A Synopsis Report
on

Smart Assistant For Doctors

by

1. Kush Hingol
2. Shreyash Kapadi
3. Sagar Renuke
4. Mitesh Gaonkar

under the guidance of

Mr. Bhavesh Panchal

MANJARA CHARITABLE TRUST

RAJIV GANDHI INSTITUTE OF TECHNOLOGY, MUMBAI

Department of Computer Engineering

University of Mumbai

October - 2018

MANJARA CHARITABLE TRUST
RAJIV GANDHI INSTITUTE OF TECHNOLOGY, MUMBAI

Juhu-Versova Link Road Versova, Andheri(W), Mumbai-53.

Certificate

Department of Computer Engineering

This is to certify that

1. Kush Hingol
2. Shreyash Kapadi
3. Sagar Renuke
4. Mitesh Gaonkar

Have satisfactory completed this synopsis entitled

(Smart Assistant For Doctor)

Towards the partial fulfillment of the

BACHELOR OF ENGINEERING
IN
(COMPUTER ENGINEERING)

as laid by University of Mumbai.

Guide | **Head of Department**

Prof. Bhavesh Panchal | **Dr.Satish Y. Ket**

Principal

Dr.Udhav Bhosle

Internal Examiner | **External Examiner**

Declaration

We wish to state that the work embodied in this synopsis titled ”Smart Assistant for Doctors” forms our own contribution to the work carried out under the guidance of ”Prof. Bhavesh Panchal” at the Rajiv Gandhi Institute of Technology.

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. we also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

(Students Signatures)
Kush Hingol (A-728)
Shreyash Kapadi (A-738)
Sagar Renuke (A-769)
Mitesh Gaonkar (A-719)

Acknowledgement

We wish to express our sincere gratitude to Dr. Udhav. V. Bhosle, Principal and Dr. Satish. Y. Ket , H.O.D of Computer Department of Rajiv Gandhi Institute of Technology for providing us an opportunity to do our project work on Smart Assistant For Doctors ".

This project bears on imprint of many peoples. We sincerely thank our project guide Mr. Bhavesh Panchal for his guidance and encouragement in carrying out this synopsis work.

Finally, we would like to thank our colleagues and friends who helped us in completing the Project Synopsis work successfully

1. Kush Hingol
2. Shreyash Kapadi
3. Sagar Renuke
4. Mitesh Gaonkar

Abstract

The project is based on development of a digital based Smart Assistant (SA) to benefit doctors and patient by reducing the treatment time. The Healthcare industry contains big and complex data that may be required in order to discover fascinating pattern of diseases makes effective decisions with the help of different machine learning techniques.

Advanced data mining techniques are used to discover knowledge in database and for medical research. This paper has analyzed prediction systems for Diabetes, Kidney and Heart disease using more number of input attributes. SA utilizes various advanced computing techniques such as Fuzzy logic, Hypothetico-Methodology that applies differential diagnosis, probability algorithm and pattern-matching algorithm to ease diagnosis. The data mining classification techniques, namely Support Vector Machine(SVM) and Random Forest (RF) will be used. The performance of these techniques is compared, based on precision, recall, accuracy, $\mathrm{f}_m easureaswellastime.$

The proposed system will take symptoms as input and give the possible diseases as output after processing the input data. The output will contain a number of diseases based on symptoms, it will be in doctor's discretion to give the final disease. Smart Assistant will therefore aid in faster diagnosis of diseases.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

1.1 Introduction Description

In the world of an active life, health is a factor of serious concern. Due to increase in various health hazards, the health of entire mankind is degrading since a last few decades. People are not receiving better means of diagnosis and analysis which is required at a higher priority.

In the traditional treatment methods, the patients need to schedule various appointments to meet doctors from a week prior till the treatment suffices. This takes at least three or four consultations, that include bringing reports to the doctor after performing the test, taking the prescribed medications and the final health feedback about patients recovery. The entire procedure takes almost a month or two. By the use of the Smart Assistant the patients can be treated approximately in three to four days.

The Smart Assistant software helps to reduce long waiting in clinical queues and also increases diagnosis speed by suggesting the doctor possible diseases based on symptoms. SA gives suggestions to the doctor about the medicines that can be prescribed.

1.2 some other point in introduction

Predictive systems are the systems that are wont to predict some outcome on the basis of some pattern recognition. Disease detection is that the method by which patients diagnosis is performed on the basis of symptoms analyzed which may causes difficulty while predicting disease affect. As an example, fever itself could be a symptom of the many disorders that doesnt tell the healthcare professional what exactly the disease is. Because the results or opinion vary from one physician to a different, there's a requirement to help a medical physician which will have similar opinion certainly symptoms and disorders.

Predictive medical diagnosis could be a net application which is able to predict a selected disorder on the basis of symptoms and supply diagnosis for same disorder which is able to be detected by rule. Healthcare professionals use their previous data and insights to reach a particular decision regarding any disease or disorder. In clustering process such as K-means, Fuzzy c-means, etc, data is partitioned into sets of clusters or subclasses. Machine learning techniques such as SVM, Nave bayes etc, can be used to classify different objects on the basis of a training set of data whose outcome value is known.

1.2.1 Clustering Techniques

The clustering process divides the data into cluster groups or subclasses. We used four clustering algorithms, namely K-Means, EM, PAM, Fuzzy C-Means. The K-Means classification algorithm works by partitioning n observations in k-subclasses defined by centroids, where k is chosen before the algorithm begins. K-Means and EM are two iterative algorithms. EM (expectation-maximization) is a statistical model that depends on the unobserved latent variables to estimate the maximum likelihood parameters. Partitioning around medoids (MAP) is similar to Kmeans that partitioning is based on the K-medoids method, which divides data into a number of disjoint clusters. In fuzzy clustering, data elements can belong to multiple clusters. This is also called soft clustering.

1.2.2 Classification Techniques

Machine learning based classification techniques can be used to classify various objects based on a series of training data whose result value is known. In this study four classification algorithms are used: KNN, SVM, Naive Bayes and C5.0. In the nearest neighbor KN, the object is classified by the majority of its neighbors, with the object being assigned to the most commonly used class among its nearest neighbors. In SVM (Support Vector Machines), data is first converted into a set of points and then classified into classes that can be separated linearly. The Naive Bayes model calculates the probability of a set of data that can belong to a class using the Bayes rule. The C5.0 algorithm is a decision tree that recursively separates observations in branches to form a tree to improve prediction accuracy. It is an improved version of the C4.5 and ID3 algorithms. It also provides the powerful gain method to increase the accuracy of this classification algorithm.

Chapter 2

Aims and Objectives

Smart Assistant for doctors is a specially designed software that would work as an aid to doctors for curing patients faster so as to eradicate the diseases. The main aim of the project is to get the accurate results based on the symptoms given as the input by the patient.

The main aim of the software is to ease doctors work. The doctor has to login using his unique doctor id, uploading his certificates that shall be verified by the admin manually. The software will provide the doctor with a list of possible diseases and the doctors job will be to study the optimized list, test the patient and get to a final conclusion. The patient will login using his patient id. The job of the patient is to choose a doctor based on his priority and then submit the list of symptoms to the software.

The software will map the symptoms into a fuzzy set using membership functions and will be clustered into different classes and subclasses with the help of fuzzy c-mean algorithm. The clustered data will be separated using linear separability algorithm which is called Support Vector Machine (SVM). After separation of data, the data should be trained. Hence to train the data we need to use Naive Bayes Algorithm which will classify the data item and will show the class of data. Using confusion matrix, we will compute the output.

2.1 Motivation & Aims

In the world of an active life, health is a factor of serious concern. Due to increase in various health hazards, the health of entire mankind is degrading since a last few decades. People are not receiving better means of diagnosis and analysis which is required at a higher priority.

Clinical test outcomes are often created on the basis of doctors perception and experience instead of the knowledge enriched data masked within the dataset and generally this procedure prompts unintended predispositions, doctors experience might not be capable to diagnose it accurately that affects the disease diagnosis system. Here, the term data mining will mean to research the clinical data to predict patients health status. Therefore discovering fascinating pattern from healthcare information, different data mining techniques are applied with statistical analysis, machine learning.

In the traditional treatment methods, the patients schedule various appointments to meet doctors from a week prior till the treatment suffices. This takes at least three or four consultations,including bringing reports to the doctor after performing the test, taking the medications and the health feedback. The entire procedure takes almost a month or two. By the use of the SA the patients can be treated approximately in three to four days.

A major hindrance in healthcare sector is the challenging the existing constraints of traditional methods. The unavailability of medical experts on time for patients, Scarcity of doctors in rural areas, variations in prescription and quality of service at affordable costs have large impact on people. Quality service implies diagnosing patients correctly and administering treatments that are effective. Poor clinical decisions can lead to consequences that are unacceptable. There is a high rate of death causes of people from the disease that could been detected at an early stage and diagnosed. Early diagnosis will help to cure disease at an early stage.

