CN Aarogya Mitra: A Multiple Disease Prediction System

Chinmay Shivratriwar[1], Rahul Agarkar[2], Yash Dharpure[3], Soham Sarde[4], Dr. K.P. Wagh[5]

[1]*Department Of Information Technology, GCOE Amravati, Maharashtra 444604, India*
[2]*Department Of Information Technology, GCOE Amravati, Maharashtra 444604, India*
[3]*Department Of Information Technology, GCOE Amravati, Maharashtra 444604, India*
[4]*Department Of Information Technology, GCOE Amravati, Maharashtra 444604, India*
[5]*Asst. Prof. Department of Information Technology GCOE Amravati, Maharashtra 444604, India*

---***---

Abstract - The rise of chronic diseases has been steadily increasing in recent years, creating a significant impact on healthcare systems worldwide. Early detection and diagnosis of chronic diseases play a critical role in improving patient outcomes and reducing healthcare costs. Machine learning (ML) algorithms have shown promise in predicting the likelihood of chronic diseases by analyzing patient data.

In this research paper, we propose a Multiple Disease Prediction System (MDPS) that uses ML algorithms like Random Forest Algorithm to predict the likelihood of four chronic diseases: diabetes, kidney, liver and cancer. The proposed system utilizes a dataset of patient's health record like blood parameters and different health related factors. The data was pre-processed, and features were extracted before being used to train four different ML models, each tailored to predict one of the four diseases. The models were then combined using an ensemble approach to make the final prediction

The proposed Multiple Disease Prediction System has several potential applications in healthcare, including early detection of chronic diseases, personalized treatment plans, and reduced healthcare costs. By identifying high-risk patients, healthcare providers can intervene early and prevent the onset of complications, resulting in improved patient outcomes. Additionally, the MDPS can help healthcare providers optimize treatment plans by predicting the likelihood of multiple chronic diseases.

The performance of the system was evaluated using various evaluation metrics, including accuracy, precision, recall, and F1-score. The results demonstrate the effectiveness of the proposed system, with an overall accuracy of 0.75, 0.96, 0.97, 0.76 for diabetes, kidney, liver and cancer respectively. The MDPS has the potential to be used as a decision-support tool for healthcare professionals in predicting the likelihood of multiple chronic diseases, facilitating early detection and treatment, and improving patient outcomes.

Key Words: Multiple Disease Prediction System, Machine Learning, Random Forest Algorithm, Diabetes, Kidney, Liver, Cancer

1. INTRODUCTION

The need for disease prediction in the healthcare industry is very crucial for several reasons. First and foremost, early detection and diagnosis of diseases are key to improving patient outcomes and reducing healthcare costs. By predicting the likelihood of a disease, healthcare providers can intervene early, preventing the onset of complications and reducing the need for more invasive and expensive treatments.

Secondly, the healthcare industry is facing significant challenges due to the growing prevalence of chronic diseases, an aging population, and the increasing cost of healthcare. Predictive analytics can help healthcare providers optimize resource allocation, reduce healthcare costs, and improve patient outcomes. By identifying high-risk patients, healthcare providers can develop personalized treatment plans that are more effective and efficient, reducing healthcare costs and improving resource allocation.

2. LITERATURE REVIEW

The development in Machine learning has become an increasingly popular as a tool for disease prediction in healthcare. Machine learning algorithms have been used to predict the onset of these diseases, identify risk factors, and develop personalized treatment plans. These algorithms can then be used to build predictive models that can be used to identify individuals who are at high risk of developing a particular disease. The use of machine learning in disease prediction has the potential to revolutionize healthcare by improving outcomes and reducing costs.
In the research "Diabetes Prediction Using Different Machine Learning Approaches," the authors have used Machine Learning approaches to predict diabetes. Diabetes is one of lethal diseases in the world. It is additional an inventor of various varieties of disorders for example: coronary failure, blindness, urinary organ diseases etc. The aim of this analysis was to develop a system which might predict the diabetic risk level of a patient with a better accuracy. Model development is based on categorization methods as Decision Tree, ANN, I Bayes and SVM algorithms.[1]

