Health Analyzer System

Nalin Bedi[1], Adarsh Singh[2], Avinab Sharma[3], Asst. Prof. Kajol Dahiya(guide)[4]

[1]Nalin Bedi, Dept. of Computer Science Engineering, Maharaja Agrasen Institute of Technology, Delhi, India
[2]Adarsh Singh, Dept. of Computer Science Engineering, Maharaja Agrasen Institute of Technology, Delhi, India
[3]Avinab Sharma, Dept. of Computer Science Engineering, Maharaja Agrasen Institute of Technology, Delhi, India
[4]Prof.Kajol Dahiya, Dept. of Computer Science Engineering, Maharaja Agrasen Institute of Technology, Delhi, India

---***---

Abstract - *Many of the existing machine learning models for health care analysis are concentrating on one disease per analysis. Like one analysis if for diabetes analysis, one for cancer analysis, one for skin diseases like that. There is no common system where one analysis can perform more than one disease prediction. In this paper, we are proposing a system which is used to predict multiple diseases by using Flask API. This paper is used to predict Diabetes, Stroke, Breast Cancer, Fetal Health, Liver disease and Heart disease. Python pickling is used to save the model behaviour whenever required. The importance of this research paper analysis is while analysing the diseases all the parameters which causes the disease are included so that it becomes possible to detect the maximum effects which the disease will cause. For example for diabetes analysis in our system few parameters have been considered like age, sex, bmi, insulin, glucose, blood pressure, diabetes pedigree function and pregnancies. Final model's file will be saved as python pickle file. Flask API is designed. When user accesses the user interface, the user has to send the parameters of the disease asked through forms created for each disease prediction. Flask API will invoke the corresponding model and returns the status of the patient.*

Key Words: Flask API, machine learning, user interface, Pickle file, python

1.INTRODUCTION

Breast cancer, diabetes, heart disease, liver disease are for the most part driving reasons for death in the present society. Heart disease is a general term and it is also called cardiovascular disease, which means heart and blood vessel disease. Arrhythmias (issues with heart rhythm), coronary artery disease, and congenital heart defects are all diseases that fall under the category of heart illness (the defects of the heart you are born with). The cardiovascular disease normally indicates heart attack, angina (heart pain), or stroke, also conditions that affect your rhythm valves or muscles of your heart also referred to as heart diseases. Around 10 lakh patients of liver cirrhosis are newly diagnosed every year in India.Liver disease is the tenth most common cause of death in India as per the World Health Organization. Liver disease may affect every one in 5 Indians. Liver Cirrhosis is the 14th leading cause of deaths in the world and could be the 12th leading cause of deaths in the world by 2020. Thus, we are concentrating on providing immediate and accurate disease predictions to the users about the symptoms they enter along with the disease predicted. In this system, we are going to analyze Diabetes, Heart disease, Liver disease, stroke, breast cancer and fetal disease analysis. To implement multiple disease prediction systems we are going to use machine learning algorithms. Python pickling is used to save the behavior of the model. This system analyses the diseases after taking as input all the possible parameters which cause the disease.

1.1 Description

The existing systems in the health care industry are concerned with considering only one disease prediction at a time. For example, one system is used to analyse diabetes, another is used to analyse breast cancer or stroke, and another system is used to predict heart disease. Maximum systems focus on one particular disease. Many models are deployed by organizations when they want to analyse their patient's health reports. The approach used in the existing systems are useful for analysing a particular disease. Moreover, patients also spend a lot of money in consulting various doctors for various diseases diagnosis in a single disease prediction system, which in turn is very expensive. In multiple diseases prediction system more than one disease can be analysed on a single website. The user doesn't need to go to or browse different places in order to predict whether he/she is disease-prone or not. In such a system, the user needs to select the name of the particular disease, enter its parameters and just click on submit. This would prove to be a cost effective solution. The suitable machine learning model will be invoked and it would predict whether user is disease-prone or not and display it on the screen for the user on the interface.

1.2 Existing System

Many of existing analysis involved 1331nalyzing particular disease. One potential problem with a single disease prediction system is that it may not be able to accurately identify all potential diseases or conditions that a person may have. This is because different diseases can have similar symptoms, and a single prediction system may not be able to accurately distinguish between them. Additionally, a single disease prediction system may not be able to account for the complexity of an individual's medical history or other factors that can affect their health. This could lead to inaccurate predictions and potentially harmful treatment plans. Another potential problem with a single disease prediction system is that it may not be able to adapt to new developments in medical research or changes in a person's health over time. This could lead to outdated or ineffective treatment recommendations.

