### ABSTRACT

# The accurate diagnosis of Mental Health (AD) plays an important role in patient treatment, especially at the disease’s early stages, because risk occurrence of brain damage.

# AD can be diagnosed-but not predicted-at its early stages, as prediction is only applicable before the disease manifests itself. Deep Learning (DL) has become a common technique for the early diagnosis of AD.

# Here, we introduce prediction on AD using Recurrent Neural Network help researchers diagnose the disease at its early stages.

### CHAPTER 1

**INTRODUCTION**

Mental Health (AD) creates memory loss and cognitive decline; it is a neurological disorder, and neuro degenerative type of dementia, Autism occurs because of brain cell death Taylor (2006). The size of brain gets shrinks with Autism - the tissue has increasingly minimum nerve cells connections (Glenner, 1990). The following figure 1.1 shows the existence of AD on neuron cells.



## Figure 1.1. Nerve cells (neurons) in the brain, in Autism, there are microscopic 'plaques' and 'tangles' between and within brain cells

While they can't be seen or tried in the living brain influenced by Mental Health, postmortem/autopsy will dependably indicate minor incorporations in the nerve tissue, called plaques and tangles:

* Plaques – it is identified among the dying cells in the brain - from the increase of a protein called beta-amyloid.
* Tangles – it is identified within the brain neurons- due to the disintegration of another protein, called tau.

The unusual protein bunches, in the cerebrum tissue are constantly present with the disease, however there could be another fundamental procedure that is really causing the AD, and researchers are not yet beyond any doubt. This kind of progression in brain nerves is seen in different scatters, and analysts need to discover something other than that there are protein variations from the norm - they additionally need to know how these grow with the goal that a cure or aversion may be found. Specialists don't completely comprehend why the progressions that prompt Autism illness. A few distinct components are accepted to be included. Hazard factors for building up the condition which incorporates aging, a family history of Autism, and carrying certain genes.

## 1.1.1. SYMPTOMS OF MENTAL HEALTH

Indications can be analyzed at any phase of Autism dementia and the movement through the phases of the infection is observed after an underlying analysis when creating the manifestations directs how mind is overseen. Obviously, the side effects can mystify both a patient and the general population around them, with various levels of seriousness. Consequently, and on the grounds that indications could flag any of various analyses, it is constantly beneficial consulting with a neuro specialist.

For neuro specialists to make an underlying conclusion of Autism sickness, they should first be fulfilled that there is an existence of dementia. It includes cognitive or behavioral symptoms that demonstrate a decrease from past levels of "working and performing" and meddle with capacity to operate at work or at common exercises (McKhann et al., 2011).

## 1.1.2. STAGES OF MENTAL HEALTH

The sequence of Autism can be analyzed into three basic stages (Hyman et al., 2012):

* Preclinical
* Mild cognitive impairment
* Dementia.

The Autism Association has separated this further, portraying seven phases along a continuum of cognitive decline in light of side effect seriousness - from a condition of no impedance, through gentle and direct decrease, and in the end achieving "exceptionally extreme decay".

The affiliation has distributed as seven phases (Glenner, 1990). It isn't for the most part until arises to the fourth phase for a clear determination - here it is known as mild or early-stage Mental Health, and “a cautious restorative meeting ought to have the capacity to distinguish obvious manifestations in different regions.”

## 1.1.3. MENTAL HEALTH RISK FACTORS

A few things are all the more ordinarily connected with Autism ailment - not seen so frequently in individuals without the turmoil. These components may along these lines have some immediate association. Some are preventable or modifiable variables for instance, reducing the danger of diabetes or coronary illness may thus cut the danger of dementia. On the off chance that analyst's acquire comprehension of the hazard factors, or logically demonstrate any “cause” association for Autism, this could discover approaches to anticipate it or create medications. Hazard factors related with Autism sickness incorporate (DeFina et al., 2013):

## Unavoidable risk factors

* Age - the turmoil is more probable in older individuals and a more prominent extent of more than 85-year-olds , than of more than 65’s (Glenner ,1990).
* Family history - having Autism in the family is related with higher hazard. This is the second greatest hazard factor after age.
* Having a specific quality puts a man, contingent upon their particular hereditary qualities, at three to eight times more hazard than a man without the quality. Various different qualities have been observed to be related with Mental Health (Tanzi, 2012).

## Potentially avoidable or modifiable factors

* Factors which maximizes the blood vessel risk - involves diabetes, high cholesterol and high blood pressure.
* Prior head injury. While a traumatic brain injury does not essentially direct to Autism, some research links have been drawn, with increasing risk tied to the harshness of trauma history.
* Sleeping Disorders.

## 1.1.4. TESTS AND DIAGNOSIS FOR MENTAL HEALTH

Mental Health isn’t easy to identify – we doesn’t have single test to confirm it. So, basically doctors will proceed to reject other issues for the individuals, before affirming whether mental signs and symptoms are rigorous to be a variety of dementia or something else. Doctors may:

* Take a history
* Do a physical tests to identify any signs of, for instance, a stroke, heart condition or kidney disease, and
* Validate the neurological function, e.g. by studying senses and reflexes.

Based on the consultation with neurologist, the other diagnosis is to proceed are:

* Testing the blood and urine samples.
* Organizing brain scans.

At some times, dementia symptoms were associated with an inherited disorder like Huntington disease, this leads to do the genetic testing. For few people, evaluations like depression can be done.

# 1.2. IMAGE PROCESSING IN MENTAL HEALTH DIAGNOSIS

In Autism the aggregate of brain size shrinks and the tissue has continuously lesser nerve cells and associations. In general brain structure also changes like body structure as we get older. Therefore it reduces thinking and remembering ability. Perception of brain structure and capacity from the level of individual particles are to the entire brain image. Many imaging methods are noninvasive and enable dynamic procedures to be checked after some time.

Imaging is nothing but empowering scientists to distinguish Neural Networks (NNs) engaged with cognitive processes; comprehend disease pathways; perceive and analyze sicknesses early, when they are most adequately treated; and decide how treatment work.

Also, as in different zones of biomedical research, opportunities are intelligent. Understanding that, imaging can help and give a superior comprehension about how the medication or a treatment functions at the molecular level, prompting a more exact comprehension of the disease procedure.

Diverse sorts of imaging methods are utilized to uncover brain structure, physiology, and biochemical activities of individual cells and of the molecules that make them, cells capacities, practices and associations. The three fundamental classes, in this manner, regularly alluded to as structural, functional and molecular imaging. While many imaging procedures are utilized all through the body, the portrayals gave here concentrate on their utilization in the nervous system, principally the cerebrum.

## 1.2.1. IMAGING’S EVOLUTION

Early Structural Imaging Techniques with numerous auxiliary and practical imaging systems are moderate , later the inception of X-ray. This innovation was the clinician's primary imaging instrument for the greater part of the 20th century. For brain imaging, various X-ray beams are gone through the head at various points. Extraordinary sensors measure the measure of radiation that is consumed by various tissues. At that point, a PC utilizes the distinctions in X-ray assimilation to shape cross-sectional images or "slices" of brain called "tomograms". Indeed, even it can't be seen or tried in the living brain of

Mental Health influenced patients, however by postmortem/autopsy examination will demonstrate the modest considerations in the nerve tissue, called as plaques and tangles.

