DISEASE PREDICTION USING SYMPTOMS

A PROJECT REPORT

Submitted by

ARYAN PRATAP (18030141CSE056)

NAMITH BABU E (18030141CSE031)

SACHIN RAMPUR (18030141CSE067)

In partial fulfilment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

Under the Supervision of

Dr CHETAN J SHELKE

Professor

Department of CSE

Department of Computer Science and Engineering

ALLIANCE COLLEGE OF ENGINEERING AND DESIGN

Alliance University - Bangalore-562106

June- 2022

Department of Computer Science and Engineering

Alliance College of Engineering and Design

Alliance University

Chikkahagade Cross, Chandapura-Anekal Main Road,

Bangalore-562106

CERTIFICATE

This is to certify that the project work entitled **"Disease Prediction Using Symptoms"** is the bonafide work done by **Mr Aryan Pratap** (18030141CSE056)**, Mr Namith Babu E** (18030141CSE031)**, Mr Sachin Rampur**(18030141CSE067) submitted in partial fulfilment of the requirements for the award of the degree **Bachelor of Technology** in **Computer Science and Engineering** during the year 2018-2022.

Dr Chetan J Shelke — [SUPERVISOR]

Dr Abraham George — HOD

External Examiners:

1. Name: Signature:

2. Name: Signature:

Declaration

This is to declare that the report titled "**Disease Prediction Using Symptoms**" has been made for the partial fulfilment of the Course Bachelor of Technology in Computer Science And Engineering, under the Supervision of **Dr Chetan J Shelke.** We confirm that this report truly represents our work undertaken as a part of our project work. This work is not a replication of work done previously by any other person. We also confirm that the contents of the report and the views contained therein have been discussed and deliberated with the faculty guide.

NAME	REG NO	SIGNATURE
NAMITH BABU E	18030141CSE031	
ARYAN PRATAP	18030141CSE056	
SACHIN RAMPUR	18030141CSE067	

ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of the task would be put incomplete without the mention of the people who made it possible, whose constant guidance and encouragement crown all the efforts with success.

We are very thankful to our guide and Project Coordinator **Dr Chetan J Shelke,** Department of Computer Science and Engineering for his sustained inspiring guidance and cooperation throughout the process of this project. His wise counsel and valuable suggestions are invaluable.

We would like to thank **Dr. Abraham George,** Head of the department and **Dr. Reeba Korah**, Dean for their encouragement and cooperation at various levels of Project.

We avail this opportunity to express my deep sense of gratitude and hearty thanks to the Management of Alliance University, for providing world class infrastructure, congenial atmosphere and encouragement.

We express my deep sense of gratitude and thanks to the teaching and nonteaching staff at our department who stood with me during the project and helped me to make it a successful venture.

We place highest regards to my parents, my friends and well-wishers who helped a lot in making the report of this project.

NAMITH BABU E

ARYAN PRATAP

SACHIN RAMPUR

ABSTRACT

Modern technologies such as data analytics and machine learning have paved the way for healthcare innovation. As a result, we propose a disease prediction system that can anticipate potential diseases based on symptoms, allowing them to be treated at an early stage. It saves time by eliminating the need for a full diagnosis of the patient, and by relying on the system's recommendations, we can only diagnose the patient for the diseases that are needed. We're utilizing machine learning methods in this research to try to predict diseases accurately. The proposed approach generates findings that are up to 92 percent accurate. The system has enormous potential in terms of more precisely anticipating potential ailments. The primary goal of this research is to assist nontechnical people and new clinicians in forming accurate opinions on diseases.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

1. Introduction

It is a system that is made by using machine learning algorithms for guessing the possible diseases based on the symptoms of the patient. The growth of technology has been improving our lives so far. It provides many tools that can save millions of lives, and machine learning is one of them. Machine Learning is used to develop systems that can help us predict so many diseases based on symptoms. It can suggest the doctors, probability of the possible diseases. And diagnosis can be done based on suggestion, thus cost could be reduced. We are living in the age of technology and nowadays humans can say that almost anything is possible with the help of technology. Today we have so many tools and methods to access information from any region of this world and Information at this age is so important that without information we would not survive. We have tools that can give us or suggest relevant information at our fingertips and the internet is one of those tools. Today billions of search queries are performed daily and sometimes there given results are relevant and sometimes they are not. In those search queries, thousands of searches are related to medical advice. People often want to know if they have any serious diseases based on their signs and symptoms. But there are no tools available to give them proper information. This project tries to give them tools so that possible disease prediction information can be provided to the end-user at their fingertip.