2.2 Objectives

The proposed system can be deployed as a mobile application and MDM server. Secure connection,single-sign on both can be used for deployment. Following objectives can be fulfilled with the help of the mobile application and MDM server:

- The Smart Assistant helps to reduce the drawbacks of traditional diagnosis method.
- Predictive medical diagnosis could be a net application which is able to predict a selected disorder on the basis of symptoms and supply diagnosis for same disorder which is able to be detected by rule.
- Reduce the doctor's workload by faster diagnosis
- Smart Assistant will help to cure patients at a faster rate.
- The software will implement multiple machine learning and data mining algorithms which will improve system accuracy and performance.

Chapter 3

Literature Review

Research has been done to provide a better solution to the traditional approach for medical diagnosis, but still the medical field does not approve such system because human coalition diagnosis is preferred more. So the diagnosis using computing is still in research phase.

3.1 Technical papers review

Any software-based system for optimistic diagnosis is difficult to implement and manage so this paper proposes a type of system which efficiently enhances the already existing systems and also brings into limelight all the benefits and the drawbacks of using various determining techniques for prediction the proposed system in the paper uses following methods and techniques 1. Fuzzy C-Mean Clustering 2. Support Vector Machine 3. Random Forest. The paper also includes various algorithms and datasets that are being used. On the basis of analysis made out of the result. This paper claims to be accurately measuring diabetes, kidney and liver diseases [2].

A traditional method of diagnosis involves interaction between doctor and patient and also some related tests, based on which diagnosis is carried out. So, in this paper a comparison is made between normal diagnosis and computational diagnosis method. An analysis is carried out on the basis of questionnaire to a group of people for the same. This analysis has resulted that the software-based systems is of a great need for today. The paper also proposes a new type of system which has all together various different methodologies such as probability algorithm, mapping algorithm, Hypothetico deductive methodology, etc. This new type of system will help in encouraging computing in medical field [1].

The paper highlights various methods to get a system based diagnosis. The medical data contains information about various diseases and disorders, symptoms and statistics of patients. Analysis methods to find associations between different medical data. If these data is analyzed with the help of different data mining techniques can be used for optimized and precision decision making process. Thus, they aim in developing a web application as a support to healthcare which take input from users as symptoms and gives the probability of disease the user or patient may have on the basis of input given by using data mining technique viz. Nave Bayesian algorithm. Also the user gets diagnosis information about the disease which is predicted. The paper intends to give detailed description about how data mining method can be used in todays research for prediction of diseases as a supporting tool for physicians, medical students and users [8].

3.2 Problem Formulation

The traditional techniques makes a lot of chaos in detection of the appropriate diseases hence the software Smart Assistant for Doctors uses various various datamining techniques for prediction the proposed system in the paper uses following methods and techniques 1. Fuzzy C-Mean Clustering 2. Support Vector Machine 3. Random Forest. This improves the efficiency and accuracy of disease detection by the performance measure results.The output of the system will be if the patient is affected by the disease or not.

3.3 Problem Statement

Many hospitals use Information systems to manage their healthcare or patient data. These systems typically generate huge amount of data which can take form of numbers, graphs, text, images etc. Decision support systems are also used but they are largely limited in their functionality. They can answer simple queries like the average age of patients for particular disease, analysis of patients data. However, they cannot answer complex queries given the patient symptoms, what will be the probability of getting disease and its diagnosis. Clinical decisions are made on doctors intuition and experience. But a support to traditional method by using an application that will help the physicians in understanding more detail perspective of disorder and decision making for diagnosis will increase the efficiency of results and system.

Chapter 4

Analysis

The development of software incorporates multiple researches and a feasibility study, based on several research papers and survey analysis.

The result of the analysis is that, the software will be implementing various machine learning algorithms to process and learn on the basis of the input.

4.1 Survey Analysis

4.1.1 Patient Survey

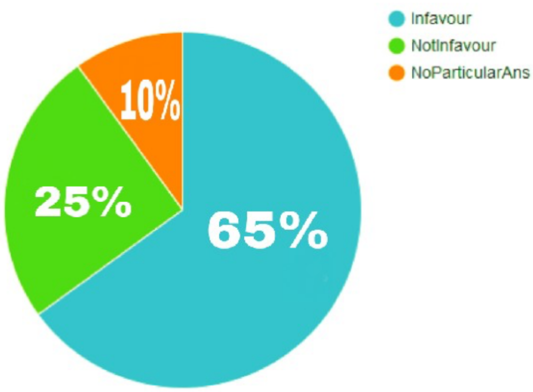

Figure 4.1: Patient Survey

A survey was carried out in the form of a questionnaire to a group of people regarding their medical experiences. The survey included people of various age groups and from different regions. This survey was carried out at public places. The questionnaire includes a set of questions such as a normal patients traditional approach to the clinician and also made a comparison of the software-based system. The Fig.2 shows the pie chart of people surveyed that are in favour of SA, had a different opinion for SA and also those people who didnt have particular answer for it. This survey focused on the following points:

- Doctor consultation frequency

- Recovery time period
- Medical document management
- Medication Assurance

Approximately 65 percent of the people were in favour of the system as they find it much simpler than the traditional approach. Around 25 percent of the people surveyed were against or not in the favour of the system and it can be probably said that they were unimpressed by the non-human software and prefer a personal human touch. Remaining 10 percent people didnt have proper answers as they were having certain agreement to the points of the software-based system and disagreed to others and so were with mixed views.

4.1.2 Doctor's Survey

The doctors survey was conducted at various medical facilities. This survey was based on the following points:

- Technical Incorporation for Treatment
- Fast Diagnosis
- Digital Doctor-Patient Interaction
- Database of Medical Record

68 percent of the doctors agreed that SA will be a great advancement in the medical field as every patient cannot just be treated faster but even their medical histories in their E-Reports can be saved in database for further references in medical history. The 27 percent doctors who voted no had a different opinion about SAs Application in medical field, while 7 percent of the doctors felt it is not presently mandatory.

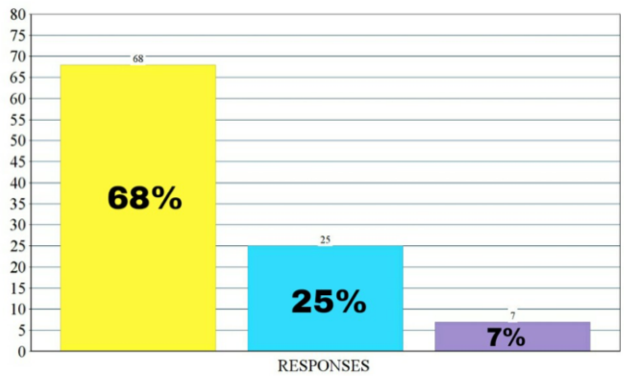

Figure 4.2: Doctor Survey

4.2 Technical Analysis

In [1] neural networks, decision tree and naive bayes machine learning approach is used to diagnose heart disease. For optimized feature selection genetic algorithm is used and obtained an accuracy of 100 percent, 99.62 and 90.74 respectively.

In [2], performed a prediction of heart disease detection using neural network with genetic algorithm based feature extraction. Back propagation based neural network weight optimization by Genetic algorithm is designed and obtained the 89 percent accuracy of prediction of heart disease.

In [3], author developed a heart disease prediction system using an approach of ANN with LVQ and achieved accuracy of about 80 percent, sensitivity of about 85 percent sensitivity as well as specificity of about 70 percent.

In [4] proposed a heart disease diagnosis is proposed using lazy data mining approach with data reduction strategies i.e. principal component analysis is used to get category association rules. The result analysis shows that J4.8 has 10.26 enhancement as well as 8.6 percent enhancement over naive bayes.

In [5] presented a heart disease prediction system using data mining approach with two additional features i.e. obesity and smoking to boost the prediction rate. Neural networks, Decision trees and Naive Bayes was used in for predicting heart disease with an accuracy of 99.25 percent, 94.44 percent and 96.66 percent respectively.

A web based application has introduced in [6] using Naive Bayesian algorithm which took symptoms from user and gave the diagnosis result to the user or patient.

In [7] Association rule mining technique was used for diagnosis of diabetes. The authors concluded that the data mining techniques when used appropriately increases the computation and also the classification performance. These rules have the potential to improve the expert system and to make better clinical decision making.

For predicting diabetes disease on weka tool, author in research work [8] had presented a comparison between Naive bayes algorithm and decision tree algorithm and achieved system accuracy of about 79.56 percent and 76.96 percent respectively.

