2023-01-24

An Undergraduate Internship/Project on “ChatBot Assistant for Healthcare solution”

Sarkar, Bapti Niloy

Independent University, Bangladesh

An Undergraduate Internship/Project on

“ChatBot Assistant for Healthcare solution”

By

Bapti Niloy Sarkar

Student ID: **1820049**

Autumn, 2022

At

Supervisor:

Ajmiri Sabrina Khan

Lecturer

Department of Computer Science & Engineering

Independent University, Bangladesh

January 24, 2023

Dissertation submitted in partial fulfillment for the degree
of Bachelor of Science in Computer Science
Department of Computer Science & Engineering

Independent University, Bangladesh

Attestation

I, Bapti Niloy Sarkar, hereby certify that none of the work that has been done in this report is plagiarized. Any resources used are mentioned in the bibliography section of the report. This report is submitted in fulfillment of the internship requirement for the Bachelor of Science in Computer Science and Engineering degree at Independent University, Bangladesh. It has been constructed to serve as documentation for the professional internship experience that I was engaged in at “Grameen Communications” during the Autumn 2022 session. The only parties that have contributed in guidance of constructing this report are my internal and external supervisors at IUB and Grameen Communications.

Bapti

Signature

01.24.2023

Date

Bapti Niloy Sarkar

Name

Acknowledgement

I would like to express my gratitude to my internal supervisor Ajmiri Sabrina Khan, Lecturer, Independent University, Bangladesh. She has provided me with the guidance and suggestions necessary to conclude all research, industrial and documentation based tasks associated with my internship at Animo.AI.

I would also like to thank my external supervisor Ataur Rahman, Assistant General Manager, Head of Training and Development, Grameen Communications, for granting me the opportunity to contribute to the state of the art projects at the company and learn from this experience.

Lastly, I appreciate the authorities and regulatory bodies at Independent University, Bangladesh that are conducting the internship program and validating our internship experience through constructive criticism.

Bapti Niloy Sarkar

1820049

January 2023

Dhaka, Bangladesh

Letter of Transmittal

January 19, 2023
Ajmiri Sabrina Khan
Internal Internship Supervisor & Lecturer
Department of Computer Science and Engineering
School of Engineering, Technology and Sciences
Independent University, Bangladesh

Subject: Documentation of my work progression on “ChatBot Assistant for Healthcare solution”, a customize-able chat bot service which can predict diseases, developed for future deployment.

Greetings,

It is a pleasure to share this documentation, it includes a record of my work and contribution in the “ChatBot Assistant for Healthcare solution” project at Grameen Communications. The agenda of this report is to discuss the activities that I was involved with in the internship period and serve as fulfillment of the internship requirement for the Bachelor of Science in Computer Science and Engineering degree at Independent University, Bangladesh.

I would like to take this opportunity to express my gratitude for your time, expertise, guidance and support. I have made a notable effort in the preparation of this report and I hope that the purpose of this report is served successfully. Lastly, I am open to any relative queries.

Thanks & Best Regards

Bapti Niloy Sarkar

1820049

Evaluation Committee

. . . . . .

Signature 

. . . . . .

Name Ajmiri Sabrina Khan

. . . . . .

Supervisor of the intern

. . . . . .

Signature 

. . . . . .

Name Sanzar Adnan Alam

. . . . . .

Panel Member-1

. . . . . .

Signature 

. . . . . .

Name Sarwar Shahidi

. . . . . .

Panel Member-2

. . . . . .

Signature 

. . . . . .

Name Dr. Mahady Hasan

. . . . . .

Head, Department of computer science & engineering

Table of contents

Chapter 1

Introduction

Internships are often closely linked to the academic and career goals of each individual examinee. The purpose of the internship is to enable candidates to focus on their career interests and possibilities throughout the country.

Internship Engineering Technology and Science and Independent University School offers undergraduate programs in Computer Science and Engineering as a student of CS. The program requires me to complete an internship with a reputable company where I work. Training in Environmental Development and Familiarity with the Industry.

I have worked in a software company called Grameen Communications where I have completed the construction of my internship. The experience of working in that organization I have learned and how to make it a professional way for me.

In this report, I have discussed my internship period at Grameen Communications, an overview of the work I have done, my experiences working for a reputed organization, what I have learned and how it has helped me to develop and grow with a professional way this report, I have discussed my internship period at Grameen Communications, an overview of the work I have done, my experiences working for a reputed organization, what I have learned and how it has helped me to develop and grow with a profession.

1.1 Overview/Background of the Work

The project of creating a chat application is a new thing for me to start because I am doing it as an internship. Creating a Chatbot for my company was a whole new challenge for me to deal with. I am constantly learning from them, and I am facing the strange challenge of implementing them correctly. This is a common misunderstanding. I am learning Data preprocessing with Python & Julia, Machine Learning, Natural Language Processing, and Intelligent Human-Computer Interaction. I am facing rising challenges to implement them properly. I am using the Anaconda platform, Jupyter Notebook to train my datasets. I am learning to use existing datasets from kaggle. I am also learning to use streamlit because streamlit is the most effective platform to deploy DATA applications.

1.2 Objectives

Project objectives are the things we hope to accomplish by the project's conclusion. A project's objective must be precise, quantifiable, adhere to schedule and financial constraints, and most significantly, satisfy the client's needs.The main objectives of this application are described below:

- This app will help the clients by answering relative questions & building an interactive conversation.
- This Chatbot will be a B2C Service system.
- This application shall be made for company & possibly client use.
- This Chatbot shall be created for the company so that they don't have to use any other third-party application.
- It is a service to minimize the company's operational costing.

1.3 Scopes

This Chatbot assistant is a data application. The Dashboard shall have a text field where clients will ask questions related to health. The Dashboard will also have some descriptions & pictures related to the chatbot. The asserted scope is to fulfill the project objective within my internship period. This would enable Grameen Communications to move forward with full scale deployment of this service within their application.

