

## SYMPTOMS BASED DISEASE PREDICTION

Mayur Gadekar<sup>\*1</sup>, Soyeb Jamadar<sup>\*2</sup>, Prajak Pachpute<sup>\*3</sup>, Sanket Shinde<sup>\*4</sup>,  
Swati Bhosale<sup>\*5</sup>

<sup>\*1,2,3,4</sup>Students, Computer Engineering Department, HSBPVT's GOI Parikrama College of Engineering, Kashti, Maharashtra, India

<sup>\*5</sup>Assistant Professor, Computer Engineering Department, HSBPVT's GOI Parikrama College of Engineering, Maharashtra, India

### ABSTRACT

Health information needs are also changing the knowledge-seeking behavior and should be observed around the globe. Challenges faced by many of us are looking online for health information regarding diseases, diagnoses, and different treatments. If a recommendation system are often made for doctors and medicine while using review mining will save plenty of time. In a system like this, the user faces many problems in understanding the core medical vocabulary because the users are laymen. The user is confused because an outsized amount of medical information on different mediums is out there.

**Keywords:** Random Forest Algorithm, Naive Bayes, Support Vector Machine

### I. INTRODUCTION

Disease Prediction using Machine Learning is a system that predicts the disease based on the information provided by the user. It also predicts the disease of the patient or the user based on the information or the symptoms he/she enters into the system and provides accurate results based on that information. If the patient is not very serious and the user just wants to know the type of disease, he/she has been through. It is a system that provides the user the tips and tricks to maintain the health system of the user and it provides a way to find out the disease using this prediction. Now a day's health industry plays a serious role in curing the diseases of the patients so this is often also some quite help for the health industry to tell the user and also it's useful for the user just in case he/she doesn't want to travel to the hospital or the other clinics, so just by entering the symptoms and every one other useful information the user can get to understand the disease he/she is affected by and therefore the health industry also can get enjoy this technique by just asking the symptoms from the user and entering in the system and in just a few seconds they can tell the exact and up to some extent the accurate diseases.

### II. LITERATURE REVIEW

Numerous disquisition factory have been carried out for the prophecy of the conditions predicated on the symptoms shown by an existent using machine knowledge algorithms:

Monto et al. [1] designed a statistical model to prognosticate whether a case had influenza or not. They included 3744 unvaccinated grown-ups and adolescent cases of influenza who had fever and at least 2 other symptoms of influenza. Out of 3744, 2470 were verified to have influenza by the laboratory. Predicated on this data, their model gave an delicacy of 79.

Colorful machine learning algorithms were streamlined for the effective prophecy of a habitual complaint outbreak by Chen et al. [2]. The data collected for the training purpose was deficient. To overcome this, a idle factor model was used. A new convolutional neural network- grounded multimodal complaint trouble vaticination (CNN-MDRP) was structured. The algorithm reached an delicacy of around 94.8%.

The DNN model performed more in terms of average performance and the LSTM model gave close prognostications when circumstances were large. Haq et al. [3] used a database that contained information about patients having any heart complaint. They pulled features using three selection algorithms which are relief, minimum redundancy, and maximum connection (mRMR), and least absolute loss and selection motorist which was cross-verified by theK-fold system. The pulled features were transferred to 6 different machine learning algorithms and also it was classified predicated on the presence or absence of heart complaint.

Maniruzza- man et al. [4] classified the diabetes complaint using ML algorithms. Logistic regression (LR) was used to identify the trouble factors for diabetes complaint. The overall delicacy of the ML- predicated system was 90.62%.

### III. METHODOLOGY

Method and analysis which is performed in your research work should be written during this section. A simple strategy to follow is to use keywords from your title in the first few sentences.

