National Research Journal of Information Technology & Information Science
Volume No: 13, (January) Year: 2026 (Special Issue)
PP: 809-821
ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

SMART HEALTHCARE BASED DISEASE DIAGNOSIS APPLICATION

Vishesh Kaushik

Dept. of CSE, UIE Chandigarh University Mohali, India

Azhar Ashraf Gadoo

Dept. of CSE, UIE Chandigarh University Mohali, India

Ayush Jyoti

Dept. of CSE, UIE Chandigarh University Mohali, India

Vansh Bansal

Dept. of CSE, UIE Chandigarh University Mohali, India

ABSTRACT—

Artificial Intelligence has been one of the most vital innovations of the last decade. It has revolutionised our problem-solving ability by using powerful algorithms and different methods to achieve what was never done before. One field where artificial intelligence has provided us with major breakthroughs is healthcare. Primarily, the healthcare system is very labour-intensive requiring doctors and nurses to work hard hours. Using artificial intelligence in disease diagnosis is one such use case where doctors can be helped and their burden lessened. Disease diagnosis is the most critical step in providing any patient with the care and medication they require. Sometimes diseases are not detected early and patients' health is compromised. Using tools like computer vision and algorithms like Random Forest, Logistic Regression can be used to effectively predict disease based on the patient's symptoms. In this paper we develop and web application solution which will have basic healthcare services like booking appointments with doctors, having one-to-one conversations with your healthcare professional and so on. This web app will also possess a feature of disease detection using machine learning algorithms. Logistic regression and the Random Forest algorithm have been used to implement the solution but the Support Vector Machine algorithm has also been used to compare results just by the developers. The paper starts with an introduction to the concept and provides an overview of the system implemented. The next section provides a comprehensive literature summary of the previous work done in the field and how it began to start with. The subsequent section deals with the methods and insight into the technologies used to develop the web app and the fine details of machine learning algorithms. Lastly, results from the prediction of the implemented system are discussed and then future works to be taken up are discussed.

Keywords—Disease detection, Machine learning, logistic regression, healthcare

I. INTRODUCTION

Being a healthcare professional is one of the most stressful jobs. Healthcare professionals face considerable pressure in every decision they make throughout the day. This constant cycle of decision-making can fatigue anyone. 39% of global healthcare professionals report mental burnout while doing their jobs[1]. A major part of this decision-making is disease diagnosis which is the cornerstone of any medical treatment that follows afterwards. Diagnosis is the most critical step in providing patients with the best treatment as early as possible. Early diagnosis of mortal disease can significantly reduce the chances of death. Individuals who are diagnosed in the early stages of breast cancer have a 5-year survival rate of 90% compared to 27% for those where it is detected at later stages [2]. Cardiovascular disease if detected in its infancy can reduce the chances of heart attacks by 30% [3]. Machine learning requires data to work on and is only as good as the data they are trained upon. In the medical ecosystem, there is an abundance of data available in the forms of medical reports, imaging data such as CT scans, X-ray scans and other forms such as ECG, EEG etc. Although these data forms are of rich quality but still machine learning algorithms still require more data which can compensate for any shortcomings in the working of the system [4]. Combining different data pipelines can enhance the accuracy of the system and provide better results than just relying upon only a single stream of data, moreover using this multimodal data makes the system more robust to errors and outliers [5]. Artificial Intelligence has tremendous potential to provide accurate diagnostics, improve utilization of resources and provide timely interventions in clinical decisions. The system implemented will use two powerful machine learning algorithms to make predictive analysis for diagnostics. Firstly, Logistic Regression is used which is a capable statistical technique used for binary classification tasks. Provided the quality data set and enough features logistic regression has an 80% accuracy rate in predicting heart disease [6]. Logistic regression provides binary classification in its result as in the context of medical diagnostic disease or no disease due to this it is the ideal tool to use in medical diagnostics. Additionally, Logistics Regression models are adaptable to new data and features. On the other hand,

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

Random Forest is an ensemble learning method which works by making layers of decision trees during its processing and outputs a class from individual trees. It is very robust to variance and overfitting making it a go-to technique where the relationship between variables is not linear. Random Forest also has a higher level of accuracy for heart disease with a diagnosis rate of 93.3% [7]. Aside from using machine learning algorithms for predictive disease diagnosis the web application also will have other features. Booking appointments with doctors will be made easy by the application and any output given by diagnosis will be shared with the patient's profile to the doctor. The web application will also have a video call feature for live doctor consultations. Moreover, the system will have a page for general information about various kinds of diseases and preventive measures. Patients will also be able to chat with doctors through the messaging capability of the web application. The overall goal of the system is to provide the healthcare professional with a helping hand in their noble endeavour to save as many lives as possible.

II. LITERATURE REVIEW

The integration of machine learning into clinical diagnosis is aimed at enhancing the accuracy and efficiency of the whole process. An abundance of research and exploration has taken place in recent times. This literature review synthesises such recent exploration and their methods of experimentation to achieve more accuracy in disease diagnostics.

Support Vector Machines and Random Forest algorithms can be used to diagnose liver disease with high accuracy based on patient demographics and biochemical signatures of patients as features for model training[8]. Similar research done by N N Anirudh et al. found classification techniques like Gaussian Naïve Bayes in the prediction of general disease which also uses patient history and symptoms for diagnosis[9]. Harsha Bhute et al. showed that Gradient boost and KNN algorithms can be used for these purposes with KNN achieving accuracies of 89.1% for diagnosis of heart disease and 77.9% for diabetes [10]. The application of machine learning in diagnosis is also used in oncology. Shimuzu & Nakayama (2020) experimented and found out that deep-learning techniques are exceptionally good for detecting cancer in patients. The ability of deep learning technology to work with complex datasets and extract valuable information from them makes them well suited for the task[11].

