from the Streamlit framework, enabling users to compatibility with the enter
- interact with machine learning models straightforwardly.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 03 March 2025
Page No: 448-455
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0062

4. Technologies Used

- **Python** is a dynamically typed, higher-order language that is used quite frequently in the domain of Data Science and Machine learning.
- **NumPy:** A package for numerical computation in Python; adds support for large, multi-dimensional arrays and matrices.
- Pandas: The Python library is used for manipulating and analysing data. It is very useful for working with structured data.
- **Scikit-Learn, sklearn:** A Python machine learning library that provides an array of tools used for classification, regression, clustering, and evaluating a model.
- **Machine Learning Algorithms:** These algorithms conduct supervised learning of models from labeled data for making effective predictions through learning of patterns.
- **Pickle:** A Python module for serializing and deserializing Python object structures. Saving and loading a trained model highly relies on it.
- **Streamlit:** An open-source framework that focuses on the interactivity of machine learning applications by making them visually appealing through high-order web applications, enabling real-time interaction with the user.

5. System Model

It involves identifying multi-diseases by using the Supervised Machine Learning Model. Supervised learning is a type of machine learning where a model is loaded on labelled data, with each input related to the training data being well known for its label or result. That gives insight into understanding the relationship between two types of inputs and outputs. The whole idea behind this approach relies on input-output pairs: it learns from the training data a function that maps inputs to outputs. Later on, the function learned from training data can be used for other, previously unseen data and allow the system to make predictions. Supervised learning works well when it comes to performance in terms of task-specific applications and also works well with labeled datasets. In this work, some supervised learning models were applied to forecast the possibility of specific diseases based on symptoms that a user may enter. For each disease, selection of different machine learning models was done because we wanted to ensure the best performance and accuracy of the model. Each of the various models predicting diseases has a base in a number of machine learning algorithms, the selection being done by the capability to explain the peculiar patterns and relations between symptoms and the presence of a disease. Logistic Regression, Support Vector Machines, and K-Nearest Neighbours are selected, since these are typical algorithms when it comes to data analyses and the delivery of precise forecasts. We will focus on the selection of different machine learning algorithms for different diseases with the intention of enabling reliable and practical prediction of diseases. This project will facilitate the early diagnosis, timely invention, and improved health outcomes by healthcare professionals and users. [6-10]

6. Experiment

6.1 Generation of Hypothesis

The hypothesis of the Multi-Disease Prediction System is that using general medical data, which is analysed by sophisticated machine learning algorithms, will make highly accurate predictions of the emergence of individuals in terms of a variety of diseases possible. It is envisioned to base the work on the assumption that there are perceptible patterns and relations within medical data, which can be used to build a robust predictive model.

6.2 Data Collection

We begin the project by gathering different types of information. A majority comes from Kaggle for these datasets. In fact, datasets are an essential thing to have for practice and research, and also to work as the base of different machine learning models. These curated datasets provide a diverse range of information, allowing for accurate training and validation of our prediction system.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 03 March 2025
Page No: 448-455
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0062

6.3 Data Pre-Processing

Data collected from diverse sources aroused a lot of irrelevant elements, such as noise, redundancy, several missing values, etc. Preprocessing clears the dataset of useless data that is irrelevant, deals with the missing values, and handles some outliers. This step ensures that only the highest-quality data gathered flows into the process and increases model accuracy and the reliability of predictions.

6.4 Feature Selection

Feature selection is an important process wherein we try to determine the most relevant and informative features of the dataset. When there is a large volume, filtering out key features raises the efficiency and performance of machine learning models trained in the data. The importance of each feature was judged using the statistical measures and correlation analysis techniques. This process helps us in focusing on those factors that would bring most impact and ensures that our predictions are truly based on meaningful variables. The selection of just essential features makes the system of multiple disease predictions much more accurate and energetic, hence bringing good health care outcomes and useful insights for medical persons.