In the research "Implementation of a Web Application to Predict Diabetes Disease: An Approach Using Machine Learning

Algorithm", authors have proposed the design and developed a web application to predict diabetes. Diabetes is caused due to the excessive amount of sugar condensed into the blood. Usual identifying process for diabetic patients needs more time and money. But with the rise of machine learning, we have that ability to develop a solution to this intense issue. As they have proposed and developed an approach for diabetes disease prediction using machine learning algorithm, it has significant potential in the field of medical science for the detection of various medical data accurately.[2]

In another research "Diabetes Disease Prediction Using Data Mining" authors have discussed about data mining as a subfield in the subject of software engineering. It is the methodical procedure of finding examples in huge data sets including techniques at the crossing point of manufactured intelligence, machine learning, insights, and database systems. The aim of the data mining methodology points to think data from a data set and change it into a reasonable structure for further use. Prediction was made with the help of two algorithms Naïve Bayes and K-Nearest Neighbor and also compared which algorithm gives better accuracy on the basis of their performance factors.[3]

In the study "Application of machine learning in Cancer prediction" the author talks about the significance of Random Forest algorithm in prediction of the disease. In the model proposed uses different ML algorithms like Random Forest, Logistic Regression, Decision tree and others for the sake of prediction of the Disease and is used for the prediction of the Heart Disease, Breast Cancer and Diabetes. All the algorithms used in the system have their own way of predicting the Disease and are used accordingly.[4]

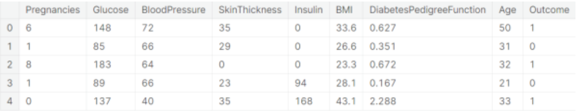
Similarly, in "Liver Disease Prediction using Machine learning Classification Techniques" author talks about the significance of Random Forest, KNN algorithm for disease related to Liver. The authors found that based on the findings of utilizing the python application to test the Naive Bayes and KNN algorithms to solve predicting issues for patients with liver illness. Different classification techniques, such as Logistic Regression, Support Vector Machine, and K-Nearest Neighbor, were utilized by authors in their study to predict liver illness. All of these algorithms were compared based on classification accuracy.[5]

In summary, the development of Multiple Disease Prediction using Machine learning has been of significant importance to the overall research community. This prediction can largely impact healthcare infrastructure by improving the healthcare facilities and providing patients services to their convenience. Several research have been made to improve the accuracy of the prediction and to increase the number of diseases that could be predicted using Machine Learning models. Overall, the advancement in this research can impact the healthcare domain helping a large number of populations.

3. METHODOLOGY

1. Dataset Selection

We have used 4 different datasets for our project. These 4 datasets are taken from 4 different sources. We have used Indian Patients database for Diabetes and Liver diseases, for other two diseases we have taken global database. Snapshot of Liver Dataset can be seen in the following figure i.e. Figure 1



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Figure 1: Liver Dataset Snapshot

2. Exploratory Data Analysis (EDA)

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. For our project we had to do EDA for all four data sets. We will be demonstrating EDA with diabetes dataset because other datasets have relatively grater number of attributes so it is very difficult to interpret heatmap

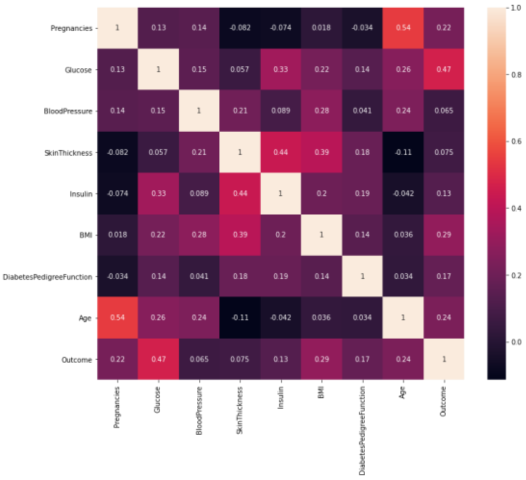


Figure 2: Heatmap Diabetes

Looking at the Figure 2 we can see that outcome and glucose are positively corelated, i.e. the more glucose a patient has, he is prone to be diabetic. Another example can be seen with Insulin and Skin Thickness, our data suggest that Insulin and Skin Thickness are positively corelated i.e. patients with high insulin levels should have a thicker skin. This helps us in determining which attributes should be used.