Dimension reduction of the data was done and then inputted into models such as Decision Trees, KNN and PCA by Ritesh Jha et al. to predict thyroid. The reduced dimension helped them have an accuracy rate of 95% making it excellent with comparable systems[12]. Rodrigo Ibarra et al. in their work on detecting cardiovascular disease implemented a system using the Random Forest algorithm. The research boasts of a huge dataset of 320,000 records, providing it with plenty of training. The system resulted in 94% accuracy and 81% area under the receiver operating characteristic curve[13]. In the field of Neurology Support Vector Machines and Random forest algorithms were implemented by V Lokesh Raju et al. to detect Parkinson's disease. Parkinson's disease is a neurological disorder which shows its symptoms very late in progression and thus is very difficult to predict in its stages. The system used voice-based bio-features from patients. It resulted in earlier detection of disease than usual[14]. Nandimandlam et al. used a convolutional neural network to classify different lesions on the skin. To predict skin disease images of the affected area can be used along with data from medical reports of patients. Using the HAM10000 dataset they were enable to enhance the capability of the VGG16 model to classify skin disease [15].

Michele Bernardini et al. used real-time records from IOT devices to diagnose an eye disease called Diabetic Retinopathy a side effect of diabetes. They used a unique way to collect data into the machine learning model. They achieved 72.43% area under the curve with this work [16]. Alzheimer's disease is an irreversible neurological condition and no treatment is available but it can be slowed down if detected early enough. Helaly et al. used the Convolutional Neural Network and VGG19 to classify data and attained an accuracy of 97% with VGG19 performing better [17]. Researchers have used the random forest algorithm in a hybrid manner to diagnose cardiovascular disease with an accuracy of 88.7% [18]. P. Krishnamoorthy et al. use Convolutional Neural Networks to automate disease diagnosis. They used biomedical images to implement advanced pattern recognition to increase the specificity and accuracy of diagnosis. They also implemented a web interface for the patients to use[19]. Prof. S.B. Pagrut et al implemented a variety of machine- learning techniques to compare the results of each. The algorithms implemented were XGBoost, Random Forest, and KNN. In the study, they achieved an accuracy of 95%. Another finding was that the XGBoost Random Classifier was more accurate than the naïve Bayes classifier. Features used were the patient's medical levels such as blood pressure, cholesterol and sugar level [20]. Vishal Prasad et al. addressed model bias and high implementation costs of machine learning algorithms do disease diagnostics specifically for chronic disease such as cancer [21]. Sushruta Mishra et al. conducted a study to find the effects of large neural networks in enhancing the accuracy of multiple disease detection systems. They found that large neural networks outperform the traditional methods in the realm of disease detection [22]. Rama Al-Momani et al. use varied machine- learning techniques to detect kidney diseases like Neural Networks, Support Vector Machines and K- Nearest Neighbors. They used a dataset of 400 samples and the comprised of 13 critical features. Artificial Neural Networks achieved the highest accuracy of 99.2% in kidney disease detection [23].

Machine learning can also have applications in the field of dermatology. Various skin images were used to detect

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

skin cancer. Two methods of artificial intelligence were used on these images namely ensemble learning and Deep Learning to get an accuracy rate of 90.1% [24]. Another way of detecting disease is using MRI imaging as the data set. Rishab Ranjal et al. use Decision trees to identify heart disease and brain tumours with a highly effective rate of accuracy of 97.7%. Researchers used MRI imaging for the model training the algorithms [25]. Paul Dhiman et al. used federated learning to diagnose disease and they found out that the model performs better with data augmentation and preprocessing datasets into different layer sizes. The federated learning also protects privacy. They also used methods such as parameter sharing [26]. Abiona Akeem Adekunle experimented with Support Vector Machines to detect Parkin's disease using the dataset comprised of voice recordings of patients, This in itself is a novel approach to detecting disease and also an effective one at that. Similar to Paul Dhiman's study they found that preprocessing data using methods such as cleaning, transformation, and feature engineering results in higher accuracy. This study attained 88.46% accuracy. They proposed this as privacy ensuring and cost-effective method of diagnosis as it uses only sound recordings of patients[27]. Ritu Aggarwal and Suneet Kumar used Multi multi-layered perceptron and chi-square methods to effectively detect heart disease by analyzing patients' medical traits like high-density lipoprotein and low-density lipoprotein levels. Together the two methods achieve high accuracy of 98.08% and 98.25% respectively [28].

Apart from diagnosing human disease, all these techniques can also be used to detect disease in plants and crops. Lots of studies and experimentation have already been done on this subject as it can be very cost-effective and certainly protect farmers and other stakeholders from large losses. Anamika Jain et al. implemented Convolutional Neural Networks to classify images of plant and crop leaves as healthy or infected. The methods require digital image processing and machine learning techniques to accurately predict disease in crops [29]. Food crops need constant inspection to check their health to see if they are not being infected by any disease. G Jayanthi et al. use machine learning techniques to inspect tomato crops for potential diseases. They made the system to be able to work all the time so they used IOT devices such as Rasberry Pi in the fields providing machine learning model constant data to keep checking [30].

The immense potential of machine learning is being identified and the growing number of research and studies on this topic substantiate the statement. In upcoming years more better tools are to be expected to be built for disease diagnosis with higher accuracy.

III. TECHNOLOGIES USED

We have used many of the popular available frameworks and libraries to build an efficient web application that will facilitate machine-learning disease diagnosis and enable communication between patient and doctor. All such frameworks and libraries are discussed below.

