Predictive Health: ML for Disease Prognosis

MINI PROJECT REPORT

Submitted by

PRAVINESH H	2116**210701194**
RAM PRAKASH L	2116**210701208**

in partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI ANNA UNIVERSITY::
CHENNAI 600 025

MAY 2024

BONAFIDE CERTIFICATE

Certified that this Report titled "**Predictive Health: ML for Disease Diagnosis**" is the bonafide work of **"PRAVINESH H (210701194) and RAM PRAKASH (210701208)"**who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr. RAKESH KUMAR M, M.E., Ph.D

Assistant Professor,
Department of Computer Science and Engineering,
Rajalakshmi Engineering College,
Chennai – 602015

Submitted to Mini Project Viva-Voce Examination held on ________________________

Internal Examiner **External Examiner**

ABSTRACT

Disease prediction using machine learning is the system that is used to predict the diseases from the symptoms which are given by patients or any other user. The system processes the symptoms provided by the user as input and gives the output as the probability of the disease. Naive bayes, KNN, Decision tree, Random forest classifiers is used in the prediction of the disease which is supervised machine learning algorithm. With an increase in biomedical and healthcare data, accurate analysis of medical data benefits early disease detection and patient healthcare. By using linear regression and decision tree we are predicting diseases like Diabetes, Malaria, Jaundice, Dengue and Hypertension etc.., In conclusion, machine learning has the potential to revolutionize disease prediction by leveraging large-scale healthcare data and advanced algorithms. By addressing the challenges and harnessing the strengths of machine learning, healthcare professionals can better predict and manage diseases, ultimately leading to more efficient and personalized healthcare. This abstract serves as a foundation for further exploration into the promising field of disease prediction in machine learning.

ACKNOWLEDGEMENT

Initially we thank the Almighty for being with us through every walk of our life and showering his blessings through the endeavour to put forth this report. Our sincere thanks to our Chairman **Mr. S.MEGANATHAN, B.E, F.I.E.**, our Vice Chairman **Mr. ABHAY SHANKAR MEGANATHAN, B.E., M.S.,** and our respected Chairperson **Dr. (Mrs.) THANGAM MEGANATHAN**, **Ph.D.,** for providing us with the requisite infrastructure and sincere endeavoring in educating us in their premier institution.

Our sincere thanks to **Dr. S.N. MURUGESAN, M.E., Ph.D.,** our beloved Principal for his kind support and facilities provided to complete our work in time. We express our sincere thanks to **Dr. P. KUMAR, Ph.D.**, Professor and Head of the Department of Computer Science and Engineering for his guidance and encouragement throughout the project work. We convey our sincere and deepest gratitude to our internal guide, **Dr. Rakesh Kumar M, PhD.,** Assistant Professor, Department of Computer Science and Engineering. Rajalakshmi Engineering College for his encouragement and guiding us throughout the project to build our project.

PRAVINESH H (210701194)
RAM PRAKASH L (210701208)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

OC	Organized Crimes
RF	RANDOM FOREST
EHR	Electronic Health Record
KNN	K Nearest Neighbours
FIR	First Information Report
JSON	Java Script Object Notation

CHAPTER 1
INTRODUCTION

Machine Learning is the domain that uses past data for predicting. Machine Learning is the understanding of a computer system under which the Machine Learning model learns from data and experience. The machine learning algorithm has two phases: 1) Training & 2) Testing. To predict the disease from a patient's symptoms and from the history of the patient, machine learning technology is struggling from past decades. Healthcare issues can be solved efficiently by using Machine Learning Technology. We are applying complete machine learning concepts to keep track of patient's health.
ML model allows us to build models to get quickly cleaned and processed data and deliver results faster.

By using this system doctors will make good decisions related to patient diagnoses and according to that, good treatment will be given to the patient, which increases improvement in patient healthcare services. To introduce machine learning in the medical field, healthcare is the prime example. To improve the accuracy of large data, the existing work will be done on unstructured or textual data. For the prediction of diseases, the existing will be done on linear, KNN, Decision Tree algorithm.

With the help of developed big data analytics technology, more attention has been paid to patients disease prediction from the perspective of big data analysis, various researches have been researched and conducted by selecting the characteristics automatically from a larger number of data to improve the efficiency and accuracy of risk classification and reduction rather than the previously selected characteristics. However, those existing works are mostly considered structured data.

