A Project Based Seminar Report

on

Studying the applications of Machine Learning methods and Algorithms in the healthcare industry

Submitted to the

Savitribai Phule Pune University

In partial fulfillment for the award of the Degree of

Bachelor of Engineering

in

Information Technology

by

Devaki Kulkarni

71828894H

Under the guidance of

Mrs. R.V. Kulkarni

Department Of Information Technology

Pune Institute of Computer Technology College of Engineering

Sr. No 27, Pune-Satara Road, Dhankawadi, Pune - 411 043.

2019-2020

CERTIFICATE

This is to certify that the project based seminar report entitled "**Studying the applications of Machine Learning methods and Algorithms in the healthcare industry**" being submitted by **Devaki Kulkarni (**71828894H**)** is a record of bonafide work carried out by her under the supervision and guidance of **Mrs. R. V. Kulkarni** in partial fulfillment of the requirement for **TE (Information Technology Engineering) – 2015 course** of Savitribai Phule Pune University, Pune in the academic year 2019-20.

Date: / /2020

Place: Pune

Mrs. R. V. Kulkarni
Guide

Dr. A.M.Bagade
Head of the Department

Dr. P. T. Kulkarni
Principal

This Project Based Seminar report has been examined by us as per the Savitribai Phule Pune University, Pune requirements at Pune Institute of Computer Technology, Pune – 411043 on ……………..

Internal Examiner

External Examiner

ACKNOWLEDGEMENT

On the very outset of this project, I would like to extend my sincere and heartfelt obligation towards all the personages who have helped me in this endeavor. Without their active guidance, help and cooperation, I would have not made headway in the project.

I express my sincere gratitude to Dr. Prahlad Kulkarni, Principal of Pune Institute of Computer Technology.

I pay my deep sense of gratitude to Dr. Anant Bagade, IT HOD, who has encouraged me to the highest peak and has provided me with this opportunity.

I thank our External Guide Prof. Naman Buradkar Sir for his valuable guidance and support for the completion of this project.

I am ineffably indebted to our Internal Guide Mrs. Radhika Kulkarni Mam for continuous evaluation, encouragement and supervision given throughout the project which shaped the present work.

I extend my gratitude to my teammates Shweta Patil, Akhil Shaji and Ruchika Pande who have given their full contribution in completion of this project.

Lastly, I acknowledge with a deep sense of reverence, my gratitude towards my parents and members of my family, who has always supported me morally as well as economically.

Devaki Kulkarni
(Signature)

ABSTRACT

With the technological advances in the healthcare industry, it has become imperative to develop an efficient healthcare management system to solve the problems of patient confidentiality breaches,theft of sensitive personal information and other medical malpractices. As the healthcare data is growing rapidly, it is difficult for traditional systems to handle this healthcare data securely and efficiently. Blockchain provides an immutable, tamper-proof mechanism for efficient handling of patient records. The use of a longitudinal system ensures all hospitals across the country are interconnected,thus providing inputs to the medical community.

Furthermore, we use machine learning to predict future health risks by combining patient data with known disease patterns,and other crucial parameters to help the medical authorities devise optimum strategies for prevention of the outbreak.This aids the medical community to take the necessary steps for preventing outbreaks.This study aims at improving the healthcare industry with optimum and consistent standards, to maintain a sustainable and reliable EHR (Electronic Health Record System), the usage of Machine Learning further increases the scope towards disease prediction and cure.

CONTENTS

V

LIST OF FIGURES

VI

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Introduction to Project

The complete deployed project aims at resurrecting a platform, on the principles of a blockchain-based infrastructure to simulate an EHR, to enhance data security and efficiency, furthermore the project aims at deploying a ML based model, to incorporate disease predictions, and out-predictive prevention methods.