## 1.2.2. COMPUTING TOMOGRAHY (CT)

Imaging is turning into an inexorably critical apparatus in both research and clinical care. A scope of imaging advances now give extraordinary affectability to Computed Tomography (CT) imaging which was the principal procedure, to demonstrate clear confirmation, amid life, of declines in the measure of brain tissue in more older contrasted with youthful individuals. CT can be utilized to indicate bone, delicate tissues and veins in similar images. Since CT is particularly helpful in emergency trauma situations, demonstrating any anomalies in brain structure including brain swelling, hemorrhagic stroke, and head injury.

## 1.2.3. MAGNETIC RESONANCE IMAGING (MRI)

Magnetic Resonance Imaging is a medical imaging technique used in radiology to form pictures of the anatomy and the physiological processes of the body. In any case, new ultrasound strategies started to develop; these procedures utilize laser innovation to consolidate data from both light and sound and have turned into an essential piece of concentrated observing of cerebral blood stream in patients with extreme head injury. These innovations and their uses are portrayed later in the area on Electrical and Doppler Ultrasonic Imaging Techniques.

## 1.2.4. POSITRON EMISSION TOMOGRAPHY (PET)

Positron Emission Tomography (PET) was considered as a first major technology to compute the physiological functioning of the brain. In PET scanning, the regional distribution of exogenously managed positron emitting tracers is computed with the help of tomographic imaging.

## 1.2.5. ELECTROENCEPHALOGRAPHY (EEG)

EEG computes the electrical movement that is delivered by neurons as recorded from electrodes set along the scalp. Magneto Encephalo Graphy(MEG) maps brain movement by measuring magnetic fields that are produced by neural action in the brain. Both EEG and MEG give data about neural action, however with MEG there is less mutilation of the electrical signs. Regularly either of these electrophysiological techniques is joined with MRI to give integral data about ordinary and exasperates functioning of brain.

## 1.2.6. MAGNETOENCEPHALOGRAPHY

MEG, by computing the magnetic fields, it is utilized to explore the premise of sensory processing and engine arranging in the brain. Magneto Encephalography (MEG) is utilized with MRI in brain tumor patients preceding their surgery to distinguish the side of the equator controlling languages and to decisively find the areas engaged with expressive and responsive languages so specialists can save these areas amid surgery.

## 1.2.7. NEAR INFRARED SPECTROSCOPY

Near Infrared Spectroscopy (NIRS) is an optical strategy for measuring blood oxygenation in the brain. It works by sparkling light in the near infrared part of the spectrum (700-900n m) and distinguishing how much the re blending light is decreasing and furthermore how much the light is weakened relies upon blood oxygenation. In this way NIRS can give a roundabout measure of brain action.

# 1.3. IMAGE PREPROCESSING

The fundamental meaning of image processing alludes to expelling the noise and any sort of inconsistencies exhibit in an image utilizing the advanced

PC’s (Geetha and Chitradevi, 2014). The noise or inconsistency may crawl into the image among its development. Different procedures have been created in image processing amid the last four to five decades. A large portion of the methods are created for upgrading images acquired from spacecrafts, space probes and military reconnaissance flights. Image Processing systems are getting popular because of simple accessibility of capable PCs, extensive size memory gadgets, designs programming and so on. To get over such blemishes and to get innovation of data, need to experience different periods of preparing methods are:

* Image preprocessing
* Image enhancement
* Image analysis
* Image segmentation

## 1.3.1. IMAGE PREPROCESSING

The initial phase in the image processing chain comprises of preprocessing images. The Pre-Processing is a common name for operations with images at the lowest level of abstraction- both input and output are intensity images. The aim of Pre-processing is an improvement of the image data that suppresses unwanted distortions or enhances some image features important for further processing. Image of preprocessing is to expand the nature of an image by reducing the measure of noise showing up in the image and highlighting features that are utilized as a part of image division.

## 1.3.2. IMAGE ENHANCMENT

Image upgrade is the adjustment of image by modifying the pixel brightness esteems to enhance its visual effect. Image upgrade includes a gathering of procedures that are utilized to enhance the visual appearance of a image to a shape which is more qualified for human. The preprocessed images will have some noise which ought to be expelled for the further handling of the image. Image noise is most obvious in image areas with low flag level, for example, shadow regions or under uncovered images.

In image enhancement, the objective is to emphasize certain image features for image display (Jahne ,2005). Cases incorporate difference and edge upgrade, pseudo-coloring, noise filtering, sharpening, and magnifying. Image improvement is helpful in feature extraction, image analysis and an image display. The upgrade procedure itself doesn't expand the intrinsic data content in the information. It essentially underlines certain predetermined image attributes. Enhancement algorithms are by and large intelligent and application subordinate. For the most part image filtering methods were utilized for image upgrade. In this research work, the hybrid median filtering method is implemented for upgrading the nature of the image and to increase the lucidity of the specimens.

## 1.3.3. IMAGE ANALYSIS

Richards and Richards (1999) proposed an image analysis for making quantitative estimations from an image to deliver a depiction. In the least difficult shape, this undertaking could read a mark on a grocery item, arranging diverse parts on an assembly line, or measuring the size and introduction of platelets in a medicinal image. Further developed image analysis systems measure quantitative data and utilize it to settle on a modern choice, for example, controlling the arm of a robot to move a question in the wake of recognizing it or exploring an air ship with the guide of images obtained along its direction. Image examination methods require extraction of specific features that guide in the recognizable proof of the object.

## 1.3.4. IMAGE SEGMENTATION

Image segmentation is the procedure that subdivides an image into its constituent parts or objects. Segmentation is a standout amongst the most broadly utilized strides in reducing images to data (John, 1999). Segmentation techniques were utilized to confine the coveted object from the scene with the goal that estimations have been made on it consequently. The level to which this subdivision is done relies upon the issue being tackled, i.e., the segmentation should end when the objects of enthusiasm for an application have been separated from the image and afterward to portion the substance of the road down to potential vehicles. The algorithms for segmentation of images depend on discontinuity or similarity in pixel esteems. In the irregularity approach an image is partitioned on the premise of sudden changes in intensity, edges, points, blobs, and lines in the image. In similar approach, an arrangement of pre-characterized criteria is utilized for partitioning the image into areas that are comparable as per the model set. Thresholding, region growing, and region splitting and merging are cases of similarity-based segmentation techniques.

## Relevance of the Project

The major problem with using AI for the diagnosis of disease is the lack of data for training predictive models. Though there is vast amount of data including mammograms, genetic tests, and medical records, they are not open to the people who can make use of them for research. Some initiatives like “100,000 Genomes Project” in the UK, the U.S. Department of Veteran Affairs’ “Million Veteran Program”, and the NIH’s “The Cancer Genome Atlas” will hopefully provide data to researchers and data scientists.

In many countries’ health records are being digitized. The adoption of EMR is also increasing. According to a data brief by The Office of National Coordinator for Health Information Technology (ONC), 3 out of 4 private or not-for-profit hospitals adopted at least a Basic EHR system in the US. In many other countries, different EHR systems exist. The Stockholm EPR corpus is a great example of such systems which consists of data from 512 clinical units with over 2 million patient records. India is thinking about setting up a National eHealth Authority (NeHA) during the Digital India program. These types of electronic health documents provide a huge amount of data for intelligence data analysis. Many research have been conducted on predicting various diseases like Liver Disease, Alzheimer Disease, Dementia etc., detecting tumours, leukaemia etc. using computer vision, assisting doctors in making efficient decisions.