3. Multiple Disease Prediction System (MDPS)

The whole project can be divided into three parts i.e. The Machine Learning Model, The User Interface and the Backend. We have developed the frontend with React.js with Node.js as the development environment.

We have proposed to use the Random Forest model in our system, since random forest is a ensemble learning technique which is capable to handle different variation of data without having significant impact on accuracy and maintaining low variance. In random forest various decision trees works collectively to predict the outcome.

The Figure 3 shows how the prediction takes place in our machine learning model.

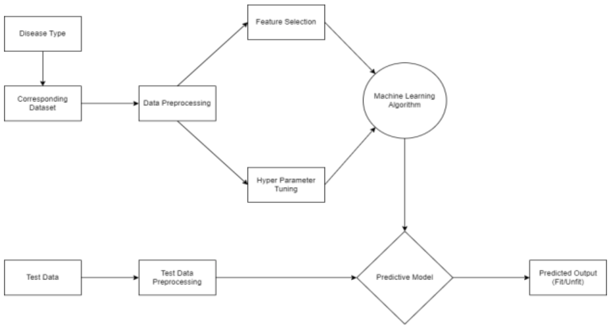


Figure 3: Working of Machine Learning Model

4. Working of Application

As discussed in point 3 we have kept the backend and frontend indigenous, this primarily helps in cross platform usage of application. Let's understand this by taking example of various food delivery applications, these companies have their websites and mobile applications, one can order food from both the sources.

So, in most of the cases the backend is the same which is connected to different frontend through API endpoints. Similarly in our application backend is hosted on a different server(port) and frontend is hosted on different server(port). Refer to the following Figure 4 to understand working of application.

As mentioned, the frontend and backend work differently and independently hence the API endpoint of the model can be used with any other frontend if required

Please refer the following Figure 4 for details of the working.

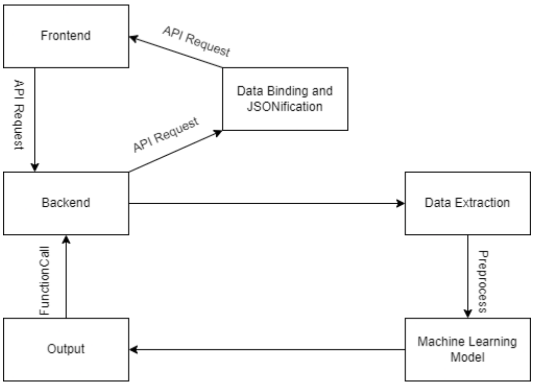


Figure 4: Working of Application

5. Network Deployment

We have deployed the application on our private network. Any devices connected to out private network will be able to access the application, any other devices not connected to our network won't be able to access the application.

4. CN AAROGYA MITRA

- Landing page has access to developer details and prediction pages of all diseases
- Each Disease Page has details regarding diseases, symptoms, model used, no of attributes trained,
- We propose a Train button, Train button acts as an end point, if the users want's to train the model by themselves then we can make use of this end button endpoint to do so
- Test button is used to activate the test window where we have to enter report data and our application will predict if he is affected or not.
- Some contacts of doctors can be added incase if the patient is affected

We have attached the snapshots of our application as you can see in the figure 4.1,4.2,4.3 respectively. The figure comprises of various pages. Landing page is attached as you can see that on a landing page we have different disease navigation.

We have attached only page for diabetes output page and test page of cancer. On Diabetes output page we can see that there is a predicted output and the output is negative i.e. the person is not infected while at cancer page we have a test console where we can see no of trained attributes.