1.2 Motivation behind project topic

Presently, there are many EHR software vendors such as Epic,Cerner,AthenaHealth that are used by certain healthcare institutions in the USA. In the USA, there is a prevalent EHR system used which must comply with certain rules and regulations of the HIPAA(Health Insurance Portability and Accountability Act). However, India does not have legal provisions for an EHR system. Considering the rapid growth of the population, dispensing healthcare becomes difficult. To introduce efficient and easy access to healthcare, we decided that the study of the topic of EHR would be of utmost importance to the nation. Furthermore, digitizing medical records and obtaining meaningful predictions from it could aid the medical community be better prepared for handling outbreaks.

1.2 Aim and Objectives of the work

The main aim of the project is to disrupt the way in which the critical medical data is managed in the current healthcare industry through decentralization. We intend to empower the patients to own their own health data thereby making a patient-centric healthcare system. We hope that we would provide a framework for individuals and organizations that is secure and a marketplace where the data owners can share the data without compromising security and ownership rights.

The objective of this project is to simply enable better collection, its use and sharing from patients, consumers and providers. We aim to design an interoperable healthcare system that is both secure and transparent. We also intend to focus on lowering healthcare transaction-related costs by improving and automating processes such as the use of smart contracts, removing the intermediaries and reducing the administrative burden.

1.3 Introduction to Studying the applications of Machine Learning methods and Algorithms in the healthcare industry

The topic involves studying data preprocessing techniques to suit the dynamic requirements of the healthcare industry. It further includes developing an in-depth understanding of features pertaining to a specific area of healthcare, for developing efficient ML models. Comparison of various ML algorithms is done to choose a best-suited algorithm. Enterprises use Machine Learning for a vast variety of domains such as finance [1], networking [2], manufacturing [3]. For example : ING, a banking and financial services corporation uses machine learning for bond trading and credit-risk management [4] .Thus, this report aims to explore the applications of machine learning in the healthcare domain.

1.5 Organisation of the report

The rest of the paper is arranged as follows. The various methods of implementation are discussed in Chapter II. The proposed solution is introduced in Chapter III. The applications in healthcare are discussed in Chapter IV. The future works are discussed in chapter V.The conclusion is introduced in Chapter VI.

CHAPTER 2

LITERATURE SURVEY

2.1 Introduction

Presently, in the USA a vast majority of hospitals use EHR systems for maintaining patient records. These institutions must comply with the rules and regulations according to the HIPAA act.

The Government of India has stated standard practices for EHR.However, there are no specific rules that all healthcare entities must comply with. Some of the areas of healthcare where machine learning can be of use are disease prediction, prescription generation.This chapter aims to study the said seminar topic with respect to technologies and its efficiencies.

2.2 Approaches to gather data and process it

2.2.1 Obtaining medical data from IoT-based sensors

In this approach, the medical data for clinical decision support is sent directly to the machine learning model from the IoT based devices [5]. This data is then analysed with reference to existing datasets using a variety of machine learning algorithms to predict the presence of certain diseases. A cloud database is used for storage of patient records. Fig. 1. shows the system architecture for this approach.

Following are the advantages and disadvantages of the said system :

Advantages :

- A continuous stream of patient data aids accurate predictions
- Fatal medical incidents can be avoided well in advance
- Reliable clinical decision support for doctors

Disadvantages :

- Patient must invest in IoT device
- Usage of cloud for storage causes security issues

The system architecture as proposed in [5] is as follows :