CHAPTER 2

LITERATURE SURVEY

(Dr C K Gomathy et al,2021)[1] In disease prediction, machine learning (ML) processes people's symptoms to predict the likelihood of various diseases. Naive Bayes classifier is a supervised machine learning algorithm that calculates disease incidence to aid in early diagnosis and patient care. Combining regression and decision trees can improve the prediction of diseases such as diabetes, malaria, jaundice, dengue fever, and tuberculosis and use the physical growth of biomedical information to improve health outcomes.

(Sethi, R. S et al ,2019)[2] Today, more mobile phones are used in the world than ever before. Cell phones are everywhere and mobile technology is growing exponentially. The functionality of mobile phones allows them to provide us with services that make people's lives better. One of the services that mobile phones can provide us is digital therapy. Additionally, mobile apps are recognized to provide cost-effective healthcare solutions. Such applications provide easy and portable healthcare services for everyone. Apps like these provide users with rich experiences where they can learn more about their health and fitness. Mobile digital health apps can use a patient's symptoms to diagnose their illness. Doctors can use this information for further consultation.

(Geluvaraj et.al., 2022)[3] Research approaches disease prediction based on the prediction models and using classifiers and classification techniques. These approaches using the KNN , RFC , NB demonstrate the high accuracy of comparing one another and making them promising for healthcare applications requiring precise predictions.

(Singh, R et.al., 2019)[5] Precise early diagnosis is crucial for improving prognosis and survival rates,spurring advancement in ai for interpreting cardiovascular data to identify risk factors and manifestation of disorders.

(S Vijayarani et al 2020)[6]Their study elaborates the data have become important for predicting clinical outcomes and extracting useful information from general medical data. Disease prediction from big medical data enables researchers to perform tasks such as classification and policies. This study focuses on the prediction of liver disease using extraction methods, specifically naive Bayes and Support Vector Machine (SVM). The performance of this method, testing time and testing time showed that the vector material was before gambling.

(S Mohan et al, 2019)[7]Cardiovascular disease is still the leading cause of death worldwide, so its prediction is important in the analysis of medical data. Machine learning (ML) has proven useful in predictive analytics using big medical data and has also been incorporated into Internet of Things (IoT) applications. Current research is only a fraction of the machine learning research into predicting heart disease. This paper presents a new method to improve the accuracy of prediction by identifying key features and using various classification methods. Our hybrid model combines random forests and methods to increase the accuracy of heart disease detection by 88.7%.

(Kumar, Y., et al,2023)[10]Artificial Intelligence (AI) improves patient care and health by using machine learning and deep learning to perform tasks such as disease diagnosis and drug discovery. The research examined AI tools used to diagnose Alzheimer's, cancer, diabetes, heart disease, tuberculosis, stroke, high blood pressure, and more, using different clinical data such as MRI and CT scans. Articles up to October 2020 are selected from archives such as Web of Science, Scopus and PubMed. The survey compares studies on metrics such as accuracy, sensitivity, specificity, AUC, precision, recall and F1 score, demonstrating the effectiveness of clinical skills and patient treatment.

CHAPTER 3

SYSTEM DESIGN

3.1 DEVELOPMENT ENVIRONMENT

3.1.1 HARDWARE SPECIFICATIONS

This project uses minimal hardware but in order to run the project efficiently without any lack of user experience, the following specifications are recommended

Table 3.1.1 Hardware Specifications

PROCESSOR	Intel Core i5
RAM	4GB or above (DDR4 RAM)
GPU	Intel Integrated Graphics
HARD DISK	6GB
PROCESSOR FREQUENCY	1.5 GHz or above

3.1.2 SOFTWARE SPECIFICATIONS

The software specifications in order to execute the project has been listed down in the below table. The requirements in terms of the software that needs to be pre- installed and the languages needed to develop the project has been listed out below.

Table 3.1.2 Software Specifications

FRONT END	Python GUI
BACK END	Python
FRAMEWORKS	TkINTER
SOFTWARES USED	Python IDLE, Google Chrome,VSCODE

3.2 SYSTEM DESIGN

3.2.1 ARCHITECTURE DIAGRAM

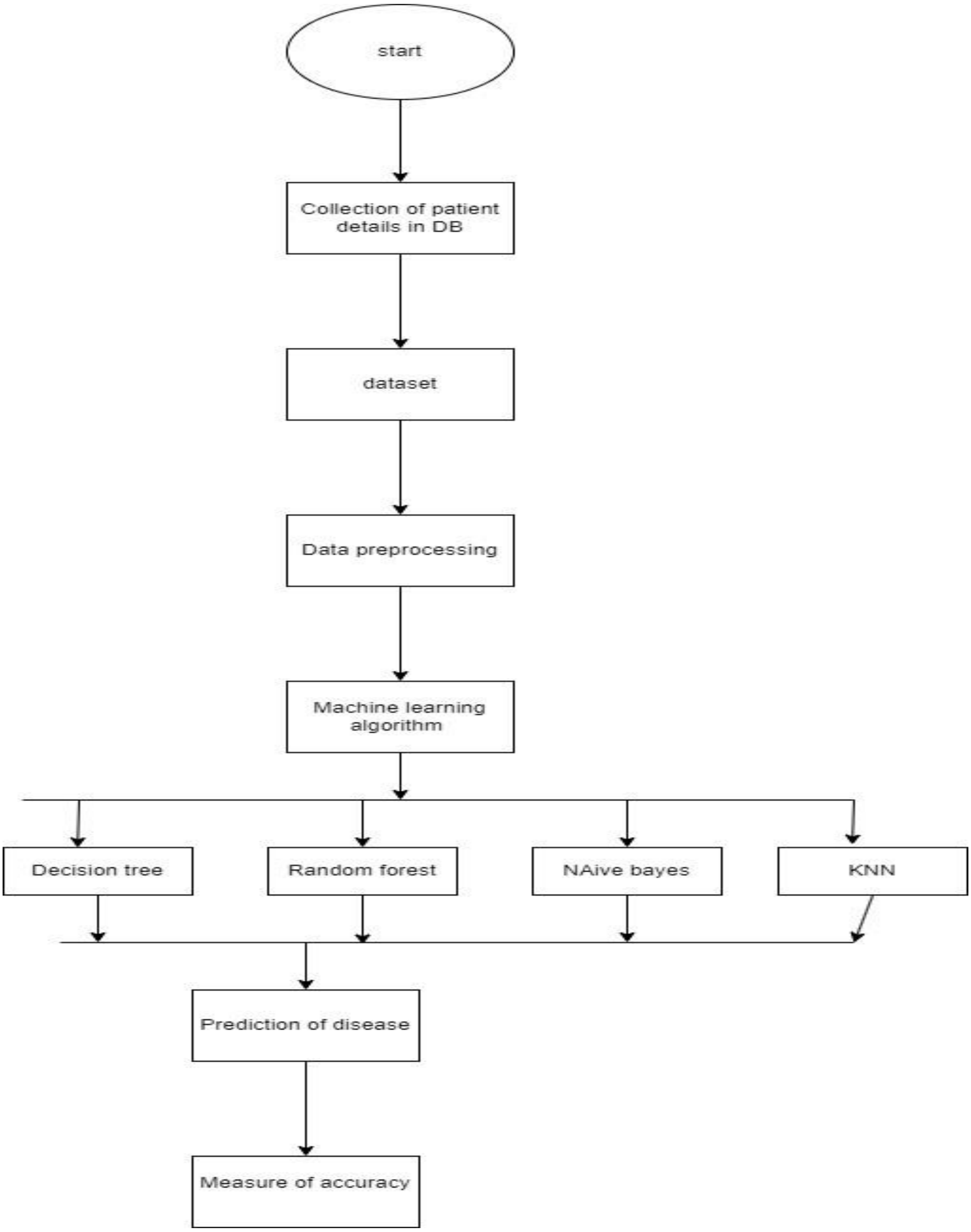


Fig 3.2.1 Architecture Diagram

ALGORITHMS

NAIVE BAYES:

Naive Bayes is an easy however amazingly powerful rule for prognosticative modeling. The independence assumption that allows decomposing joint likelihood into a product of marginal likelihoods is called'naive'. This simplified Bayesian classifier is called naive Bayes. The Naive Bayes classifier assumes the presence of a particular feature in a class is unrelated to the presence of any other feature. It is very easy to build and useful for large datasets. Naive Bayes is a supervised learning model. Bayestheorem provides some way of calculative posterior chance P(b|a) from P(b),P(a) and P(a|b).

Look at the equation :

P(b v a)= P(a v b)P(b)/P(a)Above,

P(b|a) is the posterior chance of class (b,target) given predictor (a, attributes)

P(b) is the priori probability of class.

P(a|c) is the chance of a predictor given the class.
P(a) is the priori probability of a predictor. In our system, Naïve Bayes decides which symptom is to put in the classifier and which is not. 8.3 LOGISTIC REGRESSION Logistic regression could be a supervised learning classification algorithm accustomed to predict the chance of a target variable that is Disease.

DECISION TREE

A decision tree is a structure that can be used to divide up a large collectionof records into successfully smaller sets of records by applying a sequence of simple decision tree. With each successive division, the members of the resulting sets become more and more similar to each other. A decision tree model consists of a set of rules for dividing a large heterogeneous populationinto smaller, more homogeneous (mutually exclusive) groups with respect to a particular target. The target variable is usually categorical and the decision tree is used either to: Calculate the probability that a given record belongs to each of the categories and, To classify the record by assigning itto the most likely class (or category). In this disease prediction system, the decision tree divides the symptoms as per its category and reduces the dataset difficulty.

RANDOM FOREST ALGORITHM

Random Forest is a supervised learning algorithm. It is an extension of machine learning classifiers which include the bagging to improve the performance of Decision Tree. It combines tree predictors, and trees are dependent on a random vector which is independently sampled. The distribution of all trees are the same. Random Forests splits nodes using the best among of a predictor subset that are randomly chosen from the node itself, instead of splitting nodes based on the variables. The time complexityof the worst case of learning with Random Forests is O(M(dnlogn)) , whereM is the number of growing trees, n is the number of instances, and d is thedata dimension.It can be used both for classification and regression. It is alsothe most flexible and easy to use algorithm. A forest consists of trees. It is said that the more trees it has, the more robust a forest is. Random Forests create Decision Trees on randomly selected data samples, get predictions from each tree and select the best solution by means of voti pretty good indicator of the feature importance. Random Forests have a variety of applications, such as recommendation engines, image classification and feature selection. Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML.

K NEAREST NEIGHBOR:

k-Nearest Neighbors (kNN) is a simple machine learning method. This article introduces some basic concepts of the kNN algorithm and then focuses on the use of R for kNN models. After predicting the results using the kNN algorithm, the diagnosis of the model should be checked. True mean is the most commonly used method regarding the kNN algorithm. Factors such as the k value, distance calculation, and selection of the appropriate variable value all have an impact on the performance of the model and The nearest neighbor (KNN) algorithm is a general machine learning algorithm used for classification tasks, including disease prediction.

CHAPTER 4

PROJECT DESCRIPTION

Develop a system to predict potential human diseases based on a set of symptoms using machine learning classification algorithms. The project will compare the performance of four algorithms: Random Forest, K-Nearest Neighbors (KNN), Naive Bayes, and Decision Tree.

Data:

The system will utilize a dataset containing historical medical records with patient information including symptoms and corresponding diagnoses. The data will be preprocessed to ensure quality and consistency.

<u>Methodology:</u>

Data Acquisition:

Acquire a medical dataset containing patient symptoms and diagnosed diseases.

Data Preprocessing:

Clean the data by handling missing values and outliers.
Encode categorical features (e.g., gender, blood type) into numerical values.
Feature scaling may be applied to ensure all features have a similar range.

Model Development:

Split the data into training and testing sets. Implement the following machine learning algorithms:

Random Forest:An ensemble learning method that combines multiple decision trees for improved accuracy and reduced overfitting.

K-Nearest Neighbors (KNN): Identifies the most similar instances (based on symptom data) in the training set to predict the disease for a new patient.

Naive Bayes: Classifies data based on the probability of symptoms occurring given a specific disease, assuming features are independent.

Decision Tree: A tree-like model that learns a series of rules based on symptom data to classify diseases. Train each model on the training set.