Application Snapshots

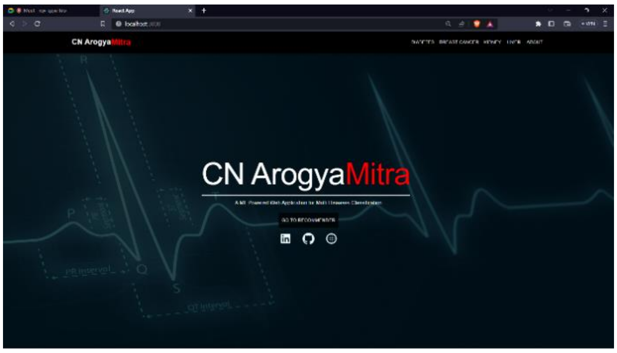
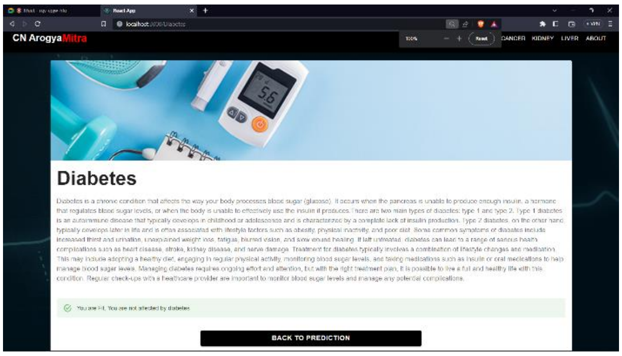


Figure 4.1: Landing Page

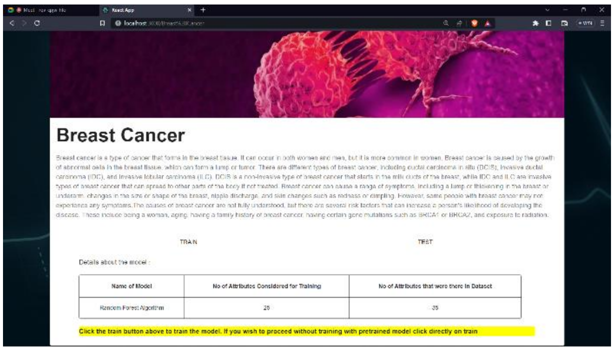


Diabetes

Diabetes is a chronic condition that affects the way your body processes blood sugar (glucose). It occurs when the pancreas is unable to produce enough insulin, a hormone that regulates blood sugar levels, or when the body is unable to effectively use the insulin it produces. There are two main types of diabetes: type 1 and type 2. Type 1 diabetes is an autoimmune disease that typically develops in childhood or adolescence and is characterized by a complete lack of insulin production. Type 2 diabetes, on the other hand typically develops later in life and is often associated with lifestyle factors such as obesity, physical inactivity, and poor diet. Some common symptoms of diabetes include increased thirst and urination, unexplained weight loss, fatigue, blurred vision, and slow wound healing. If left untreated, diabetes can lead to a range of serious health complications such as heart disease, stroke, kidney disease, and nerve damage. Treatment for diabetes typically involves a combination of lifestyle changes and medication. This may include adopting a healthy diet, engaging in regular physical activity, monitoring blood sugar levels, and taking medications such as insulin or oral medications to help manage blood sugar levels. Managing diabetes requires ongoing effort and attention, but with the right treatment plan, it is possible to live a full and healthy life with this condition. Regular check-ups with a healthcare provider are important to monitor blood sugar levels and manage any potential complications.

✓ You are Fit, You are not affected by diabetes

Figure 4.2: Diabetes Page



Breast Cancer

Breast cancer is a type of cancer that forms in the breast tissue. It can occur in both women and men, but it is more common in women. Breast cancer is caused by the growth of abnormal cells in the breast tissue, which can form a lump or tumor. There are different types of breast cancer, including ductal carcinoma in situ (DCIS), invasive ductal carcinoma (IDC), and invasive lobular carcinoma (ILC). DCIS is a non-invasive type of breast cancer that starts in the milk ducts of the breast, while IDC and ILC are invasive types of breast cancer that can spread to other parts of the body if not treated. Breast cancer can cause a range of symptoms, including a lump or thickening in the breast or underarm, changes in the size or shape of the breast, nipple discharge, and skin changes such as redness or dimpling. However, some people with breast cancer may not experience any symptoms. The causes of breast cancer are not fully understood, but there are several risk factors that can increase a person's likelihood of developing the disease. These include being a woman, aging, having a family history of breast cancer, having certain gene mutations such as BRCA1 or BRCA2, and exposure to radiation.