Comparative Analysis of Different Techniques in terms of Accuracy		
Name of Author	Technique	Accuracy
Bhatla et al.	Neural Network, Naive Bayes and Decision Tree for heart disease detection.	Neural networks = 100 pr Decision tree= 99.62pr Naive Bayes 90.74pr
Amin et al.	Optimized Neural Network for heart disease detection.	Training data was = 89pr Validation data = 96.2pr.
Chen et al.	ANN based heart disease detection.	ANN = 80pr.
Dangare et al.	Decision trees, Neural networks and Naive Bayes for heart disease detection.	Neural Networks = 99.25pr Naive Bayes = 94.44 pr Decision Tree =96.66 pr
Iyer et al.	Decision tree and Naive bayes algorithm for diabetes detection.	Decision Tree =76.96pr Naive Bayes= 79.5pr
Uma Ojha and Savita Goel	Decision Tree, SVM, FCM	Decision tree (C5.0) =81pr SVM =81pr FCM = 37pr
Chetty et al.	KNN and F-KNN for diabetes and liver disease detection.	KNN = 97.02prF-KNN = 99.25pr for diabetes data. KNN = 96.13pr F-KNN =98.95pr for liver disease data.
Kumari Deepika and Dr. S. Seema	Naive Bayes, Decision tree, Support Vector Machine (SVM) and Artificial Neural Networks (ANN) for diabetes and heart disease detection.	SVM = 95.556pr (heart disease) Naive Bayes = 73.58pr (diabetes).

table: 4.1 Comparative Analysis of Different Techniques in terms of Accuracy

By the above done research we have analyzed that: k-mean algorithm, fuzzy c-mean algorithm will be approprite for clustering of data as it has an accuracy greater than 90 percent. SVM for separability of data between two or more classes or cluster. Naive Bayes algorithm for classification and trainning.

Chapter 5

Methodology

The Research have been done to provide a software which will scale down the process of differential diagnosis by using various machine learning algorithm. From analysis done we have analyzed that following machine learning algorithms are used. 1. K-mean algorithm/ Fuzzy c-mean algorithm 2. Support Vector Machine. 3. Naive bayes algorithm.

5.1 Methodology/Procedures

5.1.1 Hypothetico - deductive methodology

Hypothetico-deductive methodology is the life-cycle method in which each phase follows a continuous cycle. The figure shows the working of the methodology. The steps involved are patient input data analysis, observation and identification of diseases according to the input.

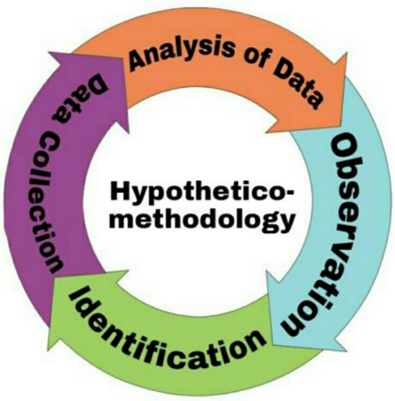

Figure 5.1: Hypothetico-dedcutive methodology

- Observation: Analyses symptom inputs for possible diseases .
- Identification: Identifying the disease using K-mean,FCM,SVM and naive bayes algorithm.
- Data Collection: Collects detailed data of finalised diseases from medical database.

- Analysis of Data: Gives optimized output list of diseases, medication.

5.1.2 Procedure

One of the interesting and important subjects among researchers in the field of medical and computer science is diagnosing illness by considering the features that have the most impact on recognitions. The subject discusses a new concept which is called Medical Data Mining (MDM). Indeed, data mining methods use different ways such as classification and clustering to classify diseases and their symptoms which are helpful for diagnosing. A disease diagnosis system is created in order to predict different diseases such as diabetes, kidney disease as well as liver disease, etc. Systems workflow is discussed below:

Step 1: Through the proposed application user (doctor, patient, physician etc.) can input the attribute values of disease and send it to the decision support system for analysis.

Step 2: At decision support system, dataset of different diseases are loaded and apply data mining algorithms to train dataset. Requested user inputs are collected and processed on server to predict the diagnosis result.

Step 3: For analyzing healthcare data, major steps of data mining approaches like preprocess data, replace missing values, feature selection, machine learning and make decision are applied on train dataset. On the decision support system end different classification algorithms would be executed on train dataset and ready to classify the test dataset.

In the proposed algorithm Support Vector Machine and Random Forest is used to give clustering level for different subspaces. The voting model will ensemble all these results and output the final classification result. Finally, the predicted results will be compared with true labels of the testing phase.

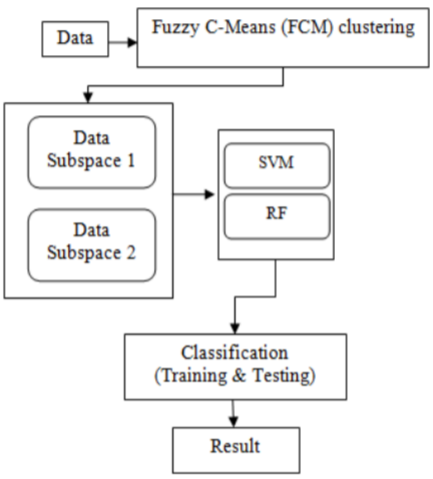

Figure 5.2: Proposed Model

5.1.3 K-mean algorithm

K-means (MacQueen, 1967) is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids shoud be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point we need to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more. Finally, this algorithm aims at minimizing an objective function, in this case a squared error function. The objective function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} |x_i^j - c_j|^2$$

where

$$\sum_{i=1}^{n} |x_i^j - c_j|^2$$

is a chosen distance measure between a data point and the cluster centre , is an indicator of the distance of the n data points from their respective cluster centres.

The algorithm is composed of the following steps:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.

2. Assign each object to the group that has the closest centroid.

3. When all objects have been assigned, recalculate the positions of the K centroids.

4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

5.1.4 Fuzzy c-mean

Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method (developed by Dunn in 1973 and improved by Bezdek in 1981) is frequently used in pattern recognition. It is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_i j^m |x_i^j - c_j|^2$$

where m is any real number greater than 1, uij is the degree of membership of xi in the cluster j, xi is the ith of d-dimensional measured data, cj is the d-dimension center of the cluster.

Following is the algorithm for fuzzy c-mean.

1. *Initialize $U=[u_{ij}]$ matrix, $U^{(0)}$*
2. *At k-step: calculate the centers vectors $C^{(k)}=[c_j]$ with $U^{(k)}$*

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

3. *Update $U^{(k)}$, $U^{(k+1)}$*

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\| x_i - c_j \|}{\| x_i - c_k \|} \right)^{\frac{2}{m-1}}}$$

4. *If $\| U^{(k+1)} - U^{(k)} \| < \varepsilon$ then STOP; otherwise return to step 2.*

Figure 5.3: FCM Algorithm

5.1.5 Support Vector Machine

Support vector machine is a machine learning approach that can be used as classifier as well as for regression. SVM classifies the data into different classes by finding hyperplane (line) which separates training data into classes. SVM does not overfit the data and gives best classification performance in terms of precision and accuracy. SVM does not make any strong assumptions on data. It shows more efficiency for correct classification of the future data. SVM is classified into 2 categories i.e. Linear and non-Linear. In linear approach, training data is separated by line i.e. hyperplane.

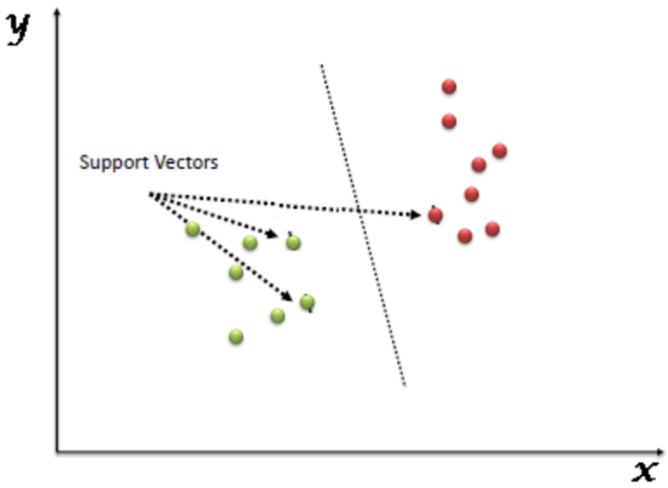

Figure 5.4: SVM output

5.1.6 Naive Bayes algorithm

It is a classification technique based on Bayes Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as Naive.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Bayes theorem provides a way of calculating posterior probability P(c—x) from P(c), P(x) and P(x—c). Look at the equation below:

$$P(c|x) = \frac{P(x|c).p(c)}{P(x)}$$

5.1.7 Proposed Algorithm

Input: D Clininal data;

Output: Label Disease Label;

Patients LabelNormal,Disease

Step1: Pre-processing and data cleansing

Step2: For each instance in D, do
Find feature vector (V)

Step 3: For each V do
Data clustering using FCM and split data in two halves and classify data using SVM algorithm

Step 4: Determine the total class label Find
True-positive (TP)
True-negative (TN)
False-positive (FP)
False-negative (FN)

Step 5: Find Performance Parameters

Step 6: Predict Disease Class as
if (class=1) Patient=Normal State
else-if(class =0) Patient=Disease State
end for

5.1.8 Performance measure

In this study, we used three performance measures: Precision, Accuracy, Recall, F-measure and Total execution time.