#### 1. The Dataset

Firstly, for getting some insights from and training our model, we want some datasets. So for that, we have made some surveys in medical field, explored some data on internet and made a raw dataset by combining all of that. So now, we have a dataset

	Disease	Symptom_1	Symptom_2	Symptom_3	Symptom_4	Symptom_5	Symptom_6	Symptom_7	Symptom_8	Symptom_9	Symptom_10	S
0	Fungal infection	itching	skin_rash	nodal_skin_eruptions	dischromic_patches	NaN	NaN	NaN	NaN	NaN	NaN	
1	Fungal infection	skin_rash	nodal_skin_eruptions	dischromic_patches	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
2	Fungal infection	itching	nodal_skin_eruptions	dischromic_patches	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
3	Fungal infection	itching	skin_rash	dischromic_patches	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
4	Fungal infection	itching	skin_rash	nodal_skin_eruptions	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

#### 2. Preprocessing of data

After collecting that data, as that data is raw data we have to make it suitable for training our machine learning model. By using some python libraries like NumPy, and pandas, we have made that data suitable for machine learning models.

#### 3. Applying machine learning algorithms

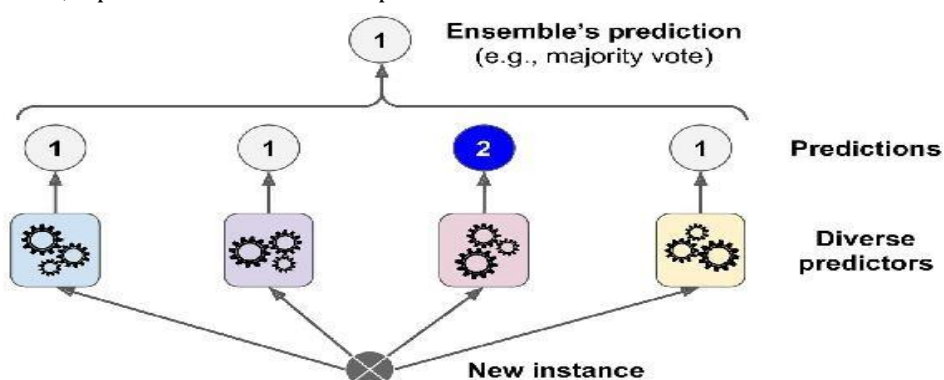
Now, our data is ready to use with machine learning algorithms to predict some output. As our problem come under unsupervised machine learning technique, we have used three algorithms viz. Support Vector Machine, Random Forest Classifier, Naive Bayes algorithm.

#### 4. Model Building

After applying these algorithms, we have to select which is most fitted with our dataset and which gives us more accuracy. So, we have used a confusion matrix for that and mapped out the accuracy of each model. And, we have found that all are giving the same 100% accuracy, so we have selected Random Forest Classifier for building our model.

##### 4.1 Random Forest Classifier

It is an ensemble classifier using many decision tree models; it can be used for regression as well as classification. As the name suggests, Random Forest may be a classifier that contains a variety of decision trees on various subsets of the given dataset and takes the typical to enhance the predictive accuracy of that dataset. Instead of counting on one decision tree, the random forest takes the prediction from each tree and supported the bulk votes of predictions, it predicts the ultimate output.



Below are some points that specify why we should always use the Random Forest algorithm:

1. It takes less training time as compared to other algorithms.
2. It predicts output with high accuracy, even for the massive dataset it runs efficiently.
3. It also can maintain accuracy when an outsized proportion of knowledge is missing.

#### 4.2 Naive Bayes:

The Naive Bayes algorithm is that the algorithm that learns the probability of an object with certain features belonging to a particular group/ class. as an illustration, if you are trying to identify a fruit grounded on its color, shape, and taste, also an orange- colored, globular, and pungent fruit would presumably be an orange. of these parcels collectively contribute to the probability that this fruit is an orange and that's why it's known as " naive". The " Bayes" part, refers to statistician and champion, Bayes and the theorem named after him, Bayes'theorem, which is that the base for Naïve Bayes Algorithm. Further formally, Bayes'Theorem is stated because the following equation