Details about the model :

Name of Model	No of Attributes Considered for Training	No of Attributes that were there in Dataset
Random Forest Algorithm	25	35

Click the train button above to train the model. If you wish to proceed without training with pretrained model click directly on train

Figure 4.3: Diabetes Output Page

5. RESULTS AND DISCUSSIONS

We have proposed as Random Forest based classification technique for multiple diseases prediction system. We have compared three algorithms.

The Algorithms Compared are Random Forest, K Nearest Neighbors and Support Vector Machine Classifier. We have compared these three algorithms for all four diseases. Random Forest Classifier, Support Vector Machine (SVM) Classifier, k-Nearest Neighbor (KNN) Classifier, these three algorithms are primarily used for classification purposes also Random Forest makes use of Decision Trees hence they are also covered

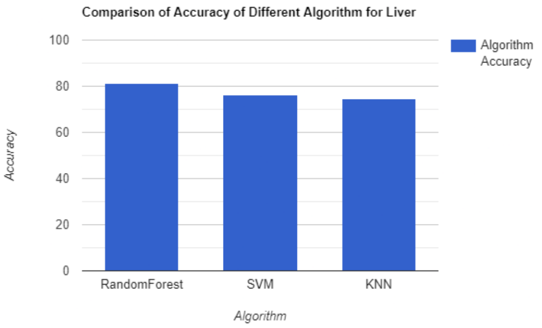


Figure 5: Liver Comparison

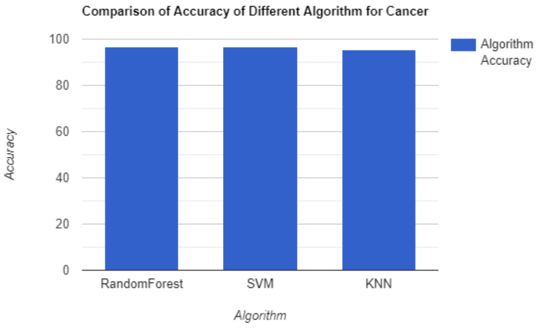


Figure 6: Cancer Comparison

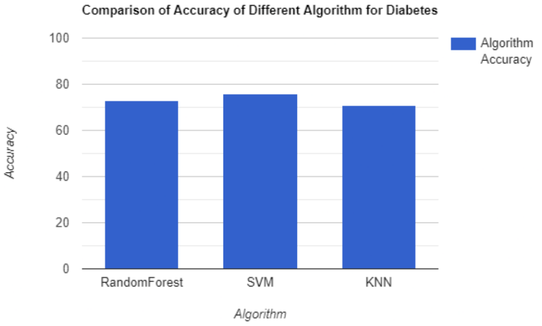


Figure 7: Diabetes Comparison

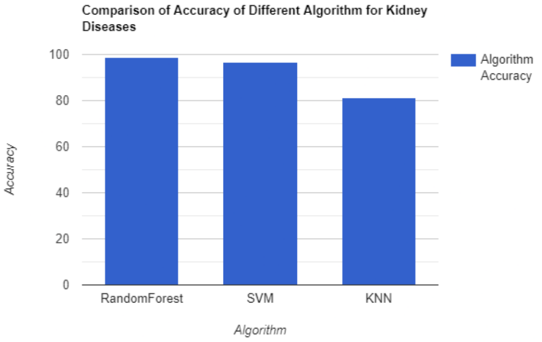


Figure 8: Kidney Comparison

Refer to Figure 5,6,7,8 we can see that for All Diseases except Diabetes the Random Forest algorithm outperforms other algorithms, in diabetes as well the Random Forest Algorithm is the second-best performer with a very narrow margin between SVM and Random Forest. Hence it can be seen that the proposed system with Random Forest Algorithm is the best performer. The Confusion Matrices for Liver, Cancer, Diabetes, Kidney is shown in the Figure 9, Figure 10, Figure 11, Figure 12