Accuracy is termed as ratio of the number of correctly classified instances to the total number of instances.
Accuracy = (TP+TN)/(TP+TN+FP+FN)

Precision is the ration of actually true predicted instances out of the total true instances.
Precision = TP/(TP+FP)

Recall is the ratio of actual true instances out of all the items which are true.
Recall = TP/(TP+FN)

F-measure is the harmonic mean of both precision and recall.
F-Measure = 2*(Precision*Recall)/(Precision + Recall)

5.1.9 Flowchart

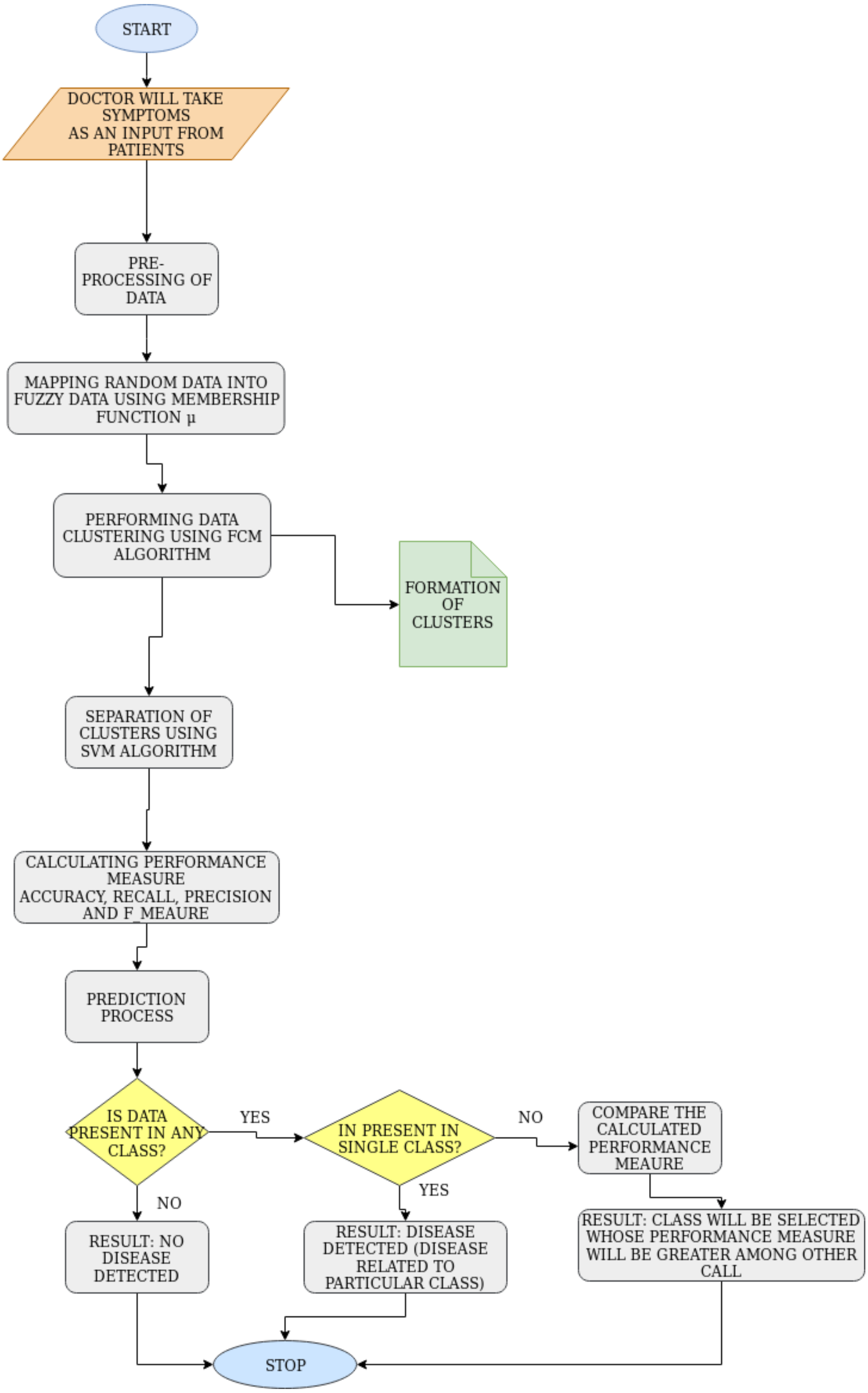

Figure 5.5: Flowchart

Chapter 6

Software & Hardware Requirements

6.1 Software Requirements

- OS: Windows/Linux
- Framework: .NET Framework, phycharm, pyhton 3.
- Server: Ubuntu 16.04-Machine Learning Server for Linux
- Database: MySql

6.2 Hardware Requirements

- Processor : Intel i5 7th gen processor with 3.60 Ghz
- RAM: Minimum Requirement of 8 GB RAM
- ROM: Minimum Requirement of 1 TB of ROM
- GPU: Tensorflow non-GPU version, GTX Nvidia.

Chapter 7

Design

The Design of Smart Assistant includes Architecture Diagram along with various UML diagrams which indicates the working and the processing of the data in the system.

7.1 Architecure

The Architecture diagram shows the various module of the system. It indicates the flow of the system i.e how data is processed by various algorithms.

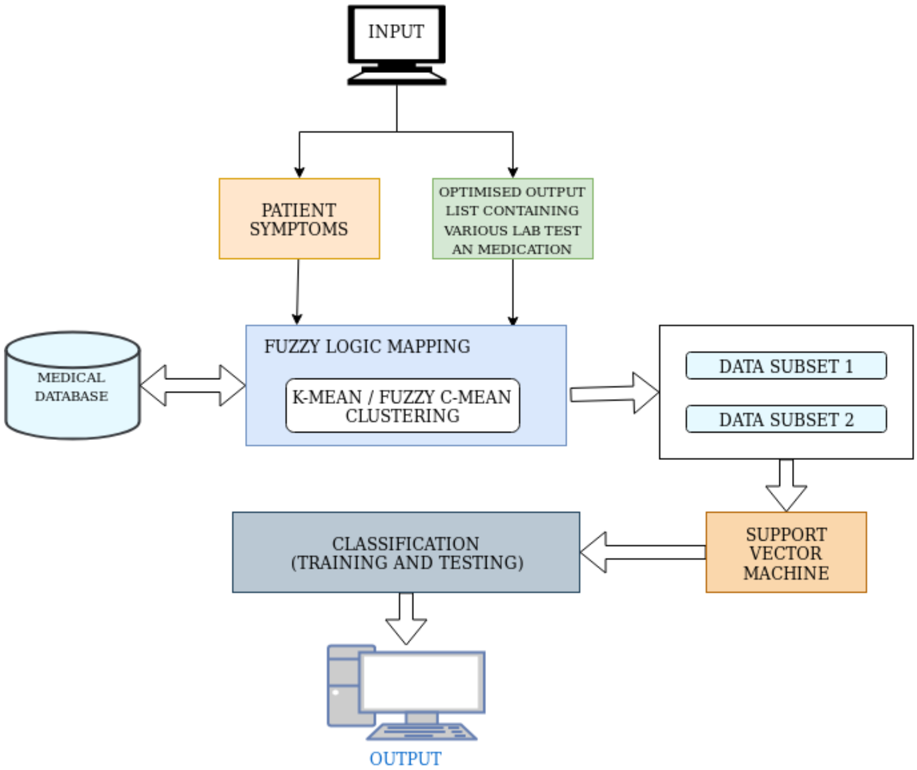

Figure 7.1: System Architecture

For implementation of the system Hypothetico-Deductive Methodology is used which contains mainly four phases they are as follows:

- Observation
- Identification
- Data Collection
- Analysis of data
- **Patient Symptoms:** Patient Symptoms will be taken as input from the doctor and will be given as an input to the system to predict whether the patient is suffering from disease or not.
- **Fuzzy Logic mapping:** The symptoms which is taken as a random input will be mapped into a fuzzy set using triangular membership function.
- **K-mean/Fuzzy c-mean algorithm:** The input which is mapped into fuzzy set will be processed using k-mean or FCM algorithm will which form various cluster or classes. Each cluster or classes indicates a disease. The fuzzy data will be plotted according to the nearest distance from the cluster. Thus various cluster will which will include various fuzzy data.
- **SVM:** SVM is supervised learning which will learn from the previous input. SVM is used to separate the clusters or classes linear which ensures that no data changes its cluster while processing.
- **Classification:** In classification naive bayes algorithm is used to predict the disease. In naive bayes algorithm probability is calculated to check which data is present in which class. If a data is present in more than on class than the by comparing performance measure the best class is selected.
- **Output:** The output will indicate whether the patient is suffering from kidney, liver or diabetes disease so with the help of this the doctor will get to know the presence of the disease at must faster rate.

