PERSONAL HEALTHCARE ASSISTANCE PROVIDING BOT - PHAB

Bhavana Sundar¹, Atishay Jain², Dr. M. Umadevi³

¹Computer Science Engineering, SRM Institute Of Science and Technology
Chennai,Tamil Nadu, bs4750@srmist.edu.in

²Computer Science Engineering, SRM Institute Of Science and Technology
Chennai,Tamil Nadu, as3845@srmist.edu.in

³Computer Science Engineering, SRM Institute of Science and Technology
Chennai, Tamil Nadu, umadevi.mu@ktr.srmuniv.ac.in

ABSTRACT

Personal Healthcare Assistance providing Bot (PHAB) is an artificially intelligent personal healthcare assistant. Due to the recent COVID19 pandemic, various social distancing norms have been imposed causing a hindrance in accessing healthcare facilities. This application aims to provide smart diagnosis while attempting to minimize COVID-19 transmission among patients and healthcare workers, by analyzing symptoms to diagnose, predict the medical conditions and generate helpful suggestions and precautions to be followed based on the inputs provided by the user. This medical care partner can serve to be exceptionally valuable during the pandemic and surprisingly post pandemic as it holds the possibility to help clients in need of supportive care without them having to anticipate their visit to healthcare facilities. As the application aims to perform accurate and comprehensive diagnosis, it also holds potential to serve as a pre-screening intermediary device to be used by healthcare officials.

Keywords: Digital Healthcare, COVID19 pandemic, Machine Learning in healthcare, Virtual assistance, Decision Trees

I. INTRODUCTION

With the Healthcare costs shooting skies with each progressing day, quality & affordable healthcare has grown into one of the major concerns & problems in the world. The accessibility of healthcare services and the hefty costs incurred in the process has caused a major number of deaths in the underlying regions of the world. Witnessing the COVID-19 pandemic at its peak, this issue has now taken its gravest form. With Digital Transformation & Artificial Intelligence (AI) capabilities at its peak, coupled with advanced medical knowledge and expertise, this paper hopes to pave a way for a breakthrough to solve one of the second world countries’ major crises. Even though maximum medical facilities have been deployed to fight the virus, and infection spreading each consecutive day, the per-patient availability of medical treatment is deteriorating drastically. And adding onto that, “social distancing” as a practice acts as an added blocker to the same. In such trying times with no hope of improvement, AI-powered Personal Healthcare Assistant can change the face of the medical industry.

Personal Assistance is the need of the hour given the unfortunate circumstances of a pandemic. It can be defined as a one-stop-shop for all the healthcare services provided on a platform with an interactive user interface. Thus, Personal Healthcare Assistants will provide timely care to the users at minimal costs. Thus, Personal Healthcare Assistants aim to provide cost-effective, easy, and instant diagnosis.

The information delivered by the National Health Mission demonstrates that a decrease in the instances of clinical hospitalization shows a deficiency to medical care access, instead of a genuine absence of access. In the present circumstance, Personal Healthcare Assistants go about as an alleviation for the average folks. By using misleadingly smart models, clinical consideration authorities can meet and break down patients without a necessity for any genuine presence or contact, while propelling social removing and lessening the peril of COVID-19 transmission. As well as defeating the market interest of medical services provisioning we likewise plan to assist convey with trip various protections, like the social removing guidelines constrained by the public
position, simultaneously recollecting the doubtful attitude of patients concerning visiting jam-pressed offices at this shocking time. Thus, with this application we expect to overcome any barrier between the admittance to medical care offices, give exact determination and interviews distantly to permit ideal conclusion and hazard free diagnosis, consequently viably helping us assembling and advancing a solid society.