6.5 Model Building

The algorithms being used in our multiple disease prediction system are supervised machine learning algorithms: Logistic Regression, SVM, and K-Nearest Neighbours. These algorithms have been chosen for their utility in the performance of classification tasks and the capability to handle multi-class predictions. Based on these different supervised learning algorithms, a strong and accurate model will be built to perform the prediction for multiple diseases. Each of them contributes differently to the field, which enables us to consider various approaches and to select an appropriate model to suit each particular type of disease. We have built models to come up with consistent and timely predictions.

6.6 Deployment

A multi-disease predictor system was deployed to Streamlit Cloud so that the users would have a web-based interface with ease of access and interaction. Streamlit Cloud allows the user to input certain disease parameters and returns accurate predictions for multiple diseases. Furthermore, this deployment serves ease of access and scalability, enabling healthcare professionals and people to make informed decisions about health with much ease.

7. Design

7.1 Architecture Design

Our system architecture is specifically designed to predict various diseases, namely Heart Disease, Diabetes, Lung Cancer, Parkinson's, and Thyroid diseases, as those conditions have been selected upon considering their high prevalence rate and the inter-related risk factors. The process starts with data extraction, as shown in Figure 6.1. Datasets for these diseases were sourced from Kaggle for a diverse and reliable base to train and test the models. We import datasets, then explore and pre-process the data by handling the missing values. The data was then prepared for analysis. We split the dataset into two subsets: one for training and the other for testing. Various machine learning algorithms were applied to the training data, which were validated on the testing data. We experimented with different classification algorithms and chose in the following Logistic Regression, SVM, and KNN for their accuracy at each disease. For every disease, a different predictive model was made and saved as a pickle file. Now these 'pickle' files—the pre-trained models of various diseases—were combined into one web application using the flask framework. This allows having a smooth interface where the user can input parameters and get predictions of multiple diseases at once. Figure 1 shows Predictions of Diabetes

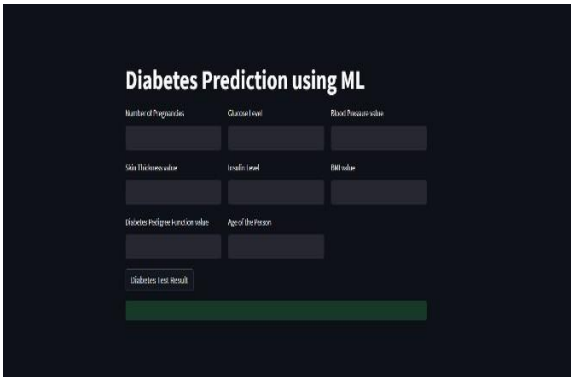

Figure 1 Predictions of Diabetes

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 03 March 2025
Page No: 448-455
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0062

7.2 Architectural Design Interface

The architectural design interface is organized in such a way that the interaction between users and the prediction models is smooth and effective. We used Streamlit, where a user selected the disease for which they wanted the model to predict; after filling in the necessary health parameters, they get the predictions of disease likelihood. The interface leverages the collective functionalities of saved models toward an appropriate and timely response to users' queries on health. Such accessible design further increases usability by both healthcare professionals and individuals using the predictive capability of the system. Figure 2 shows System Flow Process .