Chapter 2

Literature Review

2.1 Relationship with Undergraduate Studies

Independent University, Bangladesh offers advanced courses in the domain of Intelligent Visual Computing that aided in the development and deployment of the POC project. The courses are as follows:

- **CSE 303, Database Management:** An introduction to database design and the use of database management systems. The course includes detailed coverage of the development process, database architectural principles, relational algebra and SQL using Oracle or SQL Server. Other key database topics covered are data modeling (E-R model, relational data model, integrity constraints, data model operations, normalization, object oriented data modeling), database security, administration and distributed systems.
- **CSE 317, Numerical Methods:** This course introduced me to the modern numerical approximation techniques necessary for in-depth data analysis.
- **CSE 417, Data Mining and Warehouse:** This course is a combination of data analysis techniques such as cleaning, transforming, warehousing, cubing, classification, clustering and mining frequent patterns
- **CSE 421, Machine Learning:** Supervised learning: Information theoretic decision tree learner, best current hypothesis search, candidate elimination (version space) algorithm, learning in the first order Horn clause representation, inductive logic programming, applications; Unsupervised learning: hierarchical clustering, category utility, incremental and non-incremental algorithms for hierarchical clustering, applications; Connectionist learning: introduction to neural networks, Feed forward and recurrent networks, perception, multilayer feed forward networks, backpropagation algorithm for training a feed forward network, applications; Genetic algorithms: genetic operators, fitness function, genetic algorithm in supervised learning framework, applications.
- **CSE 425, Artificial Intelligence:** This unit covers the foundational concepts and programming techniques of AI: search and problem solving methods, knowledge representation, reasoning, intelligent agents and natural language processing. Additional aspects of AI discussed include logic, uncertainty, puzzle solvers, simulative and cognitive process, expert systems and data processing.

2.2 Related works

This project **“ChatBot Assistant for Healthcare solutions”** is totally a new experience for me. I have zero experience regarding Natural Language Processing, and Intelligent Human-Computer Interaction. Although I can relate my university course outline as mentioned before. During these courses I learned how data preprocessing works, Dataset visualization, approaching problems on a machine learning view. Right now I am working with text to text interactive conversational mathots. I have learned about NLTK. Intelligent Human-Computer Interaction.

“CSE 317, Numerical Methods” from this course I learned how python & libraries of this programming platform works. That study outline is helping my progress of work very effectively.

I have Proficiency in the design of database applications starting from the conceptual design to the implementation of database schemas and user interfaces. Solid foundation on database design concepts, data models, the database query language SQL, and components of a database management system. Thanks to “**CSE 303, Database Management**”. This database management system course prepared me to handle such data as I am working on now.

Chapter 3

Methodology

3.1 Agile Approach

This project was conducted in an agile process. Agile processes are incremental in nature. It is conducted in 1/2 week sprints. Here, sprint begins with defining a backlog of activities, the completion of which should provide an output close to the goal of the project. After each sprint the outcomes are reviewed and a new backlog of activities is defined to reach a better outcome then the one in the first sprint. This is a repetitive cycle which ensures periodic improvement of the project outcomes.

3.2 Data Preparation

Data integrity is a very important factor at Grameen Communications, thus they didn't have any existing data for such a project. The data was collected from some outside sources. But Also the data was verified by some expert opinions.
The initial data had 4900 entry points. From there I had to clean the data for next steps. Then as for the feature importance. Training & testing was conducted. To conduct the process I had to go through some existing references related to healthcare. That way I could understand the symptoms related to prospective diseases. After preparing the data there were 41 data points found for the final disease list. And 131 data points for symptoms. Then each of the 41 disease data points the disease description was added.

3.3 Algorithms

In this project, **Decision tree map**, **Naive Bayes**, **Random forest** algorithms used to process both structured and unstructured data collected from hospitals.

3.3.1 Decision Tree Map

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical tree structure, which consists of a root node, branches, internal nodes and leaf nodes.
Compared to several typical calculating algorithms, the scheming accuracy of the proposed algorithm reaches 94.8% with a regular speed which is quicker than that of the unimodal disease risk prediction algorithm and produces reports.

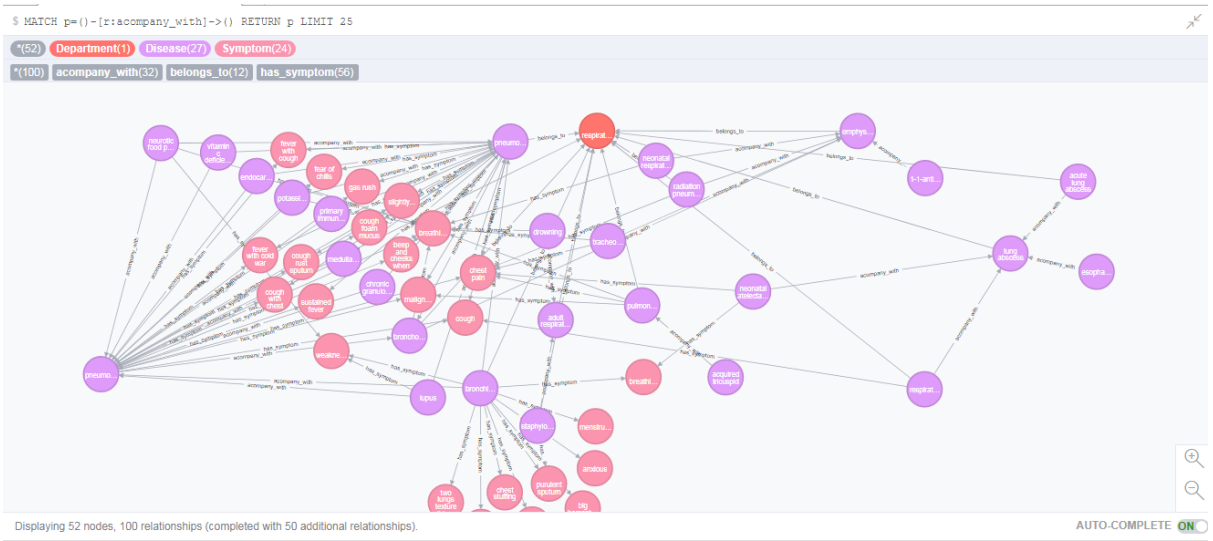


Fig-3.3.1: Decision Tree map showing predictions.

