DAYANANDA SAGAR UNIVERSITY

KUDLU GATE, BANGALORE – 560068

Bachelor of Technologyin
COMPUTER SCIENCE AND ENGINEERING

Special Topic - 1 Report

(DISEASE PREDICTION USING MACHINE LEARNING)

BY

SANDEEP KUMAR PRADHAN-ENG20CS0315
UDAY KUMAR A – ENG20CS0389
VINOD V – ENG20CS0413
SHASHIDHARA K M – ENG21CS1019

Under the supervision of

PROF. AMRUTA B
Assistant Prof. and Department of CSE

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING,
SCHOOL OF ENGINEERING
DAYANANDA SAGAR
UNIVERSITY,

(2021-2022)

DAYANANDA SAGAR UNIVERSITY

School of Engineering
Department of Computer Science& Engineering
Kudlu Gate, Bangalore –
560068Karnataka, India

DECLARATION

We, **Sandeep Kumar Pradhan (ENG20CS0315), Uday Kumar A (ENG20CS0389), Vinod V (ENG20CS0413), Shashidhara K M (ENG21CS1019)** are student's of the fourth semester B.Tech in **Computer Science and Engineering**, at School of Engineering, **Dayananda Sagar University**, hereby declare that the Special Topic 1 titled **"Disease Prediction Using Machine Learning"** has been carried out by us and submitted in partial fulfillment for the award of degree in **Bachelor of Technology in Computer Science and Engineering** during the academic year **2021-2022**.

Student **Signature**

Name1: Sandeep Kumar Pradhan

USN: ENG20CS0315

Name2: Uday Kumar A

USN: ENG20CS0389

Name3: Vinod V

USN: ENG20CS0413

Name4: Shashidhara K M

USN: ENG21CS1019

Place : Bangalore

Date :

ACKNOWLEDGEMENT

It is a great pleasure for us to acknowledge the assistance and support of many individuals whohave been responsible for the successful completion of this Special Topic 1.

First, we take this opportunity to express our sincere gratitude to School of Engineering & Technology, Dayananda Sagar University for providing us with a great opportunity to pursue our Bachelor's degree in this institution.

We would like to thank **Dr. A Srinivas. Dean**, **School of Engineering & Technology**, **Dayananda Sagar University** for his constant encouragement and expert advice. It is a matterof immense pleasure to express our sincere thanks to **Dr. Girisha G S, Chairman**, **Department of Computer Science, and Engineering**, **Dayananda Sagar University,** for providing the right academic guidance that made our task possible.

We would like to thank our guide **Prof. Amruta B, Assistant Professor**, **Dept. of Computer Science and Engineering**, **Dayananda Sagar University**, for sparing his/her valuable time to extend help in every step of our Special Topic 1, which paved the way for smooth progress and the fruitful culmination of the project.

We would like to thank our Special Topic 1 Coordinators, Dr. Savitha Hiremath, Dr. T Kumaresan and all the staff members of Computer Science and Engineering for their support. We are also grateful to our family and friends who provided us with every requirement throughout the course. We would like to thank one and all who directly or indirectly helped usin the Special Topic 1.

TABLE OF CONTENTS

FIGURES

LIST OF TABLES

Abstract

The wide adaptation of computer-based technology in the health care industry resulted in the accumulation of electronic data. Due to the substantial amounts of data, medical doctors are facing challenges to analyze symptoms accurately and identify diseases at an early stage. However, supervised machine learning (ML) algorithms have showcased significant potential in surpassing standard systems for disease diagnosis and aiding medical experts in the early detection of high-risk diseases. In this literature, the aim is to recognize trends across various types of supervised ML models in disease detection through the examination of performance metrics. The most prominently discussed supervised ML algorithms were Naïve Bayes (NB), Decision Trees (DT), K-Nearest Neighbor (KNN). As per findings, Support Vector Machine (SVM) is the most adequate at detecting kidney diseases and Parkinson's disease. The Logistic Regression (LR) performed highly at the prediction of heart diseases. Finally, Random Forest (RF), and Convolutional Neural Networks (CNN) predicted in precision breast diseases respectively