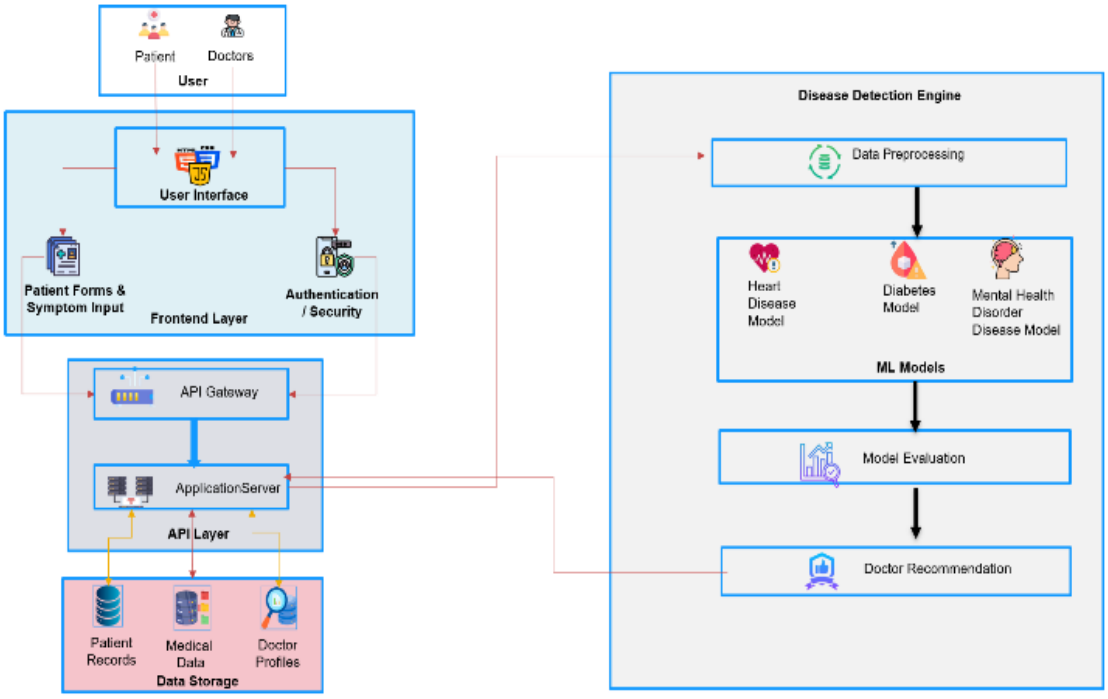

Figure 2 System Flow Process

8. Machine Learning Algorithm

8.1 Logistic Regression Algorithm

Logistic Regression is a statistical model that can be used for binary classification. It is based on a logistic-sigmoid function, which can be defined mathematically as follows:

$$F(x) = \frac{1}{1+e^{-x}} = \frac{e^{x}}{e^{x}+1}$$

Output values in the range (-∞, ∞) are taken and squeezed within the range of (0, 1). The method Logistic Regression starts with a linear equation but then uses the log-odds to transform such an equation. Further, these log-odds are moved through the sigmoid function, which returns a probability between 0 and 1. This probability may be used for classification to tell how likely a data point belongs to a particular class depending on a set boundary. Although logistic regression is mostly executed in binary classification, it is applicable in the case of multi-class classification problems, using methods such as one-vs-rest and multinomial logistic regression. Its simplicity and interpretability make it a widely used model in various tasks of classification. performance of machine learning models trained in Figure 3 shows X-axis and Y-axis, Figure 4 shows Machine Algorithm

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 03 March 2025
Page No: 448-455
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0062

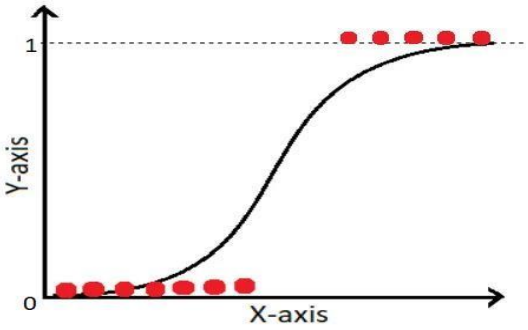

Figure 3 X-axis and Y-axis

8.2 Support Vector Machine Algorithm

The SVM algorithm is a type of supervised machine learning mainly used in binary classification problems, though it is capable of dealing with multiclass classification and regression problems. Support Vector Machines find the hyperplane that separates data best into different classes. Let's represent any observation i from n data points each with p features in a p-dimensional space as xi while its class label yi would be either +1 or -1. Given two classes, SVM tries to place a hyperplane in this feature space such that all instances of one class are on one side and the others are on the other side of the hyperplane. The hyperplane is defined in p dimensions, but intrinsically has dimensionality of p-1; for instance, in 2D space, the hyperplane is a line. SVM is particularly effective for high-dimensional data, and by maximizing the margin the distance between the hyperplane and the nearest data points that support it, hence the name support vectors it gives a robust classification. This makes the SVM a mighty tool for classification tasks where there is a clear margin of separation between classes.

