[[1]](#footnote-1)

**Disease Prediction Using Machine Learning**

Mandar Kalse, Sushil Kandhare, Rohit Jadhav, Ganesh Karode, Omkar Karpe, Abha Marathe

Department of Multidisciplinary Engineering

Vishwakarma Institute of Technology (Pune)

***Abstract*- The progress and also the use of a few outstanding data mining methods in many fields around the world such as Industrial sector, healthcare sector and many more has paved off the way to the utilization of such and many other techniques in machine learning atmospheres, to gain necessary data fragments of specific data in health sectors, medical sectors etc. Precise or exact analysis of the proof of a medical website in the diagnosis of presently infection or disease, people services and the concerned patient’s care. Machine learning algorithms, they have successfully been tested in a number of programs including Disease Predictability. The aim is to improve the classification method using techniques in machine learning to greatly contributing to solving issues related to health by helping professionals from medical sector to predict the disease as early as possible. The data which was considered had records of around 4900 patients which were facing around 40 diseases. Around 90 of the 130 symptoms, which were independent variable relating to disease were inspected and considered. In this paper it shows a disease predictor technique which was done using machine learning techniques for e.g., the SVM classification method, the Random Forest classifier, and the KNN class. The research paper presents a study about the comparison of the techniques which we have used.**

**Keywords: Data mining, SVM classification method, Random Forest classifier, KNN classifier.**

1. ***INTRODUCTION***

Emergence of Artificial Intelligence (AI) has enabled computer systems to detect, think, and function intelligently as human beings. AI is a complex concept of ML, Computer Vision, In-Depth Learning, and Natural Language Processing. ML algorithms use a variety of development techniques, statistics, and opportunities to learn from the data generated from previous experience and apply them in decision making. efficiency, credit card fraud, and disease conversion. Many of these applications are built using a supervised learning method. In this way, data sets with known labels are compensated for predictive models to predict non-labeled models This suggests the idea that medical practitioners can use supervised learning as a powerful tool for diagnosing diseases more effectively [5].

Sectors such as healthcare and related sectors are in dire requirement of data processing presently. If methods of data mining are applied correctly, important knowledge can be obtained from a website and can be helpful to the doctor to make a decision before the time and enhance health related services [8].

Certain Diseases and problems relating to health such as cholera, anaemia, brain tumor, cancer, kidney failure, dengue, diabetes, migraine, jaundice, hepatitis, aids etc., lead to serious health effects and can sometimes prove to be fatal. Sectors such as health-care and biomedical can make decisions by “digging” their major website, that is, by removing some pattern and inter-relationships from website. Data mining techniques such as SVM, RF and KNN will provide a solution to the problem. Therefore, we have made a system which can detect and take disease-related information from the historical site (disease-symptoms) [6].

1. ***Literature Review***

There are many project activities undertaken in this article and many of them are for kidney and disease prediction of heart using supervised ML methods, some of which are listed below.