1.3 Proposed system

We have proposed a system that will flaunt a simple and elegant User Interface and also be time efficient . Our proposed system closes down the gap between doctors and patients which will help both classes of users to achieve their desirable outcomes. This system is used to predict diseases according to symptoms as well as the medical records of patients. This proposed system will take down several symptoms from the users and medical record readings evaluate after applying algorithms such as Decision Tree, Random Forest, Naïve Bayes, Logistic Regression and SVM which will help in getting accurate prediction. Our system will large datasets which includes large diversity of medical parameters to get more effective results and thus our system will improve and enhance the accuracy, diversity of population to get more effective results and thus our system will improve and enhances the accuracy of the results. Along with the increased accuracy rate, we will proliferate the reliability of our system for this job and can gain the trust of patient in this system. Hence this system will contribute in easier health management with better satisfaction to the users.

2. LITERATURE REVIEW

1. G Naveen Kishore and few other authors proposed the work named Prediction Of Diabetes Using Machine Learning Classification Techniques proposed. In this work, various classification algorithms like SVM, Logistic Regression, Decision Tree, KNN, Random Forest are utilized on the 769 instances of the Pima dataset which contain features like Pregnancies, Blood pressure, body mass index, etc. They have reported the highest accuracy as 74.4 %for the classification algorithm Random Forest and the lowest accuracy in this work is attained by the KNN reported as 71.3%.

2. In the work presented by M. Marimuthu, S. Deiva Rani, Gayatri.

R described the cardio diseases in a detailed manner and also applied the classification algorithms like SVM, Decision Tree, Naïve Bayes, K-Nearest Neighbors on the Framingham dataset from Kaggle. The authors compared various machine learning algorithms for the forecast of the risk of heart disease. The highest reported accuracy in this work is 83.60% for the KNN classification algorithm.

3. Ch. Shravya, K.Pravallika, Shaik Subhani presented the work on Breast cancer prediction using Supervised machine learning techniques on the dataset and also analyzed the results with (PCA)principal component analysis and also used the dimensionality reduction and explained in a well-mannered way.

3. SYSTEM ANALYSIS

3.1 Functional Requirement

- The system has the ability to analyze the collected data to identify patterns and make predictions about a person's likelihood of developing certain diseases.

- The user inputs the symptoms and other medical details/records for a particular disease and based on the trained model of the machine learning input the output will be displayed on the user interface.

- The system has the ability to integrate with different machine learning frameworks and libraries to enable the use of state-of-the-art algorithms and technique.

3.2 Non Functional Requirement

- The system would be able to make predictions quickly and efficiently, with minimal delay or downtime.

- The system should be user-friendly and easy to use, with clear instructions and intuitive interfaces for both medical professionals and patients.

- The system would be scalable and able to handle large volumes of data without performance degradation.

4. DESIGN

4.1 Architecture

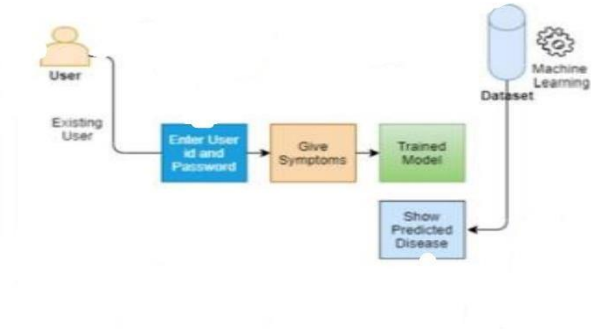


Figure No.4.1: Block Diagram

In the figure no 4.1 we have experimented on six diseases that is heart, diabetes, liver disease, stroke, fetal health and breast cancer. The first step is to gather the dataset for heart disease, diabetes disease, liver disease, stroke, fetal health and breast cancer. We have extracted the PIMA Indian Diabetes dataset, Indian liver dataset, Stroke Prediction Dataset, Fetal Health Classification and Breast Cancer Wisconsin (Diagnostic) Data Set respectively. Once dataset is imported then visualization of each input data takes place. After visualization pre-processing of data takes place where checking for outliers, missing values was done and then dataset was split into training and testing .Next is on the training dataset we had applied knn, xgboost, Logistic Regression , Naive Bayes, Decision tree and random forest algorithm and applied knowledge on the classified algorithm using testing dataset. After applying knowledge we will choose the algorithm with the best accuracy for each of the disease. Then we build a pickle file for all the disease and then integrated the pickle file for the output of the model on the webpage.

