How to Cite:
Karwa, H., Gupta, P., Agrawal, R., Virdi, G. S., Kumar, A., & Jain, S. (2022). Multi-disease prediction with machine learning. *International Journal of Health Sciences*, *6*(S2), 9477–9483. https://doi.org/10.53730/ijhs.v6nS2.7487

Multi-disease prediction with machine learning

Harsh Karwa
Student, CSE Department, Shri Ramdeobaba College of Engineering and Management, Nagpur, India
Correspondence author email: harshkarwa25@gmail.com

Pavan Gupta
Student, CSE Department, Shri Ramdeobaba College of Engineering and Management, Nagpur, India

Ram Agrawal
Student, CSE Department, Shri Ramdeobaba College of Engineering and Management, Nagpur, India

Gursewak Singh Virdi
Student, CSE Department, Shri Ramdeobaba College of Engineering and Management, Nagpur, India

Amit Kumar
Student, CSE Department, Shri Ramdeobaba College of Engineering and Management, Nagpur, India

Sweta Jain
Professor, CSE Department, Shri Ramdeobaba College of Engineering and Management, Nagpur, India

***Abstract*---**In the present era, Machine learning (ML) algorithms are extensively used in computer assisted diagnosis of the disease based on the symptoms of the disease. The widespread use of healthcare applications in the pandemic time, provides a motivation to further develop new computer assisted diagnostic application in the healthcare domain. Prevention and treatment of disease, accurate and timely diagnosis of any health-related problem is essential. In the case of a serious illness, a standard diagnostic method may not be enough. We have proposed a system for predicting the disease. There were about forty-one diseases in the data corpus that needed to be analyzed based on the symptoms. The system delivers a disease prediction that a person may have depending on the symptoms. This diagnostic program can assist a physician in diagnosing disease, allowing for timely treatment and saving lives. The disease forecasting system was developed using ML models such as the Random Forests,

International Journal of Health Sciences ISSN 2550-6978 E-ISSN 2550-696X © 2022.
Manuscript submitted: 27 March 2022, Manuscript revised: 18 April 2022, Accepted for publication: 9 May 2022

the Naive Bayes, and the Support Vector Machine Classification Algorithm. The presented work outlines an analysis of the aforementioned algorithms.

***Keywords*---**machine learning, disease prediction, random forest, naive bayes, support vector machine.

Introduction

Medicine and health are important factors in economic growth and human life. Technology assisted health care applications are significantly increasing since the past two decades. In the pandemic period, there are many remote areas that still do not have emergency health care services. To effectively cater the need of masses in heavily populated. Country like India, the online diagnosis system is the need of hour. Disease predicting systems may be a boom in many cases as it can prevent the caused life risk diseases beforehand and can suggest the individual to get immediate treatment to prevent further damage from the underlying diseases. Not just that, this strategy will cut treatment costs and reduce fear in the late stages, enabling sufficient treatment to be provided at the right moment and reducing the rate of death. Furthermore, several localized diseases have distinct features in different places, making disease outbreak prediction difficult.

Related Work

There is a lot of research into disease prediction models utilizing various machine learning algorithms, with varying results for various medical techniques. The author Chauhan et al., (2020) shown an accuracy of the machine learning models - Decision Tree, Random Forest, and Naive Bayes,as 92.4 %, 95.7 %, and 94.5 % respectively[1]. Another study published by author Chen et al. (2017) on CNN-based multimodal disease risk prediction achieved an accuracy of 94.5 % [2]. The accuracy of the research work on Fuzzy Logic, Fuzzy Neural Networks, and Decision Tree published by Leoni et al. (2017) was 58.8 %, 91 %, and 68.7 %, respectively [3]. Furthermore, the accuracy achieved by Vijayarani et al. research work on the SVM and Nayes Bayes was determined to be 79.66 % and 61.28 %, respectively [4].

Materials and Methods

Our proposed work is based on the prediction of many diseases that follow a patient's symptoms. In any medical application task the main important footstep is to get the data corpus. The data preprocessing is an essential step to clean it and prepare it for building a model during training phase. The testing of the model is carried out using unseen data/test data from corpus.User will provide the symptoms to our system. The symptoms will be provided as input / key feature to our ML model where we will be using algorithms like Random Forest, Naive Bayes, and SVM to predict disease in order to help the patient in early stages of their disease. In this work, we have used python as a platform for using machine learning algorithms. We've also built a great GUI to provide system connectivity.