2. Literature Review

There have been numerous studies done related to predicting the disease using different machine learning techniques and algorithms which can be used by medical institutions. This project reviews some of those studies done in research papers using the techniques and results used by them.

MIN CHEN et al, proposed a disease prediction system in his paper where he used machine learning algorithms. In the prediction of disease, he used techniques like CNN-UDRP algorithm, CNN-MDRP algorithm, Naive Bayes, KNearest Neighbor, and Decision Tree. This proposed system had an accuracy of

94.8%.

Sayali Ambekar et al, recommended Disease Risk Prediction and used a convolution neural network to perform the task. In this paper machine learning techniques like CNN-UDRP algorithm, Naive Bayes, and KNN algorithm are used. The system uses structured data to be trained and its accuracy reaches 82% and achieved by using Naïve Bayes.

Dhiraj Dahiwade et al, designed a model for prediction of the disease using approaches of machine learning and used techniques like KNN and CNN. This paper suggests disease prediction i.e., based on patient's symptoms. The accuracy of KNN is 95% and the accuracy of CNN is 98%.

Pahulpreet Singh Kohli et al, suggested disease prediction by using applications and methods of machine learning and used techniques like Logistic Regression, Decision Tree, Support Vector Machine, Random Forest and Adaptive Boosting. This paper focuses on predicting Heart disease, Breast cancer, and Diabetes. The highest accuracies are obtained using Logistic Regression that is 95.71% for Breast cancer, 84.42% for Diabetes, and 87.12% for heart disease.

CHAPTER 2

3. Dataset and Model Description

In our proposed system we are using structured datasets that can be created by collecting patient's symptoms and diagnosis from local hospitals and from opensource libraries available online. We are using true datasets that gives higher accuracy.

For this project the data has been taken from https://www.kaggle.com/itachi9604/disease-symptom-description-dataset

The data set required cleaning and structuring before further use. So, after preprocessing the data was used for the train test split.

```
df.tail(10)
```

Out[5]:

	Disease	Symptom_1	Symptom_2	Symptom_3	Symptom_4	Symptom_5	Symptom_6	
4910	Hypothyroidism	fatigue	weight_gain	cold_hands_and_feets	mood_swings	lethargy	dizziness	puffy_fac
4911	Hyperthyroidism	fatigue	mood_swings	weight_loss	restlessness	sweating	diarrhoea	fas
4912	Hypoglycemia	vomiting	fatigue	anxiety	sweating	headache	nausea	blurred_and_dis
4913	Osteoarthristis	joint_pain	neck_pain	knee_pain	hip_joint_pain	swelling_joints	painful_walking	
4914	Arthritis	muscle_weakness	stiff_neck	swelling_joints	movement_stiffness	painful_walking	0	
4915	(vertigo) Paroymsal Positional Vertigo	vomiting	headache	nausea	spinning_movements	loss_of_balance	unsteadiness	
4916	Acne	skin_rash	pus_filled_pimples	blackheads	scurring	0	0	
4917	Urinary tract infection	burning_micturition	bladder_discomfort	foul_smell_of urine	continuous_feel_of_urine	0	0	
4918	Psoriasis	skin_rash	joint_pain	skin_peeling	silver_like_dusting	small_dents_in_nails	inflammatory_nails	
4919	Impetigo	skin_rash	high_fever	blister	red_sore_around_nose	yellow_crust_ooze	0	

Fig 1. Data Set after preprocessing

3.1 Support Vector Machine (SVM)

SVM is well-known among data mining algorithms for its classification discriminative capacity, especially in circumstances where sample sizes are small and many features (variables) are involved (i.e., high-dimensional space). SVM is one of the most well-known classification supervised machine learning techniques. SVM training algorithm generates a model for a given collection of training data, each designated as belonging to one of two categories, by finding a hyperplane that classifies the given data as accurately as possible by maximising the distance between two data clusters.