A. Scikit -Learn

Scikit–Learn is an open-source Python library for building machine learning models and one of the most used ones. Scikit-Learn provides both supervised and unsupervised machine-learning algorithm implementation. In supervised learning commonly used algorithms have linear regression, logistic regression, support vector machines, random forest and other such models. K-means, Principal Component Analysis, Factor Analysis etc are part of unsupervised learning [31]. It also provides developers with data preprocessing methods and tools to prepare data for algorithms like Standardization, Normalization, and Missing value filler with various statistical attributes like mean, mode and median. For evaluation purposes, it has different functions to measure the accuracy of the model such as precision, recall F1 score, Mean Squared Error etc [32].

B. React.js and Node.js

React.js is a popular JavaScript library used for building dynamic and interactive frontend for web applications. React has the concept of components which are regarded as building blocks of the application and each component renders a part of the user interface. The advantage of these components is that they can be reused accordingly reducing the development time significantly. React.js has the states which are essential to keep the web application dynamic and responsive. Including these features also allows easy Third-Party integration with a vast number of libraries and APIs such as WebRTC required in this application for making video calls [33].

Node.js is a JavaScript runtime which is used to build the backend of web applications. It is open open-source event-driven runtime used for server-side programming. Node.js has a non-blocking I/O model which enables it to handle multiple operations concurrently. It boasts of Node Package Manger which is the world's largest software registry allowing developers to easily extend their application with features of database connectivity, file handling services and many more [34].

C. Flask

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

Flask is a framework for Python made to build web applications and APIs with less development time. It is lightweight and follows a micro-framework making it an ideal choice for writing REST APIs. It is used to create REST APIs to connect frontends written in React.js with the machine learning model written in Python [35].

D. WebRTC and MySQL

Web Real-Time Communication(Web-RTC) is an open- source program which provides web applications with the feature of real-time communication. It enables direct person-to-person communication between two browsers without the need for an intermediate server to handle the relaying of information through. It encompasses Secure Real- Time Transport Protocol (SRTP) which encrypts communication ensuring the security and privacy of the users. It also has cross-browser support meaning people using different browser platforms can still use the service easily [36].

MySQL is an open-source relational database management system(RDBMS) operating on structured query language(SQL) used for creating and managing with large databases. It is a preferred choice for data string and retrieving purposes in most web applications due to its ease of use and speed. MySQL adheres to ACID properties for the database making it reliable and ensuring the integrity of the data stored [37].

IV. METHODOLOGY

To implement the machine learning algorithms we have to go through several phases from data acquisition to integrating the model with the web application. The model needs to be trained and evaluated before it can be used, for which different evaluation metrics are used to measure the correctness of the diagnosis. After the machine learning model is integrated with the web application users can input their data and the results will be shared with them and doctors helping them with easy diagnosis. Below each phase of machine learning model implementation is discussed and later in this section, we elaborate a bit on the features of web applications.

A. Data Collection

Data is the first and foremost requirement of any artificial intelligence system. This phase is also most important as it is the base for the whole system as everything else will be built on this. Quality of data matters because noisy data can result in underfitting or overfitting and provide us with distorted inaccurate results. For this system, we have used a dataset available on the internet for five different diseases. Each of these datasets is taken from a quality source like the University of California, Irvine Machine Learning Repository ensuring the model is well generalized and accurate to diagnose disease. Table I Shows each dataset, its source, the number of records present in the dataset and the disease it addresses.

TABLE I - DATASET USED AND SOURCES

Dataset Name	**Source**	**No.of Records**	**Disease of Dataset**
PIMA Indian Diabetes	UCI	768	Diabetes
Heart Disease	UCI	303	Heart Disease
Breast Cancer Wisconsin	UCI	569	Breast Cancer

B. Data Preprocessing

After data is collected next we have to process that data to make it suitable to be feed into our model for predictions. Data preprocessing involves a variety of methods, some common for all types of algorithms some of which are done to cater to specific algorithms being implemented. Below are are methods used to preprocess the data for this system.

1) *Handling Missing Data* - Missing values are common in the dataset but handling these missing values is important to avoid having biases in the model. In this system, we imputed the mean of all the values into missing records ensuring the generalization of data [38]. Another way to remedy the missing values is to fill them with median or mode. The advanced method is to use K-Nearest Neighbors to fill in missing values where the relationship

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

between features is more complex.

2) *Feature Scaling* – Feature Scaling is the process of scaling the independent features in a standard range usually between 0 and 1 or -1 and 1. Feature scaling is important to ensure that each feature contributes to the model's prediction equally and any single feature doesn't dominate the model's features at their threshold values which are not affected by the magnitude of features.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad (1)$$

X = feature to be sampled

$X' = Sampled\ feature$

X_{min} = Minimum value in the feature

X_{max} = Maximum value in the feature

3) *Encoding Categorical Variable* - Here we encode our categorical data such as gender, smoking status, chest pain or not etc. Encoding is necessary for logistic regression because it cannot work directly with categorical variables. One-Hot Encoding technique was used to encode categorical variables in this study. In this, we encode the feature into binary form(0 or 1) [39]. As in the case of scaling encoding is not required for Random Forest.

4) *Data Splitting* – It is the process of dividing the available into two parts i.e. Training Set and Testing Set. As the name suggests the training set is used to train the model to diagnosis and the test set is then used to check the accuracy of the system. Dividing the data depends on the volume of data available, the split is 80% training and 20% test data but other ways are also there like cross-validation. We divide the typical split in 80-20 for this system [40].

