EAI Endorsed Transactions
on Internet of Things
Research Article EAI.EU

Early-Stage Disease Prediction from Various Symptoms Using Machine Learning Models

Devansh Ajmera[1], Trilok Nath Pandey[2,*], Shrishti Singh[3], Sourasish Pal[4], Shrey Vyas[5], Chinmaya Kumar Nayak[6]

[1,2,3,4,5]School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, Tamil Nadu
[6]Faculty of Emerging Technologies, Sri Sri University, Odisha

Abstract

Development and exploration of several Data analytics techniques in various real-time applications (e.g., Industry, Healthcare Neuroscience) in various domains have led to exploitation of it to extract paramount features from datasets. Following the introduction of new computer technology, the health sector had a significant transformation that compelled it to produce more medical data, which gave rise to a number of new disciplines of study. Quite a few initiatives are made to deal with the medical data and how its usage can be helpful to humans. This inspired academics and other institutions to use techniques like data analytics, its types, machine learning and different algorithms, to extract practical information and aid in decision-making. The healthcare data can be used to develop a health prediction system that can improve a person's health. Based on the dataset provided, making accurate predictions in early disease prediction benefits the human community.

Keywords: Data analytics, healthcare, disease, prediction, machine learning

Received on 16 December 2023, accepted on 03 March 2024, published on 11 March 2024

doi: 10.4108/eetiot.5361

*Corresponding author. Email: triloknath.pandey@vit.ac.in

1. Introduction

For any health-related issue, its root cause, diagnosis in the early stage and then the correct preventive measure to cure it is necessary and it's possible only if the detection is done at an early stage. This project aims at predicting diseases based on symptoms of the patient using machine learning algorithms. Machine Learning algorithms use a variety of statistical, probabilistic, and improvisation techniques to

learn from the past experiences from what it learns and then can be used for decision making. The supervised learning method has been used in the design of several of these applications. This method involves forcing prediction models to predict unlabelled examples from datasets with known labels. This puts forth the theory that doctors can diagnose diseases more effectively by using supervised learning as a potent tool.

The first and foremost step in this project was dataset collection followed by data pre-processing. The selected dataset consists of disease description, symptoms, severity (on a scale of 1-10) and precautions to be taken Data pre-processing involved steps like handling missing values, data cleaning and encoding the data for model building. The missing values are dealt using an imputation method. In this project, the two supervised learning algorithms used are Decision Tree Classifier and KNN (K-nearest neighbour). A decision tree is a category of tree structure. It makes use of nodes where the internal nodes tell whether a given condition based on an attribute should be taken or not. It represents a test result. At the end, all the target labels are present from the decision based on the nodes. The KNN algorithm assumes that the new case and the existing cases are comparable, and it places the new instance in the category that is most like the existing categories. The accuracy of the models has been

calculated to compare the best suited algorithm for the prediction.

2. Related works

In recent times, work on disease prediction, how machine learning algorithms are modified, and better results are obtained have been done. This section reviews some literature. In this project, for prediction of heart Disease, M. Akhil Jabbar [1] have used Genetic Algorithm in combination with KNN. From the UCI repository different datasets are used. Genetic Algorithm has 4 main properties namely Selection, cross over, Mutation and Fitness function. They have proposed that initially using genetic algorithms rank the entire dataset using the properties mentioned above. Based on those properties, train the KNN classifier and get the accuracy result.

D. Dahiwade, G. Patle and E. Meshram [2] have done a comparative study between KNN and CNN (Convolution neural Network). Using the Patient dataset from the UCL repository, hamming distance, Euclidean distance and distance metric is utilized. Among the class labels, majority vote is considered using K-nearest neighbors and the weighted factor. With this utilization the accuracy is less compared to CNN implementation. Another drawback with [2] is that in memory utilization comparison and time comparison it exceeds when compared with CNN. S. Ambesange [3] uses the KNN model in a different manner giving a hyper parameter tuning twist to it. They have used feature selection considering the features which show a strong correlation. Along with outlier checking, parameters like solver, random, state, penalty, mix_iter were hyper tunes.

These techniques in combination with KNN yielded a good metric value on the Liver dataset generated from Andhra Pradesh, India.