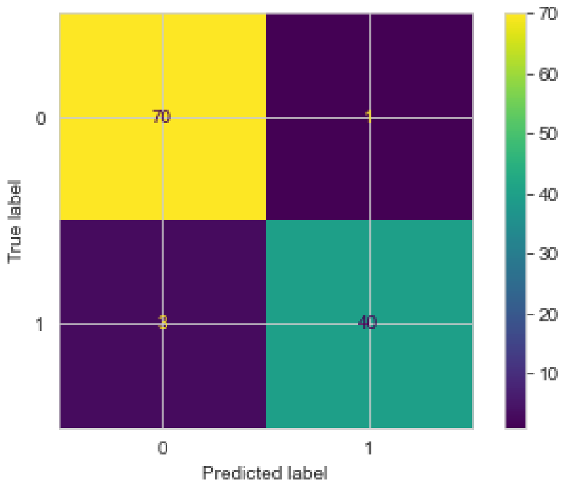


Figure 9: Liver Confusion Matrix

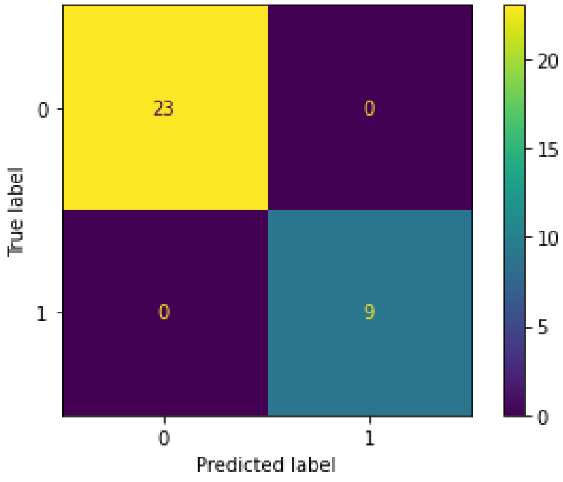


Figure 10: Diabetes Confusion Matrix

Above are Confusion matrices for Liver and Diabetes, you can see the confusion matrices for cancer and kidney below in the figure 11 and 12. True Negative and True Positives are generally compared to evaluate the performance of the application