CHAPTER 1
INTRODUCTION

1.1 INTRODUCTION

The emergence of Artificial Intelligence (AI) enabled com-pauperized systems to perceive, think and operate in an in-Telligent manner like humans. AI is a multidisciplinary concept of ML, Computer Vision, Deep Learning, and Natural Language Processing. ML algorithms apply various optimization, statistical, and probabilistic techniques to learn from data that was generate from past experiences, and deploy it in decision making

Medicaid services and centers for Medicare reported that 50% of Americans had multiple chronic diseases, which led the US health care to spend around $3.3 trillion in 2016, that amounts to $10,348 per person in the US. Moreover, the World Health Organization and World Economic Forum reported that India had a huge loss of $236.6 billion by 2015 because of fatal diseases, caused by malnutrition and morbid lifestyles. Such expenditures revealed how prone people are to a spectrum of diseases, which show cased how vital itis to detect diseases early, to consequently reduce the fatality of these maladies.

In contrast, Ismaeel argued that standard statistical techniques, the work experience and the intuition of medical doctors led to undesirable biased and errors when detecting risks associated to the disease. With the substantial surge of electronic health data, medical doctors are facing challenges to identify diseases accurately at an early stage. For this reason, advanced computational methodologies such as ML algorithms were introduced to discover meaningful patterns and hidden information from data, which can be used for critical decision making

1.2 SCOPE

The analysis accuracy is reduced when the quality of medical data in incomplete. Moreover, different regions exhibit unique characteristics of certain regional diseases, which may weaken the prediction of disease outbreaks. However, those existing work mostly considered structured data. There are no proper methods to handle semi structured and unstructured. The proposed system will consider both structured and unstructured data. The analysis accuracy is increased by using Machine Learning algorithm and Map Reduce algorithm.

CHAPTER 2
PROBLEM DEFINITION

Some of the problems the Disease prediction system will overcome are:

- Improving the medical attention given to patients.
- Decreasing the rush on OPD's.
- In case of unavailable of the doctors
- It reduces the panic movement of the user, so that proper medication can be provided at the right time.

CHAPTER 3
LITERATURE REVIEW

Author's Name	Journal year	Technology/Design	Result shared by author	What you infer
Ashish Kailash Pal	2019	Random Forest/Naive Bayes	Performances of the classifiers are compared to each other to find out accuracy	Less accurate
Kedar Pingale	2019	Naive Bayes Classifier	Accurate analysis of medical data benefits early disease detection	Accuracy is higher
Marouane Ferjani	2020	Naïve Bayes (NB), Decision Trees (DT), K-Nearest Neighbor (KNN)	accumulation of electronic data	Disease prediction is easier and Faster
LANG Van Tran	2021	UCI	Experimental results demonstrated the effectiveness of the proposed method for precise HD prediction making	Recovery rate is higher

CHAPTER 4
PROJECT DESCRIPTION

Disease Prediction using Machine Learning is the system that is used to predict the diseases from the symptoms which are given by the patients or any user. In this project, Python is used as a platform for executing Machine Learning the algorithm. The system processes the symptoms provided by the user as input and gives the output as the probability of the disease. The given data set will be divided into two parts, Training set and Testing set. 80% importance is given to the Training set and 20% importance will be given to the Testing set. Naïve Bayes classifier is used in the prediction of the disease which is a supervised machine learning algorithm. The probability of the disease is calculated by the Naïve Bayes algorithm. With an increase in biomedical and healthcare data, accurate analysis of medical data benefits early disease detection and patient care. In addition, an elegant GUI has been developed to facilitate system interaction. The result ensures that the system would be functional and user oriented for patients for timely diagnoses of diseases.

CHAPTER 5

REQUIREMENTS

5.1 Software Requirements

IDE: Visual Studio Code.
Operating System: Any windows version /MAC/Linux.
Stacks used: Python

5.2 Hardware Requirements

Processor: Intel Pentium IV / Ram above 512 MB.
Hard Disk: 40GB or more

5.3 Functional requirements

- In this system have two actors are there admin and user
- Admin has to login using his username and password
- After admin login they can upload a dataset, they can train the machine using machine learning approach.
- In user part they have to register themselves.
- User has to login using user id and password
- Then user has to input patient data, and then based on trained model user input data will check and it will give the output.