$$P(A/B) = (P(B/A) * P(A)) / P(B)$$

#### 4.3 Support Vector Machine:

SVM Support Vector Machine (SVM) is a managed AI calculation technique which can be employed for both order and relapse difficulties. It's a truly important content in the field of jargon recognition, machine learning, memoir informatics etc. SVM follows the approach of chancing a hyperplane which maximizes the geometric fringe and minimizes the type error which is predicated on given a two class direct linearly separable variable (3). In any case, it's for the utmost part employed in characterization issues. In this calculation, plot every detail thing as a point in n-dimensional space where n is number of highlights you have with the estimation of each element being the estimation of a specific organize. Bolster Vectors are just theco-ordinates of individual perception. Support Vector Machine is an outskirts which stylish isolates the two classes. The support vector machine has been chosen because it represents a frame both interesting from a machine learning perspective and from an bedded systems perspective. An SVM is a direct or non-direct classifier, which is a fine function that can distinguish two different kinds of objects .

Training a SVM can be illustrated with the following mock law:

Algorithm 1 Training an SVM

Bear X and y loaded with training labeled data, a = 0 or a partly trained SVM

1 some value (10 for illustration)

2 repeat

3 for all{ xi, yi},{ xj, yj} do

Optimize  $\alpha_i$  and  $\alpha_j$

5 end for

6 until no changes in  $\alpha$  or other resource constraint criteria met

Ensure Retain only the support vectors ( $\alpha_i > 0$ )

#### 5. Model Validation:

Confusion Matrix:

Actual Values vs. Predicted Values

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Accuracy= (TP+TN)/ Total n

Miss Classification = (FP+FN)/n

True positive rate = TP/ (FN+TP)

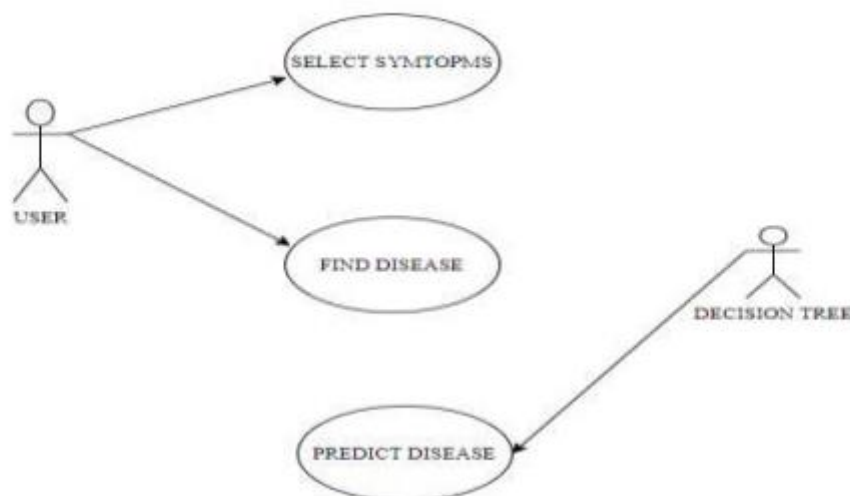
False positive rate =FP/ (TN+FP)

#### 6. Deploying our model

After creating our machine learning model, we dumped it into a pickle file. And used that binary file for our website for getting output for users. We have used Flask (Python web framework) for deploying our machine learning model as a website.