4.2 User Interface Design

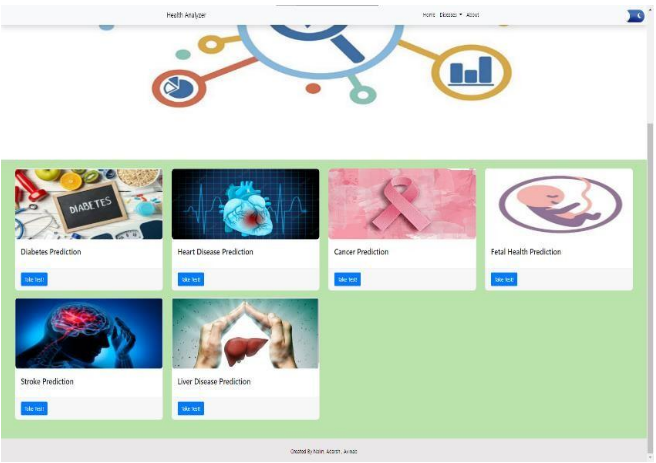


Figure No 4.2: Graphical User Interface

5. IMPLEMENTATION

5.1 Algorithm

SVM Algorithm

Support vector machine (SVM) is a supervised learning algorithm for classification and regression tasks in machine learning. It is used to find the hyperplane in an N-dimensional space that maximally separates the classes.

The working of the SVM algorithm is as follows:

Step-1: SVM uses a kernel function to transform the data into a higher-dimensional space where it can be separated by a linear boundary.

Step-2: Common kernel functions include the linear kernel, the polynomial kernel, and the radial basis function (RBF) kernel. The regularization parameter controls the complexity of the model and helps prevent over-fitting.

Step-3: Once the data is prepared and the hyper-parameters are chosen, you can train the SVM model using the training set. This involves solving a quadratic optimization problem to find the hyper-plane with the largest margin.

Random Forest Algorithm

Random forests are a type of ensemble learning algorithm, which means that they combine the predictions of multiple individual models to make a more accurate and stable prediction. In the case of random forests, the individual models are decision trees, which are trained on subsets of the training data. The working of the random forest is as follows:

Step-1: Collect and preprocess the training data.

Step-2: Select the number of decision trees to generate. This is a hyper-parameter of the random forest algorithm, and it determines the number of individual models that will be trained and used to make predictions.

Step-3: For each decision tree:

- Generate a bootstrap sample of the training data. This involves randomly selecting a subset of the training data, with replacement, to use as the training set for the decision tree.

- Train a decision tree on the bootstrap sample. This involves applying the decision tree learning algorithm to the bootstrap sample to train a decision tree model.

Step-4: To make a prediction using the random forest, feed the test data point to each of the trained decision trees, and use the majority vote of the individual decision tree predictions as the final prediction.

Linear Regression Algorithm

Linear regression is a statistical technique that is used to model the relationship between a dependent variable and one or more independent variables. The working of Linear Regression Algorithm is as follows:

Step-1: We need to choose an optimization algorithm such as gradient descent, and use it to find the coefficients that minimize the loss function. The loss function measures the difference between the predicted probability and the true value of the dependent variable.

Step-2: The predicted probability is calculated using the sigmoid function, which maps the output of the linear regression model (a continuous value) to a probability between 0 and 1. The sigmoid function has an "S" shaped curve, and the output is 1 when the input is very large, 0 when the input is very small, and 0.5 when the input is 0.

Step-3: Once the model is trained, you can use it to make predictions on new data. You can evaluate the performance of the model using evaluation metrics such as accuracy, precision, recall, and F1 score.

6. RESULT

In the system breast cancer disease prediction model used Random Forest Classifier algorithm, diabetes prediction model used Random Forest Classifier, stroke prediction model used Random Forest Classifier, liver prediction model used Random Forest Classifier, heart disease prediction model uses SVM algorithm and fetal disease prediction model uses the random forest algorithm as these gave the best accuracy accordingly.