We use the Support Vector Machine (SVM) model in the proposed system to forecast diseases based on patient symptoms.

SVM's Operation

An SVM model is simply a representation of separate classes in a hyperplane in multidimensional space. In order to minimise the error, SVM will iteratively construct the hyperplane. The goal of SVM is to divide datasets into classes in order to find the greatest marginal hyperplane **(MMH).**

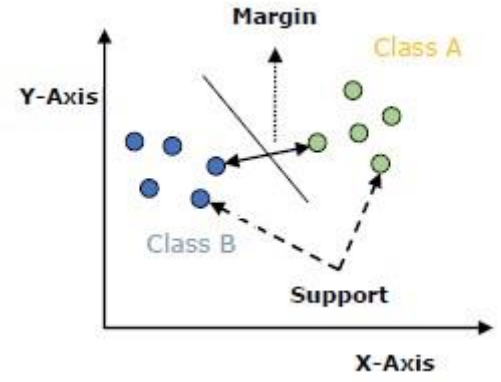

Fig 2: SVM's Operation

ThingSpeak is being used as the cloud. For simplicity, only when the diseases GERD or Hepatitis C are detected, the data is sent to cloud. The data sent to cloud are the predicted disease and the location of the user. Also, the data is later analyzed using analytical tool tableau.

4. Methodology

There are following steps involved in our proposed methodology:

First, I collected the datasets of symptoms and the diseases related to them.

Second, the data set is preprocessed for cleaning and structuring.

Next, the symptoms are encoded with their severity weights using the disease severity weight data set. The diseases and encoded symptoms were stored in separate data frames. Also, the model was checked for Accuracy (92.3%) and F1 score (93.08%).

A function using the model Support Vector Machine (SVM) was used to predict the disease from input symptoms. That may be possible for those acquired symptoms.

A GUI is implemented for ease of use which gets the symptoms from the user and predicts the disease using the model.

ThingSpeak is being used as the cloud. For simplicity, only when the diseases GERD or Hepatitis C are detected, the data is sent to cloud. The data sent to cloud are the predicted disease and the location of the user.

The data from ThingSpeak cloud is used further for analytical Tableau dashboard to visualize the disease predictions in various geographic locations along with the frequency.

The proposed model flow is shown in the fig 3.