## Problem Statement

The primary goal is to develop a prediction engine which will allow the users to check whether they have Dementia or Alzheimer Disease sitting at home. The user need not visit the doctor unless he has Dementia or Alzheimer Disease, for further treatment. The prediction engine requires a large dataset and efficient machine learning algorithms to predict the presence of the disease. Pre-processing the dataset to train the machine learning models, removing redundant, null, or invalid data for optimal performance of the prediction engine.

Doctors rely on common knowledge for treatment. When common knowledge is lacking, studies are summarized after some number of cases have been studied. But this process takes time, whereas if machine learning is used, the patterns can be identified earlier. For using machine learning, a huge amount of data is required. There is very limited amount of data available depending on the disease. Also, the number of samples having no diseases is very high compared to the number of samples having the disease. This project is about performing two case studies to compare the performance of

various machine learning algorithms to help identify such patterns in (i) and to create a platform for easier data sharing and collaboration.

## Objectives

The primary aim of this project is to analyze the “Pima Indian Dementia Dataset” and “Alzheimer Dataset” and use Recurrent Neural Network for prediction.

The secondary aim is to develop a web application that allows users to predict Alzheimer Disease utilizing the prediction engine.

## Scope of the Project

The disease diagnosis system will permit end-users to predict Alzheimer Disease .

#### Growth of AI Systems

Artificial Intelligence is one of the hottest topics today. The revenue for cognitive and artificial intelligence systems is expected to hit $12.5 billion.

#### Availability of Doctors and Chatbots

Other than disease diagnosis, artificial intelligence can be used to streamline and optimize the clinical process. There is only one doctor for over 1600 patients in India

.AI health assistants can help in covering large part of clinical and outpatient services freeing up doctor’s time to attend more critical cases. Chatbots like “Your.MD” can assist patients by understanding patients’ symptoms and suggest easy-to-understand medical information about their condition. Other assistants like “Ada” integrated with

“Amazon Alexa” provides a detailed symptom assessment report and also provides an option to contact a real doctor. Such assistants make use of Natural Language Processing and Deep Learning to understand the user and generate suggestions.

* + 1. **Internet of Things (IoT), Healthcare and Machine Learning** Increasing use of Internet of Things has promising benefits in healthcare. Dynamically collecting patient data using remote sensors can help in early detection of health problems and aid in preventive care.

## 1.5 Project Methodology

### Agile:

Agile is a process by which a team can manage a project by breaking it up into several stages and involving constant collaboration with stakeholders and continuous improvement and iteration at every stage. It promotes **continuous iteration** of development and testing throughout the software development life cycle of the project. Both development and testing activities are concurrent.

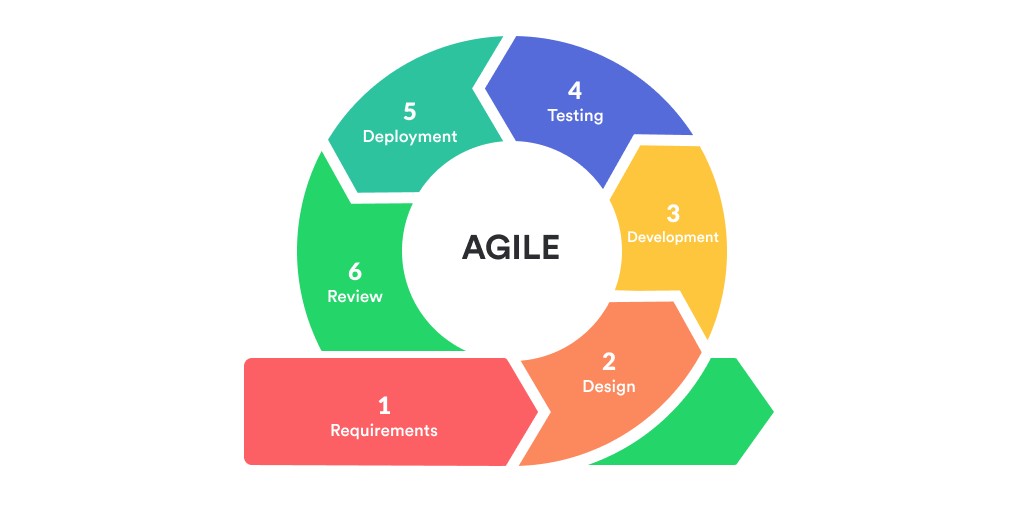


Fig 1.1 - Agile Methodology

### Scrum

SCRUM is an agile development method which concentrates specifically on how to manage tasks within a team-based development environment. Scrum encourages teams to learn through experiences, self-organize while working on a problem.

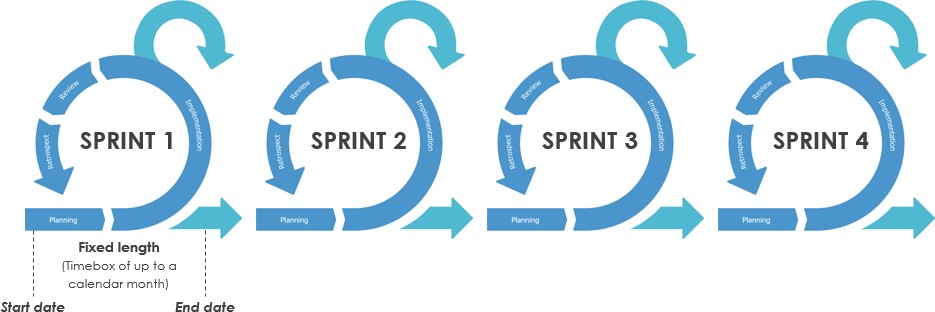


Fig 1.2 – Scrum

### The Main Artefacts

* **Product Backlog** is the master list of work that needs to get done maintained by the product owner or product manager.
* **Sprint Backlog** is the list of items, user stories, or bug fixes, selected by the development team for implementation in the current sprint cycle.
* **Increment** (or Sprint Goal) is the usable end-product from a sprint.