5) *Feature Selection* – Feature selection is the process of selecting features from the pool of available ones that are more closely related to output. This process helps to reduce redundancy in the data. It helps to increase the accuracy of the model as it generalizes the model and reduces overfitting. This process requires knowing how each available feature is relevant to the outcome [41]. There are several methods to select features such as filter methods, wrapper methods, embedded methods etc. We have used correlation matrix to measure how features are correlated with outcome variable in this system.

C. MODEL IMPLEMENTATION

1) Logistic Regression

Logistic Regression is a supervised learning algorithm which is used for binary classification work, where we want to predict the outcome in two possible states only as in this case disease or no disease. Logistic Regression works on Sigmoid Function which transforms a linear equation made from features into the probability of outcome [42]. Equation (2) describes a sigmoid function.

working [39]. Equation (1) shows the formula used to scale the feature in the system which is Min-Max Scaling.

$$\sigma(z) = \frac{1}{1+e^{-z}} \qquad (2)$$

Here:

$Z = w_1x_1 + w_2x_2 + \ldots\ldots\ldots + w_nx_n + b$

$w_1, w_2, \ldots, w_n$ are Coefficient of input features

$x_1, x_2, \ldots\ldots, x_n$ are the feature values.

b=bias

The cost function of Logistic Regression is log loss. Logistic regression is a relatively simple to implement algorithm and it also demands less computational power compared to other algorithms. It provides a probabilistic output which is easy to understand even for someone not from a technical field. This algorithm is less susceptible to the problem of overfitting and

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

is usually well-generalized if the data input is of good quality [43].

2) Random Forest

Random Forest is an ensemble learning technique which is used for classification and regression works as well. The core concept of Random Forest is that it builds multiple decision trees during the training phase of model development and then output from all these trees is combined together to make a more accurate prediction. A decision tree can be imagined as a structure where each internal node or line represents an outcome of the problem and the nodes and the end or leaf nodes are the final output. Each of these decision trees is made using a random subset taken from data and features [44].

Random Forest usually works in four stages. Firstly, Bootstrap Sampling is done, also called bagging. In this stage, multiple random subsets are taken out from the original given subset. Given a dataset A with N number of data points, eq.

(3) shows the dataset A.

A={(x1,y1),(x2,y2),…….,(xn,yn)} (3)

Xi are features or input variables and yi are the output variable. Now randomly N number of samples will be selected from this original dataset. This creates a more generalised training set for each tree and controls the variance.

Next Random Forest will choose a random subset of variables from individual decision trees at each node. This ensures that trees are not closely correlated with each other and controls overfitting further down the road making the model more robust. Then decision trees are constructed by dividing the data set again and again based on some feature subset at each node. Each of these decision tree is trained on their bootstrap samples taken before. Each time the tree is divided the aim is to increase the accuracy or decrease variance of the model [45]. For classification work as such in this system, Gini impurity denoted by Gini(p) is calculated as per eq. (4).

$$Gini(p) = 1 - \sum_{k=1}^{K} p_k^2 \qquad (4)$$

Where K is the total number of outcome variables and pk is the proportion of samples of the kth class.

Lastly, all the results are combined to get the final output. Because this is a classification task output will be a class label and it will be calculated by eq (5):

$$\hat{y}_i = argmax\left(\sum_{b=1}^{B} 1\left(\hat{y}_i^b = k\right)\right) \qquad (5)$$

Tb is the bth tree and it has predicted an output. y◎○ↄ□ □For an input variable xi, and the class output by the forest in one which gets the highest number of votes. 1 is an indicator function which equals 1 only if y◎○ↄ□ □= k and it is 0 in other cases [46].

The Random Forest algorithm is not susceptible to overfitting problems because it combines results from different decision trees which take the output from a random subset of the original dataset. It also works well with high dimensional data and handles missing data without losing its accuracy. The preprocessing requirement for data is also very low as it does not require feature scaling or normalization before. Overall it is a very robust and accurate model for disease diagnosis tasks.

D) EVALUATION

After the model has been implemented and trained now we need to test it and evaluate the result to check how much correct it is if it suffers from any problems such as underfitting or overfitting. If it does we go back and change our approach with previous stages. There are several methods which can be used to evaluate the accuracy of the model. Below are the methods used for this system.

1) Confusion Matrix - It is a basic evaluation technique used for classification tasks. It is a very simple yet very powerful way to measure the accuracy of the model. It works on the table where predicted class or variables are compared with actual variable or class and then these numbers provide us different metrics to measure different rates of the model [47].

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

Table 2 depicts a basic confusion matrix structure.

TABLE II CONFUSION MATRIX STRUCTURE

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Metrics Calculated from Confusion Metrics:

- Accuracy - It is the value which tells us the correctness of the model, basically giving us how many of the predictions were correct both positive and negative with proportion to the total predictions made. It is calculated by.

$$Accuracy = \frac{TP+T}{TP+TN+FP+FN} \quad (6)$$

Precision – It is the ratio of the total correct positive prediction made to the total positive values. It is calculated by eq (7).

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

Recall- It is the measure of how many positive predictions were made in proportion to the total true. It is given by (8)

$$Recall = \frac{TP}{TP+F} \quad (8)$$

2. Reciever Operating Characteristics(ROC) Curve and Area Under Curve (AUC) - The ROC curve is a graphical representation of the model's correctness. It graphs the Recall(True Positive Rate) on the x-axis and the False Positive Rate on the y-axis. To calculate the overall correctness of the model we check the AUC value. An AUC value of 1 is a perfect model and an AUC value of 0.5 or under means the model is not good and needs improvement [48].

RESULTS

The system was implemented with the above-discussed technologies and algorithms. The web application resulted was able to establish video call communication between patient and doctor. Most importantly machine learning algorithms were fairly accurate for each. We can see that the Random Forest works better than the Logistic Regression in almost all the cases.