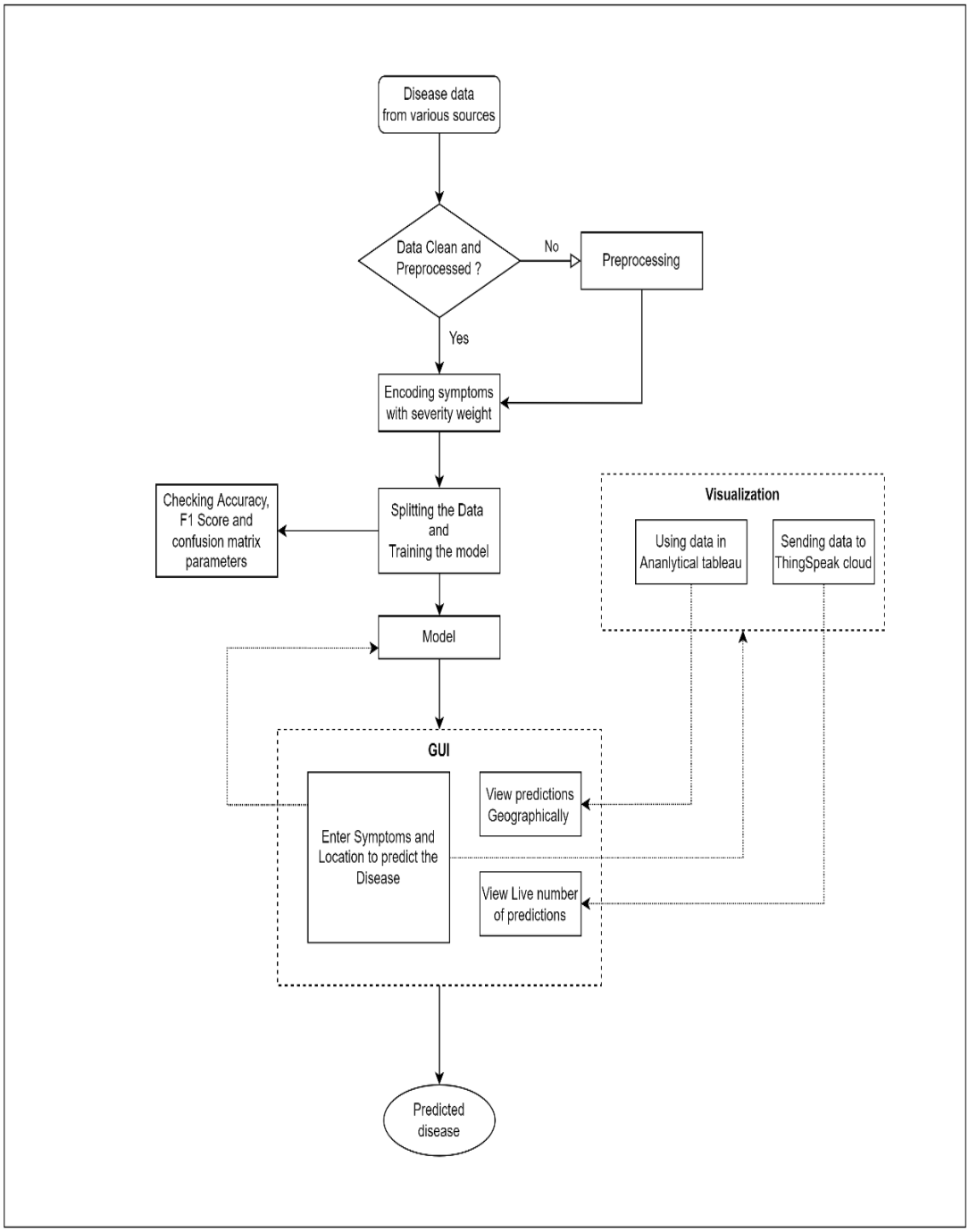

Fig 3 Flow chart of proposed method2.1 The GUI

The Disease prediction model is implemented in the form of a simple GUI for ease of use to the users where anyone can simply enter their disease symptoms and location and predict the disease. The GUI demonstration is shown below.

When the python script is executed, the following window pops up.