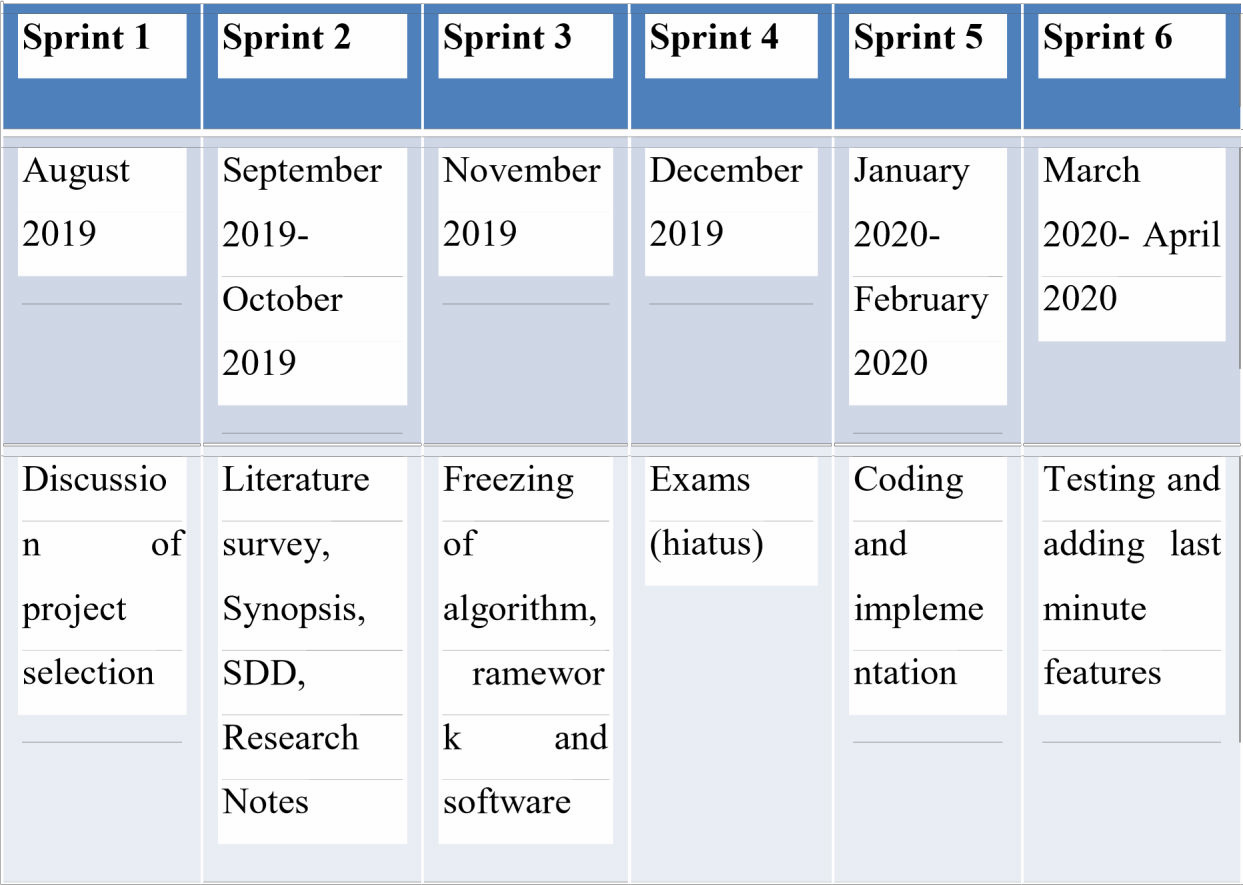


Table 1.1 – Sprint

### CHAPTER 2

**LITERATURE SURVEY**

The following chapters give an overview of the various methodologies used by various authors for disease prediction using machine learning methodologies. We can observe that there is fine comparison made between 5 major machine learning algorithms whether they are able to predict the presence of the disease with a greater accuracy, achieving optimal performance. The research efforts presented by the authors in the following papers are focused in developing and evaluating a web-based tool for disease prediction

### Dementia Prediction Using Different Machine Learning Approaches

##### Author: Priyanka Sonar, Prof. K. JayaMalini

**Published In: Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC 2019)**

**The authors have used Machine Learning approaches to predict Dementia [1]**

Dementia is one of lethal diseases in the world. It is additional an inventor of various varieties of disorders for example: coronary failure, blindness, urinary organ diseases etc. In such a case the patient is required to visit a diagnostic centre, to get their reports after consultation. Due to every time they must invest their time and currency. But with the growth of Machine Learning methods we have got the flexibility to search out an answer to the current issue, we have got advanced system mistreatment information processing that has the ability to forecast whether the patient has polygenic illness or not. Furthermore, forecasting the sickness initially ends up in providing the patients before it begins vital. Information withdrawal has the flexibility to remove unseen data

from a large quantity of Dementia associated information. The aim of this analysis is to develop a system which might predict the diabetic risk level of a patient with a better accuracy. Model development is based on categorization methods as Decision Tree, ANN, I Bayes and SVM algorithms. For Decision Tree, the models give precisions of 85%, for I Bayes 77% and 77.3% for Support Vector Machine. Outcomes show a significant accuracy of the methods.

Conclusion: SVM is very good when we have no idea on the data. Even with unstructured and semi structured data like text, images, and trees SVM algorithm works well. The drawback of the SVM algorithm is that to achieve the best classification results for any given problem, several key parameters are needed to be set correctly. Decision tree: It is easy to understand and rule decision tree. Instability is there in decision tree, that is bulky change can be seen by minor modification in the data structure of the optimal decision tree. They are often relatively inaccurate. I Bayes: It is robust, handles the missing values by ignoring probability estimation calculation. Sensitive to how inputs are prepared. Prone bias when increase the number of training dataset. ANN: Gives good prediction and easy to implement. Difficult with dealing with big data with complex models. Require huge processing time.

### Implementation of a Web Application to Predict Dementia: An Approach Using Machine Learning Algorithm,

##### Authors: Samrat Kumar Dey, Ashraf Hossain and Md. Mahbubur Rahman

**Published In: 2018 21st International Conference of Computer and Information Technology (ICCIT)**

**The authors design and develop a web application to predict Dementia [2]**

Dementia is caused due to the excessive amount of sugar condensed into the blood. Currently, it is considered as one of the lethal diseases in the world. People all around the globe are affected by this severe disease knowingly or unknowingly. Other diseases

like heart attack, paralyzed, kidney disease, blindness etc. are also caused by Dementia. Numerous computer-based detection systems were designed and outlined for anticipating and analysing Dementia. Usual identifying process for diabetic patients needs more time and money. But with the rise of machine learning, we have that ability to develop a solution to this intense issue. Therefore, we have developed an architecture which has the capability to predict where the patient has Dementia or not. Our main aim of this exploration is to build a web application based on the higher prediction accuracy of some powerful machine learning algorithm. We have used a benchmark dataset namely Pima Indian which can predict the onset of Dementia based on diagnostics manner. With an accuracy of 82.35% prediction rate Artificial Neural Network (ANN) shows a significant improvement of accuracy which drives us to develop an Interactive Web Application for Dementia Prediction.

Conclusion: In this paper, authors proposed a web-based application for the successful prediction of Dementia Diseases. From different machine learning algorithms Artificial Neural Network (ANN) provides highest accuracy with Min Max Scaling Method on Indian Pima Dataset. As they have proposed and developed an approach for Dementia disease prediction using machine learning algorithm, it has significant potential in the field of medical science for the detection of various medical data accurately. In the near future, focus is to use a deep learning model and prepare a Location based Dataset from real medical data for the successful prediction of Dementia disease

### Dementia Disease Prediction Using Data Mining

##### Authors: Deeraj Shetty, Kishor Rit, Sohail Shaikh and Nikita Patil

**Published In: 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)**

**The author talks about the significance of dataset in prediction of the disease.**

Data mining is a subfield in the subject of software engineering. It is the methodical procedure of finding examples in huge data sets including techniques at the crossing

point of manufactured intelligence, machine learning, insights, and database systems. The goal of the data mining methodology is to think data from a data set and change it into a reasonable structure for further use. Our examination concentrates on this part of Medical conclusion learning design through the gathered data of Dementia and to create smart therapeutic choice emotionally supportive network to help the physicians. The primary target of this examination is to assemble Intelligent Dementia Disease Prediction System that gives analysis of Dementia malady utilizing Dementia patient’s database. In this system, we propose the use of algorithms like Bayesian and KNN (K- Nearest Neighbour) to apply on Dementia patient’s database and analyse them by taking various attributes of Dementia for prediction of Dementia disease.

Expected Result: The goal of our project is to know whether the patient is diabetic or not, patient will be diagnosed and it will be depending on the attributes that we are going to take, such as age, pregnancy, pg concentration, tri fold thick, serum ins, body mass index (bmi), dp function, diastolic bp i.e. the factors which are majorly responsible for Dementia. So, to reduce the correctly know whether the patient is diabetic or not, we are developing a system which will be a prediction system for Dementia patients. Another best thing about the system is it is will give accurate results whether the patient is diabetic or not with the help of the knowledge base of the larger dataset that we are going to use added the recommendations we are going to provide based on the diabetic levels of the patients. Also, the prediction of the disease will be done with the help of Bayesian algorithm and K-NN algorithm.