3.3.2 Naive Bayes classifier

Naive Bayes classifiers are a collection of classification algorithms based on 'Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

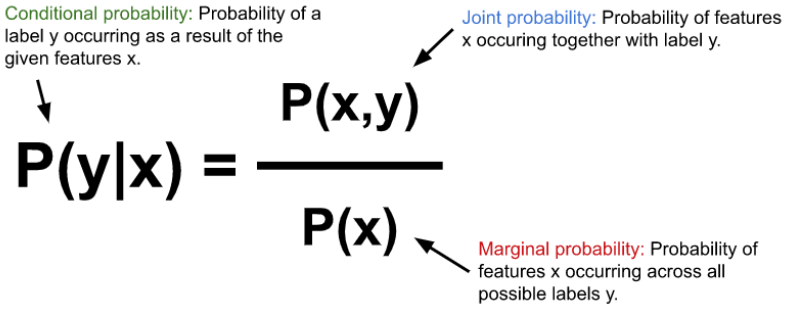


Fig-3.3.2: Naive Bayes theorem

3.3.3 Random forest classifier

Random forest is a supervised learning algorithm. The “forest” it builds is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result.

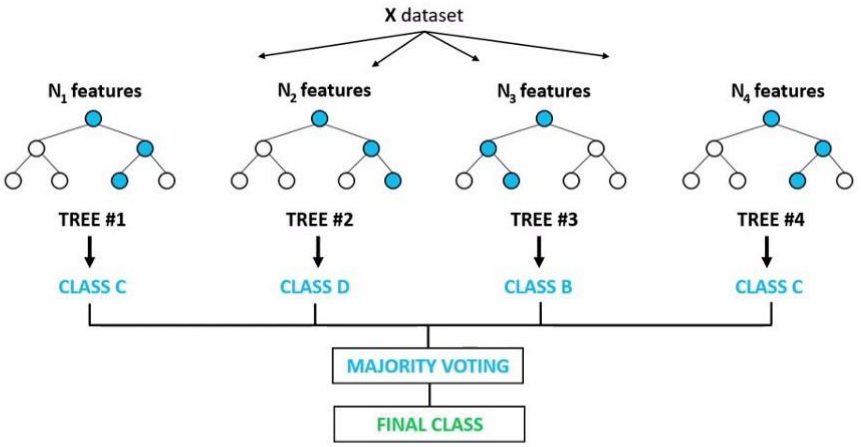


Fig-3.3.2: Random forest theorem

Chapter 4

Project Management & Financing

4.1 Work Breakdown Structure

The activities listed below were conducted on a bi-weekly basis :

- Data Acquisition & Preparation
- Model Configuration & Training
- Spill Validation
- Analyzing the Data
- Reviewing the pipeline to achieve better performance

4.2 Process/Activity wise Time Distribution

This is how the time of each agile sprint was distributed in between the backlog activities of each sprint:

- Data Acquisition & Preparation - 3 weeks.
- Model Configuration & Training - 2 weeks.
- Data Validation - 1 week.
- Configuring the statements of chatbot - 1 week.
- Analyzing the predictions and severity test result- 2 weeks.
- Reviewing the results to achieve better performance - 3 weeks.

4.2 Gantt Chart

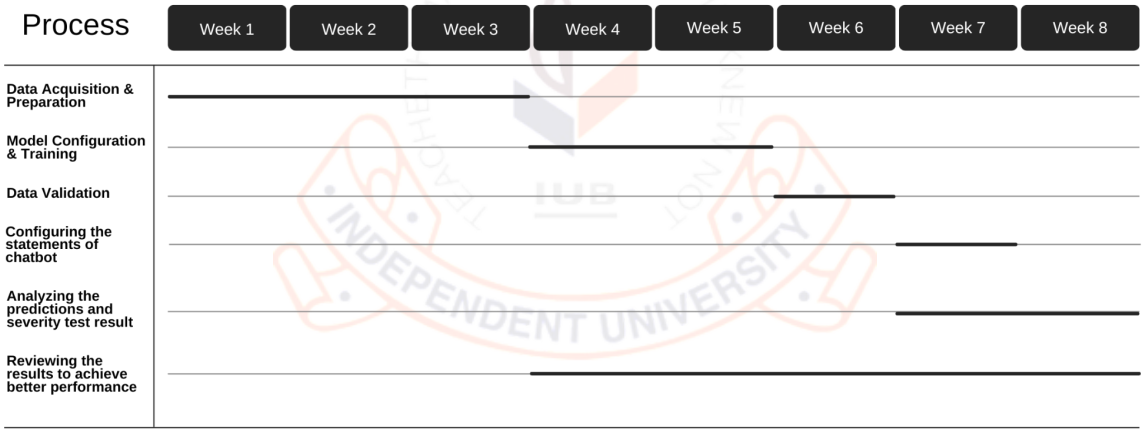


Fig-4.2: Time distribution gantt chart.

Chapter 5

Body of the Project

5.1 Requirement Analysis

5.1.1 Functional

- **Function** : Disease prediction.
- **Input** : Basic symptoms.
- **Process** : Get user info, get symptoms info, run pretrained model, target disease, target severity, target precaution.
- **Output** : Show severity result, show targeted disease description, show precaution.
- **Precondition** : Tokenizer and Decision tree model must be pretrained.
- **Postcondition** : User input must be accordingly.

5.1.2 Non-Functional

- Industry Grade Performance.
- Efficient result.
- Accurate data.
- Modular & Maintainable Solution.