Similarly, D. Rahmat [4] used a SelectKbest feature selection technique in association with 10-Fold cross validation technique. They used KNN imputation which helps in filling the blank values (Nan) if any present in the dataset. They have overcome the problem by applying cross validation techniques. Data feature selection pre-processing technique helps in reducing the attributes to be computed and fitted in the dataset. All these factors have helped them get a good accuracy and other scores.

Coming to Decision trees, P. Deepika [5] have proposed a PSO optimized decision tree algorithm for disease prediction systems. Taking advantage of the PSO model which determines complex mathematical patterns. The J48 (ID3) decision tree is also applied in this paper. The 2 models are compared in terms of accuracy and speed in which each other outperform in one category each. Naim Rochmawati et al [6] used J48 and a Hoeffding tree. Hoeffding is a type of progressive decision tree which learns from a large dataset. An important assumption is that the stream of data is constant over time. They have also used a 10-cross validation technique used in [4]. The accuracy results from both the types of trees are the same but Hoeffding takes a smaller number of nodes for classification compared to J48 resulting in a faster computation process.

As the value of cross validation technique increases, opposed changes are noticed in performance and speed. A Chi-Squared Automatic Interaction Detection (CHAID) decision tree model is implemented by [7] for Erythemato-squamous diseases (ESD) prediction generating the dataset from UCI. The main advantage is that the tree not only contains binary branching but other types of branching. Ensemble learning techniques like Bagging and Boosting are also implemented. They have divided the entire implementation in Binary step, merge step, split step and stopping step. Varied results are obtained from the above technique implemented. The design is constructed in such a way starting with the fitting the data in the classifier prepared followed by using ensemble techniques.

M. Pak and M. Shin [8] investigated factors for the cause of Type-2 Diabetes from the environment. KNIH dataset was utilized, and the values were pre-processed since majority were categorical and SVM classifiers used for prediction purposes. [9] used SVM, Adaptive Boosting, Random Forest and Decision trees and Logistic Regression for predicting the most common diseases like breast cancer, heart disease and diabetes. UCL dataset from Machine learning library was used. [10] KNN in combination with a backward neural network for clustering of information was implemented along with SVM. Accuracy, recall and other such parameters were used to evaluate their model. [11] have seen missing data in the medical sector as a hindrance in analysis and have overcome using imputation techniques and cleaning the dataset. Implemented a Naive Bayes classifier along with a KNN model and then extended their work by implementing a CNN-UDRP which shows excellent results. Also, the analysis of the classified data is necessary. For instance, [12] the government can demand a list of malnourished children in a specific area. Data must first be categorized according to a certain place, children's health reports must be triggered, children whose families fall below the poverty line must be identified, and this data must then be processed. Similarly, this data can be very useful to the organizations where they need to classify the disease based on various symptoms. Predictive [13], [16] modeling might have bad effects if it is not applied properly. For instance, carefully adhering to the recommendations for predictive modeling may lead to a reduction in the focus on patients as distinctive individuals. Implementing a predictive modeling system inefficiently can even cost money. Technology is insufficient to accomplish these goals. It can be inferred from [14] that adjustments merely at the technological level is not sufficient but at various levels including administration, management etc. There must also be a positive impact. Using the COVID-19 dataset, this research assesses the performance of classifier methods [15], [21]. The performance of any algorithm is dependent on its type and the quantity of its attributes, the

study's findings show. The Random Forest Classifier Algorithm yields greater accuracy for the COVID19 dataset than the other three classifier algorithms. In, [17] machine learning algorithms were used for diabetes prediction. Support Vector Machine (SVM) and Artificial Neural Network (ANN) was used in this research to predict whether it is diabetes positive or negative. A fuzzy logic was used for the prediction of diagnosis of diabetes to predict whether the diagnosis of diabetes is positive or negative. Decision tree model was used for prediction of heart disease in diabetic patients in [18]. The decision tree model was more fine-tuned so that it works at its best performance in prediction of heart disease in diabetic patients. [19] have used a variety of Machine learning algorithms like SVM, Logistic regression, KNN etc. It was observed that metric values generally depend on the importance of a significant feature and SVM (Support Vector Machine) and LR (Logistic regression) have performed better. The paper [20] describes an effective machine learning strategy for chronic disease prediction.
Their method focuses on feature selection using adaptive probabilistic divergence, with the goal of identifying the most relevant features in the dataset. The method improves the accuracy of illness prediction models by picking informative information. The study emphasizes the importance of early disease identification and exhibits promising results in properly predicting chronic diseases at an early stage using the feature selection approach they proposed.