5.4 Non Functional requirements

Nonfunctional necessities describe however a system should behave and establish constraints of its practicality. This type of requirements is also known as the system's quality attributes. Attributes such as performance, security, usability, compatibility are not the feature of the system, they are a required characteristic. They are "developing" properties that emerge from the whole arrangement and hence we can't compose a particular line of code to execute them. Any attributes required by the customer are described by the specification. We must include only those requirements that are appropriate for our project.

CHAPTER 6
METHODOLOGY

6.1 EXISTING SYSTEM

Machine can predict diseases but cannot predict the sub types of the diseases caused by occurrence of one disease. It fails to predict all possible conditions of the people. Existing system handles only structured data. The prediction system is broad and ambiguous. In current past, countless disease estimate classifications have been advanced and in procedure. The standing organizations arrange a blend of machine learning algorithms which are judiciously exact in envisaging diseases. However, the restraint with the prevailing systems is speckled. First, the prevailing systems are dearer only rich people could pay for to such calculation systems. And also, when it comes to folks, it becomes even higher. Second, the guess systems are non-specific and indefinite so far. So that, a machine can envisage a positive disease but cannot expect the sub types of the diseases and diseases caused by the existence of one bug. For occurrence, if a group of people are foreseen with Diabetes, doubtless some of them might have complex risk for Heart viruses due to the actuality of Diabetes. The remaining schemes fail to foretell all possible surroundings of the tolerant.

6.2 PROPOSED SYSTEM

Our application will be at affordable cost. Naive Bayes Machine Learning Algorithm predicts Diseases as well as all sub diseases. Map Reduce Algorithm is implemented to increase operational efficiency. It reduces Query retrieval time. Accuracy is improved using Machine Learning algorithm. The proposed system begins with the thought that was not executed by the ancestors. Its gadget Naive Bayes machine learning procedure for calculating diseases as well as calculating all the other thinkable sub diseases. It member Map Reduce algorithm for subdividing the data such that a request would be scrutinized only in the explicit partition, which will increase effective proficiency but cut query rescue time. In tally to that, it provides definite rations for specific clients to pattern his/her condition. Thus, making our presentation broadly open by all at cheap cost.

6.3 Naive Bayes Classification Algorithm

This approach is mostly used when the dimensionality of the provided input is high. This classification model is concerned with a simple probabilistic model based on the Bayes proposal with robust independence assumptions. It uses the Bayes statement to determine the likelihood of a result occurring by considering the likelihood of an alternative event that has already happened.

The most important aspects of ML are classification and prediction where the world is brimming with AI and ML consciousness encompassing nearly everything around. Naive Bayes is a basic yet surprisingly incredible algorithm for predictive examination. It is a classification strategy dependent on Bayes, the hypothesis with suspicion of freedom among predictors. It involves two sections such as Naive and Bayes, in straightforward terms. In the Naive Bayesian method, the classifier will accept when the nearness of a precise feature of a class is random with the presence of another attribute. Regardless of whether these features rely upon one another or upon the presence of different features, these properties autonomously add to the probability that is the reason named after seeing that Naive.