II. RELATED WORKS

The field of healthcare is known for having many ongoing research projects and their survey is extensive. While writing this paper we have considered the time frame of 2018-2019 where comprehensive research has been done. The motive is to leverage and display the analytical process & research conducted by various other papers to conclude with suitable algorithms for quick and accurate prediction. [5]. In 2019, J Clin Med presented a paper wherein, which drew a parallel between scikit-learn library and auto-sklearn on a large dataset derived of cardiovascular disease. He successfully presented that automatic machine learning produce classifiers displayed a better performance than any other machine learning algorithms. The following contributions were made by the paper: For the first time, we witnessed that AutoML is being leveraged to build classifiers of cardiovascular diseases is looked into deeply and extensive experimented on over two cardiovascular datasets provided [1]. In 2018, Mr. Chala Beyene, Prof. Pooja Kamat presented a paper in which they did a detailed study and investigation of different algorithms on heart disease predictive analysis. They have done an elaborate analysis on choosing the algorithm that can help in making a better & quicker decision. They concluded that the Decision tree algorithm, Naïve Bayes and Support Vector machine algorithm will be the right choice of algorithms that will provide quicker & more accurate analysis & diagnosis. [2]. In 2018, Harini DK and Natesh M presented a paper which discussed about a system that combined the structured and unstructured data in the medical field. Leveraging a latent factor model, they completed the analysis to survey the danger of a disease. Conclusions included that KNN, Naïve Bayesian and Decision Tree help process structured data whereas unstructured can be processes using CNN algorithm. Considering they had both structured & unstructured datasets, CNN-based multimodal disease risk prediction (CNN-MDRP) algorithm was finally proposed [3]. In 2018, Vinitha S, Sweetlin S, Vinusha H and Sajini S presented a paper that elaborated on a cost-effective system that uses Machine Learning algorithms such as Map Reduce & Decision Tree to enhance the percentage accuracy of the analysis of medical big datasets. As-is working is majorly concentrated on the structured data. While to predict diseases, Decision Tree algorithm is used, they heavily depend on Map reduce to reduce the time consumption and improve the query resolution turnaround time (TAT) [4]. In 2019, Shahadat Uddin, Arif Khan, Md Ekramlul Hossain and Mohammad Ali proposed a new paper that identified the systems that leveraged more than one supervised machine learning algorithm on single disease prediction. It was concluded that the Support Vector Machine (SVM) algorithm is applied most frequently followed by the Naïve Bayes algorithm (in 23 studies). However, the Decision Tree algorithm showed superior accuracy.

III. PROPOSED WORK

In existing frameworks, the informational index regularly comprises of patients and diseases with quite certain conditions. These applications and bots are mostly designed and implemented keeping crucial and chronic diseases at the center of it such as Cardiovascular arrests, Cancer etc. Considering the current unfortunate situation, patients would not only be hoping to skip long lines at healthcare facilities but also anxious about visiting these centers in the first place. Currently available applications are ill-managed and cause the patients to delay their treatment because the user needs to input long and lengthy questionnaires, making it incredibly user unfriendly. To summarize, the existing systems do not focus on providing diagnosis of common health conditions like the flu, common cold, etc., do not emphasize on user friendly interactions irrespective of the age and are usually not easily affordable.

With the proposed system, the centre of attention is to provide user specific analysis; get all significant data through the conversational user interface, like their age, gender, etc. The symptom-based disease classifier takes in a series of questions, where the user can choose from one of the many listed symptoms or specify problems manually, to better predict the user’s problem. The results generated are followed by general medical tips and precautions and whether or not consulting a doctor immediately is necessary. The back-end of this condition-put together classifier is prepared with respect to a broad arrangement of clinical datasets. This framework doesn't guarantee the specific finding of the ailment of the client however it plans to give an answer that is just about as precise as could really be expected. Once the prediction of the medical condition is done, the system also aims at giving users an estimate of the severity of their condition, so as to determine their next course of actions. Leveraging a well-aged and trained machine learning model on the extensive medical datasets along with other

www.turkijphysiotherrehabil.org
physical and medical records/parameters gathered from the patient. It combines and integrates these two data points to predict quicker and better.

In this manner, we mean to propose such a framework which will have a straightforward, rich and easy to use interface. Another attractive feature about our system, along with its ease of use would be its time efficiency. This system is primarily proposed with the idea of it acting like a pre-screening device, thereby speeding the entire process for both the doctor and patient. In some cases, like common cold or flu, the doctor’s intervention may not even be necessary. In such cases, the user can stick to the diagnosis and omit visiting healthcare facilities and clinicians. Accordingly, focusing on the results of the pandemic and the equilibrium shift between the interest and supply of the right now accessible medical care administrations, we propose an easy to understand individual medical care partner, which with simply a tick of a catch, give solid and exact diagnostics of regular infections, recommending master home cures and preventive estimates given that the indications and illnesses confronted are all around imparted by the client to the interface.