Ghosh et al.'s 2023 study [21] focuses on "Water Quality Assessment Through Predictive Machine Learning", highlighting the use of machine learning for analyzing and predicting water quality parameters. In "Unraveling the Heterogeneity of Lower-Grade [22] Gliomas," Rahat, Ghosh, and colleagues (2023) delve into deep learning-assisted segmentation and genomic analysis of brain MR images, offering new insights into this medical condition. Potato Leaf Disease [23] Recognition and Prediction using Convolutional Neural Networks," by Ghosh, Rahat, and team (2023), showcases the application of convolutional neural networks in accurately identifying diseases in potato leaves. Mandava, Vinta, Ghosh, and Rahat's [24]2023 research presents "An All-Inclusive Machine Learning and Deep Learning Method for Forecasting Cardiovascular Disease in Bangladeshi Population", integrating advanced AI techniques for health predictions. The 2023 study by Mandava et al., titled "Identification and Categorization of Yellow [25] Rust Infection in Wheat through Deep Learning Techniques", applies deep learning methods to detect and categorize wheat infections effectively. Khasim, Rahat, Ghosh, and colleagues' 2023 article, "Using Deep [26] Learning and Machine Learning: Real-Time Discernment and Diagnostics of Rice-Leaf Diseases in Bangladesh", explores AI-based solutions for diagnosing rice-leaf diseases. Deciphering Microorganisms through Intelligent Image Recognition", authored by Khasim, Ghosh, Rahat, and others in 2023, discusses [27] the use of machine learning and deep learning in identifying microorganisms through advanced image recognition techniques. The 2023 study by Mohanty, Ghosh, Rahat [28] and Reddy, "Advanced Deep Learning Models for Corn Leaf Disease Classification", focuses on the application of deep learning in classifying diseases in corn leaves based on a field study. Alenezi and team's 2021 research, "Block-Greedy and CNN Based Underwater Image Dehazing [29] for Novel Depth Estimation and Optimal Ambient Light", investigates novel CNN-based methods for enhancing underwater image clarity and depth estimation.

3. Methodology

The entire project is divided into different modules where each module has its own functionality and has applied a different algorithm suitable for that module.

3.1 Identify the problem Statement

As mentioned in the introduction part the need to diagnose a disease promptly and accurately is immensely important.

3.2 Generating the dataset

The dataset is generated from Kaggle. The entire data that is used comes in a set of 4 csv files. The data from these csv files were used and these were then interpreted and then the prediction model was applied in order to draw a conclusion from them according to the symptoms given to the model. The first csv file consists of various diseases along with their relevant symptoms. This csv file lists out the symptoms of various diseases. Various symptoms were listed out for any specific disease. This file was used for making the model train on which disease gives rise to which symptom or in other words which symptoms are for what disease. The second file consists of the severity of the symptom. This file gives a specific weight to each symptom and how each symptom can be weighted when predicting the disease given all its symptoms. This tells us the priority of the symptom and which symptoms should be considered or given more consideration during the prediction of a disease. This tells us for what symptom what disease should be given more priority. The next file gives us the details or the descriptions of the disease. Various descriptions of the diseases and what actually is the meaning of the particular disease is given in this file. It helps us in knowing better about the disease. The next file lists all the precautions that must be followed so as to avoid being infected from that disease.

3.3 Data Pre-Processing

There are various pre-processing methods. For our problem statement purpose 2 techniques are used.

3. 3. 1 Method 1

In the main dataset CSV file, the corresponding disease for a given set of symptoms is provided. However, some of the symptom columns in the dataset are blank. To address this, a

technique is employed where the symptom names are encoded using the Symptom Severity file. In this technique, if a symptom is not mentioned in the severity file, a lower severity value is assigned.

3. 3. 2 Method 2

In this method, after viewing the dataset and checking the percentage of Null values, only certain column attributes are taken, and the rest are deleted from the dataset. Though the dataset shrinks to some extent, that data majorly is redundant. So not considering those values and processing helps the model learn better the dataset. This helps in conducting a better analysis.

3.4 Model Building & Data Visualization

A KNN & a Decision Tree is implemented in python and R studio. Based on the architecture of Decision Tree and KNN, the data is pre-processed accordingly and fitted in the model. Also from the pre-processed data, some plots and graphs are generated, which helps us to understand the data well. From the model trained, metric score values will be generated which is discussed in the later section.