This can be attributed to the fact that Logistic Regression is a linear model and does not handle complex relationships between input and output variables very well. It is also noted that the Random Forest algorithm handles missing data and outliers better than the Logistic regression as it uses the concept of Bootstrap Samplin to split the dataset into multiple subsets. Other features of the web application also worked pretty well resulting in a comprehensive healthcare solution. Table III shows the results for various metrics for each disease.

TABLE III RESULTS OF VARIOUS METRICS FOR BOTH MODELS CORRESPONDING TO EACH DISEASE-PREDICTED

Dataset	**Model**	**Accuracy**	**Precision**	**Recall**	**ROC AUC**
PIMA Indian Diabetes	Logistic Regression	79.2%	76.5%	80.5%	0.81
	Random Forest	82.5%	81.2%	85.3%	0.86
Heart Disease	Logistic	83.5%	81.7%	84.0%	0.87

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

	Regression				
	Random Forest	88.1%	86.5%	89.3%	0.91
Breast Cancer	Logistic Regression	91.0%	90.1%	925%	0.94
	Random Forest	95.2%	94.7%	96.1%	0.97

Figure 1 is the diagnosis page where the patient inputs their data and the web application sends data to our model and it outputs the diagnostic result to the patient. Patient have to manually input their data and then the output will be visible.

For each disease, different parameters need to be considered to make the diagnosis. These parameters can be taken from medical test reports and entered into the system.

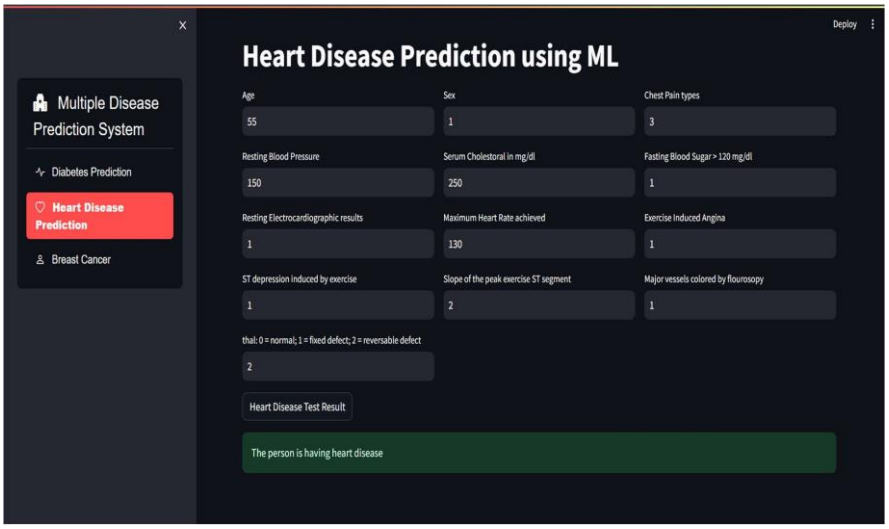

Fig 1. Heart Disease Diagnosis based on given parameters

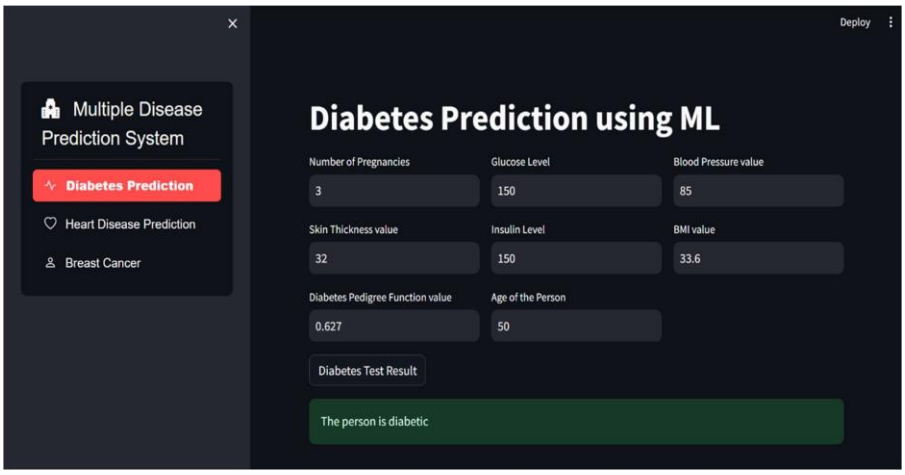

Fig. 2 – Diabetes disease diagnosis using different parameters

Figure 2 shows the Diabetes disease diagnosis by the system It can be seen that the parameters vary according to the disease. Figure 3 shows the web application appointment booking page where doctors can see all their appointments in one place. It has all the relevant details and a button to start the consultation when the time comes.