Proposed System

IV. METHODOLOGY

For the ease of understanding and implementing the entire system, we divide it into various modules.

A. Creation of the User and Doctor Profile:

The planning and design of the User Interface (UI) comprises of fundamental data like the username, email, password, age and the user’s sexual orientation. In case of the doctor’s profile that is created, the application takes in fundamental information like the doctor’s name, specialization and lets them add further diseases or symptoms to our database. The adding of diseases and symptoms is only encouraged if they are not already present in the database and the doctors can classify the severity of the disease by marking it as High, Low or Medium risk. This will further be taken into consideration while computing the risk or severity factor of the diagnosis. Thus this module mainly implements the validation so that only approved and authorized clients enter the application. The UI is made interactive and extremely user friendly without any fancy features with the aim to present a simple, elegant and easy This then produces a userid interfacing the individual client to the sqlite database which stores all the information to be questioned and queried. This module, thus mainly focuses on taking personal details as input from the user related to their age, gender, etc. All the other options for answering the questionnaire to understand the user’s condition better are listed in simple words where the user can either agree to having the symptom, mark it as don’t know or deny it. This questionnaire is generated after the user enters the symptoms from the drop-down dialog box. If the user doesn’t know the scientific name of the symptoms, he or she is facing then they can easily choose the problem area from the human body image. Once they choose the problem area then the list of possible symptoms associated with that area are generated in a drop-
down box from which they can choose their problem they are facing. This is done, keeping the user’s gender in mind, making the process very interactive and user friendly.

B. Data Pre-Processing and Predictive Analysis:

The fundamental focal point of PHAB is to help the clients by giving as exact analysis as conceivable relying upon the manifestations indicated before. This stage incorporates some essential inquiries which when posed to the clients they answer back to the intuitive interface by choosing or physically giving their answer. These questions help in the process of understanding their medical condition, and thus providing them a with a set of possible medical ailments followed by the treatment procedures according to the answers filled along with the decision tree model powered predicted disease. This can additionally help in generating a medical transcript which physicians and doctors can use as a pre-screening report. The decision tree model is trained on two datasets. The first dataset is an information data set of sickness side effect affiliations created by a computerized technique dependent on data in text based release outlines of patients at New York Presbyterian Hospital concede during 2004. This dataset is uncleaned so pre-processing is done and then model is trained and tested on it. The information pre-handling procedure encodes the information to a structure which comprises of helpful fields rather than invalid and trash esteem so information can be effectively deciphered by the calculation. It is a preliminary stage which is employed to ensure accurate results. Thus, the dataset is cleansed by employing the process of filling in missing values, ensuring there are no inconsistencies present in the dataset. The dataset taken into consideration had non-ascii character entries (UTF-8 encoded values) for example: cushingoid\xa0habitus. Thus, all the three columns present in the dataset namely - ‘Source’, ‘Target’ and ‘Weight’ were decoded from the UTF-8 format to ASCII format. The second dataset that we have considered, was scraped and created manually. This dataset is clean and extensive and hence the learning was more accurate. It consists of a total of 405 symptoms, which after detailed analysis and computation leads to approximately 149 diseases. In view of roughly 30,000 records of patients, we mean to build up a forecast model that takes in the indications from the client and predicts the illness as precisely as could really be expected. When the choice tree is prepared with the preparation set, a standard set is framed and when the client indications are given as a contribution to the model, those side effects are handled by that standard set, hence making groupings and foreseeing the sickness.