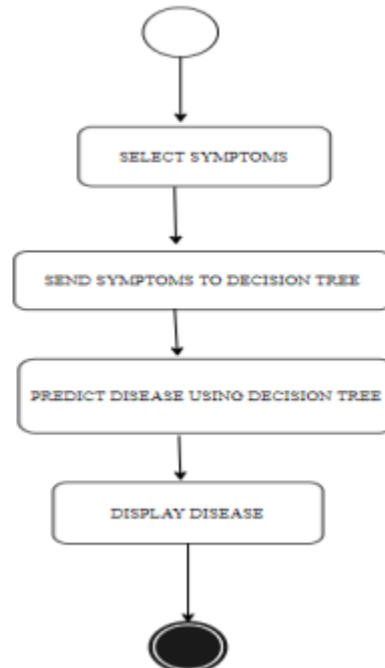
#### IV. USE CASE DIAGRAM

A Unified Modeling Language (UML) use case diagram is a type of behavioral diagram described from and generated from a Use-Case study. Its aim is to provide a graphical description of a system's functionality in terms of actors, their roles (represented as use cases), and any dependencies between those use cases. A usage case diagram has the key task of demonstrating which machine functions are executed by which person. Roles of the actors can be portrayed in the method.

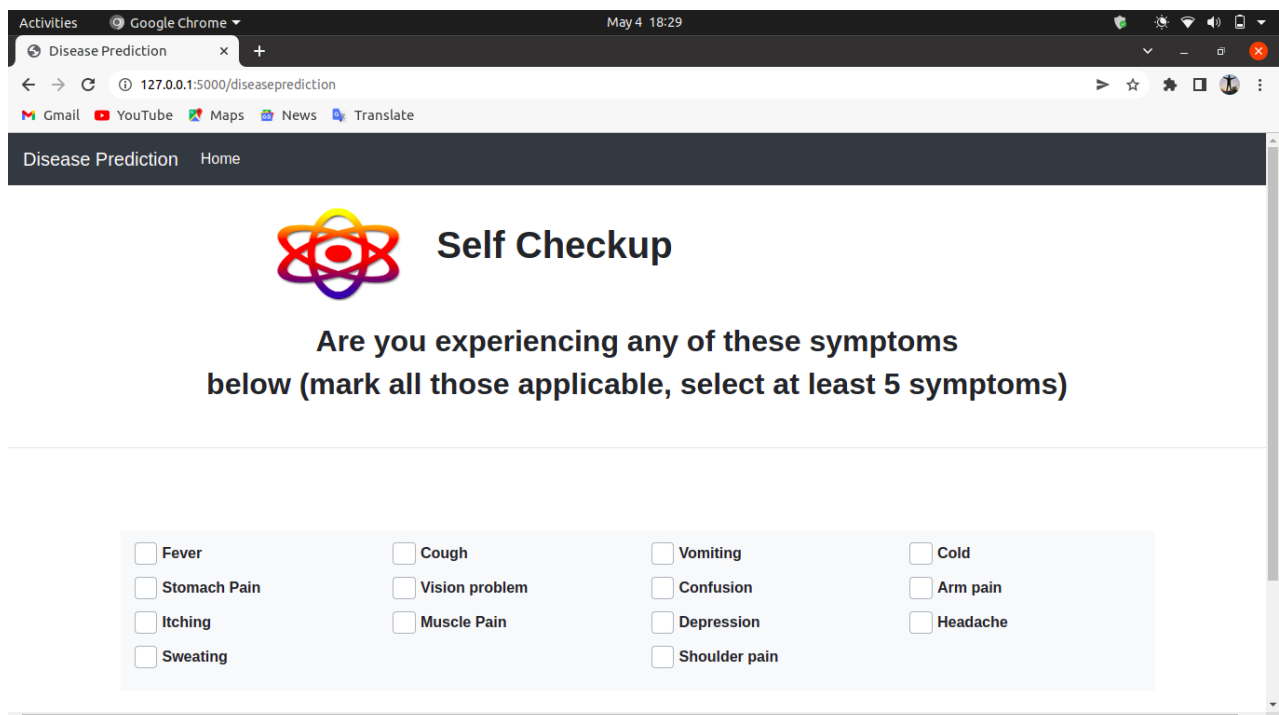


## V. ACTIVITY DIAGRAM

Activity diagrams are schematic descriptions of stepwise task workflows and activities with support for preference, repetition, and rivalry. In the Universal Modeling Vocabulary, task diagrams are structured to model both numerical and operational processes (i.e., workflows), as well as data flows that interact with the associated operations. While operation diagrams represent mainly the total control flow, they may also contain elements that display the data flow between operations across one or more data stores.