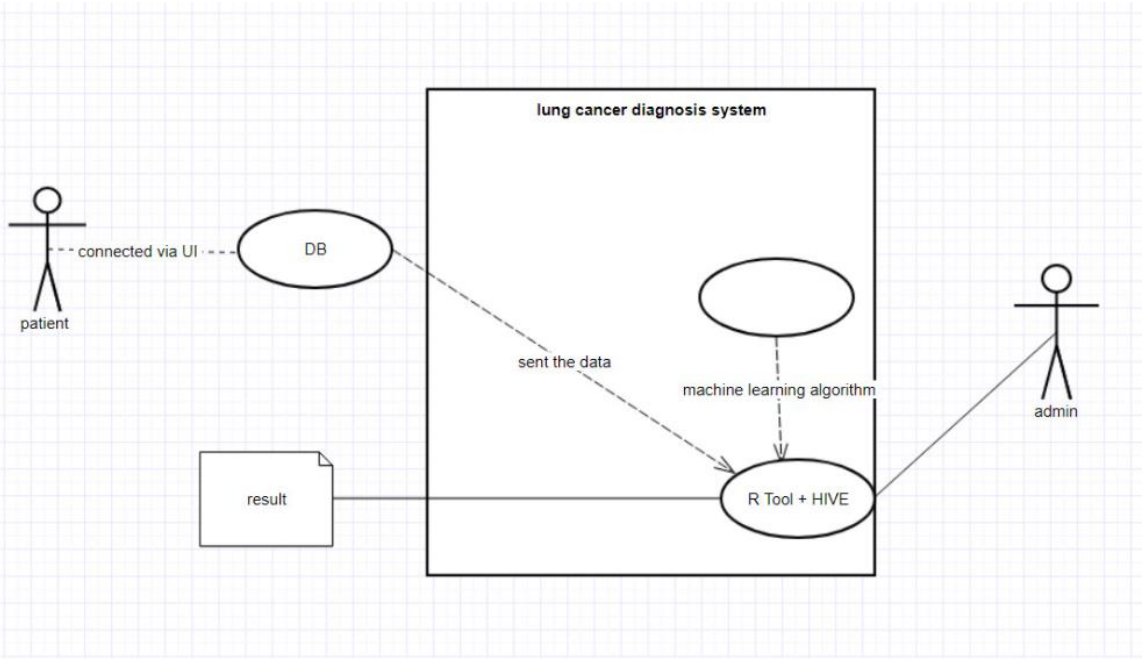

Fig 6.1 Case Diagram

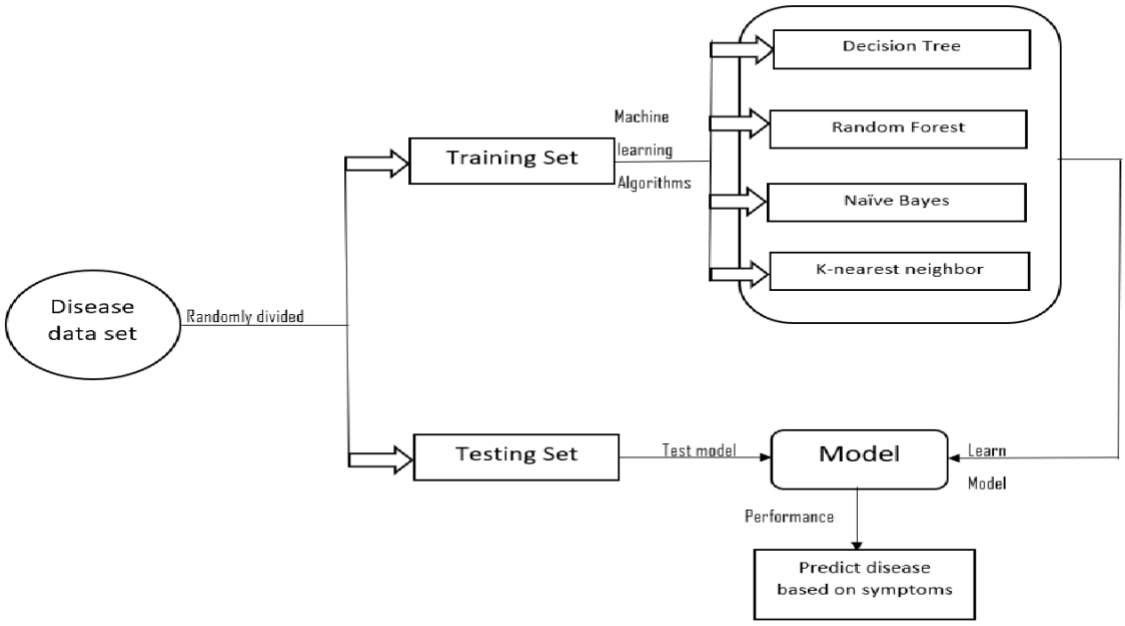

Fig 6.2 Flow Chart

CHAPTER 7
EXPERIMENTATION

- **Gathering the Data:** Data preparation is the primary step for any machine learning problem. We will be using a dataset from Kaggle for this problem. This dataset consists of two CSV files one for training and one for testing. There is a total of 133 columns in the dataset out of which 132 columns represent the symptoms and the last column is the prognosis.