Conclusion: By our in-depth analysis of literature survey, we acknowledged that the prediction done earlier did not use a large dataset. A large dataset ensures better prediction. Also, what it lacks is recommendation system. When they predict they give some recommendations to the patient on how to control or prevent Dementia in case of minor signs of Dementia. The recommendations would be such, that when followed it will help the patient. Thus build up a system which will anticipate diabetic patient with the assistance of the Knowledge base which we have of dataset of around 2000 Dementia

patients and furthermore to give suggestions on the premise of the nearness of levels of Dementia patients. Prediction will be done with the help of two algorithms Naïve Bayes and K-Nearest Neighbour and also compare which algorithm gives better accuracy on the basis of their performance factors. This system which will be developed can be used in HealthCare Industry for Medical Check of Dementia patients.

### An Intelligent System for Dementia Prediction

##### Authors: Zhilbert Tafa, Nerxhivane Pervetica and Bertran Karahoda

**Published In: 4thMediterranean Conference on Embedded Computing MECO –**

**2015 Budva, Montenegro**

**The authors develop an intelligent system for Dementia prediction.**

With the emerging increase of Dementia, that recently affects around 346 million people, of which more than one-third go undetected in early stage, a strong need for supporting the medical decision-making process is generated. A number of research have focused either in using one of the algorithms or in the comparisons of the performances of algorithms on a given, usually predefined and static datasets that are accessible through the Internet. This paper focuses on the joint implementation of the support vector machine (SVM) and Naïve Bayes statistical analysing, in the dataset acquired from the medical examinations of 402 patients, to improve the computer-supported diagnosis reliability. The dataset contains some attributes that have not been previously used in computer-based evaluations. The results show that the joint implementation of two algorithms significantly improves the overall reliability of the system outcome, which is crucial in the computer-supported Dementia diagnostic process.

Conclusions and Future Work: The research efforts presented in this paper are focused in developing and the evaluation of a computer-based support tool for the Dementia detection. The presented approach is based on the joint implementation of two algorithms in Matlab that have been executed on the newly acquired dataset with the different attributes as compared to the previous work in this field. The algorithms are

executed and evaluated independently but the decision making is based on the joint outcomes from both algorithms. The aim of this approach is to make the decision more reliable. As shown in the paper, both SVM and naïve Bayes algorithm have individually shown high overall classifier performances of 95, 52% and 94, 52%, respectively. The joint implementation on the same, newly added record leads to one of the three answers:

a) the patient is diagnosed with Dementia (or pre-diabetic condition), b) the patient is not diagnosed as having the mentioned condition, and c) the patient is further directed to the additional clinical examinations. If two algorithms show different results, the answer is classified as condition c). Otherwise, the accuracy of the answers a) or b), as shown in the paper, is improved up to the value of 97,6%. The presented methodology minimizes the false negative answers, which is a crucial issue in medical diagnoses. Finally, the construction approach, the architecture, and the evaluation of a Dementia classification tool presented in this paper, should provide an important guideline to further construction of the similar applications on improving and helping the decision making process in disease detection. The development of a user-friendly and widely accessible application would enable the personal self-screening on diabetic or pre- diabetic condition which is crucial to the disease treatment performance. The future work will focus on further quantitative evaluations of the developed tool regarding the extensive clinical examinations and results. Also, other methods should be involved in finding the best fit in the sense of accuracy, processing time, etc. The influence of cultural-related biases (such as those related to the nutrition structure and habits) should also be minimized for the results to get more generalized.

## Algorithms and Results across Publications

In 2016, Dey et al. used Principal Component Analysis (PCA) for selection of features and analysed the performance of Alzheimer Disease prediction using Naïve Bayes, Decision Trees and Support Vector Machines. After reducing the correlated variables to linearly uncorrelated variables using PCA, SVM outperformed Naïve Bayes and Decision

Trees. The authors recommend building a desktop web application using SVM to predict Alzheimer Disease.

After analysing multiple studies in the area of Alzheimer Disease prediction Darne et al. (2016) suggests the following:

* + - Neural Networks and Fuzzy techniques help in better prediction.
    - Ant Colony Optimization, Support Vector Machine can be used for better prediction.
* K-means clustering algorithm can be used to resolve data overfitting and overgeneralization.

In 2016, Purushottam et al. proposed an efficient Alzheimer Disease prediction system using decision trees. They presented the rules generated through the experiment which were further pruned and duplicates were removed and class wise rule sets were prepared. The results were tested using 10-fold cross validation. This system outperformed SVM, MLP and various other algorithms with an accuracy of 86.7%.

In 2016, Srikanth & Deverapalli performed a critical study of classification algorithms namely Decision trees and Naïve Bayes. They used the Pima Indian Dementia dataset and found the results to be accurate. They suggest enhancing the dataset using Linear regression and Logistic regression models.

In 2016, El-Baz et al. proposed a classification system using combined MLP and combined CFBN. They used the Pima Indian Dementia Dataset. At first, they used MLP and CFBN individually and obtained an accuracy of 80.21% and 81.77% respectively. Later, when they used combined techniques, the accuracy was higher than both the individual classifiers – 95.31% and 96.88% respectively.

### CHAPTER 3

**SYSTEM REQUIREMENTS SPECIFICATION**

* 1. **Functional Requirements – Dataset preparation and pre processing**

#### Data Collection

Data collection is defined as the procedure of collecting, measuring and analysing accurate insights for research using standard validated techniques. A researcher can evaluate their hypothesis based on collected data. In most cases, data collection is the primary and most important step for research, irrespective of the field of research. The approach of data collection is different for different fields of study, depending on the required information. The most critical objective of data collection is ensuring that information-rich and reliable data is collected for statistical analysis so that data-driven decisions can be made for research.

#### Data Visualization

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Example -

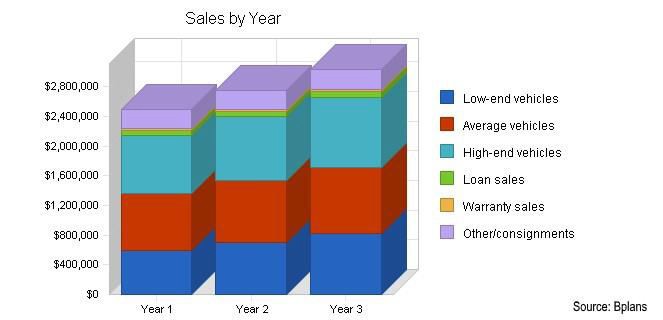


Fig 3.1 – Data Visualization

#### Data Labelling

Supervised machine learning, which we’ll talk about below, entails training a predictive model on historical data with predefined target answers. An algorithm must be shown which target answers or attributes to look for. Mapping these target attributes in a dataset is called labelling. Data labelling takes much time and effort as datasets sufficient for machine learning may require thousands of records to be labelled. For instance, if your image recognition algorithm must classify types of bicycles, these types should be clearly defined and labelled in a dataset.