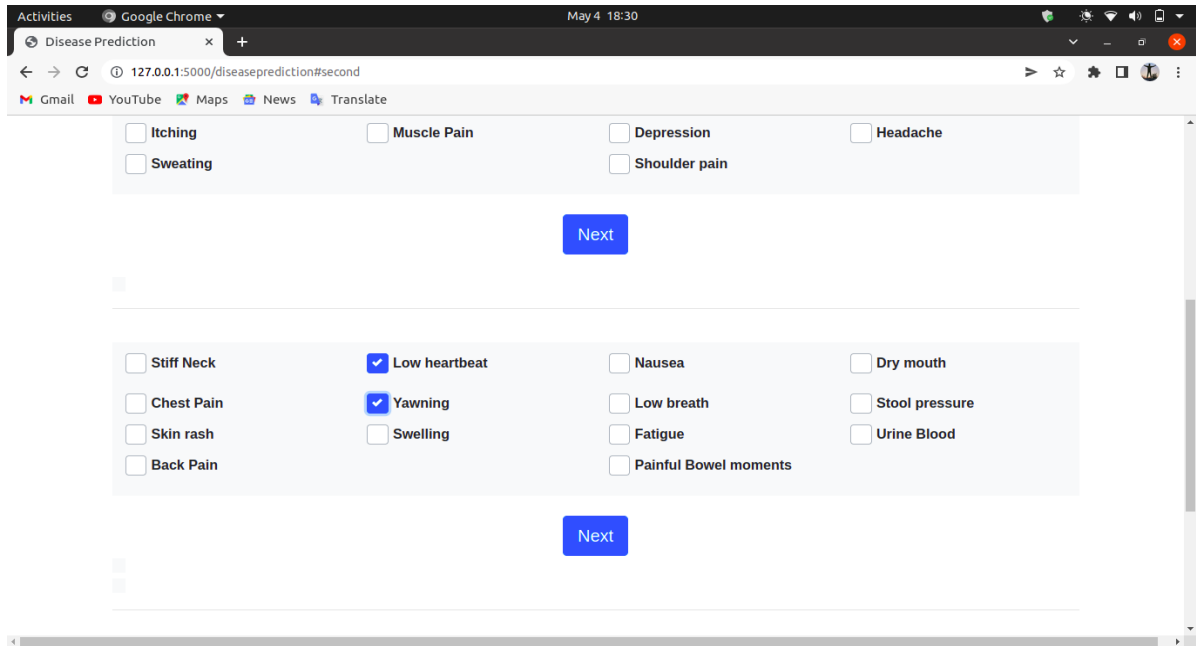


## VI. RESULTS AND DISCUSSION



The screenshot shows a web browser window with the URL `127.0.0.1:5000/diseaseprediction`. The page has a dark header with "Disease Prediction" and "Home" links. Below the header is a logo consisting of a stylized atom with the text "Self Checkup" next to it. The main content area contains the text: "Are you experiencing any of these symptoms below (mark all those applicable, select at least 5 symptoms)". Below this text is a light gray box containing a grid of 16 checkboxes, each followed by a symptom name:

<input type="checkbox"/> Fever	<input type="checkbox"/> Cough	<input type="checkbox"/> Vomiting	<input type="checkbox"/> Cold
<input type="checkbox"/> Stomach Pain	<input type="checkbox"/> Vision problem	<input type="checkbox"/> Confusion	<input type="checkbox"/> Arm pain
<input type="checkbox"/> Itching	<input type="checkbox"/> Muscle Pain	<input type="checkbox"/> Depression	<input type="checkbox"/> Headache
<input type="checkbox"/> Sweating		<input type="checkbox"/> Shoulder pain	