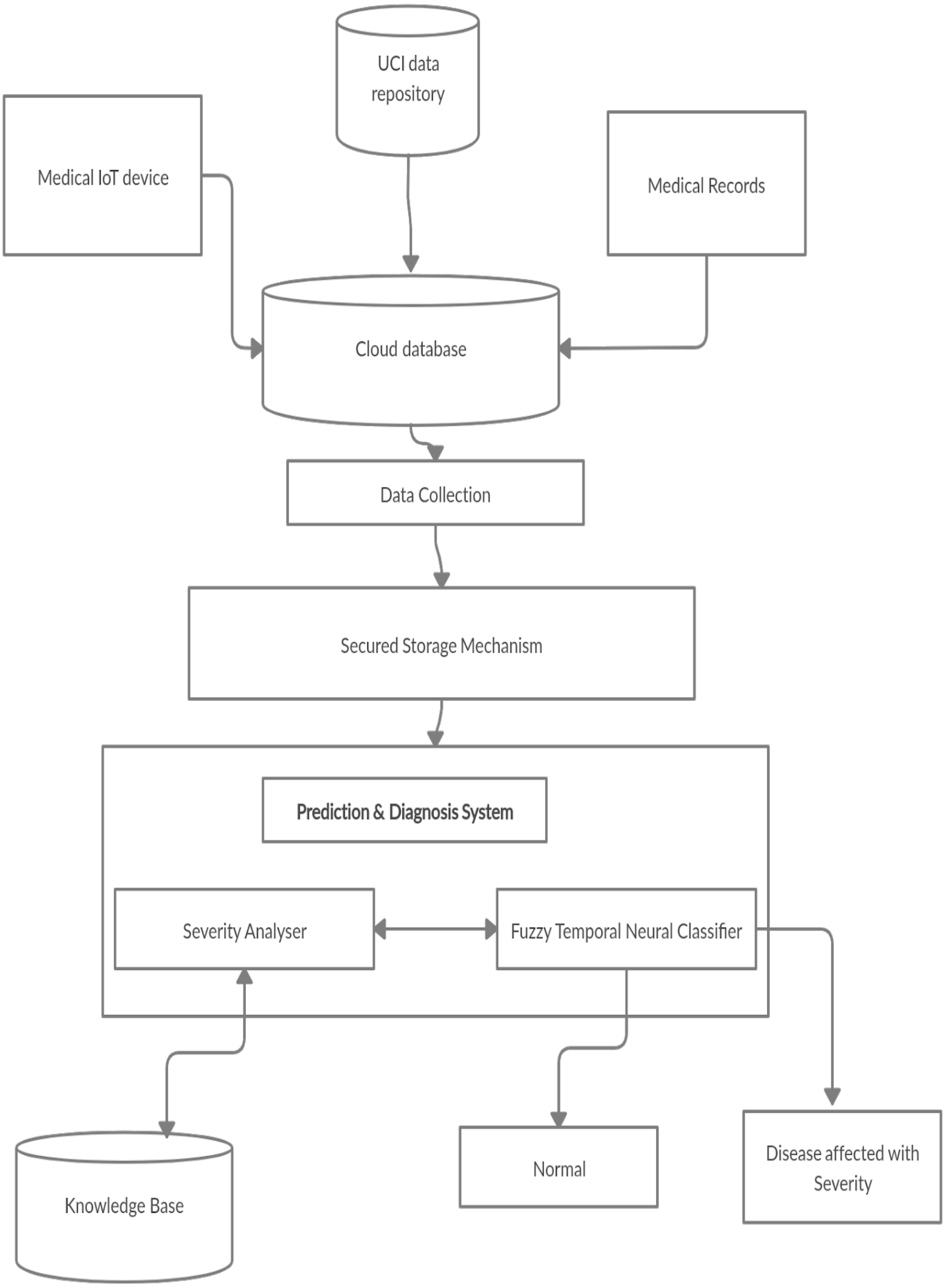

Fig. 1. System Architecture of IoT based prediction system

2.2.2 Obtaining medical data via reports and using hybridised algorithms

In this approach, we refer to the usage of PDF classification to extract data from publication reports [6] using a multi-pass sieve algorithm and apply the same methodology to medical reports in PDF format. Information extraction is to be carried out from the digital PDF generated by the doctor. Fig. 2. is the method used by multi-pass algorithm as proposed in [6] :