5.2 System Analysis

5.2.1 Analysis of proposed utilization of the six elements

	Proposed System
Process	- Disease Prediction
Human	- No need for Doctors or Nurses for basic diagnosis. Patients can major the severity without taking the help of a human. Patients can also get the description of the predicted disease.
Non Computing Hardware	- Datasets were created based on the Doctor's report & previous history of symptoms.
Computing Hardware	- XYZ Industries Limited computer present in the server room is used to store and upload necessary data to kaggle. - My Personal computer used to download necessary data, prepare data, train model on Jupyter Notebook & spyder,fetch model, run tokenizer & decision tree to predict disease, target disease description, target precaution using Spyder and save the model locally. - A GTX 1070 TI is used for gpu acceleration on my personal computer.
Software	- Datasets are created on Excel Sheets. - Jupyter Notebook is used for better data visualization. - Spyder Is used for creating the python based Console application. - Django & Flask frameworks are used for creating the UI/UX.
Database	- SQL database is used to store all the data in the local server.
Network & Communication	- N/A

Fig-5.2.1: Analysis of current and proposed utilization of the six elements.

5.3 System Design & UI/UX

5.3.1 Proposed System Design

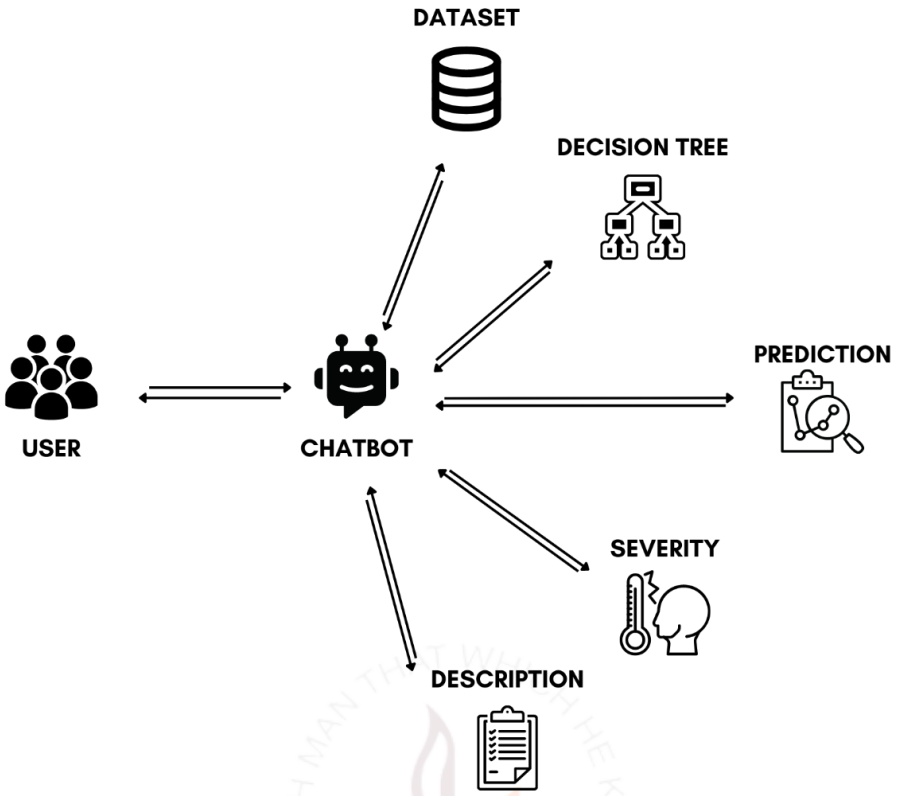
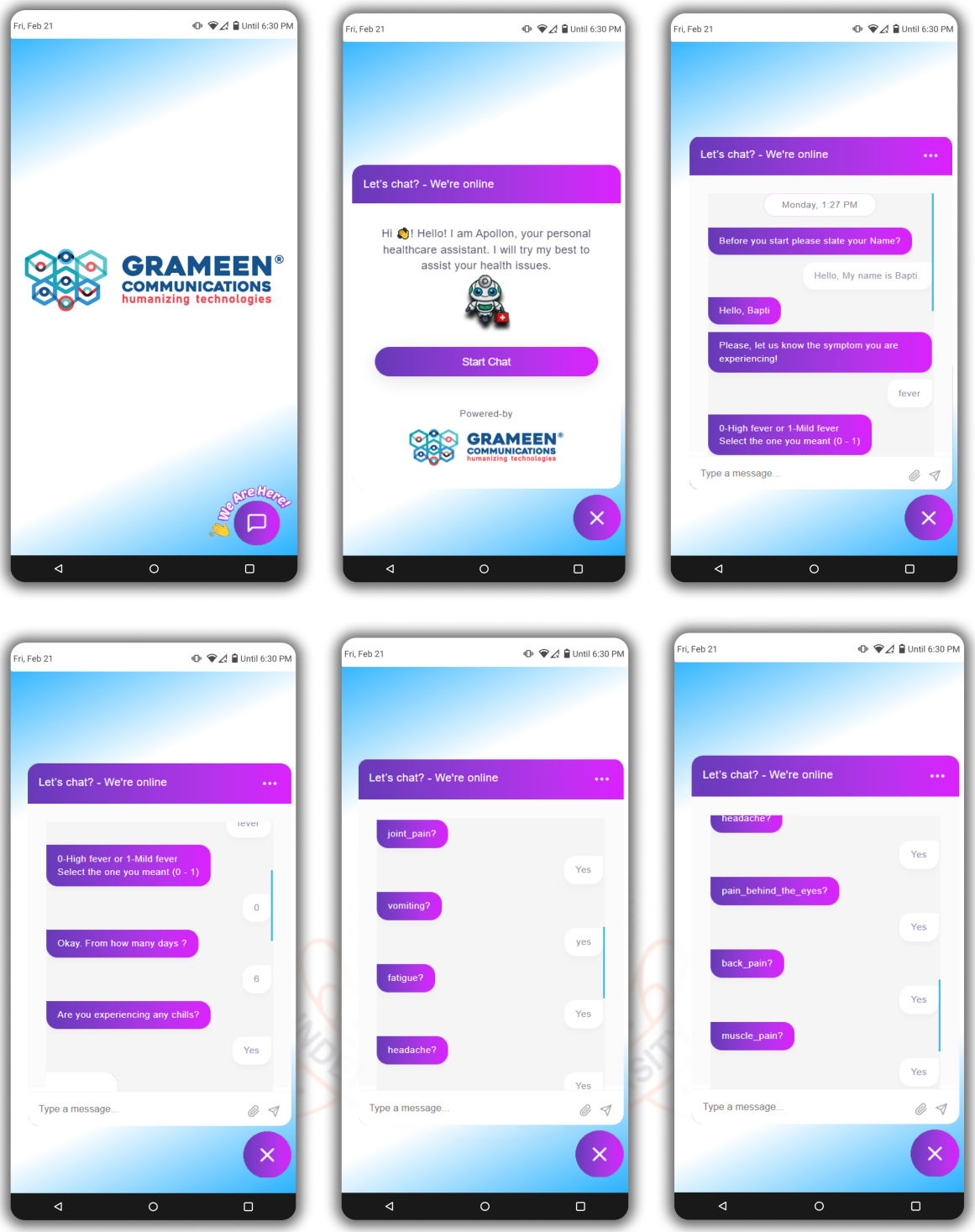