Activities Google Chrome May 4 18:30

Disease Prediction

127.0.0.1:5000/diseaseprediction#second

Gmail YouTube Maps News Translate

☐ Itching ☐ Muscle Pain ☐ Depression ☐ Headache

☐ Sweating ☐ Shoulder pain

Next

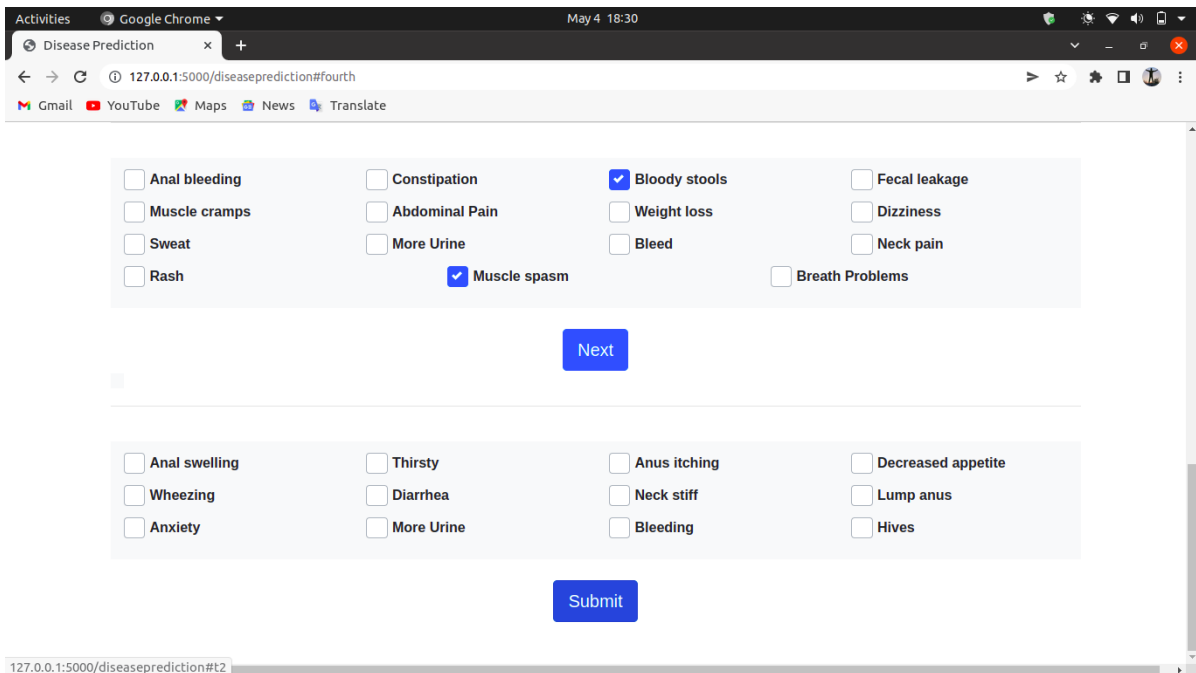
☐ Stiff Neck ☒ Low heartbeat ☐ Nausea ☐ Dry mouth