7.2 UML Design

For Designing the software UML diagrams are included. Usecase diagram and sequence diagram is used to explain the interface of various actors and the flow of the system.

7.2.1 Usecase diagram

The usecase diagram of smart assistant consist of two actors namely doctor and patient. The usecases are distributed among them according to their roles. The functions are as follows:

- Input symptoms: Symptoms is taken as an input from the patient.
- Disease Prediction: Expert System will predict the disease.
- Prescribed test: Expert system will Prescribe test related to the disease detected.
- Prescribed Medication: Expert system will Prescribe Medication related to the disease detected.

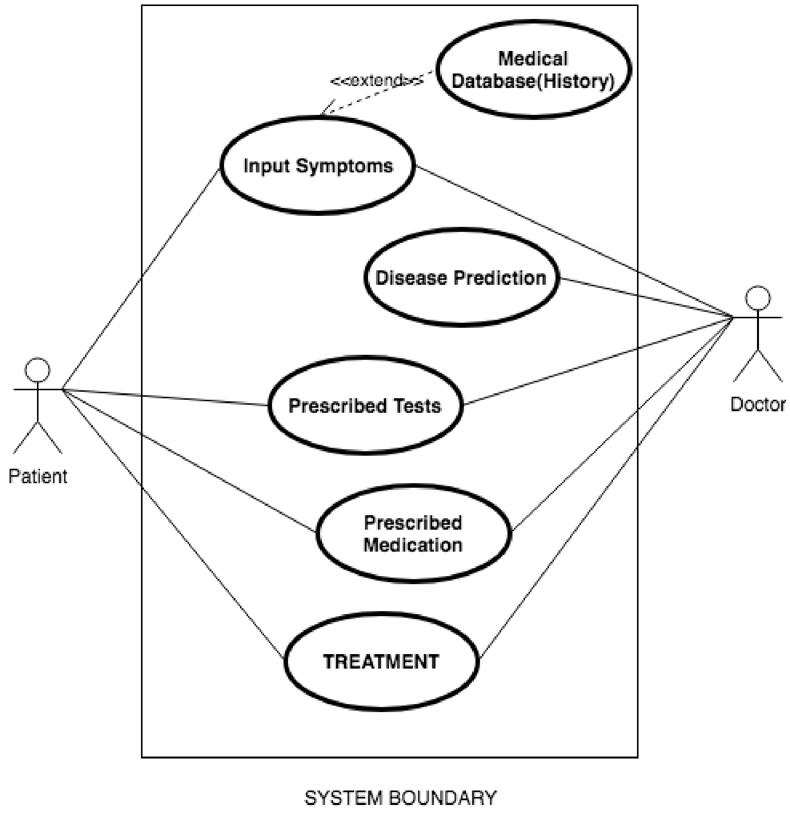

Figure 7.2: use case

7.2.2 Medical database

Medical database contains two table namely:

- Disease table
- Medicine table

Table Disease:

Disease_id	Disease_name
1	Kidney Disease
2	Liver Disease
3	Diabetes

Table Medication:

Medicine_id	Disease_id	Medicine_name
1	1	Nefrogard
2	1	Nefita
3	1	Fabustat 80
4	2	Met xl
5	2	Nicardia
6	3	Biguanides
7	3	Acarbose

Figure 7.3: Database Design

7.2.3 Dataset Description

- Pima Indians Diabetes Database: This study used data sets from the Pima Indians Diabetes Database of National Institute of Diabetes. This dataset consists of 768 samples with 8 numerical valued attribute where 500 are tested negative and 268 are tested positive instances.

- Chronic Kidney Disease Dataset: This study used data sets from the university of California Irvine (UCI) repository. The data set contains 400 patients, where 250 patients were positively affected by kidney disease and as many as 150 patients do not suffer from kidney disease.

- Liver Disorders Data Set: This study used data sets from the university of California Irvine (UCI) repository. This data set contains 416 liver patient records and 167 non liver patient records collected from North East of Andhra Pradesh, India.

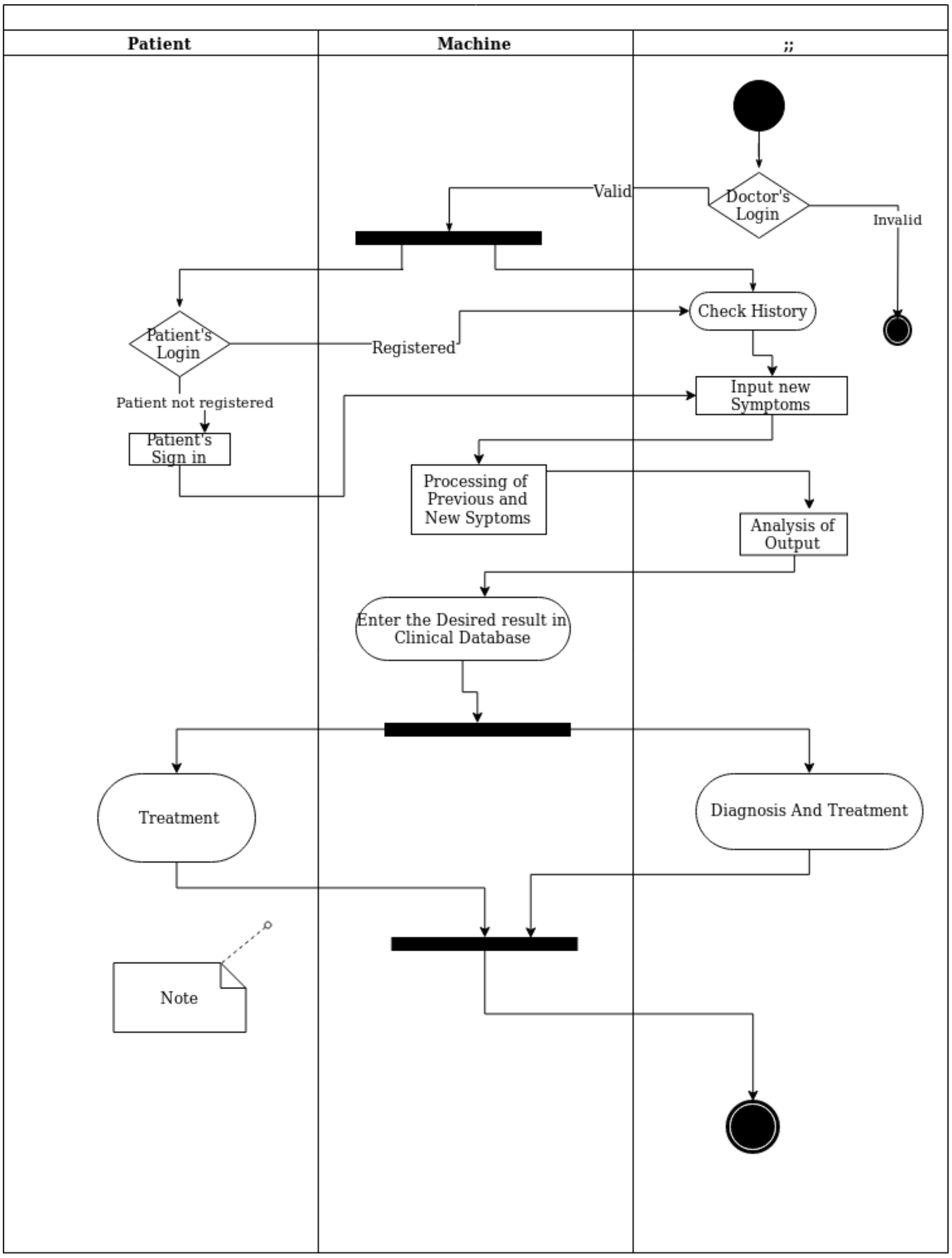

Figure 7.4: Activity Diagram

Chapter 8

Implementation

8.1 About Implementation

Implementation of fuzzy c-mean algorithm is done. We have developed the algorithm in python.