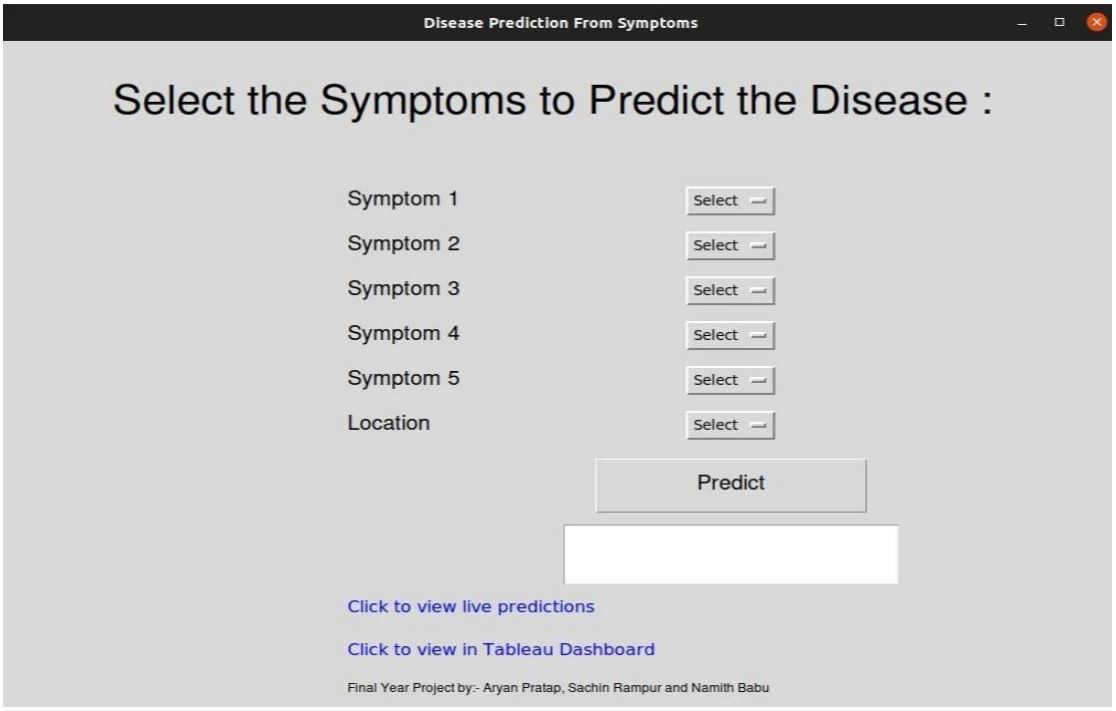

Fig 4. Symptoms selection.

The user needs to select each of the five symptoms from the drop-box selector. The user is also asked to enter their location which is further used in analytical purposes.

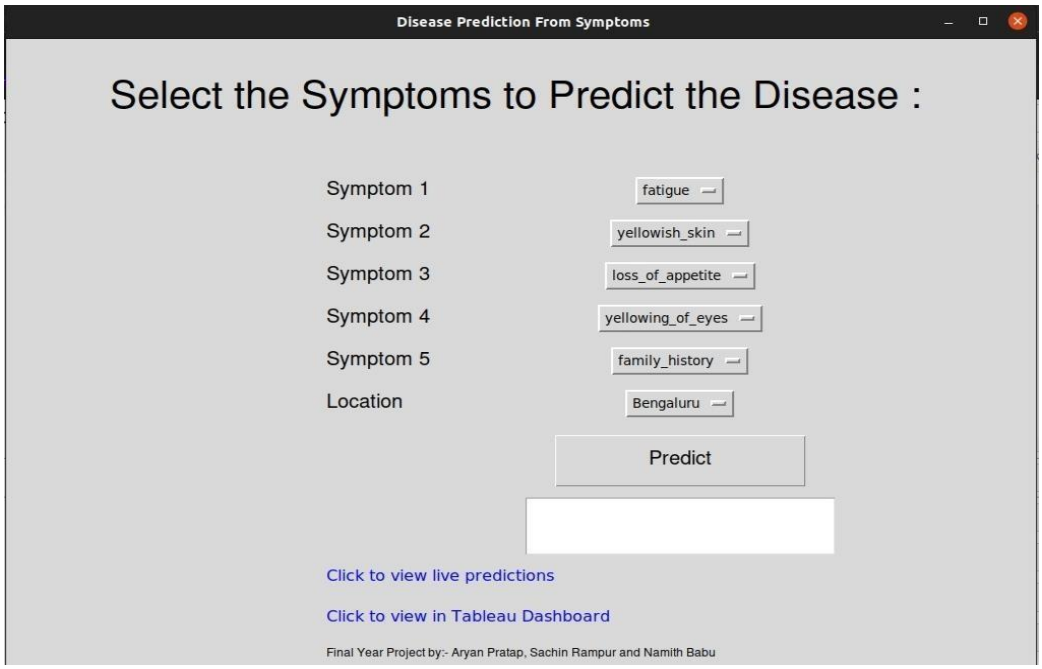

Fig 5. Symptoms and Location input

After entering the symptoms and location inputs when the Predict button is clicked, the symptoms are input to the model which predicts the disease from the given symptoms and the blank box below the predict button, as show in the figure below.