☐ Chest Pain ☒ Yawning ☐ Low breath ☐ Stool pressure

☐ Skin rash ☐ Swelling ☐ Fatigue ☐ Urine Blood

☐ Back Pain ☐ Painful Bowel moments

Next



Activities Google Chrome May 4 18:30

Disease Prediction

127.0.0.1:5000/diseaseprediction#fourth

Gmail YouTube Maps News Translate

☐ Anal bleeding ☐ Constipation ☒ Bloody stools ☐ Fecal leakage

☐ Muscle cramps ☐ Abdominal Pain ☐ Weight loss ☐ Dizziness

☐ Sweat ☐ More Urine ☐ Bleed ☐ Neck pain

☐ Rash ☒ Muscle spasm ☐ Breath Problems

Next

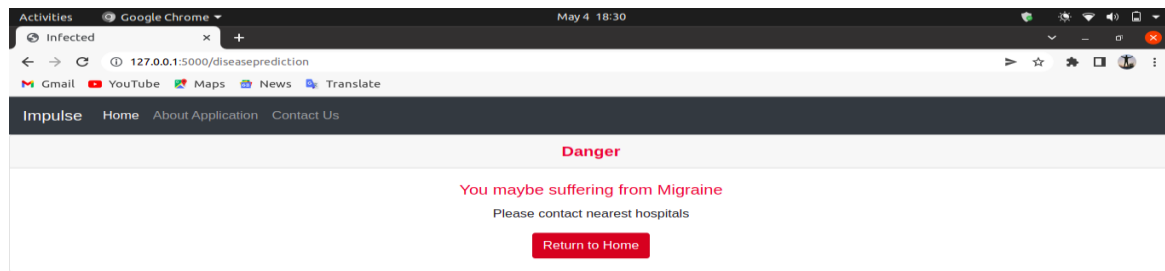
☐ Anal swelling ☐ Thirsty ☐ Anus itching ☐ Decreased appetite

☐ Wheezing ☐ Diarrhea ☐ Neck stiff ☐ Lump anus

☐ Anxiety ☐ More Urine ☐ Bleeding ☐ Hives

Submit

127.0.0.1:5000/diseaseprediction#t2



127.0.0.1:5000/home

## VII. CONCLUSION

Therefore, I have concluded that machine learning can be effectively used for tracking our health. From time to time we can check our health free of cost and be healthy. After making the machine learning model, I have deployed it with Flask (Python web framework) and in the future by creating that domain as a website, it will be available for anyone free of cost. The user just has to go to the respective website and have to select 5 to 8 diseases, so that our model can predict the best result. After getting the prediction user will get an idea about their health and if there are some serious situations they can contact their respective doctors. In this way, anyone in this world becomes a healthy person.

## VIII. FUTURE WORK

Every one of us would like to have a good medical care system and croakers are anticipated to be medical experts and take good opinions all the time. But it's largely doubtful to study all the knowledge, patient history, and records demanded for every situation. Although they've a massive quantum of data and information; it's delicate to compare and dissect the symptoms of all the conditions and prognosticate the outgrowth. So, integrating information into a case's individualized profile and performing in- depth exploration is beyond the compass of a croaker. So the result is ever heard of as a substantiated healthcare plan – simply drafted for an existent. Prophetic analytics is the process to make prognostications about the future by assaying literal data. For health care, it would be accessible to make the stylish opinions in the case of every existent. Prophetic modeling uses artificial intelligence to produce a vaticination from once records, trends, individualities, and conditions, and the model is stationed so that a new existent can get a vaticination incontinently. Health and Medicare units can use these prophetic models to directly assess when a case can safely be released.

## IX. REFERENCES

- [1] Maniruzzaman, M., Rahman, M., Ahammed, B. and Abedin, M., 2020. Classification and prediction of diabetes disease using machine learning paradigm. *Health information science and systems*, 8(1), pp.1-14.
- [2] Chen, M., Hao, Y., Hwang, K., Wang, L. and Wang, L., 2017. Disease prediction by machine learning over big data from healthcare communities. *Ieee Access*, 5, pp.8869-8879.
- [3] Haq, A.U., Li, J.P., Memon, M.H., Nazir, S. and Sun, R., 2018. A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. *Mobile Information Systems*, 2018.
- [4] Keniya, R., Khakharia, A., Shah, V., Gada, V., Manjalkar, R., Thaker, T., Warang, M. and Mehendale, N., 2020. Disease prediction from various symptoms using machine learning. Available at SSRN 3661426.
- [5] Dahiwade, D., Patle, G. and Meshram, E., 2019, March. Designing disease prediction model using machine learning approach. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1211-1215). IEEE.
- [6] Kaushik, K., Kapoor, D., Varadharajan, V. and Nallusamy, R., 2014. Disease management: clustering-based disease prediction. *International Journal of Collaborative Enterprise*, 4(1-2), pp.69-82.