- code:import pandas as pd

  import numpy as np

  import random

  import operator

  import math

  df$_{f}ull = pd.read_{c}sv("SPECTF_{N}ew.csv")$

  columns = list(df$_{f}ull.columns)$

  features = columns[:len(columns)-1]

  class$_{l}abels = list(df_{f}ull[columns[-1]])$

  df = df$_{f}ull[features]$

  Number of Attributes

  num$_{a}ttr = len(df.columns) - 1$

  Number of Clusters

  k = 2

  Maximum number of iterations

  MAX$_{I}TER = 100$

  Number of data points

  n = len(df)

  Fuzzy parameter

  m = 2.00

  def accuracy(cluster$_{l}abels, class_{l}abels)$:

  county = [0,0]

countn = [0,0]

tp = [0, 0]

tn = [0, 0]

fp = [0, 0]

fn = [0, 0]

for i in range(len(df)):

Yes = 1, No = 0

if cluster$_labels[i] == 1andclass_labels[i] ==' Yes' :$

tp[0] = tp[0] + 1

if cluster$_labels[i] == 0andclass_labels[i] ==' No' :$

tn[0] = tn[0] + 1

if cluster$_labels[i] == 1andclass_labels[i] ==' No' :$

fp[0] = fp[0] + 1

if cluster$_labels[i] == 0andclass_labels[i] ==' Yes' :$

fn[0] = fn[0] + 1

for i in range(len(df)):

Yes = 0, No = 1

if cluster$_labels[i] == 0andclass_labels[i] ==' Yes' :$

tp[1] = tp[1] + 1

if cluster$_labels[i] == 1andclass_labels[i] ==' No' :$

tn[1] = tn[1] + 1

if cluster$_labels[i] == 0andclass_labels[i] ==' No' :$

fp[1] = fp[1] + 1

if cluster$_labels[i] == 1andclass_labels[i] ==' Yes' :$

fn[1] = fn[1] + 1

a0 = float((tp[0] + tn[0]))/(tp[0] + tn[0] + fn[0] + fp[0])

a1 = float((tp[1] + tn[1]))/(tp[1] + tn[1] + fn[1] + fp[1])

p0 = float(tp[0])/(tp[0] + fp[0])

p1 = float(tp[1])/(tp[1] + fp[1])

r0 = float(tp[0])/(tp[0] + fn[0])

r1 = float(tp[1])/(tp[1] + fn[1])

accuracy = [a0*100,a1*100]

precision = [p0*100,p1*100]

recall = [r0*100,r1*100]

return accuracy, precision, recall

def initializeMembershipMatrix():

membership$_{m}at = list()$

for i in range(n):

random$_{n}um_{l}ist = [random.random()foriinrange(k)]$

summation = sum(random$_{n}um_{l}ist)$

temp$_{l}ist = [x/summationforxinrandom_{n}um_{l}ist]$

membership$_{m}at.append(temp_{l}ist)$

return membership$_{m}at$

def calculateClusterCenter(membership$_{m}at)$:

cluster$_{m}em_{v}al = zip(*membership_{m}at)$

cluster$_{c}enters = list()$

for j in range(k):

x = list(cluster$_{m}em_{v}al[j])$

xraised = [e ** m for e in x]

denominator = sum(xraised)

temp$_{n}um = list()$

for i in range(n):

data$_{p}oint = list(df.iloc[i])$

prod = [xraised[i] * val for val in data$_{p}oint]$

temp$_{n}um.append(prod)$

numerator = map(sum, zip(*temp$_{n}um))$

center = [z/denominator for z in numerator]

cluster$_{c}enters.append(center)$

return cluster$_{c}enters$

def updateMembershipValue(membership$_{m}at, cluster_{c}enters)$:

p = float(2/(m-1))

for i in range(n):

x = list(df.iloc[i])

distances = [np.linalg.norm(map(operator.sub, x, cluster$_{c}enters[j]))forjinrange(k)]$

for j in range(k):

den = sum([math.pow(float(distances[j]/distances[c]), p) for c in range(k)])

membership$_{m}at[i][j] = float(1/den)$

return membership$_{m}at$

def getClusters(membership$_{m}at)$:

cluster$_{l}abels = list()$

for i in range(n):

$\max_val, idx = max((val, idx) for (idx, val) in enumerate(membership_mat[i]))$

$\text{cluster}_labels.append(idx)$

return cluster_labels

def fuzzyCMeansClustering():

Membership Matrix

$\text{membership}_mat = initializeMembershipMatrix()$

curr = 0

while curr ¡= $\text{MAX}_ITER:$

$\text{cluster}_centers = calculateClusterCenter(membership_mat)$

$\text{membership}_mat = updateMembershipValue(membership_mat, cluster_centers)$

$\text{cluster}_labels = getClusters(membership_mat)$

curr += 1

print(membership_mat)

return $\text{cluster}_labels, cluster_centers$

labels, centers = fuzzyCMeansClustering()

a,p,r = accuracy(labels, class_labels)

print("Accuracy = " + str(a))

print("Precision = " + str(p))

print("Recall = " + str(r))

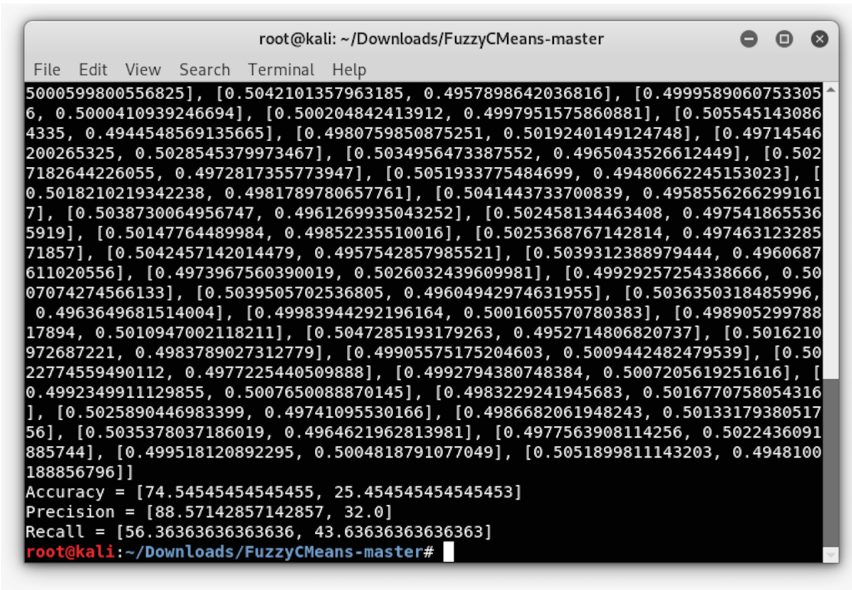

Figure 8.1: Implementation Plan

Bibliography

[1] Sharmila Gaikwad, Shreyash Kapadi, Kush Hingol, Sagar Renuke, Mitesh Gaonkar, "SMART ASSISANT FOR DOCTORS". *International Conference on*

Topical Transcends in Science, Technology and Management ICTTSTM-17 ,

[2] Shakuntala Jatav and Vivek Sharma, "AN ALGORITHM FOR PREDICTIVE DATA MINING APPROACH IN MEDICAL DIAGNOSIS" *International Journal of Computer Science Information Technology (IJCSIT)*, Vol 10, No 1, February 2018.

[3] Nidhi Bhatla, Kiran Jyoti, An Analysis of Heart Disease Prediction using Different Data Mining Techniques*IJERT*, Vol 1, Issue 8, 2012.

[4] P. Bhandari, S. Yadav, S. Mote, D.Rankhambe, Predictive System for Medical Diagnosis with Expertise Analysis" *IJESC*, Vol. 6, pp. 4652-4656, 2016.

[5] Syed Umar Amin, Kavita Agarwal, Rizwan Beg, Genetic Neural Network based Data Mining in Prediction of Heart Disease using Risk Factors *IEEE*, 2013.

[6] A. Iyer, S. Jeyalatha and R. Sumbaly, Diagnosis of Diabetes using Classification Mining Techniques *IJDKP* Vol. 5, pp. 1-14, 2015.

[7] A H Chen, S Y Huang, P S Hong, C H Cheng, E J Lin, HDPS: Heart Disease Prediction System *IEEE* 2011.

[8] AChaitrali S. Dangare, Sulabha S. Apte, Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques *IJCA* Volume 47 No.10, June 2012.

[9] Uma Ojha and Savita Goel, A study on prediction of breast cancer recurrence using data mining techniques

International Conference on Cloud Computing, Data Science Engineering, IEEE 2017.

[10] Naganna Chetty, Kunwar Singh Vaisla, Nagamma Patil, An Improved Method for Disease Prediction using Fuzzy Approach

International Conference on Advances in Computing and Communication Engineering, IEEE pp. 568-572, 2015.