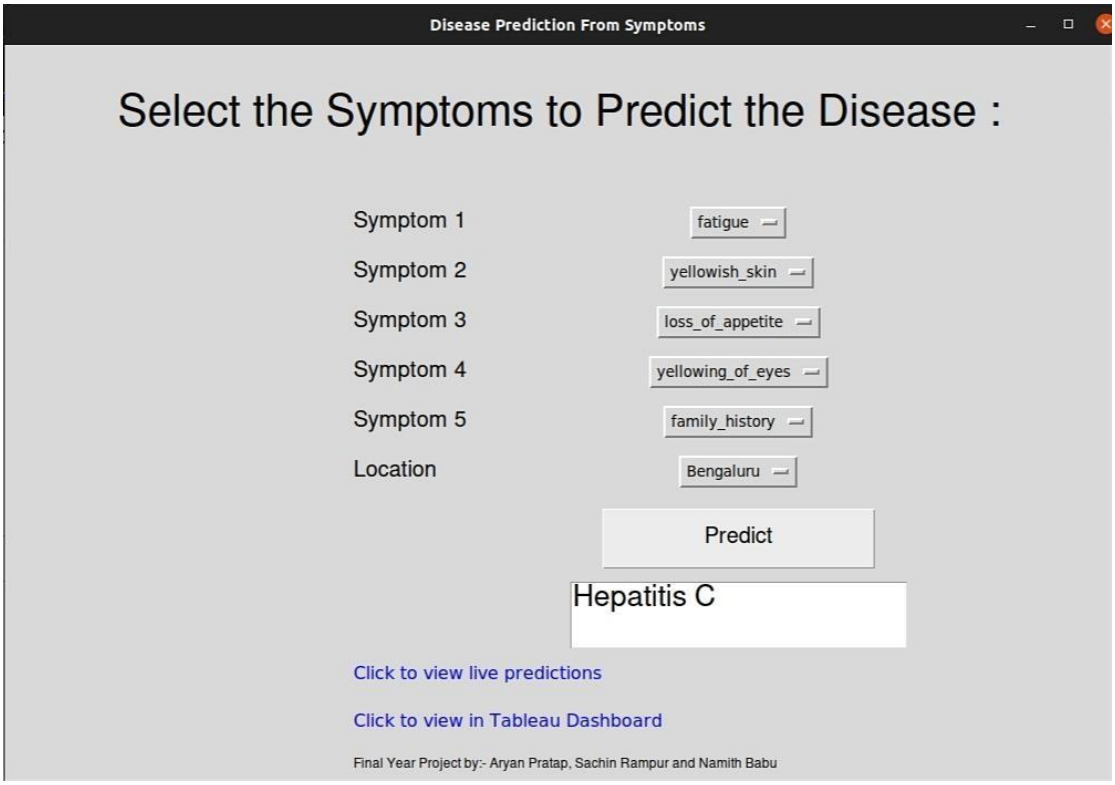

Fig 6. Predicted disease

For demonstration purpose, only the data of the diseases GERD and Hepatitis C is being sent to the cloud via ThingSpeak API which shows the live number of predicted cases of those diseases.

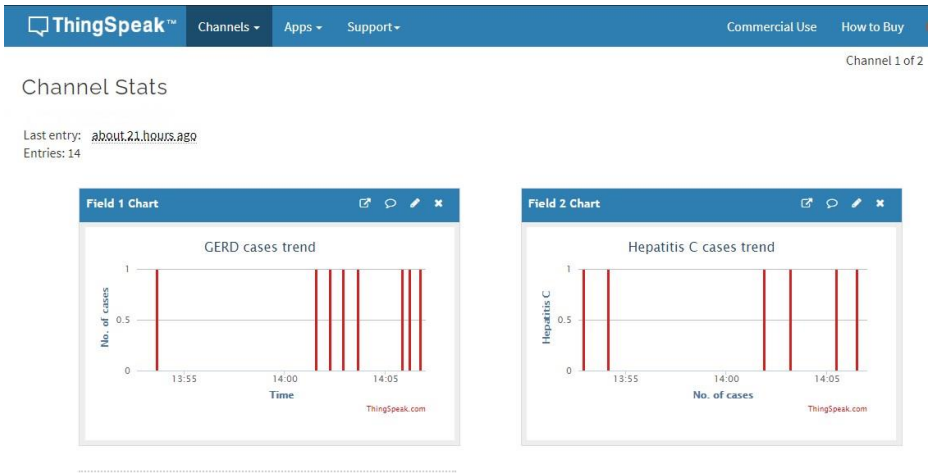

Fig 7. Live predicted diseases.