ACCURACY FOR EACH DISEASE:

Table No 6.1: Diabetes Disease

ALGORITHM	Accuracy
Random Forest	81.81%
Naive Bayes	79.22%

Table No 6.2: Heart Disease

ALGORITHM	Accuracy
SVM	82.41%

Table No 6.3: Liver Disease

ALGORITHM	Accuracy
Random Forest	74.35%
Gaussian Naive Bayes	68.37%

Table No 6.4: Breast Cancer

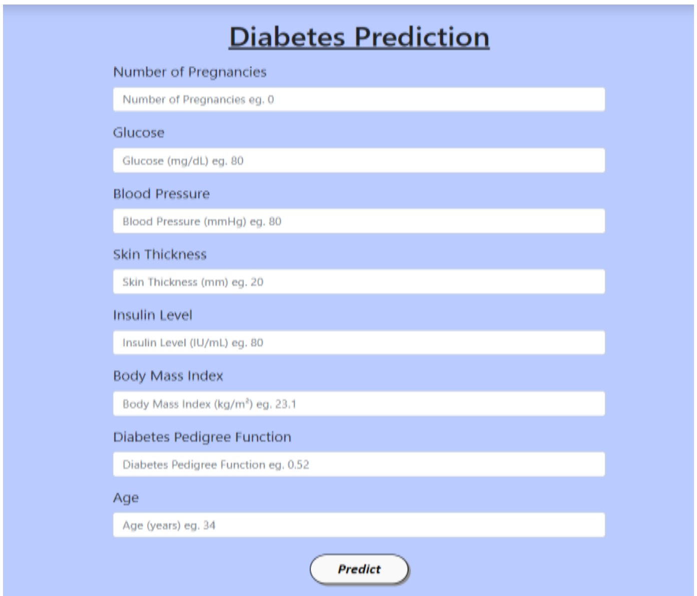
ALGORITHM	Accuracy
Random Forest	97.30%
Decision Tree	90.35%

Table No 6.5: Fetal Health

ALGORITHM	Accuracy
Random Forest	94.36%
Decision Tree	91.72%

Table No 6.6: Stroke

ALGORITHM	Accuracy
Random Forest	96.21%
Decision Tree	96.16%



Diabetes Prediction

Number of Pregnancies
Number of Pregnancies eg. 0

Glucose
Glucose (mg/dL) eg. 80

Blood Pressure
Blood Pressure (mmHg) eg. 80

Skin Thickness
Skin Thickness (mm) eg. 20

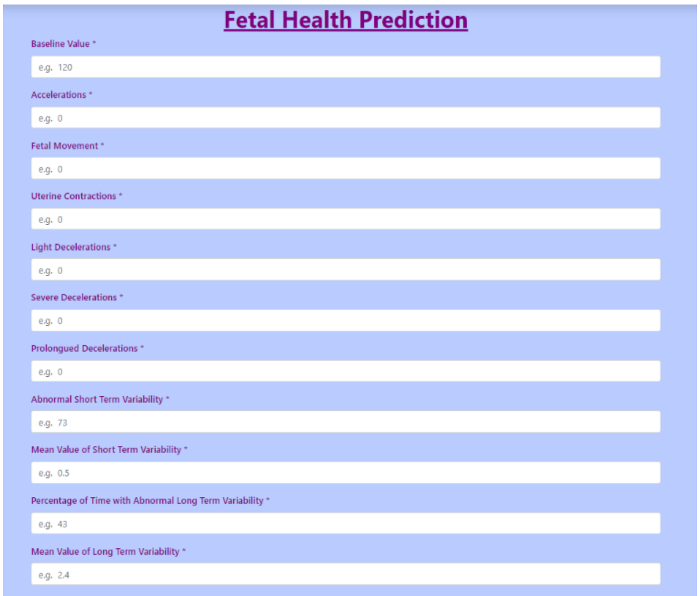
Insulin Level
Insulin Level (IU/mL) eg. 80