- **Cleaning the Data:** Cleaning is the most important step in a machine learning project. The quality of our data determines the quality of our machine learning model. So it is always necessary to clean the data before feeding it to the model for training. In our dataset all the columns are numerical, the target column i.e. prognosis is a string type and is encoded to numerical form using a label encoder.

- **Model Building:** After gathering and cleaning the data, the data is ready and can be used to train a machine learning model. We will be using this cleaned data to train the Support Vector Classifier, Naive Bayes Classifier, and Random Forest Classifier. We will be using a confusion matrix to determine the quality of the models.

- **Inference:** After training the three models we will be predicting the disease for the input symptoms by combining the predictions of all three models. This makes our overall prediction more robust and accurate.

At last, we will be defining a function that takes symptoms separated by commas as input, predicts the disease based on the symptoms by using the trained models, and returns the predictions in a JSON format.

CHAPTER 8
TESTING AND RESULTS

In the experimental study, the UCI machinery disease data set is employed. Many previous studies have made use of databases in two ways. Some studies used the entire set of 75 traits, while others focused on the 13 or 14 most important traits for analysis and prediction. To identify the necessary qualities, a thorough investigation is necessary. The dataset under consideration for implementation in this study is described.

The Cleveland database containing 303 records is used as input in Naive Bayes, ID3, C4.5, and SVM algorithms. The accuracy and error rate achieved are listed in Table 1 and are graphically shown in Figures 2 and 3. Also, the error rate results of classification algorithms are listed in Table 3 and are graphically represented in Figures 4 and 5.

Table 1
Accuracy results of classification algorithms.

Machine learning algorithms	Accuracy (%)
Naive Bayes	90
C4.5	85
ID3	78
SVM	70

Overall, we can say Naive Bayes classification algorithms have better results and fewer errors than ID3, C4.5, and that of SVM classification algorithms. Therefore, Naive Bayes classification algorithms are used by most algorithm developers.

REFERENCES

[1] Khurana, Sarthak . , Jain, Atishay ., Kataria ,Shikhar. ,Bhasin ,Kunal . , Arora ,Sunny . ,& Gupta , Dr.Akhilesh . Das. (2019). Disease Prediction System.International Research Journal Of Engineering and Technology , 6(5) , 5178-5184.

[2] Kamboj ,Mgha. (2020).Heart Disease Prediction with Machine Learning Approaches. International Journal Of Science and Research , 9(7) , 1454-1458.

[3]Ware,Miss.Sangya . , Rakesh,Mrs.Shanu. K.,&Choudhary,Mr.Bharat . (2020). Heart Attack Prediction By Using Machine Learning Techniques. International Journal Of Recent Technology and Engineering , 8(5), 1577-1580.

[4] Shirsath ,Shraddha.Subhash .,& Patil , Prof. Shubhangi . (2018).Disease Prediction Using Machine Learning over Big Data .International Journal Of Innovative Research in Science and Technology , 7(6), 6752-6757

[5] Battineni ,Gopi. , Sagaro,Getu.Gamo. ,Chinatalapudi, Nalini. ,&Amenta,Francesco. (2020). Application Of Machine Learning Predictive Models in the Chronic Disease .International of Personalised Medicine , 10(21), 1-11

[6]Marimuthu , M. , Abinaya, M. ,Hariesh,K.S., Madhan,K.,& Pavithra, Kumar. V.(2018).A Review of Heart Disease Prediction Using Machine Learning and Data Analytics Approach .International Journal of Computer Application , 181(18), 20-25.

[7] Bindhika,Galla Siva Sai., Meghana,Munaga., Reddy Manchuri Sathvika. & Rajalakshmi. (2020). Heart Disease Prediction Using Machine Learning Techniques. International Research Journal of Engineering and Technology, 7(4) , 5272-5276.