Dataset

The data corpus for this application is collected from Kaggle, which includes attributes containing diseases, and their symptoms. The user needs to understand related features in the dataset. This dataset can be easily found on Kaggle for the same link that has been provided in references [5].Workflow diagram of the system is shown in Figure1.

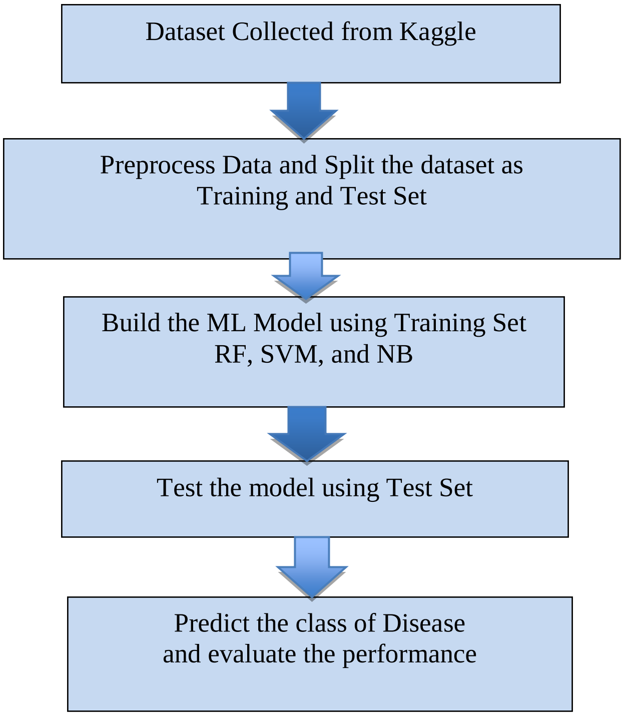

Dataset Collected from Kaggle
Preprocess Data and Split the dataset as Training and Test Set
Build the ML Model using Training Set RF, SVM, and NB
Test the model using Test Set
Predict the class of Disease and evaluate the performance

Figure1: Workflow diagram of the system

Machine Learning

The ML algorithms are heavily dependent on the amount of the data presented to them for learning / training the target variable. In this work we have implemented multiclass classification of the diseases. The dataset is publically available. Finding a subset of the relevant features is the major task in design of any healthcare application. As, it is possible that many disease may have some of the overlapping symptoms. The result obtained may not be generalized to real world diagnosis applications, unless it is examined by the specialist doctor.

Naive Bayes

It's a classification technique that utilizes Bayes' Theorem as well as the predictor independence condition. This model is straightforward and is extremely advantageous for big datasets. Naive Bayes is believed to exceed even the most sophisticated classification algorithms due to its simplicity. This is rooted in the Bayes theorem, which enables us to evaluate the conditional probabilities, say P (F|G) using P (F), P (G), and P (G|F).Thus the Bayes Theorem can be represented as

P (F|G) =P (G|F)*P (F)/P (G)

The conditional probabilities and the class probabilities P (Yi) are computed using the training dataset by the Naive Bayes classifier. Although connected features are voted twice in the model, the Naive Bayes classifier works perfectly when they are omitted. This yields an overemphasis on the value of the associated features.

Random Forest

Random forest is a supervised learning technique that can be used to classify and predict data. However, it is mostly employed to solve categorization issues. A forest, as we all know, is made up of trees, and more trees equals a more healthy forest. It's an ensemble method that's superior to a single decision tree because it averages the results to reduce over fitting of the model. It chooses the best voting solution. Random Forest produces better results than real problems mainly due to noise incompatibility in the database and is not based on overload. It works great too and shows excellent performance over other tree-based algorithms. To read the tree, bootstrap is widely used for merging or wrapping.

Support Vector Machine

SVM is a popular method of classification. It is widely used in Machine Learning for differentiating the differences in any given dataset. Linear SVM is used for linear data, which means that if a database can be divided into two categories using one straight line, such data is called linear data. Non-Linear SVM is used for non-linear data, which means that if the database cannot be categorized using a straight line. There are different types of kernels used in Non-Linear SVM, some of them are Gaussian radial basis function (RBF), Polynomial kernel, Hyperbolic Tangent kernel sigmoid kernel, ANOVA radial basis kernel. In our SVM, we have utilized the RBF, which is the Gaussian radial basis function. The RBF kernel say for a and a' can be represented as

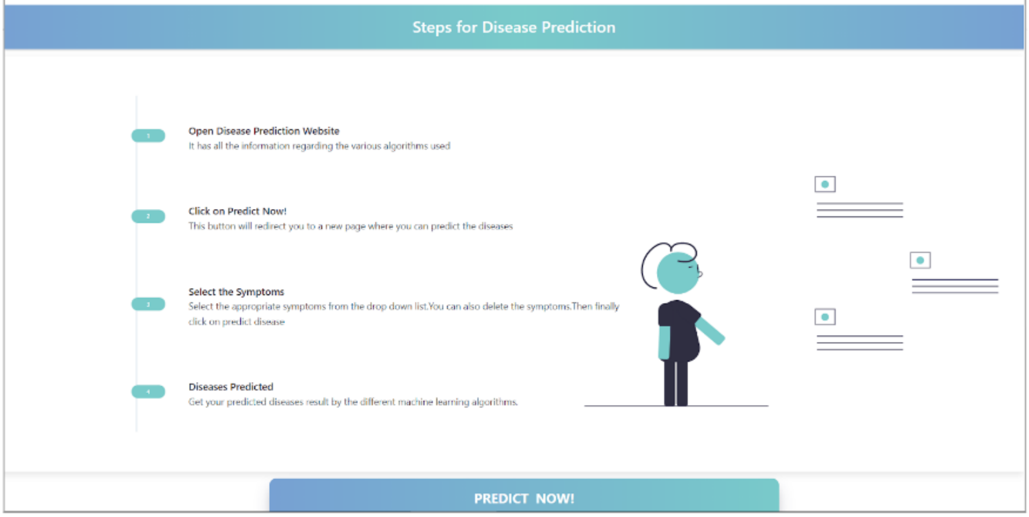
$$K(a,a') = exp(-\gamma ||a - a'||^2)$$

Here,
$||a - a'||^2$ = Square of the Euclidean distance between the two feature vectors
$\gamma = 1 / 2\sigma^2$ (For positive values)

Result and Discussion

This section indicates the results of a developed system that can predict disease faster, accurately with high fidelity than the existing system. Results are obtained with random forest, Naïve Bayes and SVM using Python. When a user accesses the Disease Prediction Website, he or she is directed to the homepage. On the homepage, there are specific procedures for predicting one's diseases, as seen in figure 2.

Figure 2: Home Page with detailed steps on how to predict diseases.



Steps for Disease Prediction

1 Open Disease Prediction Website
It has all the information regarding the various algorithms used

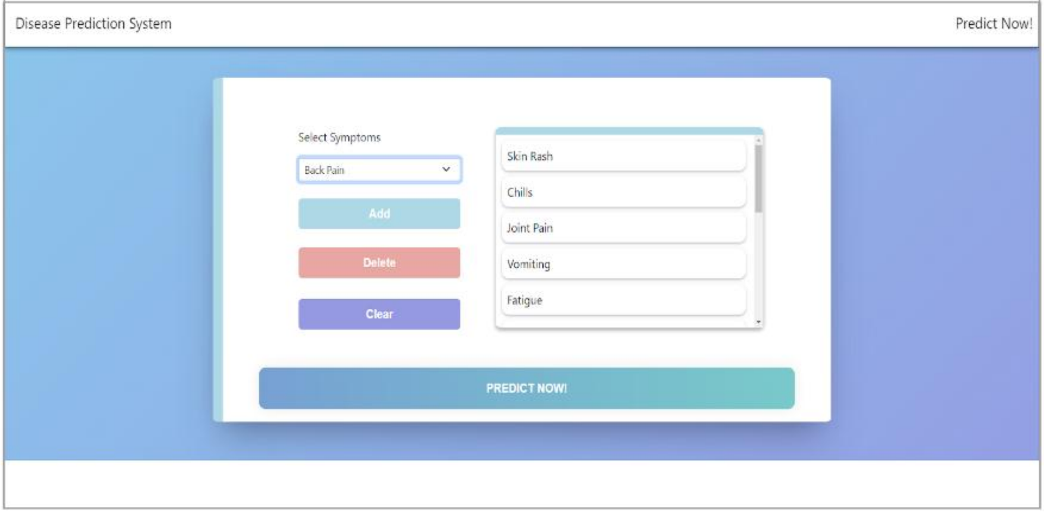
2 Click on Predict Now!
This button will redirect you to a new page where you can predict the diseases

3 Select the Symptoms
Select the appropriate symptoms from the drop down list.You can also delete the symptoms.Then finally click on predict disease

4 Diseases Predicted
Get your predicted diseases result by the different machine learning algorithms.

The application also has a good visual interface that handles all of the inputs required for prediction. The user will select symptoms from the drop down menu and add them by clicking the add button; if the user wishes to remove a specific symptom, he or she can do so by clicking the delete button and by clicking the clear button, all symptoms are removed, as illustrated in figure 3.

Figure 3: Simple and Precise GUI to add symptoms

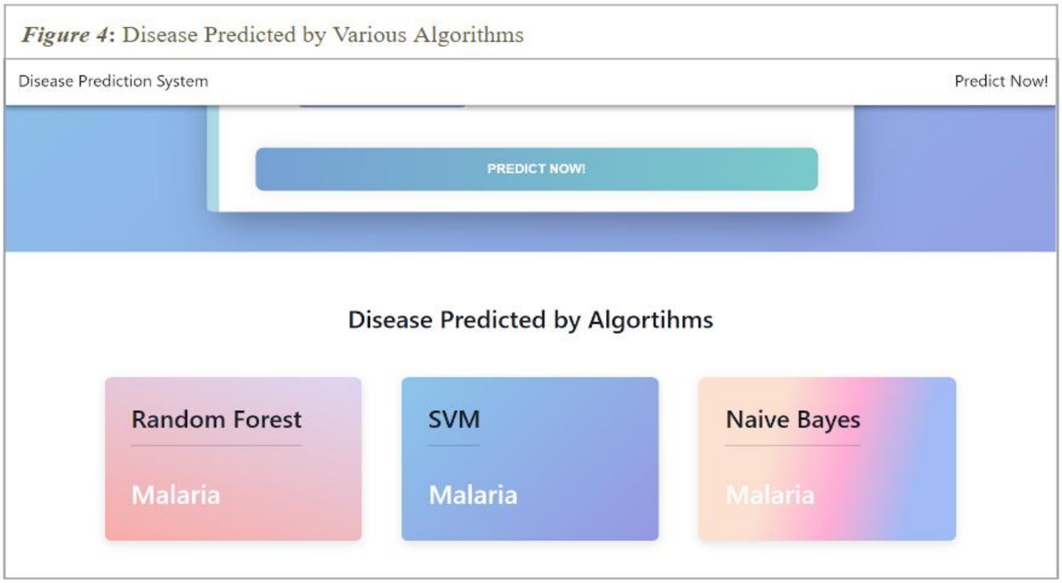


Select Symptoms: Back Pain

Skin Rash
Chills
Joint Pain
Vomiting
Fatigue

By clicking on predict now button, the user can find all of the probable diseases predicted by the various algorithms, as seen in figure 4.

Figure 4: Disease Predicted by Various Algorithms



Disease Predicted by Algortihms

Random Forest

Malaria

SVM

Malaria

Naive Bayes

Malaria

Conclusion

In this paper, we used three machine learning algorithms to predict and achieve a desirable result for the user, as well as making the system more efficient than the existing one and thus providing a better user experience than other available systems. The present study focused only on the structured dataset of symptoms. Most of the disease prediction system needs multimodal data as input for correct diagnosis of the disease. There are no standard ways for dealing with semi-structured and unstructured data.

References

1. “Disease Prediction using Machine Learning” Raj H. Chauhan, Daksh N. Naik, Rinal A. Halpati, Sagarkumar J. Patel, Mr. A.D.Prajapati, International Research Journal of Engineering and Technology (IRJET). Volume: 07 Issue: 05 | May 2020.
2. “Disease prediction by machine learning over big data from healthcare communities”, M. Chen, Y. Hao, K. Hwang, L. Wang, *IEEE Access*, vol. 5, no. 1, pp. 8869-8879, 2017. DOI: 10.1109/ACCESS.2017.2694446 https://ieeexplore.ieee.org/abstract/document/7912315
3. “Disease Classification Using Machine Learning Algorithms - A Comparative Study”, S. Leoni Sharmila, C. Dharuman and P. Venkatesan, , *International Journal of Pure and Applied Mathematics*, vol. 114, no. 6, pp. 1-10, 2017.
4. “Liver Disease Prediction using SVM and Naive Bayes Algorithms” - Dr. S. Vijayarani, Mr.S.Dayananda, International Journal of Science, Engineering and Technology Research (IJSETR), 2015.Volume 4, Issue 4, April 2015.

5. Disease and symptomsDataset, Kaggle Dataset Link: https://www.kaggle.com/itachi9604/disease-symptom-description-dataset?select=dataset.csv