The predicted disease data from the project can be further used by analytical Tableau to visualize the prediction according to the location of predictions, as shown in the figure below.

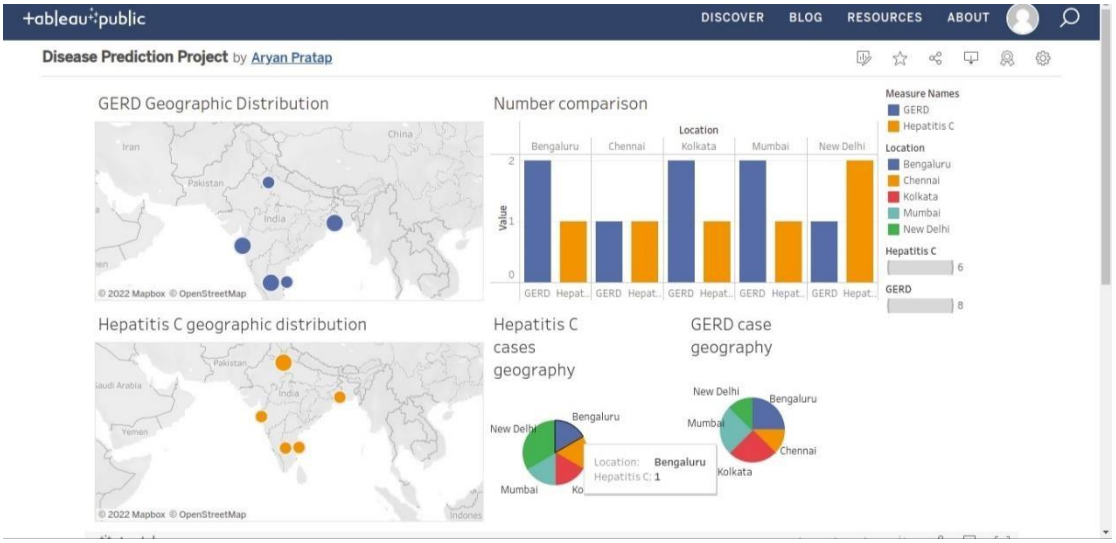

Fig 8. Analytical Tableau visualization

The data show is in the above image is not live and is updated on a period basis.

5. Performance Evaluation

5.1 Confusion Matrix

To evaluate the robustness of the estimates from the SVM models, Confusion matrix was performed. A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