#### Data Selection

Data selection is defined as the process of determining the appropriate data type and source, as well as suitable instruments to collect data. Data selection precedes the actual practice of data collection. This definition distinguishes data selection from selective data reporting (selectively excluding data that is not supportive of a research hypothesis) and interactive/active data selection (using collected data for monitoring activities/events, or conducting secondary data analyses). The process of selecting suitable data for a research project can impact data integrity. After having collected all information, a *data analyst* chooses a subgroup of data to solve the defined problem. For instance, if you save your customers’ geographical location, you don’t need to add their cell phones and bank card numbers to a dataset. But purchase history would be necessary. The selected data includes attributes that need to be considered when building a predictive model.

#### Data Pre-processing

Data pre-processing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviours or trends, and is likely to contain many errors. Data pre- processing is a proven method of resolving such issues. The purpose of pre-processing

is to convert raw data into a form that fits machine learning. Structured and clean data allows a data scientist to get more precise results from an applied machine learning model. The technique includes data formatting, cleaning, and sampling.

Data formatting. The importance of data formatting grows when data is acquired from various sources by different people. The first task for a data scientist is to standardize record formats. A specialist checks whether variables representing each attribute are recorded in the same way. Titles of products and services, prices, date formats, and addresses are examples of variables. The principle of data consistency also applies to attributes represented by numeric ranges.

Data cleaning. This set of procedures allows for removing noise and fixing inconsistencies in data. A data scientist can fill in missing data using imputation techniques, e.g. substituting missing values with mean attributes. A specialist also detects outliers — observations that deviate significantly from the rest of distribution. If an outlier indicates erroneous data, a data scientist deletes or corrects them if possible. This stage also includes removing incomplete and useless data objects.

Data anonymization. Sometimes a data scientist must anonymize or exclude attributes representing sensitive information (i.e. when working with healthcare and banking data).

Data sampling. Big datasets require more time and computational power for analysis. If a dataset is too large, applying data sampling is the way to go. A data scientist uses this technique to select a smaller but representative data sample to build and run models much faster, and at the same time to produce accurate outcomes.

#### Data Transformation

Data transformation is the process of converting data from one format or structure into another format or structure. Data transformation is critical to activities such as data

integration and data management. Data transformation can include a range of activities: you might convert data types, cleanse data by removing nulls or duplicate data, enrich the data, or perform aggregations, depending on the needs of your project.

Scaling. Data may have numeric attributes (features) that span different ranges, for example, millimetres, meters, and kilometres. Scaling is about converting these attributes so that they will have the same scale, such as between 0 and 1, or 1 and 10 for the smallest and biggest value for an attribute.

Decomposition. Sometimes finding patterns in data with features representing complex concepts is more difficult. Decomposition technique can be applied in this case. During decomposition, a specialist converts higher level features into lower level ones. In other words, new features based on the existing ones are being added. Decomposition is mostly used in time series analysis. For example, to estimate a demand for air conditioners per month, a market research analyst converts data representing demand per quarters.

Aggregation. Unlike decomposition, aggregation aims at combining several features into a feature that represents them all. For example, you have collected basic information about your customers and particularly their age. To develop a demographic segmentation strategy, you need to distribute them into age categories, such as 16-20, 21-30, 31-40, etc. You use aggregation to create large-scale features based on small- scale ones. This technique allows you to reduce the size of a dataset without the loss of information.

#### Data Splitting

A dataset used for machine learning should be partitioned into three subsets — training, test, and validation sets.

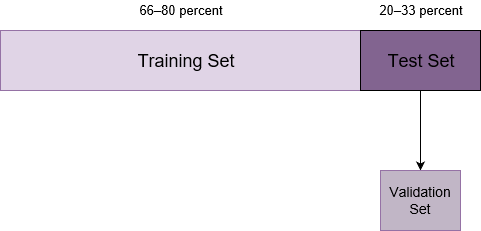
Training set. A *data scientist* uses a training set to train a model and define its optimal parameters — parameters it must learn from data.

Test set. A test set is needed for an evaluation of the trained model and its capability for generalization. The latter means a model’s ability to identify patterns in new unseen data after having been trained over a training data. It is crucial to use different subsets for training and testing to avoid model overfitting, which is the incapacity for generalization we mentioned above.

Validation set. The purpose of a validation set is to tweak a model’s hyperparameters

— higher-level structural settings that cannot be directly learned from data. These settings can express, for instance, how complex a model is and how fast it finds patterns in data.

The proportion of a training and a test set is usually 80 to 20 percent, respectively. A training set is then split again, and its 20 percent will be used to form a validation set. At the same time, machine learning practitioner Jason Brownlee suggests using 66 percent of data for training and 33 percent for testing. A size of each subset depends on the total dataset size



.

Fig 3.2 – Data Splitting

The more training data a data scientist uses, the better the potential model will perform. Consequently, more results of model testing data lead to better model performance and generalization capability.

#### Modelling

After pre-processing the collected data and split it into three subsets, we can proceed with a model training. This process entails “feeding” the algorithm with training data. An algorithm will process data and output a model that is able to find a target value (attribute) in new data — an answer you want to get with predictive analysis. The purpose of model training is to develop a model.

Two model training styles are most common — supervised and unsupervised learning. The choice of each style depends on whether you must forecast specific attributes or group data objects by similarities.

Model evaluation and testing: The goal of this step is to develop the simplest model able to formulate a target value fast and well enough. A data scientist can achieve this goal through model tuning. That is the optimization of model parameters to achieve an algorithm’s best performance. One of the more efficient methods for model evaluation and tuning is cross-validation.

Cross-validation: Cross-validation is the most used tuning method. It entails splitting a training dataset into ten equal parts (folds). A given model is trained on only nine folds and then tested on the tenth one (the one previously left out). Training continues until every fold is left aside and used for testing. As a result of model performance measure, a specialist calculates a cross-validated score for each set of hyperparameters. A data scientist trains models with different sets of hyperparameters to define which model has the highest prediction accuracy. The cross-validated score indicates average model performance across ten hold-out folds.

#### Model Deployment

Deployment is the method by which you integrate a machine learning model into an existing production environment to make practical business decisions based on data. It is one of the last stages in the machine learning life cycle and can be one of the most cumbersome. Often, an organization’s IT systems are incompatible with traditional model-building languages, forcing data scientists and programmers to spend valuable time and brainpower rewriting them.

## Non-Functional Requirements

#### Usability

The system should be easy to use. The system also should be user friendly for users because anyone can use it instead of programmers.

#### Reliability

This software will be developed with machine learning, feature engineering and deep learning techniques. So, in this step there is no certain reliable percentage that is measurable. Also, user provided data will be used to compare with results and measure reliability. With recent machine learning techniques, user gained data should be enough for reliability if enough data is obtained.

#### Performance

Processing time and response time should be as little as possible providing the result at a faster rate when compared to other methods.

#### Supportability

The system should require Python knowledge to maintenance. If any problem acquire in user side and deep learning methods, it requires code knowledge and deep learning background to solve.

## Hardware Requirements

* + - OS: Windows 10
    - RAM: Minimum of 4GB

## Software Requirements

* + - Basic Text-Editor: Visual Studio Code, Pycharm.
    - Jupyter Notebook: Development of Python Scripts.
    - Flask Framework: API Development.
    - Angular Framework: UI Development.

#### Why Python?

* + - * General purpose programming language
      * Increasing popularity for use in data science
* Easy to build end-to-end products like web applications
* Since the goal of this project is to build a web application, Python is a better choice. Though frameworks like Shiny can be used with R to create web applications, it is extremely slow.

There are two major libraries for machine learning in python: TensorFlow, ScikitLearn. The main differences are discussed below.