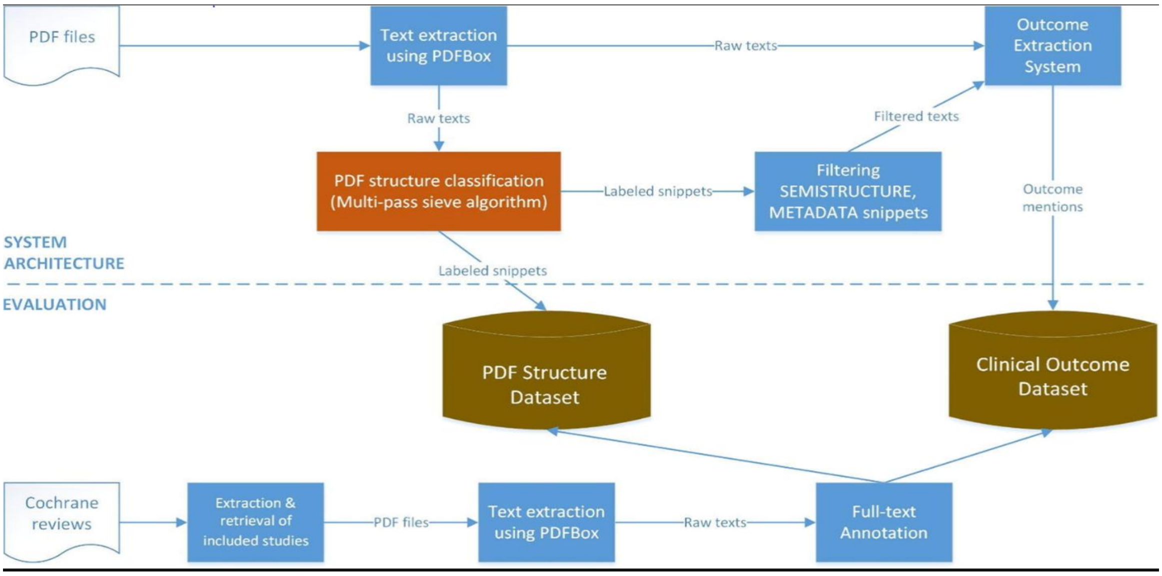
PDF files
Text extraction using PDFBox
Raw texts
Outcome Extraction System
Raw texts
Filtered texts
PDF structure classification (Multi-pass sieve algorithm)
Labeled snippets
Filtering SEMISTRUCTURE, METADATA snippets
Outcome mentions
SYSTEM ARCHITECTURE
Labeled snippets
EVALUATION
PDF Structure Dataset
Clinical Outcome Dataset
Cochrane reviews
Extraction & retrieval of included studies
PDF files
Text extraction using PDFBox
Raw texts
Full-text Annotation

Fig. 2. Multi-Pass Sieve Algorithm [6]

The dataset generated by the above method is then passed to a hybridied Machine Learning model.The use of genetic algorithms with Naive Bayes and SVM with regression analysis results than when a single algorithm is used [7]. Fig. 3. demonstrates the hybridised approach as proposed in [7] :

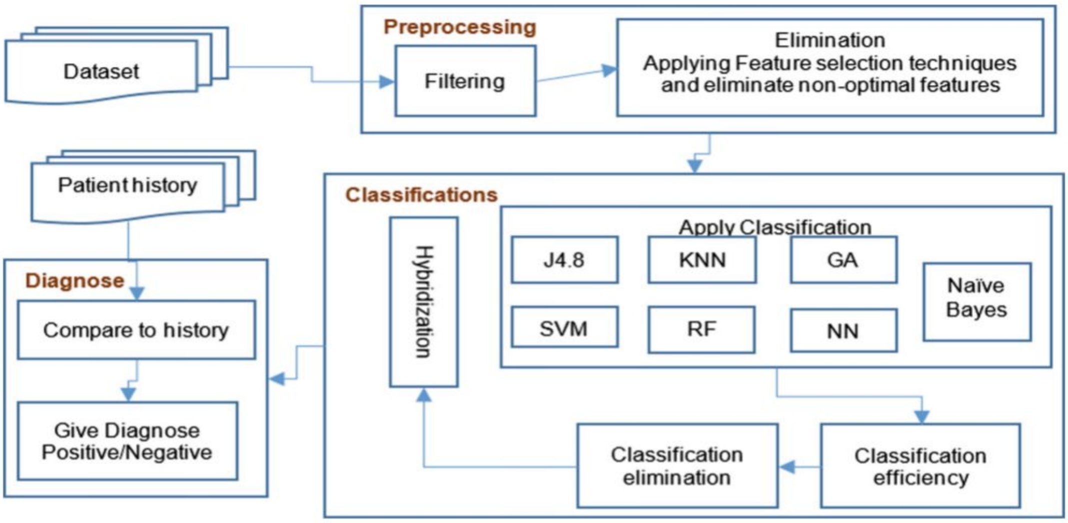

Fig. 3. Hybridised Approach[7]

Advantages :

- The doctors may proceed to generate reports electronically as before without a need to adapt to a new system
- No need for development of a new device
- Produces better efficiency than when a single Machine Learning algorithm is used.