C. Generating the Diagnosis:

The final step is to compute the degree of risk i.e., to let the user know how critical his or her condition really is. The degree of risk is computed depending on various factors and is primarily used to recommend the user to consult a doctor based on the severity of their condition which is indicated by this. The resulting diagnosis can be classified into three categories namely – low risk, medium risk and high risk or critical condition. To compute the risk factor, we have categorised all the diseases stored in our database after obtaining it from the database and mapped it to a respective degree. This degree is denoted by High (H), Low (L) or Medium (M). This degree has been assigned after thorough analysis of the disease and the organ it affects. Doctors specify this degree while they register themselves with our application in case of new diseases that aren’t common and already present in the database. The low risk factor assures the user that their condition isn’t severe and can be treated by following the suggested remedial measures at home. The medium risk factor recommends the user to get an appointment with the suggested specialist whenever possible and also aids the user with their treatment at home by providing helpful tips. The critical condition is when the user is showing symptoms of possible chronic illnesses and needs immediate professional medical attention. Thus, PHAB not only provides with the user with a list of possible diagnosis but also displays a link for home medication or remedies with each possible disease classified. Each disease diagnosis when displayed, explains the user their possible condition in lay-man terms, thereby making it easier for them to comprehend and understand their situation. The results being displayed are fetched from the disease diagnosis database, which contains an accurate description of the condition followed by the severity of the disease and an online link to certified online healthcare websites like webMD and Mayo Clinic, mentioning helpful home-remedies that the user can take into consideration. If the user still finds it appropriate to visit a clinic, PHAB also shows recommended hospitals near the user’s location. Thus, the user can choose to generate a report of the diagnosis and show it to the recommended doctor in the recommended hospital. Thus, a web service, which we propose as PHAB, is created in Flask, for disease prediction using the trained and tested decision tree model and is deployed to be accessed with ease whenever essential.
The disease prediction system, performing predictive analysis, is deployed using the data mining algorithm - Decision tree, which acts as a classifier here. Input to the model is given in the form of textual symptoms. The arrangement models worked by the trees take after the design of a real tree. By learning the game plan of the iterative assuming rules on select segment regards which for our circumstance will be the symptoms, it isolates the dataset into an ever increasing number of unobtrusive subsets that in the end will achieve expecting a target worth and end i.e the infection. A decision tree consists of the Root Nodes and Leaf Nodes. A node that is neither a root node nor a leaf node is called the Decision Node. Thus, A Decision node can be defined as the node that has two or more branches. For our situation, every one of the side effects are considered as choice hubs. The Leaf Node addresses the grouping that is, an official choice or result of the individual branch. Here the diseases compare to the leaf hubs.

The calculation that we have carried out in our work is the ID3 calculation. The calculation is a top down calculation which executes the avaricious pursuit method to look through the sections, where every segment addresses a characteristic which are side effects for our situation. Each hub is tried and the trait or indication that is best for order of a given set, is chosen. To pick which manifestation is ideal to fabricate a choice Tree, ID3 essentially utilizes the use of two ideas in particular Entropy and Information Gain

Entropy can be characterized as the measure of vulnerability which assists the model with characterizing the consistency of a specific occasion. Consequently, Entropy $E(C)$ is characterized as:

$$E(C) = \sum_{h \in H} -P(h) \log_2 P(h)$$  \hspace{1cm} (1)

Data acquired is addressed as Information Gain $IG(C, A)$ when a state C estimates the general change or delta(decrease) in entropy concerning the symptoms:

$$IG(C, A) = E(C) - E(C,A)$$  \hspace{1cm} (2)

Where $E(C,A)$ is the entropy determined for two ascribes C and A where C is the present status with characteristic A and A is the quality getting looked at.

Consider the accompanying illustration of 12 patients who were inclined to Dehydration and Diarrhea. The side effects appeared by them were: Fever, Vomiting, Chills and Dizziness.

Architecture Diagram
Sample Medical Records

<table>
<thead>
<tr>
<th>Fever</th>
<th>Vomiting</th>
<th>Chills</th>
<th>Dizziness</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Dehydration</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Diarrhea</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Diarrhea</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Dehydration</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Dehydration</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Dehydration</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Diarrhea</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Dehydration</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Dehydration</td>
</tr>
</tbody>
</table>

The first step employed is to calculate the $E(C)$:

<table>
<thead>
<tr>
<th>Dehydration</th>
<th>Diarrhea</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>4</td>
<td>12</td>
</tr>
</tbody>
</table>

Using formula (1)

$$E(C) = \sum_{h \in H} -P(h) \log_2 P(h)$$

$$= -\frac{8}{12} \log_2 \left( \frac{8}{12} \right) - \frac{4}{12} \log_2 \left( \frac{4}{12} \right)$$

$$= 0.91822$$

The following stage includes the recognizable proof of the root node. The node having the most elevated Information Gain is chosen as the root node.