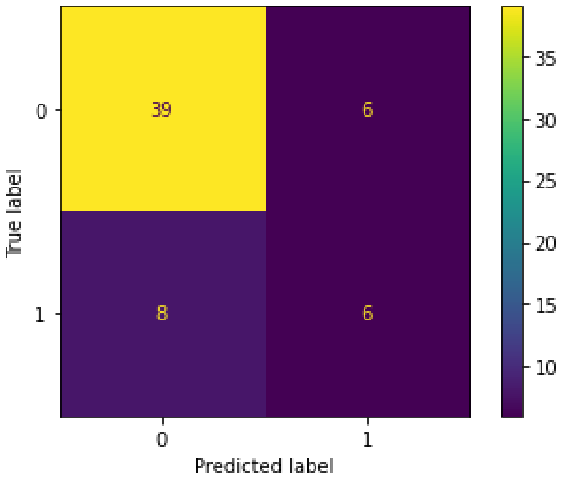


Figure 11: Cancer Confusion Matrix

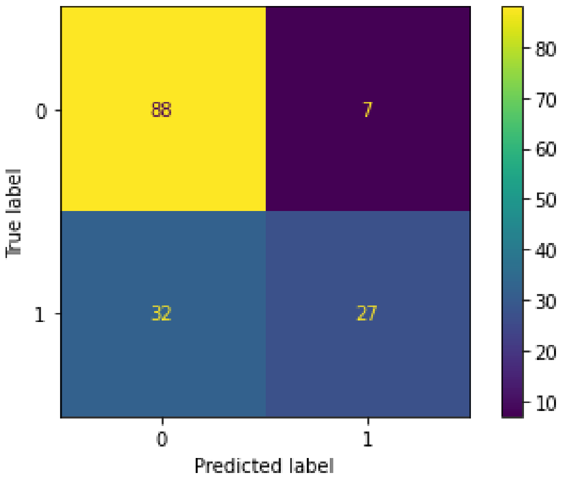


Figure 12: Kidney Confusion Matrix

So, we can see that the Random Forest Algorithm works fine for most of our diseases.

6. CONCLUSION AND FUTURE SCOPE

This article in depth makes use of Machine Learning (ML) to predict the likelihood of various multiple disease. The accuracy of prediction largely depends upon the choice of algorithm for the model and dataset that is being used This application is built such that it benefits the healthcare industry creating a significant impact. This multiple disease prediction system helps in ease of healthcare services and reduce costs. This largely helps a wide range of population to get an access to important facilities to their convenience. The rectification of dangerous chronic disease is achieved with this application. The primary goal of creating this application was to help patients for early detection and treating of chronic disease and also reducing healthcare cost.

As a part future scope of this work, large number of diseases could be added to this application, the predictions could be made more accurate without a scope of error. Further the patients could be provided with the reference to prevention or treatment of the disease. This application could be linked to various other healthcare platforms in order to create a virtual healthcare infrastructure at click.

REFERENCES

[1] P. Sonar and K. Jaya Malini, "Diabetes Prediction Using Different Machine Learning Approaches," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019, pp.367-371, doi:10.1109/ICCMC.2019.8819841.

[2] Samrat Kumar Dey, Ashraf Hossain, Md. Mahbubur Rahman, “Implementation of a Web Application to Predict Diabetes Disease: An Approach Using Machine Learning Algorithm”, 2018 21st International Conference of Computer and Information Technology (ICCIT)

[3] Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, <Diabetes Disease Prediction Using Data Mining=, 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)

[4] Kohli, P. S. and Arora, S. “Application of machine learning in Cancer prediction.= 2018 4th International Conference on Computing Communication and Automation(ICCCA)”

[5] Ketan Gupta, Nasmin Jiwani and Neda Afreen “Liver Disease Prediction using Machine learning Classification Techniques= 2022 IEEE 11th International Conference on Communication Systems and Network Technologies (CSNT)”

[6] Pratik Sharad Maratkar , Pratibha Adkar "React JS - An Emerging Frontend JavaScript Library" Iconic Research and Engineering Journals Volume 4 Issue 12 2021 Page 99-102

[7] Aslam, Fankar & Mohammed, Hawa & Lokhande, Prashant. (2015). Efficient Way Of Web Development Using Python and Flask. International Journal of Advanced Research in Computer Science. 6.

[8] Bokonda, Loola & Khadija, Ouazzani Touhami & Souissi, Nissrine. (2020). Predictive analysis using machine learning: Review of trends and methods. 10.1109/ISAECT50560.2020.9523703.

[9] Mahesh, Batta. (2019). Machine Learning Algorithms -A Review. 10.21275/ART20203995.

[10] Akinsola, J E T. (2017). Supervised Machine Learning Algorithms: Classification and Comparison. International Journal of Computer Trends and Technology (IJCTT). 48. 128 - 138. 10.14445/22312803/IJCTT-V48P126.

[11] Chen, RC., Dewi, C., Huang, SW. et al. Selecting critical features for data classification based on machine learning methods. J Big Data 7, 52 (2020). https://doi.org/10.1186/s40537-020-00327-4

[12] Patel, Harsh & Prajapati, Purvi. (2018). Study and Analysis of Decision Tree Based Classification Algorithms. International Journal of Computer Sciences and Engineering. 6. 74-78. 10.26438/ijcse/v6i10.7478.

[13] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[14] Yang, Li & Shami, Abdallah. (2020). On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice.