HUMAN DISEASE PREDICTION AND DOCTOR BOOKING SYSTEM

Mr.Joel Roy[1], Mr. Reeju Koshy[2], Mr. Roshan Roy[3]
Dept. of CSE, Mangalam College of Engineering, Ettumanoor, Kerala, India
Ms. Anjumol Zachariah[4]
Asst. Prof Dept. of CSE, Mangalam College of Engineering, Ettumanoor, Kerala, India
joelroy18@gmail.com , reeju085@gmail.com , roshanellickal@gmail.com

***Abstract*—Today, data mining is more important than ever in the healthcare and medical industries. When specific data mining techniques are applied correctly, important information can be obtained from large databases, enabling medical professionals to make quicker decisions and enhance patient care. The idea is to help the doctor by using the classification. Understanding how to correctly diagnose patients through clinical examination and assessment is essential. The health care industry generates a lot of information about clinical evaluation, patient reports, treatments, follow-up meetings, medications, and other topics. It takes careful orchestration to execute it well. Due to poor information management, the quality of the data association has been impacted. A legal method must be found to focus and process information in a viable and effective manner as data volumes increase. A classifier that can divide the data into groups based on attributes is built using one of the various machine learning applications. The data set is divided into at least two classes. These classifiers are used to analyse medical data and forecast diseases.This project aims to develop a portal for predicting disease according to the symptoms which is given by the user and an option for consulting doctor.**

I. Introduction

The creation and use of a number of well-known data mining techniques in a variety of real-world application fields (such as industry, healthcare, and bioscience) has resulted in the use of these techniques in machine learning environments to extract useful information from the target data in healthcare communities, biomedical fields, etc. The accurate analysis of medical databases aids in the early diagnosis of illnesses, patient treatment, and social services. There are numerous applications where machine learning techniques have been successfully used, including the forecast of disease. The goal of creating a classifier system utilising machine learning algorithms is to significantly aid in the resolution of health-related problems by supporting doctors in early disease prediction and diagnosis. But for a doctor, making an accurate forecast based on symptoms is too challenging. The hardest task is making an accurate diagnosis of a condition. Data mining is crucial in predicting the sickness in order to solve this issue.

The annual data increase in the medical sciences is substantial. The accuracy of medical data analysis, which has benefited from early patient care, has increased as a result of the rise of data in the medical and healthcare fields. Data mining uses disease data to uncover patterns that are hidden in the vast volume of medical data. 1 The most prevalent health conditions frequently have certain fundamental signs that people typically display. For instance, a person with a headache may also display a number of other diseases' symptoms. We rely on doctors in situations where we demand an immediate diagnosis. Based on the symptoms, a machine learning model can be created to predict the disease's type. Early disease identification and quicker diagnosis may be made possible by the model's predictions. Malaria, dengue, impetigo, diabetes, migraines, jaundice, chicken pox, and other ailments have a substantial impact on a person's health and can even be fatal if left untreated. The healthcare sector can make smart decisions by "mining" the massive database they already have or by identifying its hidden links and patterns. This problem can be solved using data mining methods like decision trees, random forests, and naive bayes. Therefore, using the rule set of the appropriate algorithms, we can create an automated system that can find and extract secret information about diseases from a history (diseases-symptoms) database.

II. LITERATURE SURVEY

Disease prediction by machine learning over big data from healthcare communities ,in this paper, they streamline machine learning methods to accurately forecast the onset of chronic diseases in areas with a high incidence of those diseases. They tested the updated prediction models using actual hospital data from central China that was gathered between 2013 and 2015. They use a latent component model to fill in the gaps in the data to overcome the challenge of incomplete data. They test various treatments for a localised, persistent cerebral infarction. They suggest a fresh multimodal illness risk prediction algorithm based on convolutional neural networks (CNNs), which uses organised and unstructured hospital data.

A relative similarity based method for interactive patient risk prediction, this study examines the patient risk prediction issue within the framework of active learning, with comparatively similar results. Active learning has been thoroughly investigated and successfully used to address practical issues. Active learning techniques are typically used to explore absolute questions.

Disease and symptoms dataset, this study identifies diseases based on symptoms and provides more information about the most frequently occurring diseases, including treatment recommendations. This research combines unique machine learning and IR techniques. The models that performed the best were DT (Decision Tree), KNN (K-Nearest Neighbour), and LR (Logistic Regression), with an accuracy of 91.29% and cross-validation accuracy of 89.1%, respectively. The method is simple to use and can be helpful in the early detection and diagnosis of diseases. It can even be used by someone with limited medical understanding. Users who are hesitant to go to hospitals when they first experience minor symptoms may also find it helpful. This will provide them with a fundamental understanding of the disease's severity.

Intelligent heart disease prediction System using data mining techniques,this research discusses a data mining method that locates CAD instances using noninvasive clinical data. Over the course of 2012–2013, 335 subjects' clinical data were gathered at the cardiology division of the Indira Gandhi Medical College in Shimla, India. Only 48.9% of the subjects who underwent coronary angiography had coronary stenosis and were identified as having CAD.

III. PROPOSED SYSTEM

Developing the classification model In this study, we first construct a straightforward classification model using Sklearn library ideas. One of the machine learning libraries for Python is called Scikit-learn. It also makes use of numerous techniques, including support vector machines, random forests, and k-neighbours, and it supports Python's scientific libraries, which include NumPy and SciPy. B. Build the model. The model must first be defined before being compiled. The performance of various algorithms was then evaluated based on the accuracy scores of the various procedures they were used to conduct. .Another important step that came into picture was feature selection. The first step is to assemble a substantial dataset with a wide range of components known to be associated with the disease. For example, the dataset for heart disease might include information on age, gender, blood pressure, cholesterol levels, smoking habits, family history of heart disease, etc. The data is preprocessed after collection to remove any discrepancies, errors, or missing numbers. The data is also standardised to guarantee that all the variables are measured on the same scale.The process of selecting the most pertinent features from the dataset that are most likely to have a substantial impact on the prediction is known as feature selection.A suitable machine learning algorithm is then picked based on the type of disease being predicted and the characteristics of the dataset once the pertinent features have been chosen.

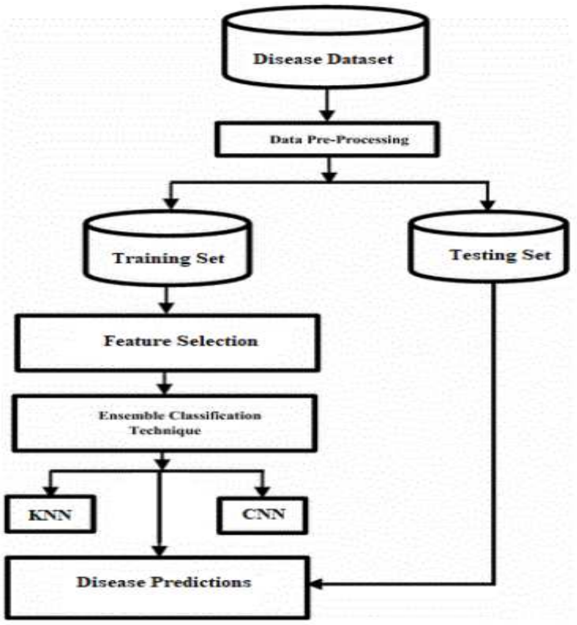

Fig 1:conceptual model

A. *disease prediction using machine learning algorithm*

The data set considered consists of 132 symptoms, the combination or permutations of which leads to 41 diseases. Based on the 4920 records of patients aim to develop a prediction model that takes inthe symptoms from the user and predicts the disease he or she is more likely to have.

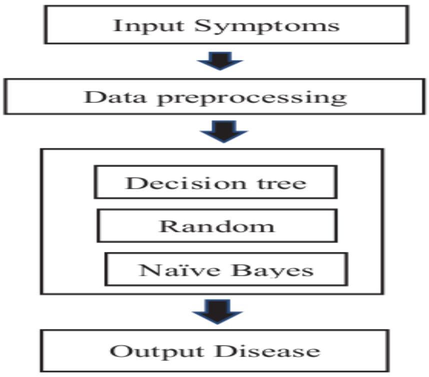

Fig 2: health prediction using data mining

B. *Symptom Based Health Prediction using Data Mining*

a) Decision Trees

• The first step is to classify the data to fit the model of decision trees for the given dataset.
• split the data into test data and training data for the model. Construct a node table to
assign the different classifiers and Gini for splitting the nodes.
• Classify the model using Decision Tree Classifier.
• The only extra advantage of using this is to apply it for both numerical and categorical
data set classification.

b) *Random forest*

• It is a supervised algorithm in which multiple decision trees are built with the help of
bagging method.

• It does not rely on the most important features rather it uses random subset of features
is considered while splitting a node.

• There is very low or no bias since it relies on the power of the crowd.

c) *Naïve bytes*

• Used a built in function called Multinomial Naïve Bytes which is mainly used for discrete features such as text classification.

• It requires a feature count parameter which helps in determining each class while fitting the sample with appropriate weights.

d) Text processing

• For the initial step, consider the symptoms data and perform count-vectorization. This is done to sort the words in the corpus into a bag of words.

• The model is simple in that it ignores the order of words and relations rather focus on the occurrence of words the dataset.

• For all the 3 algorithms, use the tf-idf parameter which helps in keyword extraction to
help in faster computation.

C. Designing Disease Prediction Model Using Machine Learning Ap

• Initially takes disease dataset from UCI machine learning website and that is in the
form of disease list with its symptoms.

• After that preprocessing is performed on that dataset for cleaning that is removing
comma, punctuations and white places. And that is used as training dataset.

• The feature extracted and selected. Then we classify that data using classification
techniques such as KNN and CNN. Based on machine learning we can predict accurate
disease.

• KNN is a slow-learning algorithm that is non-parametric. Its goal is to employ a specific database, where the data will refer to several classes, to predict how a fresh sample will be categorised. When we refer to a method as nonparametric, we mean that it does not rely on any presumptions about the underlying data. In other words, only the data is used to establish the structure.

• Convolutional neural networks, or CNNs, are a powerful deep learning technique that has been applied to a variety of tasks, including disease prediction. CNNs' capacity to automatically learn features from raw input, like images or time-series data, eliminates the need for manual feature engineering and is their main advantage. Before employing CNNs to forecast diseases, the data must first be collected and prepared. This can require getting medical imaging or other relevant data from the patient, preprocessing it to remove noise or artefacts, and then tagging it with the correct disease diagnosis or outcome.

IV. IMPLEMENTATION REQUIREMENT

A. Software Requirements

• Pandas -A strong Python library for data manipulation and analysis is called Pandas. It offers a range of data structures for representing and interacting with data, such as series (1-dimensional) and data frame (2-dimensional). Pandas is a data manipulation, cleaning, merging, and filtering tool that is developed on top of NumPy.

• Django-Model-view-controller (MVC) is a design approach used by the high-level Python web framework Django. It provides a selection of frameworks and tools to help web developers make web apps efficiently and quickly. Because of Django's adaptable and reliable ORM, developers may connect to databases using Python code rather than writing SQL queries directly. Django automatically develops an admin interface for your application based on your models that can be customised to meet your needs.

• Seaborn- A Matplotlib-based Python data visualisation library is called Seaborn. It provides an advanced user interface for creating informative and practical statistical graphics. Seaborn is frequently used to create statistical model visualisations and explore complex datasets. You may create many other visual representations using Seaborn, including scatterplots, line plots, bar plots, histograms, kernel density plots, and heatmaps. Pandas, a different Python data analysis library, is widely used in conjunction with Seaborn.

• Scikit learn - Scikit-learn, sometimes known as sklearn, is a popular open-source machine learning library for Python. It provides a number of techniques for machine learning applications like classification, regression, clustering, and dimensionality reduction through a standard interface. Scikit-learn was built on top of NumPy, SciPy, and Matplotlib and integrates well with other Python tools.

B. Hardware Requirements

- Intel I3 2.0 GHz or higher processor
- 4 GB RAM
- 512 GB SSD space

V . EXPERIMENT AND RESULTS

With the system, several experiments are conducted. The suggested random forest model's achieved precision, recall, and F1-score values are contrasted with the performance metrics of the Naive Bayes, decision tree, and logistic regression methods. The findings are presented in Table 1. Since the patient heavily depends on the outcome of the prediction, accuracy is a crucial factor because getting it wrong could harm the patient. The additional variables, such as precision, recall, and F1-score, are used to assess the effectiveness of the model, as shown in Table 1.

	Accuracy	Precision	Recall	F1-score
Naïve bytes	78	72	80	65
Decision tree	65	60	75	80
Logistic regression	86	84	88	82
Random forest	96	93	99	97

Table 1: performance evaluation comparison

The graphical comparison of the suggested and alternative algorithms' accuracy results is shown in Figure 3. This graph shows the changes in the four algorithms' prediction accuracies, which are 78%, 65%, 86%, and 96%, respectively, for Naive Bayes, Decision Tree, Logistic Regression, and the proposed Random Forest algorithms. This demonstrates that, when compared to the other machine learning methods, the suggested system gets the maximum accuracy of 96%.

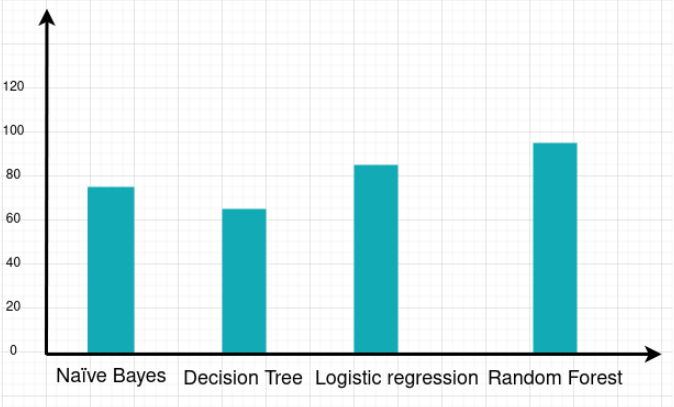

Figure 3: Comparison of accuracies of proposed and other algorithms

The comparison precision, recall, and F1-score values of the suggested and alternative algorithms are represented graphically in Figure 4. The Naive Bayes, Decision Tree, Logistic Regression, and Proposed Random Forest algorithms are represented by the three performance evaluation parameters in this graph as 72%, 60%, 84%, and 96%, respectively, for precision; 80%, 75%, 88%, and 99%, respectively, for recall; and 65%, 80%, 82%, and 97%, respectively, for F1-score. These findings demonstrate that the proposed model, created using the Random Forest algorithm, is preferred over the other three algorithms, with precision, recall, and F1-score values of 93%, 99%, and 97%, respectively.

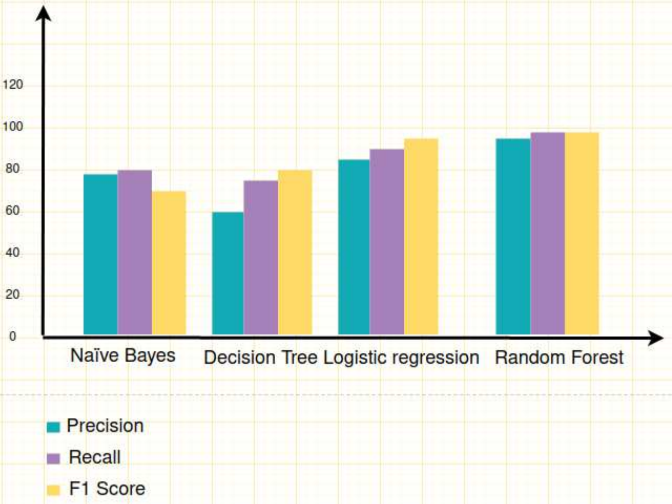

Fig 4 : Comparison of other performance evaluation metrics of proposed and other algorithms.

VI CONCLUSION

Because of the ongoing pandemic, we have realised the true value and importance of health and life. A healthy life is the greatest form of wealth. Our guiding principle throughout the development of this project was to help the common people and address their health-related problems. This webapp was created by our team to analyse and predict diseases using a variety of machine learning algorithms based on user-provided symptoms, and it has proven to be quite effective in offering the highest accuracy and precise prediction. In order to enable our users to receive proper medical treatment from home during an ongoing epidemic, we also provided them with live doctor consultations. This project is a carefully designed combo for generic healthcare with the best ideal solutions, whether it be in forecasting an illness or analysing it. receiving information about the dangers that a disease prediction poses to other body organs as well. For communicating and obtaining treatment as quickly as possible, this project also suggested seeing a specialist doctor. This project required a tremendous amount of work and rigour on our part to develop and overcome obstacles that people using the Smart Health Disease Prediction System face on a daily basis with regard to their health. In the future, we hope to create more projects to help the less fortunate as well as improve society

REFERENCES

[1] M. Chen, Y. Hao, K. Hwang, L. Wang, and L.Wang,"Disease prediction by machine learning over big data from healthcare communities", ," *IEEE Access,* vol. 5, no. 1, pp. 8869–8879, 2017.

[2] B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang, "A relative similarity based method for interactive patient risk prediction," *Springer Data Mining Knowl. Discovery,* vol. 29, no. 4, pp. 1070–1093, 2015.

[3] IM. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, and C. Young "Wearable 2.0: Enable human-cloud integration in next generation healthcare system," *IEEE Commun.* , vol. 55, no. 1, pp. 54–61, Jan. 2017.

[4] Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, and A. Alamri, “HealthCPS: Healthcare cyberphysical system assisted by cloud and big data,” *IEEE Syst.* J., vol. 11, no. 1, pp. 88–95, Mar. 2017.

[5] L. Qiu, K. Gai, and M. Qiu, “Optimal big data sharing approach for telehealth in cloud computing,” *in Proc. IEEE Int. Conf. Smart Cloud (Smart Cloud),* Nov. 2016, pp. 184– 189.

[6] Disease and symptoms Dataset *–www.github.com.*

[7] Heart disease Dataset-*WWW.UCI Repository. com*

[8] Ajinkya Kunjir, Harshal Sawant, Nuzhat F.Shaikh, “Data Mining and Visualization for prediction of Multiple Diseases in Healthcare,” *in IEEE big data analytics and computational intelligence*, Oct 2017 pp.2325.

[9] Shanthi Mendis, Pekka Puska, Bo Norrving, World Health Organization (2011), Global Atlas on Cardiovascular Disease Prevention and Control, PP. 3– 18. ISBN 978-92-4-156437-3.

[10] Amin, S.U.; Agarwal, K.; Beg, R., “Genetic neural network based data mining in prediction of heart disease using risk factors*”, IEEE Conference on Information & Communication Technologies (ICT),* vol., no.,pp.1227-31,11- 12 April 2013.

[11] Palaniappan S, Awang R, “Intelligent heart disease prediction System using data mining techniques,” *IEEE/ACS International Conference on Computer Systems and Applications*, AICCSA 2008., vol., no., pp.108115, March 31 2008-April 4 2008