Publication

Paper entitled "Smart Assistant For Doctor" is presented at *International Conference on Topical Transcends in Science, Technology and Management ICTTSTM 2018,17th and 18th August, 2018 Sai Vidya Institute of Technology, Bengaluru Karnataka, India*, In collaboration with Scopus Indexed Journal, USA with approval of IEAE.

Annexure

Smart Assistant for Doctors

Sharmila Gaikwad[1]
Professor: Computer Department
MCT's Rajiv Gandhi Institute of Technology
Mumbai, India
[1]sharmila.gaikwad@mctrgit.ac.in

Kush Hingol[2], Shreyash Kapadi[3], Sagar Renuke[4], Mitesh Gaonkar[5]
Student: Computer Department
MCT's Rajiv Gandhi Institute of Technology
Mumbai, India
[2]kushhingol.kh@gmail.com, [3]shretzz@gmail.com

ABSTRACT

The analysis is on development of a digital based Smart Assistant (SA) to benefit doctors and patient by reducing the treatment time. This paper tries to analyse why a Smart Assistant for doctors would be preferable over the traditional treatment methodology. SA utilises various advanced computing techniques such as Fuzzy logic, Hypothetico-Methodology that applies differential diagnosis, probability algorithm and pattern-matching algorithm to ease diagnosis. It will incorporate an online patient handling technique for an agile approach to medication and also create a medical database for doctors' and patients' references.

Keywords—Smart Assistant (SA), Fuzzy logic, Hypothetico-Methodology, differential diagnosis, probability algorithm, pattern-matching algorithm.

I. INTRODUCTION

In the world of an active life, health is a factor of serious concern. Due to increase in various health hazards, the health of entire mankind is degrading since a last few decades. People are not receiving better means of diagnosis and analysis which is required at a higher priority.

In the traditional treatment methods, the patients need to schedule various appointments to meet doctors from a week prior till the treatment suffices. This takes at least three or four consultations, that include bringing reports to the doctor after performing the test, taking the prescribed medications and the final health feedback about patient's recovery. The entire procedure takes almost a month or two. By the use of the Smart Assistant the patients can be treated approximately in three to four days.

Fig.1. Smart Assistant System Overview

The Smart Assistant software helps to reduce long waiting in clinical queues and also increases diagnosis speed by suggesting the doctor possible diseases based on symptoms. SA gives suggestions to the doctor about the medicines that can be prescribed.

=210mm =297mm

The main idea of this paper is the need to provide an automated clinical software to the doctors so that they can perform differential diagnosis. By this assistance the doctors will be able to find an approach quickly to the diseases from patient's symptoms. With the help of the advanced computing methods, it will give the doctor a suggesting output which will contain a list including a number of pathological tests and the medications which should be given to the patient. All the interactions performed by the doctor and the patients is done online hence it is speedy.

The next section of this paper describes the analysis for proposing the system and the need of the system based on the survey conducted by the authors of the paper. Section III introduces the details about the actual proposed system. Section IV comprises of methodologies used for the implementation of Smart Assistant. In Section V the conclusions of the paper are presented.

II. ANALYSIS FOR SMART ASSISTANT

The Smart Assistant system is analysed by the authors for its needs in the market and how useful it will be for the doctors as well as the patients. Two surveys were carried out one for the doctors and other for the patients. The survey results were positive.

A. Patient Survey

A survey was carried out in the form of a questionnaire to a group of people regarding their medical experiences. The survey included people of various age groups and from different regions. This survey was carried out at public places. The questionnaire includes a set of questions such as a normal patient's traditional approach to the clinician and also made a comparison of the software-based system. The Fig.2 shows the pie chart of people surveyed that are in favour of SA, had a different opinion for SA and also those people who didn't have particular answer for it.

This survey focused on the following points:

- Doctor consultation frequency
- Recovery time period
- Medical document management
- Medication Assurance

From the Fig.2. approximately 65% of the people were in favour of the system as they find it much simpler than the traditional approach. Around 25% of the people surveyed were against or not in the favour of the system and it can be probably said that they were unimpressed by the non-human software and prefer a personal human touch. Remaining 10% people didn't have proper answers as they were having certain agreement to the points of the software-based system and disagreed to others and so were with mixed views.

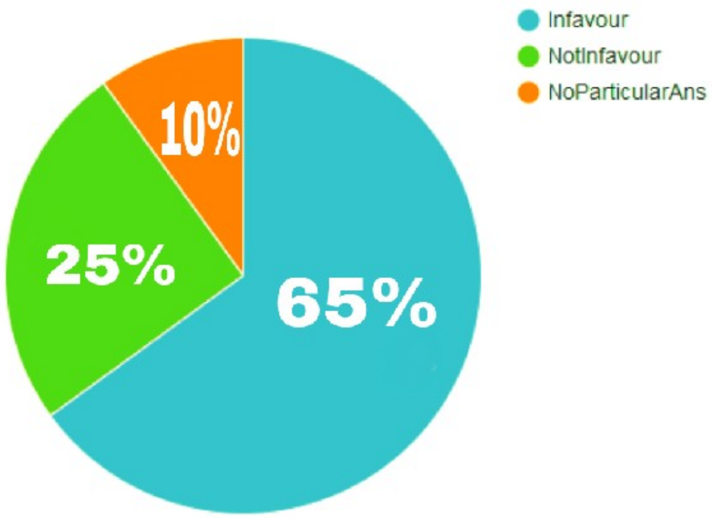

Fig. 2. Pie Chart of the Survey

=210mm =297mm

B. Doctor's Survey

The doctor's survey was conducted at various medical facilities. This survey was based on the following points:

- Technical Incorporation for Treatment
- Fast Diagnosis
- Digital Doctor-Patient Interaction
- Database of Medical Record

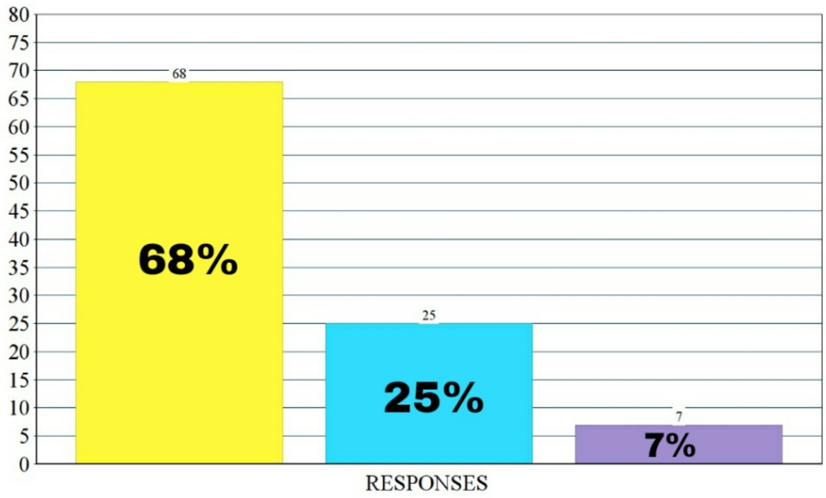

Fig.4. Doctor's Survey Bar Chart

68% of the doctors agreed that SA will be a great advancement in the medical field as every patient cannot just be treated faster but even their medical histories in their E-Reports can be saved in database for further references in medical history. In the following Fig.4. survey results are conveyed. The 27% doctors who voted no had a different opinion about SA's Application in medical field, while 7% of the doctors felt it is not presently mandatory.

III. PROPOSED SYSTEM

The Smart Assistant is a software-based Application where users are Doctors and Patients. The rights and data availability of the users will be in accordance with the login type. The medical history of the patients will be uploaded to the software. Doctors will prescribe a medical prescription on the basis of the optimised output generated by SA. (List of tests for deduced diseases, medication available). The system is comprised of a number of elements that are shown in Fig.3. Smart Assistant Architecture Diagram.

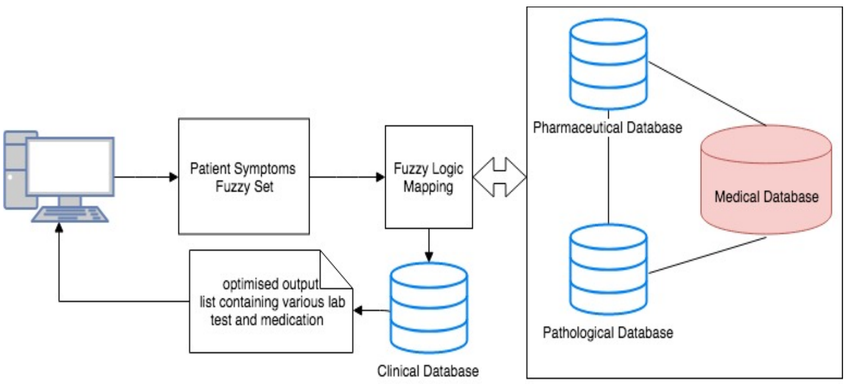

Fig. 3. Smart Assistant Architecture Diagram

A. Fuzzy Logic Input

The patient enters the symptoms faced, as an input. This set of input is a fuzzy set, which does not clearly indicate the knowledge required for illness identification. The underlying table shows the features and patient's response attributes, that give the symptom output. The Mapping algorithms match the given symptom to the disease [1] [4] [10].