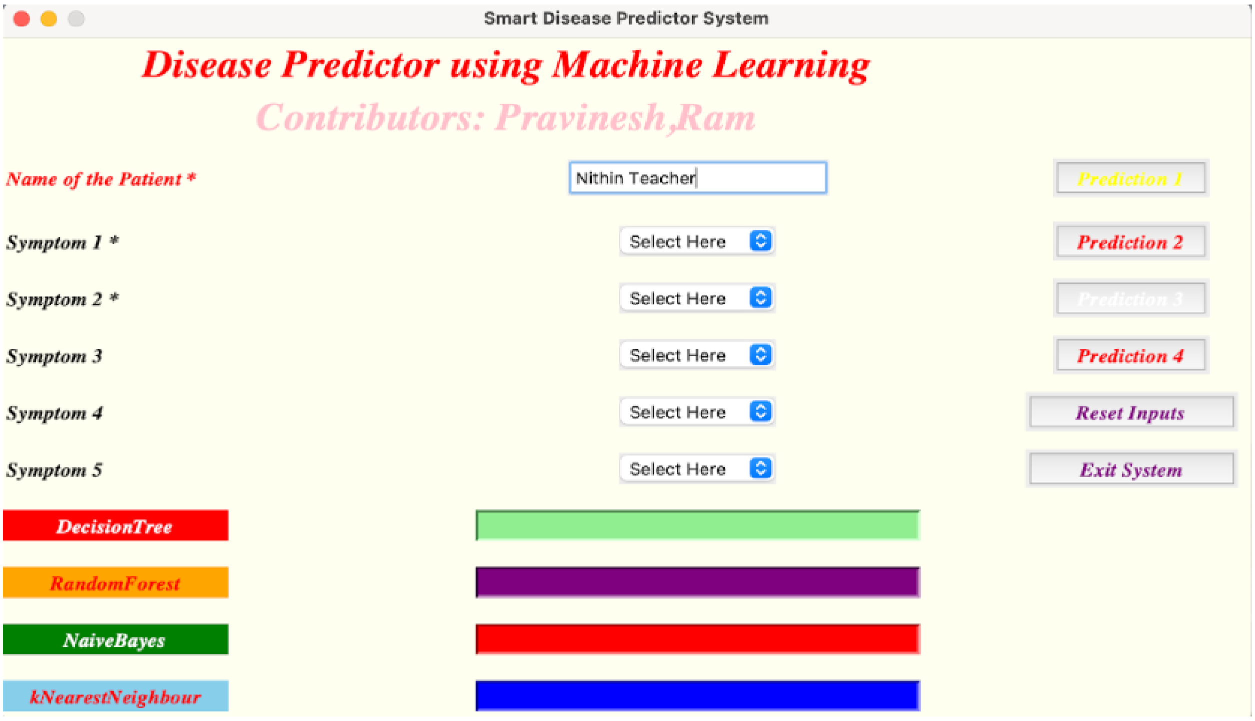
Model Evaluation:

Evaluate the performance of each model on the testing set using metrics like accuracy, precision, recall, and F1-score. Compare the performance of the different algorithms and identify the one with the highest prediction accuracy for disease classification.

System Development:

Develop a user interface for the disease prediction system.
Users can input their symptoms, and the system will predict the most likely disease(s) based on the chosen model.