Fig-5.3.1: System design (proposed).

5.3.2 Proposed System UI/UX



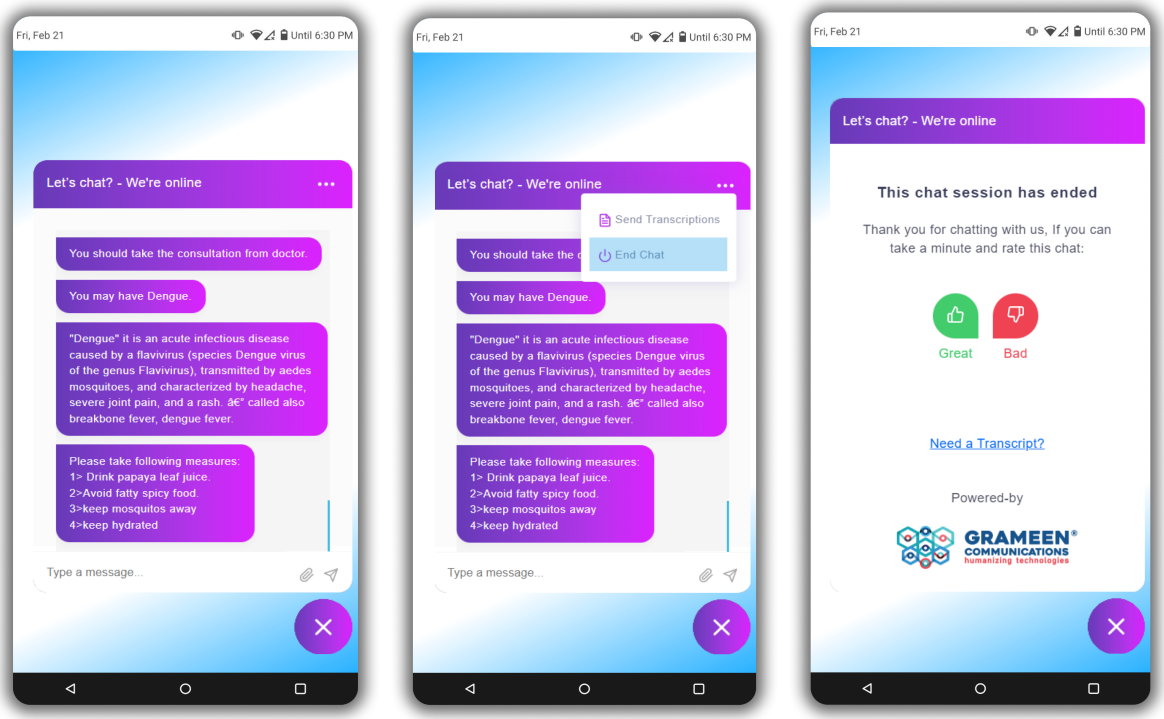


Fig-5.3.2: UI/UX Design.

5.4 Work Description

In the sprint run i worked with the following 41 categories of disease:

1. Arthritis 0.02439
2. Typhoid 0.02439
3. Malaria 0.02439
4. Dengue 0.02439
5. Hypoglycemia 0.02439
6. Osteoarthritis 0.02439
7. Dimorphic hemorrhoids(piles) 0.02439
8. Chronic cholestasis 0.02439
9. Hepatitis B 0.02439
10. Alcoholic hepatitis 0.02439
11. Hepatitis C 0.02439
12. Migraine 0.02439
13. GERD 0.02439
14. Gastroenteritis 0.02439
15. Common Cold 0.02439
16. (vertigo) Paroxysmal Positional Vertigo 0.02439
17. Drug Reaction 0.02439
18. Fungal infection 0.02439
19. Bronchial Asthma 0.02439
20. Chicken pox 0.02439
21. Hyperthyroidism 0.02439
22. hepatitis A 0.02439

23. Urinary tract infection 0.02439
24. Hepatitis E 0.02439
25. Psoriasis 0.02439
26. Acne 0.02439
27. Allergy 0.02439
28. Hypothyroidism 0.02439
29. Diabetes 0.02439
30. Peptic ulcer disease 0.02439
31. Hypertension 0.02439
32. Pneumonia 0.02439
33. Heart attack 0.02439
34. Cervical spondylosis 0.02439
35. Varicose veins 0.02439
36. Jaundice 0.02439
37. Paralysis (brain hemorrhage) 0.02439
38. AIDS 0.02439
39. Impetigo 0.02439
40. Hepatitis D 0.02439
41. Tuberculosis 0.02439