Disadvantages:

- Flow of data is not continuous as when an IoT-based device is used

2.3 State-of-Art Algorithms

Machine Learning has 2 two types of learning methods :

- Supervised Learning
- Unsupervised Learning

In this study, we will be focusing on Supervised Learning Algorithms.

2.3.1 Support Vector Machine

In the SVM algorithm, each data item is plotted as a point in n-dimensional space (where n is number of features present) with the value of each feature being the value of a particular coordinate. Furthermore, classification is performed by finding the hyper-plane that differentiates between the two classes well.

2.3.2 Naive Bayes

Naive Bayes Classifier algorithm is based on the Bayes theorem.

Bayes theorem states the Equation 1.:

$$P(A|B) = P(A \cap B)/P(B)$$

Equation 1. Bayes Theorem

This algorithm does not take into account the real-life scenario of dependence among variables. Thus, it is called Naive [8]. Prediction of membership probabilities for each class such as the probability that a given record or data point belongs to a particular class is made possible. The class having the highest probability is considered as the most likely class. This is also known as Maximum A Posteriori (MAP).In this study,class could be the presence or absence of the disease.

2.3.3 K-Nearest Neighbour

The KNN algorithm assumes that similar things exist in close proximity.

It functions as follows :

- Load the dataset
- Initialize K to your chosen number of neighbors
- For each example in the data
  - ➔ Calculate the distance between the query example and the current example from the data.
  - ➔ Add the distance and the index of the example to an ordered collection
- Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances
- Pick the first K entries from the sorted collection
- Get the labels of the selected K entries
- If regression, return the mean of the K labels
- If classification, return the mode of the K labels

CHAPTER 3

PROPOSED MODEL FOR HEART DISEASE PREDICTION

3.1 Introduction

By studying the various approaches that are currently in use and research papers, I propose the following Model for Disease prediction specifically heart disease.This model can further be extended for Cancer detection, glaucoma detection.

3.2 Data Pre-processing

Pertaining to a specific disease such as Heart Disease,Cancer,Glaucoma; the test result must be entered in a fillable PDF which can be created using various open source tools such as LibreOfficeWriter and OpenOffice. This PDF can be then extracted using the following method of Multi-pass sieve algorithm to generate a dataset. Refer to Fig. 1.

The dataset is then cleaned and missing values are imputed using K-Nearest Neighbour Algorithm.

3.3 Feature Selection

Feature selection is of utmost importance. The presence of irrelevant or partially relevant features can negatively impact the performance of the model. The various methods used for feature selection are:

- Univariate Selection
- Feature importance
- Correlation Matrix with heatmap

Of these, we shall be considering the use of correlation matrix with a heatmap. Considering the Cleveland Heart Disease dataset, there are 75 attributes of which 14 are generally considered for performing prediction pertaining to heart disease.A supervised correlation method can be used to identify relationships between variables [9].On generating a heatmap, we can drop unrelated variables.

3.4 Dimensionality reduction

The process of reducing the number of features for reducing model complexity is called dimensionality reduction.. The method employed is called Principal Component Analysis. PCA aims to identify directions of maximum variance in high dimensional data and projects it onto a new subspace with equal or lesser dimensions than the original one.

3.5 Applying a hybridized Machine Learning Model

A single classifier is weak and does not perform well. Thus using various Ensemble Machine Learning methods such as boosting,stacking,bagging, we use multiple classifiers to improve the efficiency of the system. One of the best examples is the use of SVM(Support Vector Machine) with multiple classifiers to produce accuracy with an increase from 84.15% to 84.81% when working on the dataset [9].