The confusion matrix obtained is shown below

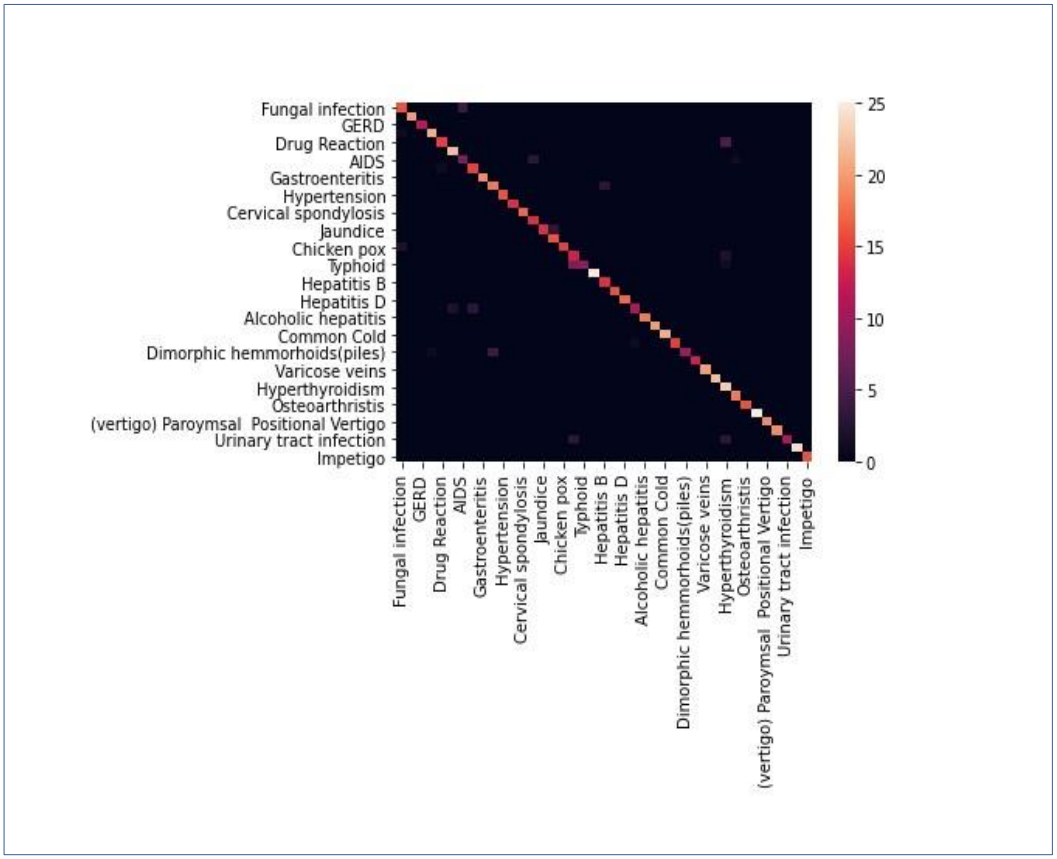

Fig 9. The Confusion matrix

5.2 Accuracy

Accuracy, which describes the number of right predictions over all predictions, is frequently used as the basis metric for model evaluation. The percentage of accuracy gained is listed below.

```
Accuracy% = 93.08943089430895
```

5.3 F1-Score

Precision and recall have a harmonic mean. It considers both false positives and false negatives. As a result, it works well with an unbalanced dataset.

Recall and precision are given equal weighting in the F1 score.

The percentage of the F1 Score attained is listed below.

```
F1-score% = 92.59780426441392
```

CHAPTER 3

6. Conclusion and Future Work

In our project, we have used a support vector machine algorithm to predict diseases. Despite being available and testing many algorithms I have found that using the support vector machine gives higher accuracy than other algorithms. The purpose of this project was to provide medical diagnosis information based on symptoms to normal people, fresher doctors, medical students, and anyone who wants to know about a set of symptoms and associated diseases. In this project, I have found that possible disease prediction can go up to 93% for some diseases and minimum 68% for some diseases but if we can feed the system humongous amount of data set then the accuracy of a disease prediction system can reach 95%. Obtaining a tremendous amount of data set related to diseases and their symptoms is very time consuming and it cannot be done within one or two years it requires multiple years to collect those data sets and train the system using those data searches. This system can be used by Ph.D. scholars to do further project. With the use of a disease forecasting system, it is possible to diagnose people based on symptoms. Disease prediction system provides only possible outcomes it does not guarantee that the disease will be predicted Accurately. But it has significantly higher accuracy for predicting possible diseases.

7. References

[1] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, “Disease prediction by machine learning over big data from healthcare communities” IEEE Access, vol. 5, no. 1, pp. 8869–8879, 2017.

[2] Sayali Ambekar, Rashmi Phalnikar, “Disease Risk Prediction by Using Convolutional Neural Network” IEEE, 978-1-5386-5257-2/18, 2018.

[3] Dhiraj Dahiwade, Gajanan Patle and Ektaa Meshram, “Designing Disease Prediction

Model Using Machine Learning Approach” IEEE Xplore Part Number: CFP19K25ART; ISBN: 978-1-5386-7808-4, pp. 1211-1215, 2019

[4] Pahulpreet Singh Kohli and Shriya Arora, “Application of Machine Learning in Disease Prediction” IEEE, 978-1-5386-6947-1/18, pp. 1-4, 2018.