Using the formula,

$$IG(C, A) = [E(C) - P(C_{\text{present}}) * E(C_{\text{present}}) - P(C_{\text{absent}}) * E(C_{\text{absent}})]$$

we compute the data acquire $IG(C, A)$ for every one of the symptoms. In this way, we have:

$$IG(C, \text{Fever}) = 0.3483$$

$$IG(C, \text{Vomiting}) = 0.2515$$

$$IG(C, \text{Chills}) = 0.0102$$

$$IG(C, \text{Dizziness}) = 0.0102$$

Thus, the information gain value for the symptom fever is the highest and thus the symptom fever is elected as the root node for this decision tree. Since we have now utilised the symptom fever as our root node, what structures the sub-tree is the presence or the absence of this symptom now. Our next step would be to calculate the information gain value with respect to the other symptoms.
Consequently, for our information, we saw that the loss_of_smell property is the top symptom that has the most noteworthy gini impurity score of 0.976. At that point comes internal_itchiness with a score of 0.975 and the others follow. This suggests that the loss_of_smell symptom has the most potential to separate different indications into individual classes and consequently is chosen as the root node. These features are extracted to help us understand the most important attributes or symptoms that will be involved in the classification process.

The entire database comprises of 5 tables namely users, diseases, symptoms, tips and feedback. The users table keeps a list of all registered users. The diseases table contains information about the disease name, description and its severity. The symptoms table lists out all possible symptoms associated with a respective disease. The tips table contains web links for home remedies and recommendations associated with each disease and finally the feedback table focuses on gathering feedback to improve the application from the users.

Once the disease diagnosis is completed the possible results are displayed to the user with weak, moderate and strong evidence percentage. The evidence is calculated as:

\[
evidence\% = \frac{Number\ of\ Matching\ Symptoms}{Total\ Number\ of\ Symptoms} \quad (4)
\]

Besides displaying the diagnosis, the application also fetches home remedial recommendations from the tips database for each possible diagnosed disease and also describes the condition to the user in layman terms for their better understanding.

VI. RESULTS

The accuracy score and the confusion matrix are used as performance metrics to further analyze and calculate the overall precision of the diagnosis.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy on Training Data</th>
<th>Accuracy on Testing Data</th>
<th>Correctly Classified</th>
<th>Incorrectly Classified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Tree</td>
<td>0.9512 19</td>
<td>0.9756 09</td>
<td>40</td>
<td>1</td>
</tr>
</tbody>
</table>

Thus, the accuracy score of the decision tree model is 0.975609 or in terms of percentage can be given as 97.56%

VII. CONCLUSION

The coronavirus pandemic disturbed and shattered the very balance of supply and demand of the health care services and provision. In our very attempt to fill the very same gap, we created the Personal Healthcare Assistance providing bot- PHAB. With regular on-site consultations, it exhibits itself with a very high scope to channel the gap between the availability of healthcare consultation and the people who need it.

This is a one-make personalized healthcare bot takes into consideration the needs and understanding of the Indian citizens and provides general and easy to comprehend healthcare consultation along with a few preventive
measures for some of the most common and household diseases and ailments. It has additional features including precautions and recommendations, age and gender-specific consultation. Our web service is quite robust in structure and reliable in detecting various common diseases, suggesting treatment corrections with the only bottleneck around the detailed entry of the symptoms and the problems faced by the user/patient are answered with precision. With this project, the goal is to make the healthcare accessible to more and more people, irrespective of how sophisticated and time-sensitive the circumstances are. Thus, we hope to have achieved the aim of delivering cost effective and reliable diagnosis with ease.

VIII. FUTURE WORK

After the whole COVID-19 crisis tapers down, this assistant will be as relevant as ever. The data-driven decisions and consultation provided by the system topped with an interactive user interface will continue to make health care readily and easily accessible to the masses. The project has a lot of scope of scaling-up. Such virtual assistants can be installed in hospitals which patients can reach out to and get a firsthand. With limitless ability to feed training data, these bots can consult for any specialty of medical science. These technologies can be leveraged to conduct regular medical tests and can also be integrated with fitness bands to track their health in real-time and suggest measures based on them. To make the experience more immersive, the assistant can also be integrated with social media applications like WhatsApp and Telegram and can also be deployed as an application of its own. Since majority of the rural population interacts in regional languages, the virtual assistant can also be upgraded to such languages to make the interface more intuitive and reachable to the people in need.

REFERENCES