CHAPTER 4

APPLICATIONS OF STUDYING APPLICATIONS OF MACHINE LEARNING METHODS AND ALGORITHMS IN THE HEALTHCARE INDUSTRY

4.1 Application in Clinical Decision Support System

In the healthcare scenario, there are many departments where the need for performing predictions is extremely useful for providing better healthcare. Some of the examples include cancer detection,glaucoma detection,heart disease prediction etc. Out of these, the author has selected heart disease as the principal object of study.

4.1.1 Heart Disease Prediction

Of the most commonly used Machine learning datasets, pertaining to the Healthcare industry is the Cleveland Heart Disease dataset. After applying various machine learning algorithms on the Cleveland Heart Disease dataset, following are the accuracies observed according to [10] :

Table 1 : Comparison of various Algorithms [10]

Classifier	Precision	Accuracy (%)
Decision Tree(DT)	0.774	77.55
Naive Bayes(NB)	0.836	83.49
K Nearest Neighbor (K=1)	0.782	76.23
K Nearest Neighbor (K=3)	0.821	81.18
K Nearest Neighbor (K=9)	0.848	83.16
K Nearest Neighbor (K=15)	0.847	82.83
MultiLayer Perceptron(MLP)	0.824	82.83
Radial Basis Function(RBF)	0.845	83.82
Support Vector Machines(SVM)	0.827	84.15

The use of single classifiers produces an acceptable accuracy, however when such a model is used for practical applications such a low accuracy makes the usage of machine learning for predictions impractical and futile.

In a further attempt to make this model suitable to be used in practical scenarios, following approaches have been tried according to [10]:

1. Bagging

   Bagging involves averaging predictions of various independent models to obtain a model with lower variance. Fig. 4. is a comparison of the results obtained after applying bagging against the use of a single classifier according to [10].

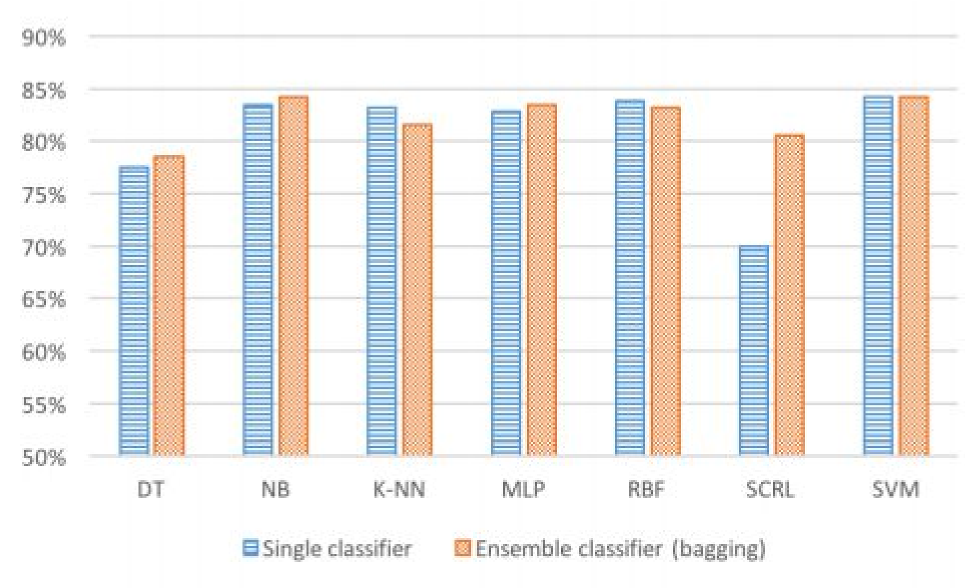

Fig. 4. Bagging vs Single Classifier

2. Boosting

   Boosting involves applying a weak classifier and running it multiple times on the training data and then allowing the learned classifier to vote. Fig. 5. is a comparison of a single classifier against the use of boosting.