TABLE 1: Fuzzy Logic Set for Mapping Different Symptoms in SA.

Attributes / **Features**	Very Low	Low	Moderate	High	Very High
Runny nose	0.1	0.2	0.5	1.0	1.0
Fever	0.1	0.5	0.5	0.1	0.2
Cough	0.3	0.5	0.5	0.5	0.5
Body aches strong	0.2	0.2	0.5	0.2	0.3

B. Mapping Algorithm

Different mapping algorithms are used to match the given input fuzzy set to the possible diseases. SA analyses the input and suggests the pathological tests and medicines. The Doctor can directly send the list of tests to the patients. Mapping Algorithm implements Hypothetico-deductive methodology. This is discussed in the Section IV of the paper.

C. Databases

The Smart Assistant utilises a number of databases to manage the data of the patients and doctors.

- **Clinical Database**

  Clinical Database is a local database managed by each clinic differently. These databases store the details of the patient's appointments, patient's medical history. This database also contains the E-Reports of the patients. All the databases are connected to the clinical database. The medical bills and the medical prescriptions are also stored here.

- **Pathological Database**

  Pathological Database is responsible to store the reports and the laboratory related data of the patient. It also stores the names and expenses of the tests.

- **Pharmaceutical Database**

  Pharmaceutical Database is responsible to store medicine details corresponding to its relative disease.

- **Medical Database**

  Medical database is an integrated database that comprises of all the databases with respect to each patient's data and disease related tests.

II. PROPOSED METHODOLOGY

The previous chapter introduces, why the system is required and many reasons to implement the concept. This section deals with methods to detect the disease of the patient at a swift rate by eliminating the traditional diagnosis method with the help of computing and engineering. The proposed system can be implemented with the concept of various methodology and mapping algorithms.

A. Differential Diagnosis

A differential diagnostic procedure is a systematic diagnostic method which is used to identify the presence of a particular disease where multiple alternative diseases are possible with the given set of symptoms.

This method is an elimination process which eliminates the disease which are not matching with the symptoms by using the information such as patient history, symptoms, and medical knowledge (or, for computerized or computer-assisted diagnosis, the software of the system). Differential diagnosis can be implemented by the concept of the Hypothetico-deductive method. In this methodology the presence of candidate diseases or conditions can be viewed as hypotheses that physicians or computing system further determine as being true or false [2] [3].

B. Probability Algorithm

Probability algorithm is used to calculate various probabilities from a given input fuzzy set which contains random data of symptoms. With the help of this algorithm the system will short list the possible diseases from a given fuzzy set of data. It will sort the data by giving preference to that data which has a likelihood of matching with medical data.

If a patient whose symptoms are similar to a specific disease which may have the relevant data analyzed against existing information. On the basis of the available matches, his physician will be able to determine which disease is the best possible match. This allows the doctor to create a careful treatment plan for the patient for a better chance of receiving the right treatment on time.Fig.5. shows the working of the probability algorithm in the system [7][8].

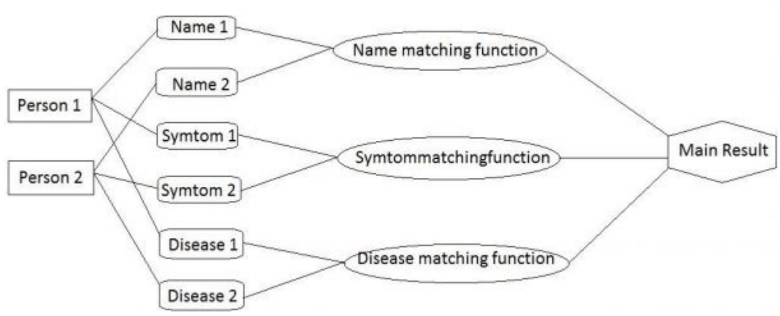

Fig.5. Probability Algorithm working

C. Hypothetico-Deductive mwthodology

Hypothetico-deductive methodology is the life-cycle method in which each phase follows a continuous cycle. The Fig.4 shows the working of the methodology. The steps involved are patient input data analysis, observation and identification of diseases according to the input, collection of the similar data for further references of a similar case [6] [11].

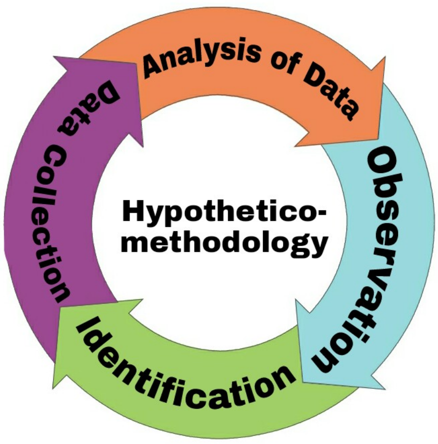

Fig.4. Hypothetico-Methodology Working Diagram

The Fig.4. Represents the lifecycle of Hypothetico-deductive methodology, which comprises of following phases:

- **Observation:**
  Analyses symptom inputs for possible diseases.
- **Identification:**
  Finalises diseases by Probability algorithm (Mapping).
- **Data Collection:**
  Collects detailed data of finalised diseases from medical database.
- **Analysis of Data:**
  Gives optimised output list of diseases, tests and medication.

D. Pattern Matching Algorithm

A pattern matching algorithm is used to map various symptoms of the patient over the corresponding diseases using computational formulation or data tables which can be used for diagnosis in healthcare. Decision tree path approach is used by pattern matching algorithm for healthcare treatment (e.g., if symptoms V, X, Y, and Z are detected, then use treatment X).

According to this algorithm, for every user input (Symptoms for disease) the software will search the disease and give a precise output as the number of tests to be performed for the cause. It will match the user's input symptoms to the diseases and also store it for future references [9].

V. CONCLUSION

The rapid lifestyle leads to degeneration of human health at a higher rate compared to its treatment. By the surveys and the analysis of the paper, a system that can be of doctor's aid (the Smart Assistant) is highly viable and needed for betterment of the treatment methods. As researched above SA not only speeds up the treatment but also creates an information database of the users for the ease of medical report handling. The implementation of the Smart Assistant is required for a healthier future as the world moves forward with digitalization.

ACKNOWLEDGMENT

We are thankful to our Management of MCT's Rajiv Gandhi Institute of Technology, Mumbai, India for giving us an opportunity and encouragement. We are also thankful to Dr. Udhav Bhosle (principal) and Dr. S.Y. Ket (HOD. Department of Computer) for all the guidance and mentoring.

REFERENCES

[1] Barna Iantovics, "The CMDS Medical Diagnosis System", Romania, Ninth International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, IEEE , pp 243-253, 2008. (reference).

[2] Anthony Farrugia, Dhiya Al-Jumeily, Mohammed Al-Jumaily, Abir Hussain, David Lamb, "Medical Diagnosis: are Artificial Intelligence systems able to diagnose the underlying causes of specific headaches?", 2013 Sixth International Conference on Developments in eSystems Engineering, IEEE Conference, pp 376-382, 2013.

[3] N. Conigliaro, A. Di Stefano and O. Mirabella, "An expert system for Medical diagnosis", IEEE, pp 75-81, 1998.

[4] Hongxu Yin, "A Health Decision Support System for Disease Diagnosis based on Wearable Medical Sensors and Machine Learning Ensembles", IEEE Transactions on Multi-Scale Computing Systems, pp 228 – 241, DEC-2017, in press.

[5] Farzad Mahootian and Tinothy E. Eastman, "Complementary frameworks of scientific inquiry: Hypothetico-Deductive, Hypothetico-Inductive, and Observational-Inductive", World Futures, pp 61–75, 2009.

[6] Vue Zhang and lie Liu, "An Outlier Mining Algorithm Based on Probability", 2009 2nd International Conference on Power Electronics and Intelligent Transportation System, IEEE, pp 209-212, 2009.

Websites references:

[7] Inductive reasing: https://en.m.wikipedia.org/wiki/Inductive_reasoning

[8] Algorithmic probability: https://en.m.wikipedia.org/wiki/Algorithmic_probability

[9] Pattern matching: https://www.geeksforgeeks.org/searching-for-patterns-set-2-kmp-algorithm/

[10] Fuzzy logic: https://www.kau.edu.sa/Files/0052366/Researches/29055_Med%20Fuzzy%20ES.doc

[11] Hypothetico-model: https://en.wikipedia.org/wiki/Hypothetico-deductive_model