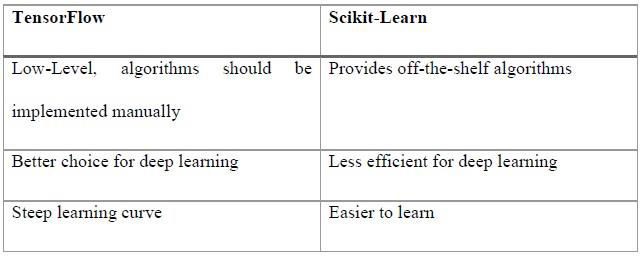


Table 3.1 – Machine Learning Libraries Scikit-learn was chosen for the project for the following reasons:

* Off-the-shelf algorithms
* Shallow learning curve compared to TensorFlow

**CHAPTER 4**

**SYSTEM ANALYSIS AND DESIGN**

## 4.1 Initial Engine Design

Before writing any code, the initial design of the algorithm for the prediction engine was created using Microsoft Visio. The initial design can be seen in Figure below.

The below algorithm is repeated for all classification algorithms. Furthermore, the accuracy score is calculated using 10-fold cross validation.

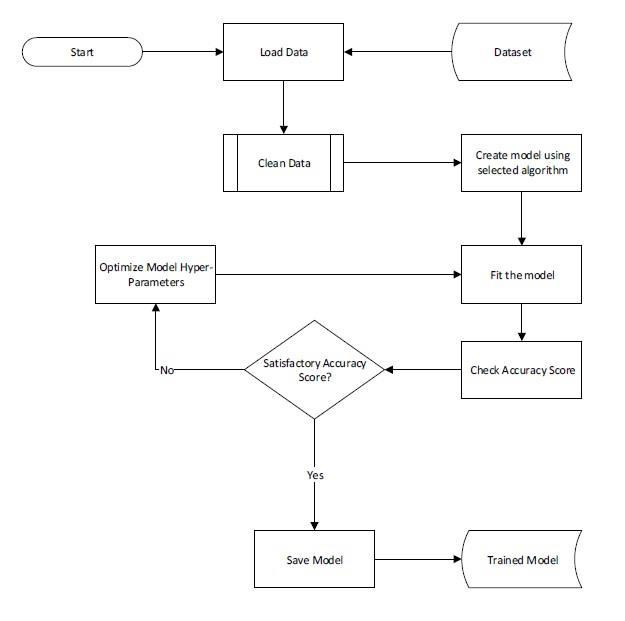


Fig 4.1 – Training Model Process

### 4.2 Use Cases

Though use cases are not recommended by agile development or feature driven development, use cases were created to better understand the requirements of the system. The use cases are packaged according to domain areas. Detailed description of only some of the use cases is present in this section

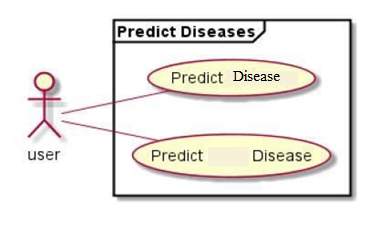


Fig 4.2 – Use Case

The description of use cases for predicting diseases shown in Figure 4.2 is given below. Use Case Number: 1

Title: Predict Alzheimer Disease

Description: The user submits the form with the values obtained after performing various medical tests. Upon submission, the user receives message indicating whether they have a risk of developing Alzheimer Disease.

Actors: User Priority: Essential

Postconditions: Patient successfully receives the prediction results.

Main Success Scenario:

1. The user performs all the tests and enters valid test results
2. The system performs prediction using the prediction engine API.
3. The user receives the prediction results. Use Case Number: 2

Title: Predict Dementia

Description: The user submits the form with the values obtained after performing various medical tests. Upon submission, the user receives a message indicating whether they have a risk of developing Dementia.

Actors: User Priority: Essential

Postconditions: Patient successfully receives the prediction results. Main Success Scenario:

1. The user performs all the tests and enters valid test results.
2. The system performs prediction using the prediction engine API.
3. The user receives the prediction results Alternative Scenario:

If the user enters invalid value for any of the provided form fields: the system will display an error message.

If the system is unable to communicate with the prediction engine API: the system will display an error message.

**CHAPTER 5**

**System specification**

* 1. **Hardware requirements**
     + **RAM** : 2 - 4GB
     + **HARD DISK** : 500 GB
     + **SYSTEM** : INTEL CORE I3,I5,I7

## Software requirements

* + - **Operating System** : Windows 7,8,10
    - **Programming Language** : Python
    - **Front End** : HTML,CSS
    - **Back End** : Python
    - **Data Base** : MySQL

### 9

**CHAPTER 6**

**IMPLEMENTATION**

* 1. **Naïve Bayes**

Naive Bayes is a set of supervised learning algorithms based on the Bayes’ theorem with the “naïve” assumption of independence between every pair of features. Despite its simplicity, it often outperforms more sophisticated classification methods.

If there are input variables x and output variable y, Bayes’ theorem states the following

relationship.

p(y|x) = p(y).p(x|y)/ p(x)

In this project, Gaussian Naïve Bayes algorithm has been implemented. In case of Gaussian Naïve Bayes, the likelihood of the features us assumed to be Gaussian i.e. all continuous values x associated with class y are distributed according to Gaussian distribution.

Given a continuous attribute x in training data, the data is first segmented by the class

y. Then, the mean and variance of x in each class is computed. If μ be the mean of the values in x associated with class y, then let d2 be the variance of the values in x associated with class y. Suppose there is some observation value v then, the probability distribution of v given by class y, p(x=v | y), can be computed by plugging into the equation for a normal distribution

## Support Vector Machine

Support vector machines (SVMs) are a set of supervised learning methods used for

classification, regression and outliers’ detection. It is effective when the number of

attributes is greater than the number of samples but when that number increases significantly, this method is likely to give poor performance.

For given labelled training data, SVM outputs an optimal hyper plane to categorize new examples. The optimal hyper plane maximizes the margin of the training data as shown below in Figure

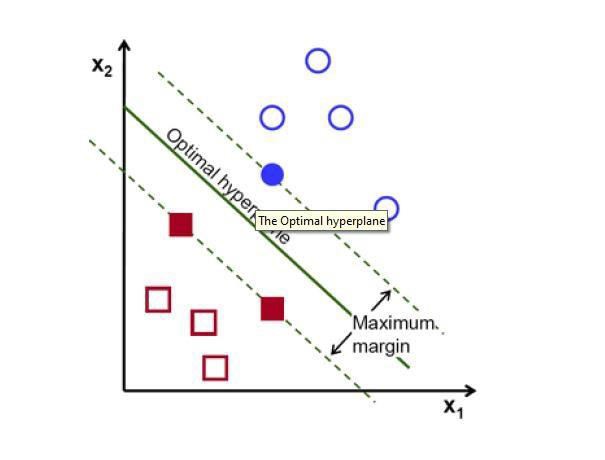


Fig 6.1 - SVM

## K Nearest Neighbours

K-Nearest Neighbours Classification is a type of instance-based learning or non- generalizing learning which stores instances of the training data. It predicts a new point by finding a predefined number of training samples closest in distance.

Weights can also be given for classifying a new point. In scikit-learn, each neighbour can either be given uniform weight or can be given weights proportional to the inverse of the distance from the query point or a user defined weight.

An example of K-Nearest Neighbours classifier is given below in Figure using uniform weight and distance weight in scikit-learn.