5.5 Output Demonstration

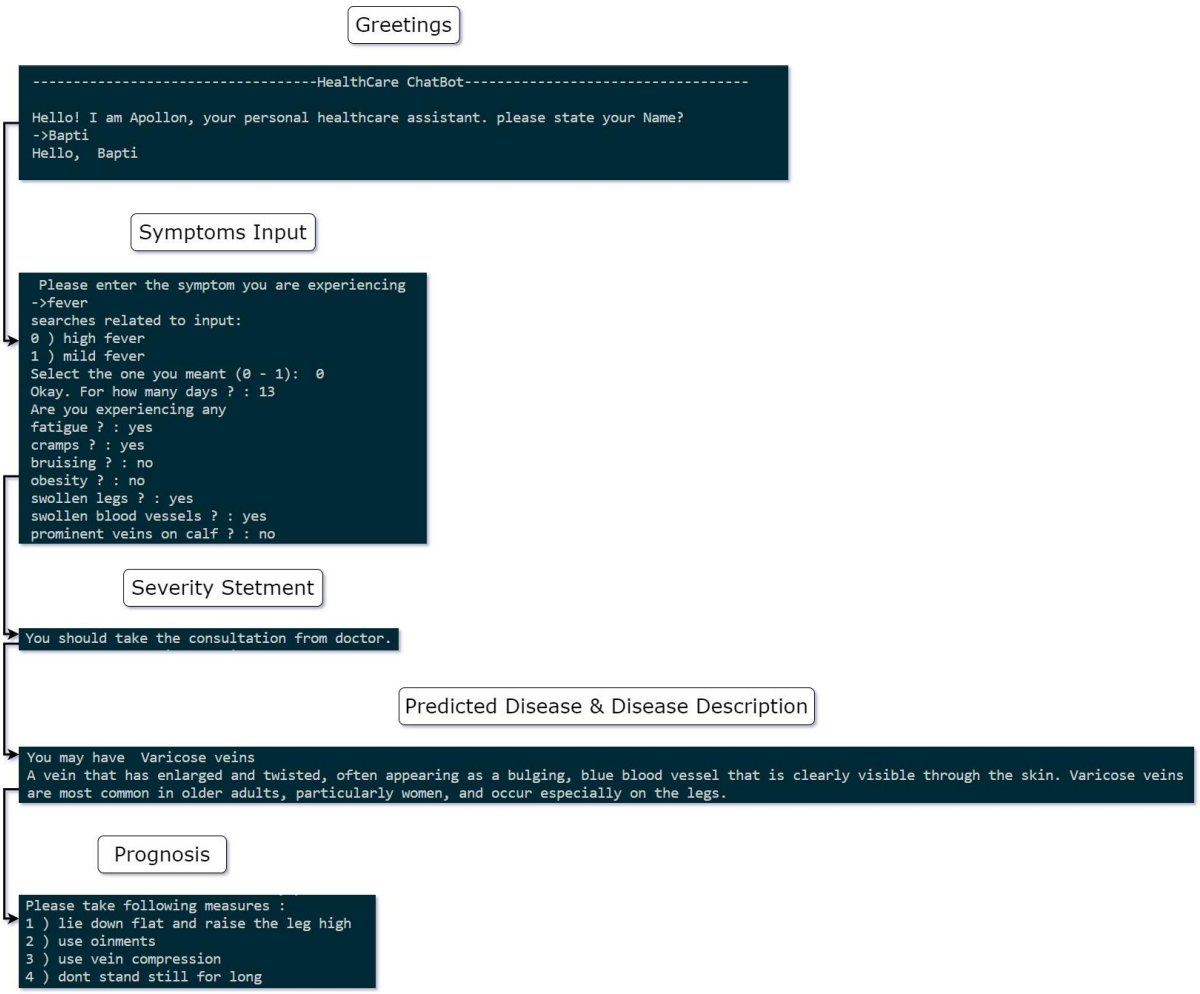


Fig-5.5: Console application demo.

Chapter 6

Results & Analysis

6.1 Sprint

6.1.1 Data Analysis

Features	Score	Features	Score
increased_appetite	0.00841	pus_filled_pimples	0.010377
burning_micturition	0.008418	toxic_look_(typhos)	0.011144
extra_marital_contacts	0.008467	high_fever	0.011179
depression	0.008549	lack_of_concentration	0.011237
internal_itching	0.008578	dark_urine	0.011283
movement_stiffness	0.008664	nausea	0.011498
knee_pain	0.008675	abdominal_pain	0.011662
stomach_bleeding	0.008678	weight_loss	0.011677
spotting_ urination	0.008726	malaise	0.012099
loss_of_balance	0.008732	bladder_discomfort	0.012194
continuous_feel_of_urine	0.00875	breathlessness	0.012449
rusty_sputum	0.008752	altered_sensorium	0.012662
coma	0.008796	loss_of_appetite	0.012747
watering_from_eyes	0.008814	mild_fever	0.01341
cough	0.009212	stomach_pain	0.013602
headache	0.009352	fatigue	0.013703
receiving_unsterile_injections	0.009424	red_sore_around_nose	0.013941
abnormal_menstruation	0.009441	unsteadiness	0.014127
mucoid_sputum	0.009516	joint_pain	0.014156
yellowish_skin	0.009761	sweating	0.014185
chills	0.009789	chest_pain	0.01437
diarrhea	0.009834	itching	0.014835
neck_pain	0.00987	yellowing_of_eyes	0.014897
sunken_eyes	0.0102	family_history	0.015274

Table 6.1: Number of train data points (symptoms) in each category.