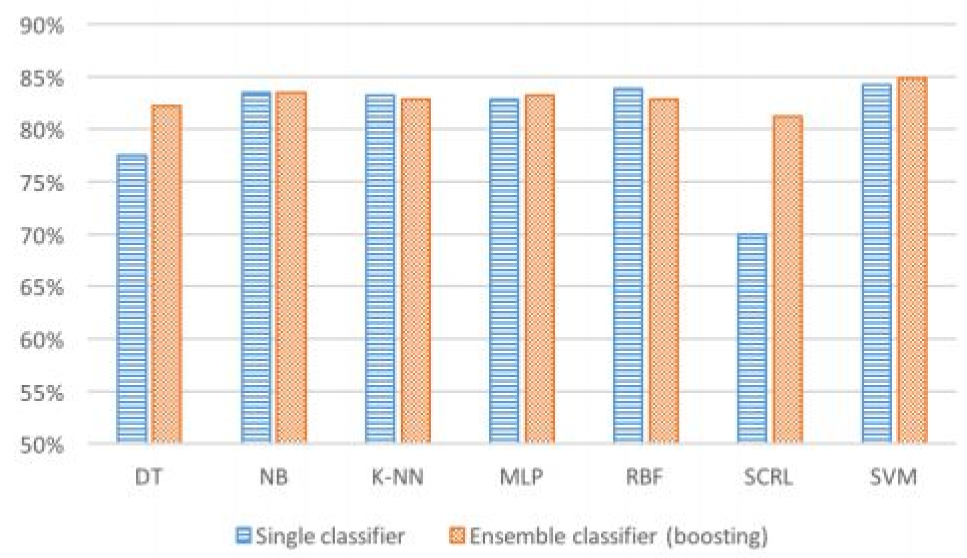

Fig. 5. Boosting vs Single Classifier

4.2 Application in Personalised Medicine

The use of generalised medicine for a particular disease for a particular patient does not guarantee that it will work on the patient. The genetics, medical history of the patient is not considered. Hence, various factors necessary for determining patient-compatibility with medicine is not taken into account thereby producing incorrect prediction.

Rapid learning is a methodology to take into account other factors such as medical history,genetics. Fig. 6. shows the 4-phase process which is as follows according to [11] :

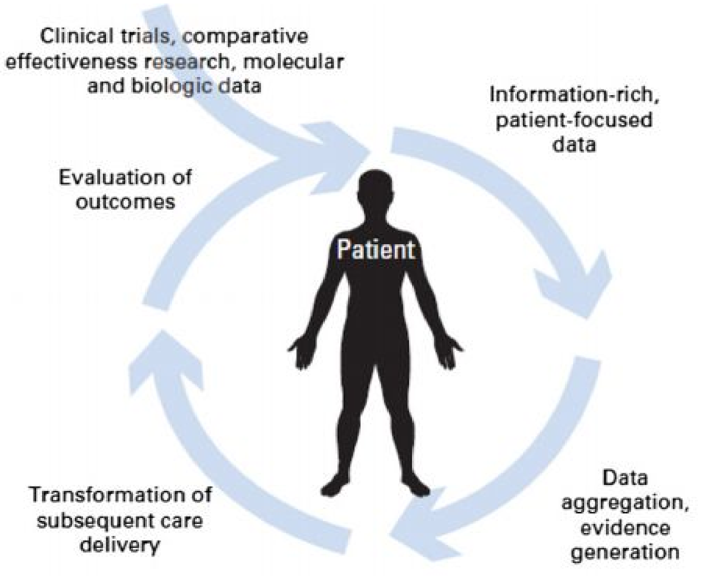
Clinical trials, comparative effectiveness research, molecular and biologic data
Information-rich, patient-focused data
Evaluation of outcomes
Patient
Data aggregation, evidence generation
Transformation of subsequent care delivery

Fig. 6. Phases of Rapid Learning

Thus, Personalised treatment ensures not only that patients receive an optimal treatment, but also that the right resources are being used for the right patients.