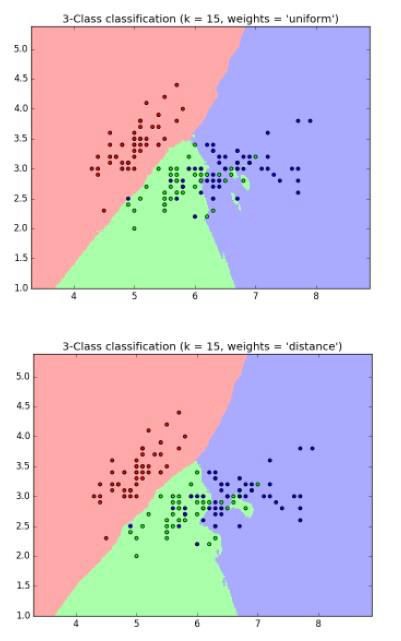


Fig 6.2 - K Nearest Neighbours

## Proposed System

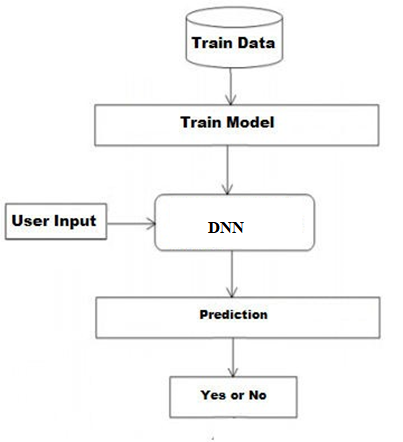


Fig 6.3 - Prediction Engine Proposed Model

* + - The user input is received by the web application using HTML forms.
    - The Web application makes use of HTTP modules to send and receive the data

to the API’s.

* + - The API’s will receive the user input in the form of a JSON object (key-value pair).
    - Trained RNN Models in the form of pickle files are consumed by the flask file housing at the local system
    - The trained models are called by passing the user input JSON object
    - Prediction result is sent as a response to the API calls

Trained RNN Models were used to predict the presence of Dementia or Alzheimer Disease. Pre-processed standard datasets were used to train the models post normalizing the dataset using Standard Scalar.

Post training the models, these models where extracted as pickle files and are stored at a local location which is used by the flask framework to call the trained model by passing in the user input.

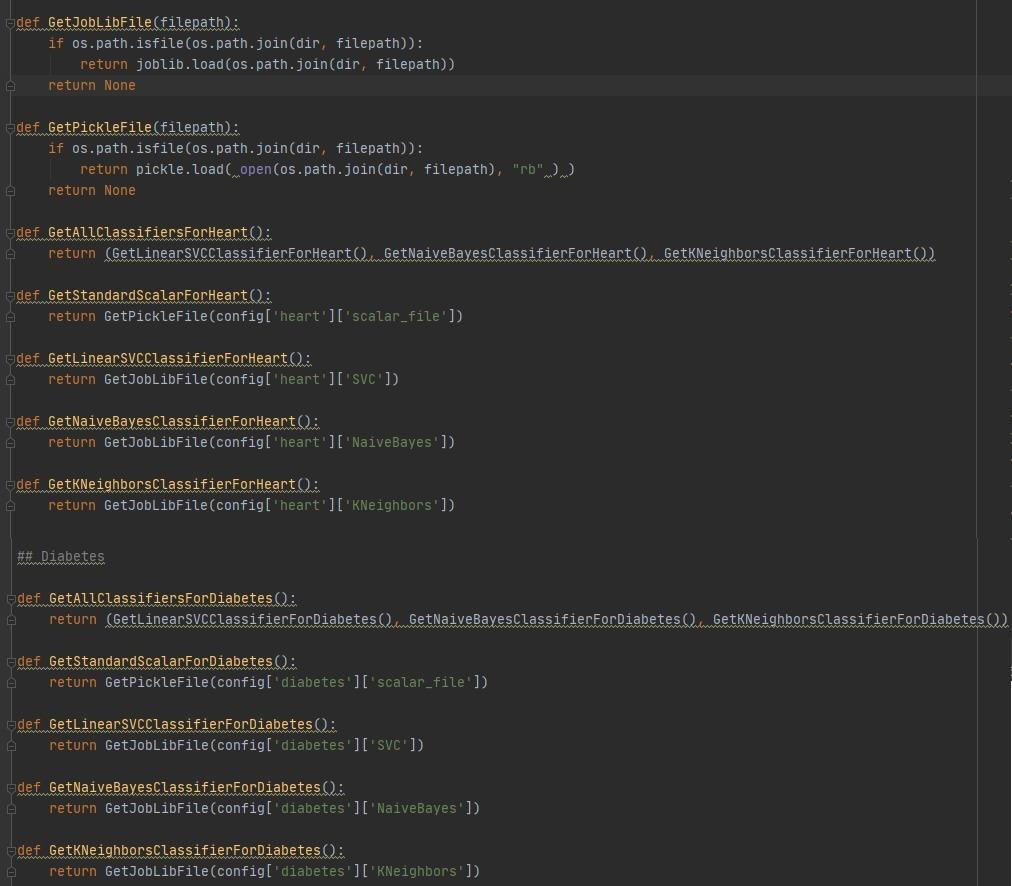


Fig 6.7 - Returning trained models to the API’s on function calls

# CHAPTER 7

**RESULTS AND DISCUSSION**

Use your screenshot

### CHAPTER 9

**CONCLUSION AND FUTURE SCOPE**

* 1. **Conclusion**

We developed a Prediction Engine which enables the user to check whether he/she has Dementia or Alzheimer Disease. The user interacts with the Prediction Engine by filling a form which holds the parameter set provided as an input to the trained models. The Prediction engine provides an optimal performance compared to other state of art approaches. The Prediction Engine makes use of three algorithms to predict the presence of a disease namely: RNN reason to choose these three algorithms are:

* + - They are effective, if the training data is large.
    - A single dataset can be provided as an input to all these 3 algorithms with minimal or no modification.
* A common scalar can be used to normalize the input provided to these 3 algorithms.
  1. **Future Work**
* To enhance the functionality of the prediction engine providing the details of 5 nearest hospitals or medical facilities to the user input location.
* Provide a user account which allows the user to keep track of their medical test data and get suggestions or support to meet the right specialists or the tests to be taken
* Provide admin controls to upload, delete the dataset which will be used to train the model.
* Automate the process of training the model and extracting pickle files of the trained models which will be consumed by the API’s to predict the disease.
* Mail the detailed report of the prediction engine results along with the information of 5 nearest medical facilities details having location and contact information.

# REFERENCES

[1] Priyanka Sonar, Prof. K. JayaMalini, “Dementia Prediction Using Different Machine Learning Approaches”, *Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC 2019)*

*IEEE Xplore Part Number: CFP19K25-ART; ISBN: 978-1-5386-7808-4*

[2] Samrat Kumar Dey, Ashraf Hossain, Md. Mahbubur Rahman, “Implementation of a Web Application to Predict Dementia Disease: An Approach Using Machine Learning Algorithm”, *2018 21st International Conference of Computer and Information Technology (ICCIT)*

[3] Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, “Dementia Disease Prediction Using Data Mining”, *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*

[4] Zhilbert Tafa, Nerxhivane Pervetica, Bertran Karahoda, “An Intelligent System for Dementia Prediction”, *4thMediterranean Conference on Embedded Computing MECO – 2015 Budva, Montenegro*