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

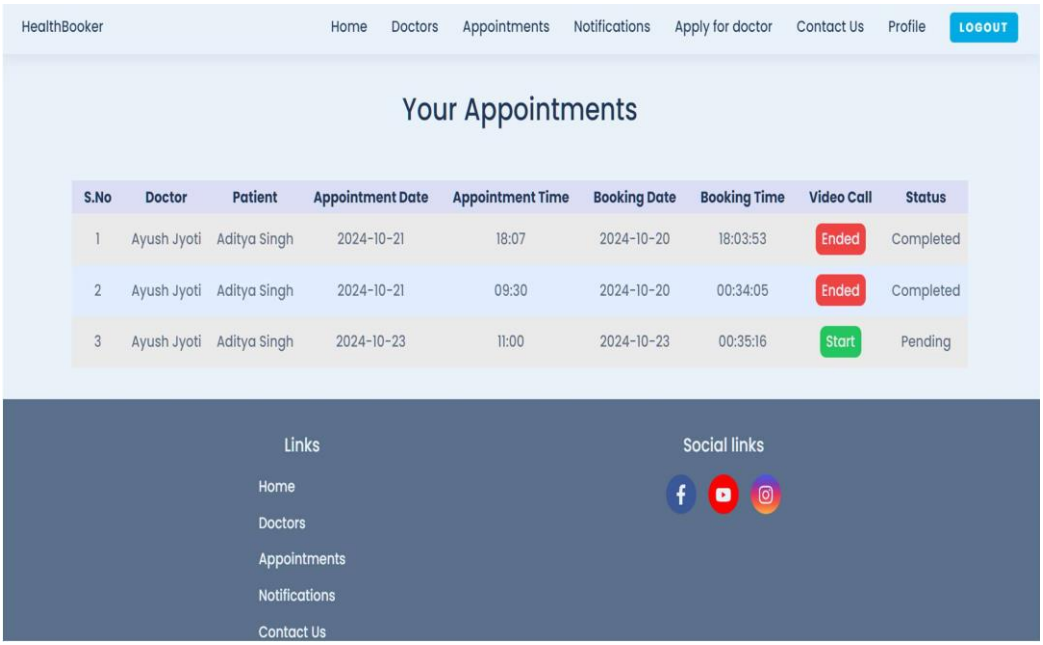

Fig. 3 Appointment booked for a doctor

patients who are unable to visit a doctor to have an easy way of consultation or in cases where it is not possible to visit as was in the case during the COVID-19

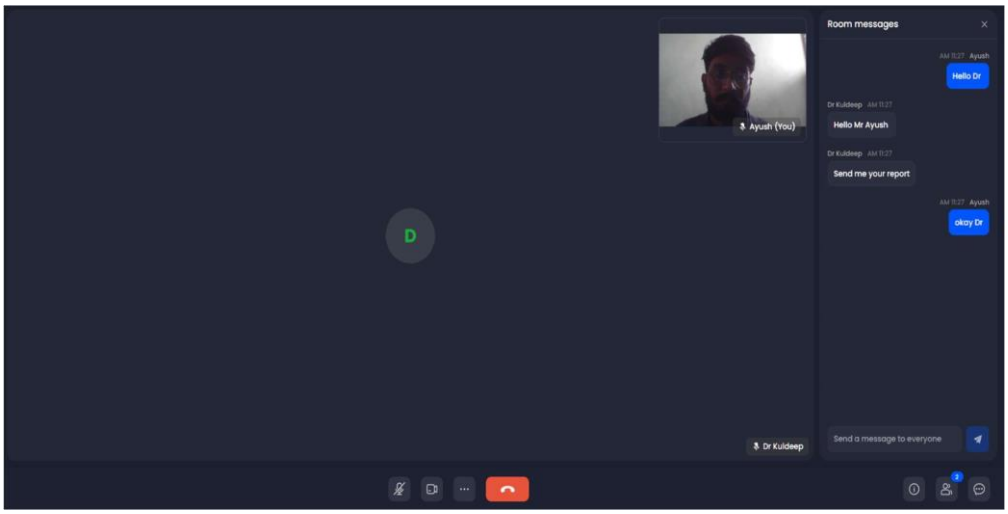

Fig 4. Video Call Consultation between doctor and patient

Figure 5 is the recommendation page where patients can check various info about particular diseases like precautions, workouts for them, medication etc.

Users just have to type in their symptoms and the system will give them the result they want according to the input.