"Use of Disease Database Data Distribution" Information Mining Technology The use of Disease Predictability System is well known and emerges successfully in this area. Disease related to heart are most common diseases in public with more death rates in today's world. Dividing the heart patient data into appropriate subcategories is key to reducing mortality rates. Information Mining Technology The use of the Disease Predictability System is well known and emerges successfully in this area [1]. “Disease Predicting Using Machine Learning Algorithms” The prediction system of diseases was created with ML procedures namely, Naïve Bayes category, random forest classifier and decision tree separator. Advantage of Naïve Bayes is that they work faster even on large databases as they require less integration capacity. A sample data of records of approximately 4920 patients diagnosed with up to 41 sickness was selected for study. Then 95 of the 132 separate variables was related to disease chosen or adjusted. Once the system has been trained through a set of training using the above-mentioned algorithms, use the K-cross (with value of k is 5) method for testing execution of above-mentioned algorithms in database [2]. "Chronic Kidney Disease Prediction and Proper Nutrition Program Recommendation through Mechanical Learning" The data set has 25 key features, a valid mathematical method such as extracting the most effective reductions in the CKD diagnosis. The proposed system incorporates 4 key modules, namely preliminary data processing, feature extraction, descriptive sites based on blood potassium levels, diet module [3]. “Research and Analysis of Diseases Predictability in Healthcare Data Mining” Data mining techniques are useful for data analysis in a number of different areas and for identifying relationships. The effectiveness of different methods of dividing a broad database. Similar algorithms include Naïve Bayes, K's closest neighbor, Artificial Neural Network, tree of decision [4]. "Data Mining use for Methods in Predicting Disease related to heart: Exploration" Analyze various excavation techniques used in recent years to diagnose heart disease using NN, Naïve Bayes, genetic algorithm, and decision tree. [5]. "Classifications Procedures for Disease Prediction related to Kidney with data mining" kidney diseases prediction with various classification methods and obtaining an effective classification algorithm Using Naïve Bayes and SVM Algorithm [6]. Marimuthu et al. [7] aimed at predicting heart disease using ML surveillance techniques. The authors classified data attributes such chest aching, sex and objective, etc. [7]. The Machine Learning algorithms used are decision tree, Naïve Bayes, KNN and logistic regression. In an analysis, the logistic regression algorithm delivered best (86.89%) accuracy, which is better than any other algorithms stated. Dahiwade et al. The proposed solution to this was supplemented by additional information on the patient's living habits of the tested patient, which proved to be useful in understanding the level of risk associated with the predicted disease [8] comparing the results between KNN and CNN algorithm in terms of time analysis and accuracy. CNN processing time and accuracy were 84.5% and 11.1 seconds, respectively. Statistics proved that the KNN algorithm was less efficient compared to the CNN algorithm [8]. "Predicting Heart Disease Using Machine Learning Algorithms" (2018). In this paper, a two-way data mining algorithm was used in the database to predict the likelihood of a patient’s heart disease, analyzed by the Naïve Bayes Classifier classification model and the Decision Tree segment. Decision tree model predicts heart disease patient rate of 91% and Naïve Bayes class predicts heart disease patient rate of 87% [9]. Prediction Algorithm threat of disease with formal and informal info from the clinic with the help of new proposed system of CNN. In clinical data study no work is done in both sector of this type only separate are existed. 94.8% is the accuracy of this system at the junction, comparatively is better than CNN disease prediction algorithm [10]. Compared to a few standard measurement algorithms, the precise calculation of our proposed algorithm is 94.8% better than that of CNN disease predictional algorithm [11].

# ***Methodology***

**Datasets:** The Dataset is Downloaded from Kaggle in a csv (comma separated values) format. This dataset has 133 total columns, 132 of them being symptoms experienced by patient and last column in prediction for that.

The dataset is imported into the main file as panda Dataframe using pandas and other dependency such as:

Pandas: To format and structure data so it can be passed into the ML model for training.

Joblib: To save the trained model.

Sklearn: To import useful functions and ML models.

The data is separated into features and label and passed into the models imported from sklearn library.

The following classifiers are used to train:

* KNN (K- nearest Neighbors) classifier
* Random Forest classifier
* SVM (Support vector machine) classifier

Graphical user interface, application, table, Excel

Description automatically generated

Fig.1 Tested data

Graphical user interface, application, table, Excel

Description automatically generated

Fig.2 Trained data

We have used Support Vector Classifier, KNN, and Random Forest Classifier to verify the opposite. Before moving on to the introductory section let's get acquainted with the machine learning models. Also, for GUI we will use Tkinter Interface.

**KNN**: K-Nearest Neighbor or K-NN is an algorithm for non-line monitoring categories. K-NN is a non-parameter algorithm that does not make any assumptions about its basic data or its distribution. It is one of the simplest and most widely used algorithm based on its k value (Neighbors) and finds its use in many industries such as the financial industry, healthcare industry etc.

**Support Vector Classifier:**The Support Vector Classifier is a discriminatory category i.e., when given the training data labelled, the algorithm attempts to find the correct hyperplane that accurately separates samples into different categories in hyperspace.

**Random Forest Classifier:**Random Forest is a compact algorithm learning machine integrated together that uses internally multiple decision trees to create a class. In the random section of the forest, all the internal decision trees are weak students, the results of these weak decision trees are summarized i.e., the mode of all predictions is like the last prediction.

**Tkinter Interface**: Tkinter is a Python binding to the Tk GUI tool kit. It is the standard Python interface in the Tk GUI tool kit and is the standard Python GUI. Python offers many options for improving GUI (Graphical User Interface). Of all the GUI methods, tkinter is the most widely used method. It is a standard Python interface for the Tk GUI toolkit delivered via Python. Python with scanner provides a quick and easy way to create GUI applications.

# ***Results and Discussions***

After trainig the model we found each algorithms used, KNN (K-nearest neighbour), SVM (Support Vector machine) and Random forest in this case perform exceptionally well on the training data provided. The model we got from training this was saved in a folder with “.joblib” extension, so it can be easily reused without training again. We then tested this trained models to see how it reacts to new data and we found that it performs extremely well there aswell. The confusion matrix of algorithms can be seen in following figures.

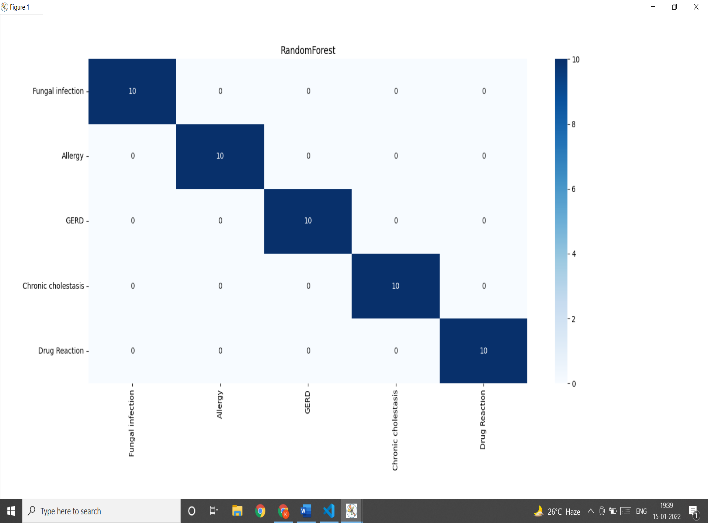


Fig 3: confusion matrix Random forest

For simplicity we have taken only first five disease and algorithm we have chosen is Random Forest. As we out first ten occurrences of each disease namely ‘fungal infection, Allergy, Gerd, Chronic Cholestasis, Drug Reaction” our model was able to predict 10 out of 10 of this disease’s occurrences.

Now that we confirmed that our model performs well, we created a gui using python. For checking purposes, we added symptoms of dengu in gui, and all our models were able to successfully predict dengue and also the predicted disease was also shown on gui for user to see.

Actions that our system can perform:

a. Entering Symptoms

b. Disease Prediction

Entering Symptoms: Once user entered into the system then he/she has to select the symptoms from the given drop-down menu.

Disease prediction: The predictive model predicts a person may have an infection, based on the user's included symptoms. which shown in Fig.4

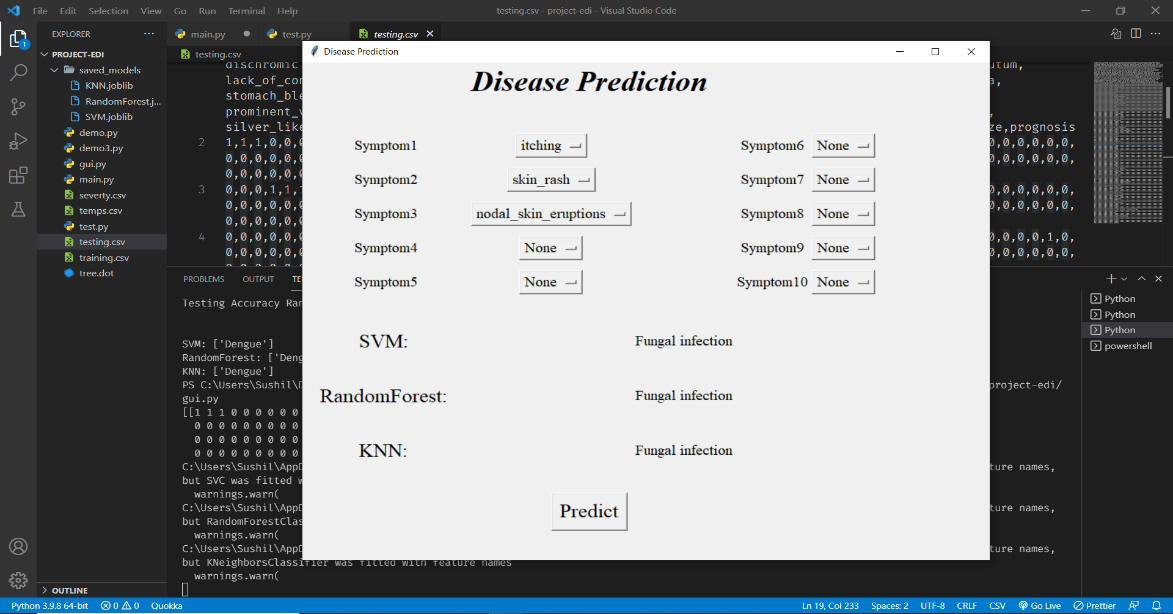


Fig.4 GUI page of system

The results obtained from our model are summarized in following Fig.5:

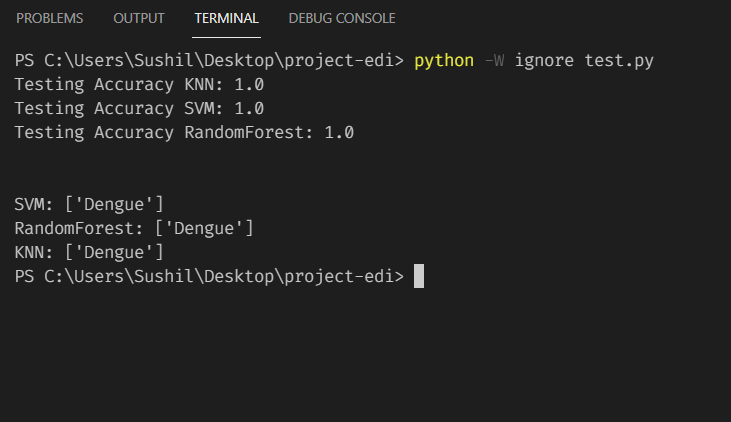


Fig.5 predicted output of our system.

It can be noted that in the pre-processing process, discretization improved the performance of all three algorithms.

# ***Conclusion***

Machine Learning algorithms of different types has led to the rapid detection diseases like breast cancer, brain disease, heart disease, kidney disease. In all literature, SVM, RF and KNN algorithms were widely used in prediction, while accuracy was the most widely used performance metrics. The KNN model has proven to be very effective in predicting common diseases. In addition, the SVM model has shown a high degree of accuracy in most cases in kidney disease due to its reliability in handling high-volume, single and unstructured data. In the prediction of breast cancer, RF has shown an increase in the likelihood of a correct diagnosis of disease due to its ability to measure well on large databases and its risk of overcrowding. In the end, the whole algorithm proved to be very reliable in predicting heart disease. In future work, the development of complex ML algorithms is much needed in order to increase the effectiveness of predictors. In addition, learning models should be evaluated several times after the training phase for best performance. In addition, data sets should be expanded to different demographics to avoid overcrowding and to increase the accuracy of the models used. Finally, the most appropriate trait selection methods should be used to make the system more effective and efficient.

1. ***References***

[1] Divya Jain, Vijendra Singh, "Use of Data Specification of Predictive Mining Data" 2016.

[2] Sneha Grampurohit, Chetan Sagarnal, "Disease Predicting Using Machine Learning Algorithms" 2020.

[3] Maurya, A., Wable, R., Shinde, R., John, S., Jadhav, R., & Dakshayani, “Chronic Kidney Disease Prediction and Commendation of Suitable Diet Plan through Machine Learning” 2019.

[4] R. Tamilarasi, Dr. R. Porkodi, "Research and Analysis of Disease Predictability Methods in Data Mining for Healthcare" 2015.

[5] Ritika Chadha, Shubhankar Mayank, Anurag Vardhan and Tribikram Pradhan, "Use of Data Mining Methods on Heart Disease Predictability: Examination" 2016.

[6] S. Vijayarani, S. Dhayanand, “Data Mining

Classification Algorithms for Kidney Disease Prediction ”2015.

[7] M. Marimuthu, M. Abinaya, K. S., K. Madhankumar, and V. Pavithra, “A Cardiovascular Predictability Review using Mechanical Learning and Data Mathematics,” International Journal of Computer Applications, vol. 181, no. 18, pages 20-25, 2018.

[8] D. Dahiwade, G. Patle, and E. Meshram, “Designing a Predictive Demonstration Model using a Mechanical Learning Program,” Procedures of the Third International Conference on Computer and Communication Systems, ICCMC 2019, no. Iccmc, pages 1211–1215, 2019.

[9] “Predicting Heart Disease Using Data Mining Strategies” H. Benjamin Fredrick David and S. Antony Belcy (2018).

[10] Min Chen; Yixue Hao; Kai Hwang; Lu Wang; Lin Wang "Big Data for Health Care" (2019)

[11] S, Vinitha and S, Sweetlin and H, Vinusha and S, Sergeant, "Disease Predictability Using a Big Data Data Machine". (February 2018).

1. [↑](#footnote-ref-1)