Body Mass Index
Body Mass Index (kg/m²) eg. 23.1

Diabetes Pedigree Function
Diabetes Pedigree Function eg. 0.52

Age
Age (years) eg. 34

Fig No. 6.1 Diabetes Disease prediction UI



Fetal Health Prediction

Baseline Value *
e.g. 120

Accelerations *
e.g. 0

Fetal Movement *
e.g. 0

Uterine Contractions *
e.g. 0

Light Decelerations *
e.g. 0

Severe Decelerations *
e.g. 0

Prolongued Decelerations *
e.g. 0

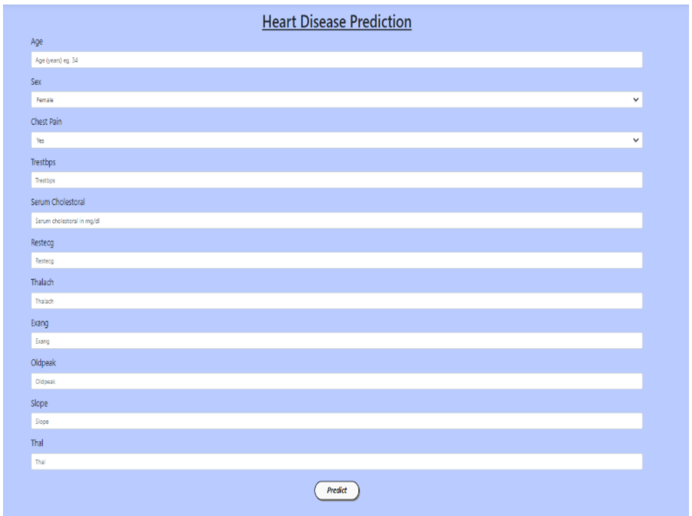
Abnormal Short Term Variability *
e.g. 73

Mean Value of Short Term Variability *
e.g. 0.5

Percentage of Time with Abnormal Long Term Variability *
e.g. 43

Mean Value of Long Term Variability *
e.g. 2.4

Fig No. 6.2: Fetal Health Disease prediction UI



Heart Disease Prediction

Age
Age (years) eg. 34

Sex
Female

Chest Pain
Yes

Trestbps
Trestbps

Serum Cholestoral
Serum cholestoral in mg/dl

Restecg
Restecg

Thalach
Thalach

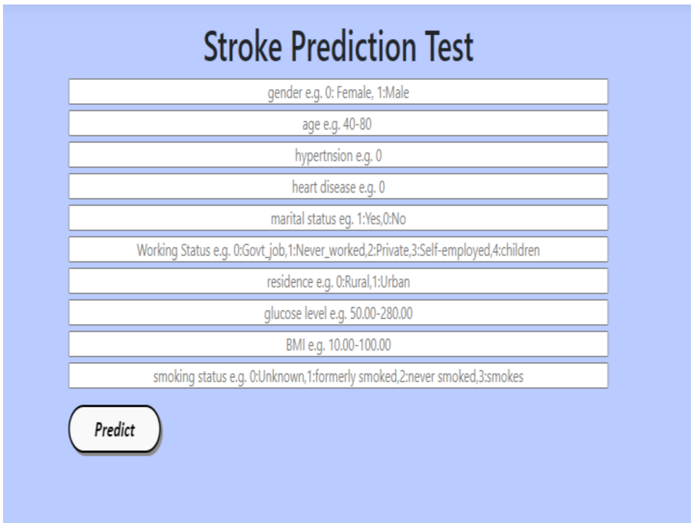
Exang
Exang

Oldpeak
Oldpeak

Slope
Slope

Thal
Thal

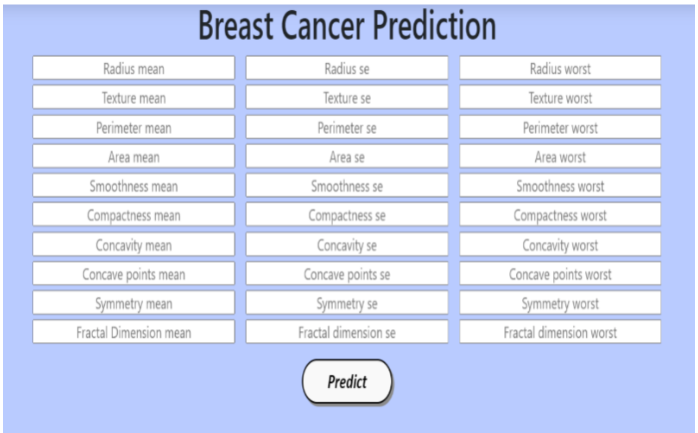
Fig No. 6.3: Heart Disease prediction UI



Stroke Prediction Test

gender e.g. 0: Female, 1:Male
age e.g. 40-80
hypertnsion e.g. 0
heart disease e.g. 0
marital status eg. 1:Yes,0:No
Working Status e.g. 0:Govt_job,1:Never_worked,2:Private,3:Self-employed,4:children
residence e.g. 0:Rural,1:Urban
glucose level e.g. 50.00-280.00
BMI e.g. 10.00-100.00
smoking status e.g. 0:Unknown,1:formerly smoked,2:never smoked,3:smokes