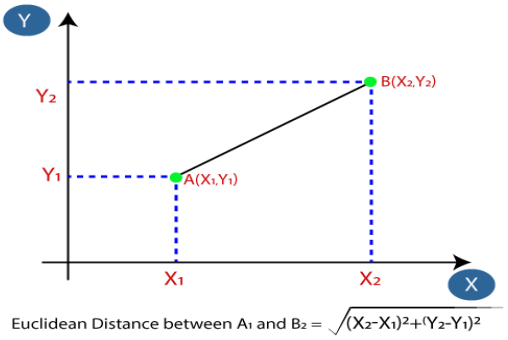

Figure 4 Machine Algorithm

vector p is represented by w and real number is represented by b. For simplicity, we assume -w=1 making the expression x*-w+-bx Shows the distance of point x to plane.

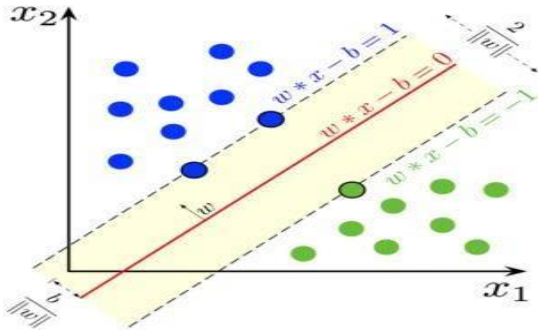

Figure 5 Distance of Point X to Plane

So we can write our class as y = +1/-1 and the general plane's class division requirement would be:

$y_i\,(xi \cdot w' + b') \geq 0$

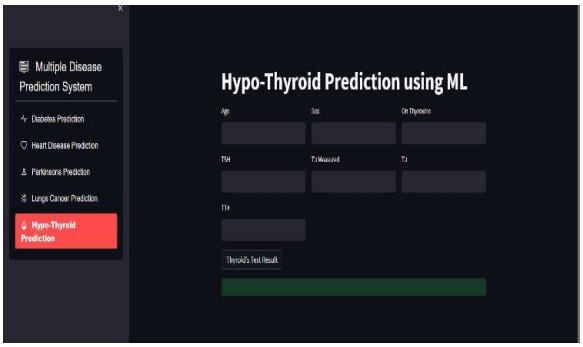

Figure 6 Predictions of Hypo-Thyroid

8.3 K-nearest neighbours (K-NN) algorithm

K-NN is an easy-to-implement yet powerful supervised machine learning model used for classification. The algorithm works based on the principle that similar data points will likely belong to the same class. Implementation of K-NN proceeds as follows:

- Choose the Number K of Neighbours: Select the value of K—the number of nearest neighbours to consider. For this project, we have chosen K=5.
- Calculate Euclidean Distance: Calculate the Euclidean distance between the new data point and all other points in the dataset.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 03 March 2025
Page No: 448-455
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0062

- Identify the K Nearest Neighbours: Sort these distances and identify the K closest data points to the new data point.
- Count the Number in Each Category: Count how many data points belong to each category among these K nearest neighbours.
- Assign New Data Point to a Category: Assign the new data point to the class with the highest number of neighbours.

The K-NN model is now prepared to classify each new data point based on the majority category among its K most similar neighbours. For instance, if three of the nearest neighbours belong to Category A and two to Category B, the data point will be classified into Category A, using a majority voting system to provide accurate classification results.

9. Result

In our multiple disease prediction system:

- Diabetes and Parkinson's Disease: Prediction models for these diseases use the Support Vector Machine (SVM) algorithm, known for its high precision in classification tasks.
- Heart Disease and Lung Cancer: Predictions for these diseases utilize the Logistic Regression algorithm, well-suited for binary classification tasks.
- Thyroid Disease: The K-NN algorithm is employed for predicting thyroid disease, selected for its accuracy in managing complex categorical data.

Upon entering specific parameters related to a disease, the system provides a prediction indicating the likelihood of a patient having that disease. The system also helps guide the user by displaying the expected range for each parameter. If a value is invalid, out of range, or left empty, a warning prompts the user to input valid data. By employing tailored algorithms for each disease and establishing clear data entry criteria, the system enhances accuracy and reliability in prediction. This approach allows users to access timely, accurate health predictions, supporting early diagnosis and proactive healthcare management.