OUTPUT SCREENSHOTS



Smart Disease Predictor System

Disease Predictor using Machine Learning

Contributors: Pravinesh,Ram

Name of the Patient * Nithin Teacher

Symptom 1 * Select Here

Symptom 2 * Select Here

Symptom 3 Select Here

Symptom 4 Select Here

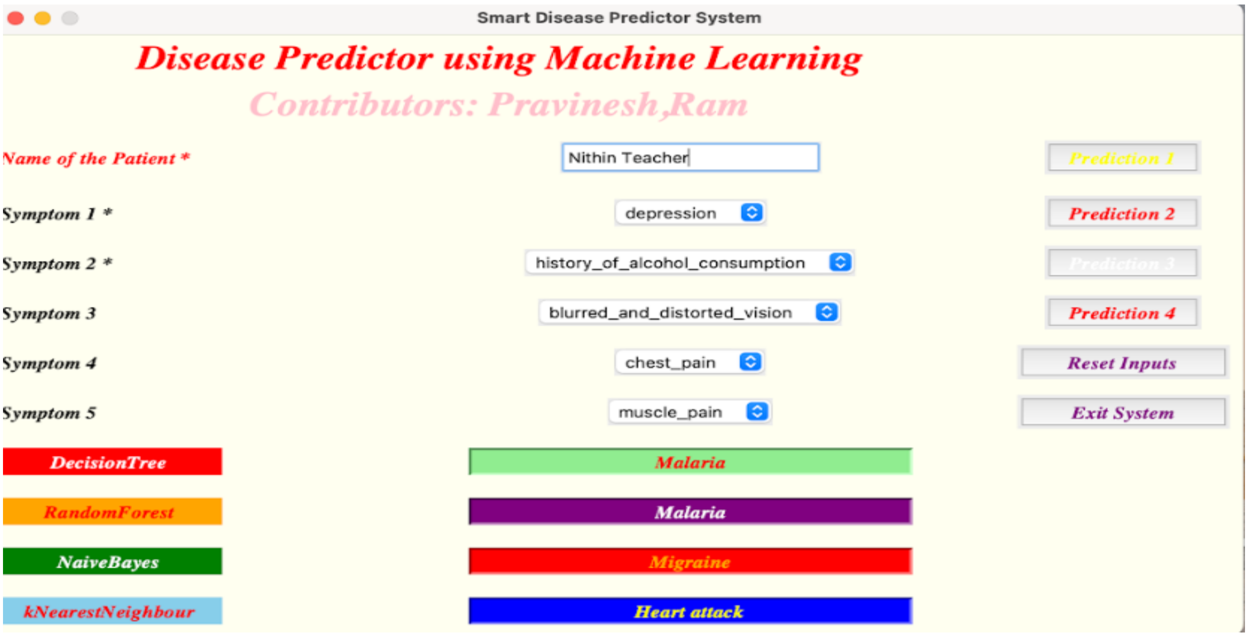
Symptom 5 Select Here

DecisionTree

RandomForest

NaiveBayes

kNearestNeighbour



Smart Disease Predictor System

Disease Predictor using Machine Learning

Contributors: Pravinesh,Ram

Name of the Patient * Nithin Teacher

Symptom 1 * depression

Symptom 2 * history_of_alcohol_consumption

Symptom 3 blurred_and_distorted_vision

Symptom 4 chest_pain

Symptom 5 muscle_pain

DecisionTree Malaria

RandomForest Malaria

NaiveBayes Migraine

kNearestNeighbour Heart attack

CHAPTER 6

CONCLUSION AND FUTURE ENHANCEMENTS

The main aim of this disease prediction system is to predict the disease on the basis of the symptoms. This system takes the symptoms of the user from which he or she suffers as input and generates final output as a prediction of disease. Average prediction accuracy probability of 100% is obtained. Disease Predictor was successfully implemented using the grails framework. This system gives a user-friendly environment and is easy to use. As the system is based on the web application, the user can use this system from anywhere and at any time. In conclusion, for disease risk modeling, the accuracy of risk prediction depends on the diversity feature of the hospital data. This systematic review aims to determine the performance, limitations, and future use of Software in healthcare. Findings may help inform future developers of Disease Predictability Software and promote personalized patient care. The program predicts Patient Diseases. Disease Prediction is done through User Symbols. In this System Decision tree, Random Forest, the Naïve Bayes Algorithm is used to predict diseases. For the data format, the system uses the Machine Learning algorithm Process Data on Database Data namely, Random Forest, Decision Tree, Naive Bayes. System accuracy reaches 98.3%. machine learning skills are designed to successfully predict outbreaks.

Integrate additional patient information (e.g., age, medical history) for improved prediction.Explore the use of more advanced machine learning models (e.g., Support Vector Machines, Neural Networks).Implement techniques to handle imbalanced datasets, where certain diseases might be less frequent.Focus on a specific disease category for a more in-depth analysis.

REFERENCES

[1]. The prediction disease using machine learningß || Dr C K Gomathy , Mr.A.Rohith Naidu,"Volume:05" "Issue:10"||Oct 2021||IJSREM.

[2]. Sethi, R. S., Thumar, A., Jain, A. V., & Chavan, S. (2019). Disease prediction application based on symptoms. International Journal of Scientific Research in Computer Science Engineering and Information Technology, 641-646.

[3]. Geluvaraj , B., Santhosh, K., Prabhu, N., Reddy, A., Sandhya, T., & Bhaskar, S. V. (2022, January). A hybrid approach for predicting diseases using clustering and classification techniques. In 2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI) (pp. 1-6). IEEE\

[4]. S Vijayarani , S Dhayanand Liver disease prediction using SVM and Naive bayes algorithms International Journal of Science, Engineering and Technology Research (IJSETR) , volume 4 , issue 4 Posted: 2015

[5]. S Mohan , C Thirumalai , G Srivastava Effective heart disease prediction using hybrid machine learning techniques IEEE Access , volume 7 Posted: 2019

[6]. Kumar, Y., Koul, A., Singla, R., & Ijaz, M. F. (2023). Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. Journal of ambient intelligence and humanized computing, 14(7), 8459-8486