CHAPTER 5

PRECISE PROBLEM STATEMENT

5.1 Problem Statement

This seminar has resulted in the complete study of the project topic and understanding of its feasibility. We have identified feasible areas of potential work. Thus, a potential problem statement for the B.E. project is :

“ Building a Blockchain based Electronic Health Record Management System for file verification and credential authentication and using a Machine Learning model for disease prediction with special emphasis on heart disease”.

We intend to apply Blockchain for the following purposes :

- For verifying the authenticity of the reports generated by the doctors in the form of PDF files.
- For verifying the identities of the doctors who are generating the reports

We intend to apply Machine learning for the following purposes :

- For providing clinical decision support to the doctors to help identify early signs of heart diseases
- Ease the workload of the doctors
- Help doctors focus on critical cases and provide better healthcare

CHAPTER 6

CONCLUSION

This seminar was undertaken to understand the methodologies of data preprocessing, analyse the features essential for disease predictions and compare the various Machine learning models.This study shall help us to implement the EHR system in India, thereby aiding doctors to tend to greater number of patients in record amount of time.

Thus, I conclude the following from the reputed papers published in the domain of machine learning :

- The use of hybridised machine learning algorithms for training the dataset of medical records produces better accuracy as compared to the use of a single algorithm. Specifically, the use of Naive Bayes coupled with SVM,produces better accuracy of diagnosis as compared to individual algorithms.
- The methodology used for extraction of information from PDFs can be used in the medical domain for dataset preparation.

REFERENCES

[1] Culkin, Robert, and Sanjiv R. Das. "Machine learning in finance: the case of deep learning for option pricing." *Journal of Investment Management,* vol. 15, no. 4, pp. 92-100, 2017.

[2] M. Wang, Y. Cui, X. Wang, S. Xiao and J. Jiang, "Machine Learning for Networking: Workflow, Advances and Opportunities", *IEEE Network*, vol. 32, no. 2, pp. 92-99, 2018.

[3] Q. Xie, "Machine learning in human resource system of intelligent manufacturing industry", *Enterprise Information Systems*, pp. 1-21, 2020.

[4] I. Lee and Y. Shin, "Machine learning for enterprises: Applications, algorithm selection, and challenges", *Business Horizons*, vol. 63, no. 2, pp. 157-170, 2020.

[5] P. Kumar, S. Lokesh, R. Varatharajan, G. Chandra Babu and P. Parthasarathy, "Cloud and IoT based disease prediction and diagnosis system for healthcare using Fuzzy neural classifier", *Future Generation Computer Systems*, vol. 86, pp. 527-534, 2018.

[6] D. Bui, G. Del Fiol and S. Jonnalagadda, "PDF text classification to leverage information extraction from publication reports", *Journal of Biomedical Informatics*, vol. 61, pp. 141-148, 2016. Available: 10.1016/j.jbi.2016.03.026.

[7] Tarawneh, Monther, and Ossama Embarak. "Hybrid Approach for Heart Disease Prediction Using Data Mining Techniques." *International Conference on Emerging Internetworking, Data & Web Technologies*. Springer, Cham, 2019.

[8] M. Saritas, "Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification", *International Journal of Intelligent Systems and Applications in Engineering*, vol. 7, no. 2, pp. 88-91, 2019.

[9] Dubey, Vimal Kumar, and Amit Kumar Saxena. "Hybrid classification model of correlation-based feature selection and support vector machine." *2016 IEEE International Conference on Current Trends in Advanced Computing (ICCTAC)*. IEEE, 2016.

[10] Pouriyeh, Seyedamin, et al. "A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease." *2017 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 2017.

[11] P. Lambin et al., "'Rapid Learning health care in oncology' – An approach towards decision support systems enabling customised radiotherapy'", *Radiotherapy and Oncology*, vol. 109, no. 1, pp. 159-164, 2013.

APPENDIX

Tarawneh, Monther, and Ossama Embarak. "Hybrid Approach for Heart Disease Prediction Using Data Mining Techniques." *International Conference on Emerging Internetworking, Data & Web Technologies*. Springer, Cham, 2019.

Note : The base paper has been attached along with this document.