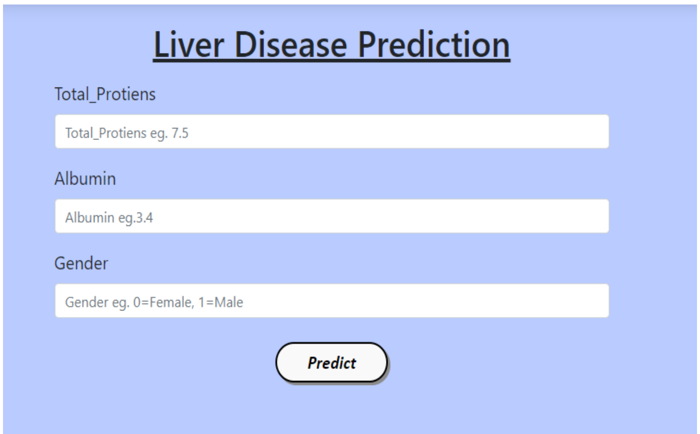
Fig No. 6.4: Stroke Prediction prediction UI



Breast Cancer Prediction

Radius mean	Radius se	Radius worst
Texture mean	Texture se	Texture worst
Perimeter mean	Perimeter se	Perimeter worst
Area mean	Area se	Area worst
Smoothness mean	Smoothness se	Smoothness worst
Compactness mean	Compactness se	Compactness worst
Concavity mean	Concavity se	Concavity worst
Concave points mean	Concave points se	Concave points worst
Symmetry mean	Symmetry se	Symmetry worst
Fractal Dimension mean	Fractal dimension se	Fractal dimension worst

Fig No. 6.5: Breast Cancer prediction UI



Liver Disease Prediction

Total_Protiens
Total_Protiens eg. 7.5

Albumin
Albumin eg.3.4

Gender
Gender eg. 0=Female, 1=Male

Fig No. 6.6: Liver disease prediction UI

7. CONCLUSION

The objective of multiple disease prediction is to identify individuals who are at high risk of developing multiple diseases. This information can help doctors to take a more proactive approach to treating and preventing disease, and can also help individuals to make lifestyle changes and take other preventative measures to reduce their risk of developing multiple diseases. By predicting multiple diseases, doctors can also develop more effective treatment plans for their patients, which can improve overall health outcomes.

7.1 Future Scope

- In the future we can add more diseases in the existing API.
- We can try to improve the accuracy of prediction in order to decrease the mortality rate.
- Try to make the system user-friendly and provide a chatbot for normal queries.

8. ACKNOWLEDGEMENT

We sincere thank to our college Maharaja Agrasen Institute of Technology for giving us a platform to prepare a project on the topic "Health Analyzer System" and would like to thank our HOD Dr. Namita Gupta for giving us the opportunities and time to conduct and research on the subject. We are sincerely grateful for Prof. Kajol Dahiya (CSE Department), as our guide for providing help during our research, which would have seemed difficult without their motivation, constant support, and valuable suggestion. Moreover, the complication of this research paper would have been impossible without the co-operation, suggestion, and help of our friends and family.

REFERENCES

[1] Archana Singh ,Rakesh Kumar, “Heart Disease Prediction Using Machine Learning Algorithms”, 2020 IEEE, International Conference on Electrical and Electronics Engineering (ICE3)

[2] A.Sivasangari, Baddigam Jaya Krishna Reddy,Annamareddy Kiran, P.Ajitha,” Diagnosis of Liver Disease using Machine Learning Models” 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)

[3]Yang, G.; Pang, Z.; Deen, M.J.; Dong, M.; Zhang, Y.T.; Lovell, N.; Rahmani, A.M. Homecare robotic systems for healthcare 4.0: Visions and enabling technologies. IEEE J. Biomed. Health Inform. 2020, 24, 2535–2549

[4] Multi Disease Prediction System:- By- Divya Mandem, 1PG Scholar, Dept. of computer science and System Engineering(A), Andhra University College of Engineering Vishakhapatnam, Andhra Pradesh B. Prajna ,Professor, Dept. of computer science and System Engineering(A), Andhra University College of Engineering Vishakhapatnam, Andhra Pradesh.