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

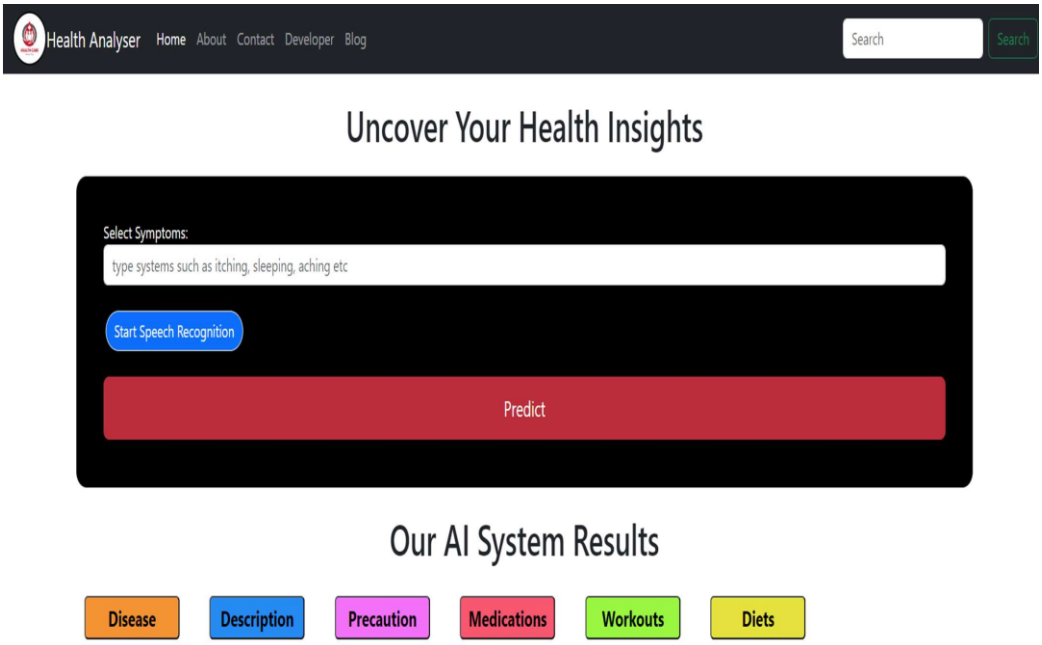

Fig. 5 - Disease information and recommendation to fight it

V. CONCLUSION AND FUTURE WORK

This project demonstrated a web application with disease diagnosis using Logistic regression and Random Forest machine learning algorithms. The web application also can hold online video consultation facilities for patients, book appointments online for visits and also chat options with doctors. All these features are packed in one single application working efficiently and robustly. The results from the prediction of both algorithms were very accurate and consistent. Between both the algorithms Random Forest performs better than Logistic Regression in every disease diagnosis across all the evaluation metrics. This superiority comes from the inherent nature of Random forest algorithms in handling complex non-linear relationships between input and output variables better than the other algorithm. The system is a comprehensive healthcare solution that provides all. These facilities are in a very cost-effective manner. The system is user-friendly and scalable according to need and has the potential to cater to a large populace effectively.

The current implementation of the concept works reasonably well to serve its purpose but yet has room for improvement is always there. More advanced machine learning algorithms in the Deep Learning field can be used to increase the accuracy with an increase in computational cost. The system asks the user to input data manually which can be further eased by only uploading PDF format or JPEG format images of medical reports into the system and having it scan the data automatically and input to the model. Moreover, real-time data from wearables can also be incorporated into the data pipeline to make better diagnoses. Overall the web application shows the immense potential hailed in the merging of Artificial Intelligence in the field of healthcare and also integrating it with the internet to make remote services possible. The role of machine learning will tend to grow further in healthcare in upcoming years as in every other field and these technologies can help us make life-saving decisions in less time and with more accuracy in the future.

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

REFERENCES

1. R. Nagarajan, P. Ramachandran, R. Dilipkumar, and P. Kaur, "Global estimate of burnout among the public health workforce: a systematic review and meta-analysis," *Human Resources for Health*, vol. 22, no. 1, May 2024.
2. JD. P. French, S.E. Scott, and R. Powell, "Promoting early detection and screening for disease," in *Springer eBooks*, pp. 533–563,2018.
3. P. Croft *et al.*, "The science of clinical practice: disease diagnosis or patient prognosis? Evidence about 'what is likely to happen' should shape clinical practice," *BMC Medicine*, vol. 13, no. 1, Jan. 2015.
4. M. Mirbabaie, S. Stieglitz, and N. R. J. Frick, "Artificial intelligence in disease diagnostics: A critical review and classification on the current state of research guiding future direction," *Health and Technology*, vol. 11, no. 4, pp. 693–731, May 2021.
5. K. Yan, T. Li, J. A. L. Marques, J. Gao, and S. J. Fong, "A review on multimodal machine learning in medical diagnostics," *Mathematical Biosciences & Engineering*, vol. 20, no. 5, pp. 8708–8726, Jan. 2023.
6. Y. Zhang, L. Diao, and L. Ma, "Logistic Regression Models in Predicting Heart Disease," *Journal of Physics: Conference Series*, vol. 1769, no. 1, p. 012024, 2021.
7. M. Pal and S. Parija, "Prediction of Heart Diseases using Random Forest," *Journal of Physics Conference Series*, vol. 1817, no. 1, p. 012009, Mar. 2021.
8. Md. Bushra, "Liver Disease Detection using Machine Learning," *International Journal for Research in Applied Science and Engineering Technology*, vol. 12, no. 8, pp. 156–162, Aug. 2024.
9. N. N. Anirudh, Joshi, V. S. Rawat and F. Rashid, "A Remote Diagnostic System Using Machine Learning for General Disease Detection," 2024 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS), pp. 1546-1551, July 2024.
10. H. Bhute, R. Wani, N. Patil, and V. Naik, "Smart Healthcare in Smart Cities: Leveraging Machine Learning for Disease Detection," *4th International Conference on Intelligent Technologies*, pp. 1–7, Jun. 2024.
11. H. Shimizu and K. I. Nakayama, "Artificial intelligence in oncology,"
12. *Cancer Science*, vol. 111, no. 5, pp. 1452–1460, Mar. 2020
13. R. Jha, V. Bhattacharjee, and A. Mustafi, "Increasing the prediction accuracy for thyroid Disease: A step towards better health for society," *Wireless Personal Communications*, vol. 122, no. 2, pp. 1921–1938,
14. Aug. 2021.
15. R. Ibarra, J. León, I. Ávila, and H. Ponce, "Cardiovascular disease detection using machine learning," *Computación Y Sistemas*, vol. 26, no. 4, Dec. 2022.
16. V. L. Raju *et al.*, "Parkinson's disease detection using Machine Learning," *International Journal of Research Publication and Reviews*, vol. 5, no. 4, pp. 1814–1823, Apr. 2024.
17. N. V. S. Greeshmanth, "Human skin disease detection using machine learning," *International Journal for Research in Applied Science and Engineering Technology*, vol. 12, no. 1, pp. 986–991, Jan. 2024.
18. M. Bernardini, L. Romeo, A. Mancini, and E. Frontoni, "A clinical decision support system to stratify the temporal risk of diabetic retinopathy," *IEEE Access*, vol. 9, pp. 151864–151872, Jan. 202
19. H. A. Helaly, M. Badawy, and A. Y. Haikal, "Deep learning approach for early detection of Alzheimer's disease," *Cognitive Computation*, vol. 14, no. 5, pp. 1711–1727, Nov. 2021.
20. S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, Jan. 2019.
21. P. Krishnamoorthy, D. Swetha, P. S. Geetha, K. Karunambiga, R. K. Ayyasamy, and A. Kiran, "Revolutionizing Medical Diagnostics: Exploring Creativity in AI for Biomedical Image Analysis," *2024 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication*, pp. 1–7, Jul. 2024.
22. N. Prof. S. B. Pagrut, N. M. A. Ozair, N. S. Ingle, N. R. Dharangaonkar, and N. A. Mundhada, "Heartcare: Heart Disease Detection using Machine Learning," *International Journal of Advanced Research in Science Communication and Technology*, pp. 249–254, Apr. 2023.

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

23. N. V. Prasad, N. U. Raj, and N. U. Dobhal, "Using machine learning for chronic disease diagnosis and prediction," *International Journal of Advanced Research in Science Communication and Technology*, pp. 554–558, Apr. 2024.

24. S. Mishra, A. Dash, and L. Jena, "Use of deep learning for disease detection and diagnosis," in *Studies in Computational Intelligence*, pp. 181–201, 2020.

25. R. Al-Momani, G. Al-Mustafa, R. Zeidan, H. Alquran, W. A. Mustafa, and A. Alkhayyat, "Chronic kidney disease detection using machine learning technique," *2022 5th International Conference on Engineering Technology and Its Applications (IICETA)*, May 2022.

26. Y. Gal, R. Islam, and Z. Ghahramani, "Deep Bayesian active learning with image data," *International Conference on Machine Learning*, pp. 1183–1192, Nov. 2017.

27. R. Ranjai and S. Gupta, "MRI Based Health Detection Using Machine Learning Techniques," *2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Jan. 2024.

28. Dhiman, S. Wadhwa, and A. Kaur, "Diseases Detection System using federated learning," in *CRC Press eBooks*, pp. 191–198,2024.

29. A. A. Adekunle, O. B. Joseph, and A. M. Olalekan, "Early Parkinson's disease detection using Machine learning approach," *Asian Journal of Research in Computer Science*, vol. 16, no. 2, pp. 36–45, Jun. 2023.

30. R. Aggarwal and S. Kumar, "Classification model for meticulous presaging of heart disease through NCA using machine learning," *Evolutionary Intelligence*, vol. 16, no. 5, pp. 1689–1698,

31. Mar. 2023.

32. A. Jain, A. Langhe, H. Choudhary, and A. Mishra, "Plant Disease Detection Using Machine Learning," *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems*, Jan. 2024.

33. G. Jayanthi, S. Brindha, S. Vijayalakshmi, V. Dharshini, J. A. Freeda, and S. Sahana, "Tomato Leaf Disease Detection Using Machine Learning," *2024 International Conference on Communication, Computing and Internet of Things (IC3IoT)*, vol. 167, pp. 1–6, Apr. 2024.

34. D. Chary, "Review on advanced machine learning model: SciKit- Learn," *SSRN Electronic Journal*, Jul. 2020.

35. PedregosaFabian *et al.*, "SciKit-Learn: Machine Learning in Python," *Journal of Machine Learning Research*, Nov. 2011.

36. V. A. M and P. Sonpatki, *ReactJS by Example - Building Modern Web Applications with React*. 2016.

37. X. Huang, "Research and Application of Node.js Core Technology," *2020 International Conference on Intelligent Computing and Human- Computer Interaction (ICHCI)*, Dec. 2020.

38. F. A. Aslam and P. S. L. H. N. Mohammed, "Efficient way of web development using Python and Flask," *International Journal of Advanced Research in Computer Science*, vol. 6, no. 2, pp. 54–57, Mar. 2015.

39. J. Cui and Z. Lin, "Research and implementation of WebRTC signalling via WebSocket-based for real-time multimedia communications," *Advances in Computer Science Research*, Jan. 2016.

40. P. Gao, Q. Chen, X. Xie, and C. Wang, "Research on Performance Optimization of MySQL Database," *2023 IEEE 3rd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA)*, May 2023.

41. K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre- processing and data augmentation techniques," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 91–99, Apr. 2022.

42. V. Agarwal, "Research on Data Preprocessing and Categorization Technique for Smartphone Review Analysis," *International Journal of Computer Applications*, vol. 131, no. 4, pp. 30–36, Dec. 2015.

43. H. S. Obaid, S. A. Dheyab, and S. S. Sabry, "The Impact of Data Pre- Processing Techniques and Dimensionality Reduction on the Accuracy of Machine Learning," *2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON)*, Mar. 2019.

44. C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building Operational data," *Frontiers in Energy Research*, vol. 9, Mar. 2021.

45. C.-Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," *The*

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

Journal of Educational Research, vol. 96, no. 1, pp. 3–14, Sep. 2002.

46. E. Y. Boateng and D. A. Abaye, "A Review of the Logistic Regression Model with Emphasis on Medical Research," *Journal of Data Analysis and Information Processing*, vol. 07, no. 04, pp. 190–207, Jan. 2019.

47. J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random forests and decision trees," *International Journal of Computer Science Issues*, vol. 9, no. 5, pp. 272–278, Sep. 2012.

48. Q. Ren, H. Cheng, and H. Han, "Research on machine learning framework based on random forest algorithm," *AIP Conference Proceedings*, Jan. 2017.

49. M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *The Stata Journal Promoting Communications on Statistics and Stata*, vol. 20, no. 1, pp. 3–29, Mar. 2020

50. M. Hasnain, M. F. Pasha, I. Ghani, M. Imran, M. Y. Alzahrani, and R. Budiarto, "Evaluating trust prediction and confusion matrix measures for web services ranking," *IEEE Access*, vol. 8, pp. 90847–90861, Jan. 2020.

51. A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, Jul. 1997.