Features	Score	Features	Score
Arthritis	0.02439	hepatitis A	0.02439
Typhoid	0.02439	Urinary tract infection	0.02439
Malaria	0.02439	Hepatitis E	0.02439
Dengue	0.02439	Psoriasis	0.02439
Hypoglycemia	0.02439	Acne	0.02439
Osteoarthritis	0.02439	Allergy	0.02439
Dimorphic hemorrhoids(piles)	0.02439	Hypothyroidism	0.02439
Chronic cholestasis	0.02439	Diabetes	0.02439
Hepatitis B	0.02439	Peptic ulcer disease	0.02439
Alcoholic hepatitis	0.02439	Hypertension	0.02439
Hepatitis C	0.02439	Pneumonia	0.02439
Migraine	0.02439	Heart attack	0.02439
GERD	0.02439	Cervical spondylosis	0.02439
Gastroenteritis	0.02439	Varicose veins	0.02439
Common Cold	0.02439	Jaundice	0.02439
(vertigo) Paroxysmal Positional Vertigo	0.02439	Paralysis (brain hemorrhage)	0.02439
Drug Reaction	0.02439	AIDS	0.02439
Fungal infection	0.02439	Impetigo	0.02439
Bronchial Asthma	0.02439	Hepatitis D	0.02439
Chicken pox	0.02439	Tuberculosis	0.02439
Hyperthyroidism	0.02439		

Table 6.2: Number of train data points (Disease) in each category

6.1.2 Visualizing Important Features

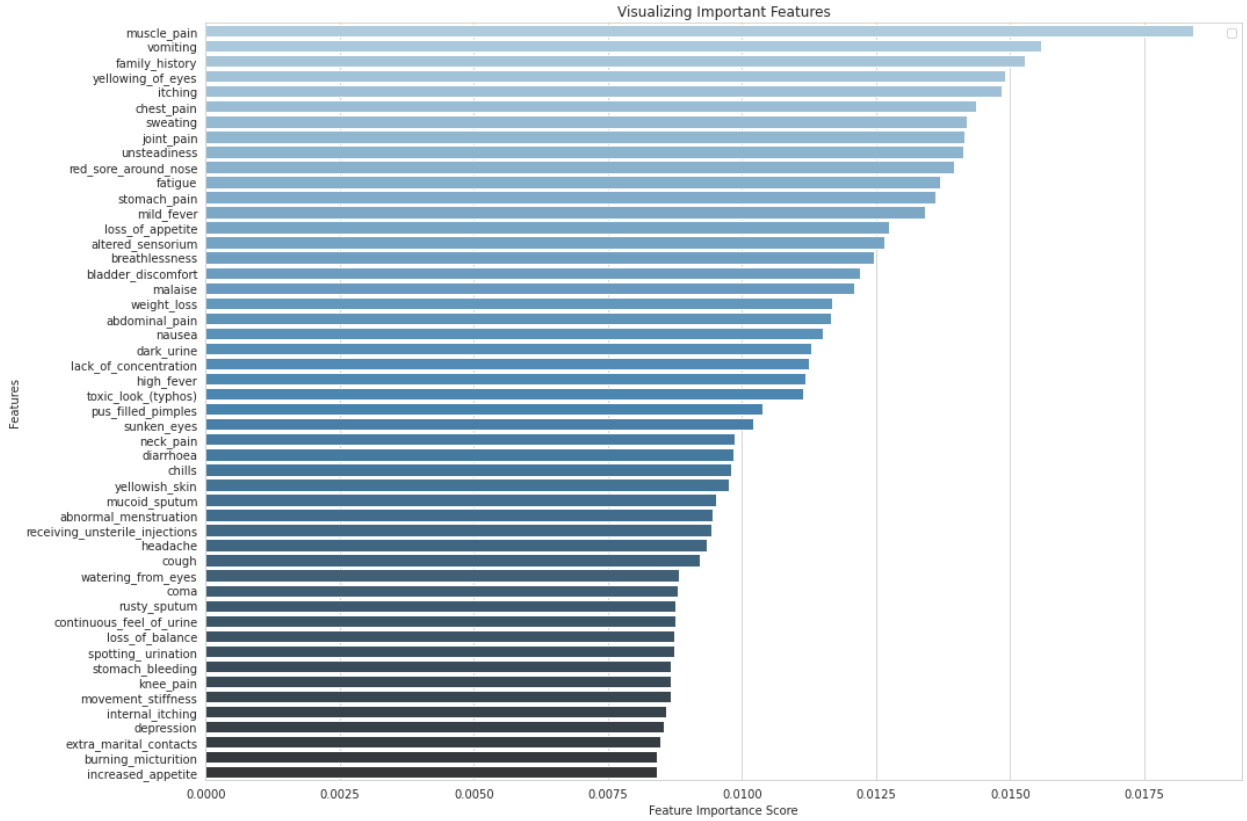


Table 6.3: Visualization of train data points (symptoms) in each category.