Conclusion

The main aim of the project was to build a system which can predict multiple diseases, consolidating several prediction models onto an easily accessible platform. It saves time against visiting different websites for each type of prediction and allows quick, essential health insights. Early predictions for various diseases can help improve life expectancy and help people avoid significant financial burdens by enabling preventive healthcare. We implemented multiple machine learning algorithms, such as Logistic Regression, Support Vector Machines, and K-Nearest Neighbours, each selected for specific disease contexts where those algorithms perform optimally. Our system makes these disease predictions reliable, supporting proactive health measures. Early diagnosis, combined with timely intervention, plays a crucial role in managing and treating many conditions, making this tool a valuable contributor to improved healthcare outcomes. In summary, our project represents meaningful progress in healthcare by utilizing machine learning for accurate, multi-disease predictions. The system's efficiency and accuracy enhance healthcare accessibility, promoting better patient outcomes and potentially saving lives.

Future Scope

There are several directions that could continue to enhance the effectiveness and efficiency of our multiple disease prediction system:

- **Addition of More Diseases:** Future iterations could integrate additional diseases onto the platform enabling the prediction and longer range of conditions to add relevance and utility in the healthcare market.
- **Improvements in Prediction Accuracy:** Continuous research activities could improve the accuracy of predicting diseases. This may be achieved through the improvement of certain machine-learning algorithms, evaluation of possible features for prediction, and increasing variables with more enhanced parameters in training datasets leading to a reduction in false predictions and improvement in model precision in less time leading to lower mortality rates through timely interventions.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 03 March 2025
Page No: 448-455
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0062

- **Integration with the Electronic Health Records:** Blackwood Deed-system integrated with EHR would provide a total spectrum of personalized health care experiences. Employing the patient's electronic hospital records could increase prediction accuracy and allow health professionals to base patient decisions on the medical history.
- **Mobile Application Development:** An optimised mobile application can make it easier for users to use the platform from their smartphones. The possibility of predicting diseases in a real-time environment can empower the patients by allowing them to take proactive actions with respect to health 24 * 7.

Each of these future pathways promises not only to enhance disease prediction and healthcare outcome efficiency capacities, but it also will have a great economic and health impact on human welfare.

References

[1]. Gopisetti, L D Kumera, SKL Pattamseti, S R Kunam, Kodali. Multi disease prediction using machine learning and streamlit. In: 4th Conference on Inventive Technology (CIT), pp. 924-831. IEEE.

[2]. Keniya, Khakharia, A. Shah, Gada, Manjalkar, R., Thaker, (2020). Disease detection from various symptoms using ML. Available at: SSRN 3231426.

[3]. Srivastava, Haroon, M. & Bajaj, A. (2013, September). Web document extraction using class attribute method. In: 3rd International Conference on Computer and Communication (ICCC), pp. 15-22. IEEE.

[4]. Raja, M., Reddy, C. P. & Sirisala, (2020, January). Machine learning based Multi disease prediction system. In 2021 International Conference on Computer Communication (ICC), pp. 14. IEEE.

[5]. Khan, W. & Haroon, M. (2021). An supervised machine learning model for anomaly detection in static attributed social networks. 1219-1353. IEEE

[6]. Vijayalaxmi, Sridevi, Sridhar, S. (2020, June). Multi-disease detection with AI from core health parameters measured through non-invasive technique. 4th Conference on Computing (ICIC), pp. 2345-1398. IEEE.

[7]. Sudha S. Kamali Priya S, & Prathiksha P. N. (2023). Multi-Disease Prediction and Classifier: A Comprehensive Approach for disease Decision. Journal of Healthcare Engineering.

[8]. Siddiqui, M. (2023). Research on significant factors affecting ML technology for applications based on integrated Deep Learning, 118, 105699.

[9]. Tripathi, Haroon, M., Khan, Z. & Husain, sM. S. (2022). Security in digital.

[10]. Sangjin Jeong; Chan-Hyun Youn; Moonjung Kim; Limei Peng: An Integrated Healthcare System for Personalized Chronic Disease Care in Home Hospital Environments.