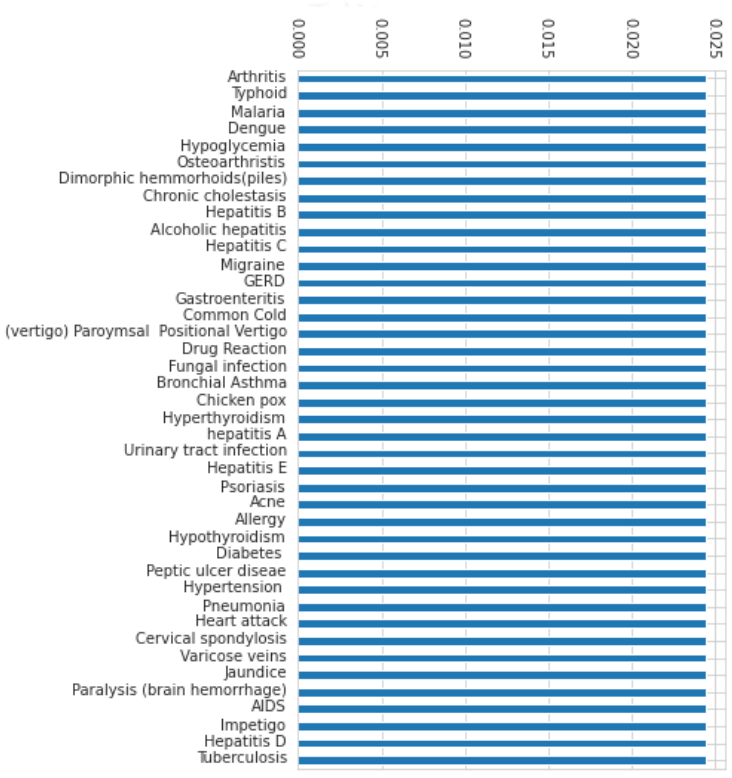


Table 6.2: Visualization of train data points (Disease) in each category

6.1.3 Accuracy

This project used both structured and unstructured hospital data to develop a machine learning **decision tree** map, **Naive Bayes**, and **random forest** algorithm. In order to segment the data, it also uses a machine learning technique. To the best of our knowledge, no recent work in the area of remedial big data analytics has focused on all data kinds simultaneously. Our proposed method's planning accuracy is 94.8%, which is faster than the unimodal illness risk prediction algorithm and provides a report when compared to various common calculating algorithms.

6.1.4 Sprint Analysis

After analysis it was clear that the data points were giving quite good results. The main reason for this was that we had over 4900 data entries & the ratio of Symptoms & Disease were (3.21:1). We implemented this dataset to train our chatbot. But for some reason this was not enough, According to some expert opinion. We need more research to find better datasets. They believe that the ratio of Symptoms & Disease is not efficient.

Chapter 7

Lesson Learned

7.1 Problems Faced During this Period

- Data acquisition & validation was a time consuming task due to lack of datasets from Grameen communications.
- As the project was followed by a supervised learning approach annotating ground truth for all objects in the sprint was a tedious task.
- The dataset had lots of invalid entries which needed to be cleaned.
- In the beginning the data was not clearly readable.
- Implementation was quite a difficult task because I had to study & do more research for such a project.
- Finding and implementing technologies for UI/UX was quite challenging as I didn't have any experiences.

7.2 Solution of those Problems

- As from the expert opinions Grameen communications must acquire more data to have much better results.
- Further research is needed with the agenda of predicting such diseases.
- A representative from the authorities at Grameen Communications should be briefed about the scenarios and the model. So, the future studies can carry on smoothly.
- As for the selecting technology part for UI/UX more python expert opinion is needed.

Chapter 8

Future Work & Conclusion

8.1 Future Works

Chat bots are a thing of the future which is yet to uncover its potential but with its rising popularity and craze among companies, they are bound to stay here for long. Machine learning has changed the way companies were communicating with their customers. With new platforms to build various types of chatbots being introduced, it is of great excitement to witness the growth of a new domain in technology while surpassing the previous threshold.

8.2 Conclusion

I would like to mention that this project has acquired the proof of concept accreditation from Grameen Communications authorities and this project will be continued to be deployed on full scale. For full scale deployment they will be performing future activities. They are now willing to find more scopes & develop this concept for future uprisings.

Concluding this report by once again acknowledging this enlightening experience for what it was and expressing my gratitude towards my internal and external supervisors for their guidance throughout this journey.

Bibliography

1. M. M. Rahman, R. Amin, M. N. Khan Liton and N. Hossain, "Disha: An Implementation of Machine Learning Based Bangla Healthcare Chatbot," 2019 22nd International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 2019, pp. 1-6, doi: 10.1109/ICCIT48885.2019.9038579.

2. L. Athota, V. K. Shukla, N. Pandey and A. Rana, "Chatbot for Healthcare System Using Artificial Intelligence," 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2020, pp. 619-622, doi: 10.1109/ICRITO48877.2020.9197833.

3. M. Bates, "Health Care Chatbots Are Here to Help," in IEEE Pulse, vol. 10, no. 3, pp. 12-14, May-June 2019, doi: 10.1109/MPULS.2019.2911816.

4. Bhirud, Nivedita, et al. "A literature review on chatbots in healthcare domain." International journal of scientific & technology research 8.7 (2019): 225-231.

5. M. Jovanović, M. Baez and F. Casati, "Chatbots as Conversational Healthcare Services," in IEEE Internet Computing, vol. 25, no. 3, pp. 44-51, 1 May-June 2021, doi: 10.1109/MIC.2020.3037151.

6. NVIDIA. (n.d.). Technical Blog. https://developer.nvidia.com/. October, 2022, from https://developer.nvidia.com/blog/faster-text-classification-with-naive-bayes-and-gpus

7. Kaggle. (n.d.). start_ML. https://www.kaggle.com/. October, 2022, from https://www.kaggle.com/getting-started/176257.

8. Regenerative. (n.d.). Machine Learning. https://regenerativetoday.com/. October, 2022, from https://regenerativetoday.com/explanation-on-how-decision-tree-algorithm

9. Adiwardana, D., Luong, M. T., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., ... & Le, Q. V. (2020). Towards a human-like open-domain chatbot. arXiv preprint arXiv:2001.09977.

10. Kaggle. (n.d.). DISEASE PREDICTION. https://www.kaggle.com/. October, 2022, from https://www.kaggle.com/getting-started/176257.

An Undergraduate Internship/Project on

“ChatBot Assistant for Healthcare solution”

By

Bapti Niloy Sarkar

Student ID: **1820049**

Autumn, 2022

At

Consent from Supervisor

The student modified the internship final report as per the recommendations made by his academic supervisor and/or panel members during and/or before final viva, and the department can use this version for archiving as well as the OBE course material for CSE499.

(Signature of the Supervisor)

Ajmiri Sabrina Khan
Lecturer
Department of Computer Science & Engineering
Independent University, Bangladesh