3.5 Prediction and Results

After model building, the model is tested to obtain accuracy and other results. If the results obtained are not satisfactory, then changes in the model will be made and again will be tested. Once the desired output from the model is achieved, the model can be used for prediction.

3.6 Machine Learning Algorithm Description and Approach

Machine Learning is a technique in computers where applications or programs are developed in order to train the machine or make the machine learn in a proper way. Then this learning can be used to predict the future outcomes. This works based on assumptions that whatever it is predicting, the same thing it will predict in the future. This also means, it is also assumed that whatever way it is
working now, it will continue to work in a similar fashion in the future. The machine is trained by following a data driven approach. For this purpose, a dataset is required based on which the machine is trained by creating a model. Then this model is used to predict the future outcomes of the project.

There are two parts in creation of a model in Machine Learning:

3.6.1 Training Part

This is the most important part in terms of Machine Learning. Training is the term which is used to define that the model is built using a dataset. These algorithms help us in finding trends or patterns in the past data. The algorithms built are supplied with an appropriate dataset which can then be used to classify which values contain the best values within it. For this purpose of training and testing, the dataset is first split into two parts. Then the part which will be used for training purposes is supplied to the appropriate model which then learns a pattern type from it. Then this model is ready to make predictions based on the pattern it has learnt from the training dataset. Basically, in this part the model is created.

3.6.2 Testing Part

With the rest of the data points in the dataset, the model is tested to see how well it is performing in real life conditions based on the data it has been trained. In the model testing part several types of performance metrics are calculated based on which we can say that the model is accurate by how much percentage. The dataset in this part of model creation must be such that most of the data points in it are not known to the algorithm or it is not predefined in the dataset. Basically, in this part of model creation the model is checked and evaluated, if it can work by how much percentage.

3.6.3 Algorithm Description

Decision Tree

Decision tree is a powerful machine learning algorithm in which a tree is built based on the dataset and based on which attribute takes the most probability. Each internal node in a decision tree determines the attribute test and each leaf of the decision tree determines the result of performing the test on the particular attribute. It classifies the attributes based on some predefined classification instance and then tests the attribute for moving down.

In the dataset, after all the cleaning process, we did a splitting on the dataset into two halves, one is the training dataset, and another one is the training dataset. Using the first half that is the training dataset, we trained the model by giving proper attributes with their appropriate and corresponding class labels. Then we have encoded the values in each attribute to convert them from string to integer. For encoding the values and by using the technique called Label Encoding, the appropriate libraries were imported. Then the tree was trained with the suitable part of the dataset. After successful training of the dataset, the model was tested for calculating its metrics.

$$\text{Entropy } E(f) = \sum_{i=1}^{n} (\text{-pi} * log_2(pi)) \quad (1)$$

$$\text{Gini}(f) = 1 - \sum_{i=1}^{n} (pi^2) \quad (2)$$

where pi is the probability. The architecture for which can be seen in Fig. 1.

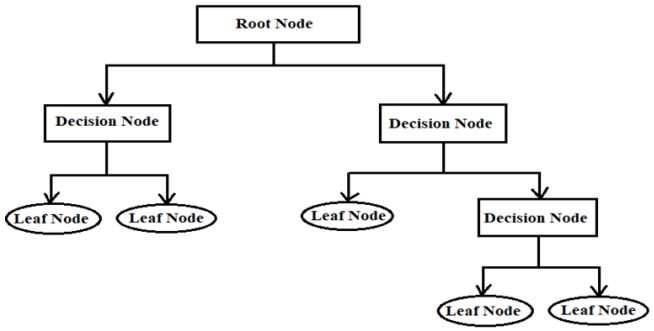

Fig. 1. Decision Tree Architecture

KNN

KNN is a supervised algorithm. The algorithm is itself facile in nature and has been used for multiple applications like classification, regression etc. The output in this algorithm is given when the input data is fed which is labeled. The learning process in this algorithm happens by using a function which makes use of the input data The methodology which KNN follows is nonparametric based which most of the times gives proper accurate results. Let us consider t records to be classified. This is done by creating k close points. The data record t is generally classified using a majority method from the local points which were collected from above step. It can either take weight-based distance into consideration. The value of k which is selected in these cases is critical since it has a great influence on the classification which will be done. A proper k value is must and selected carefully.
The mechanism deployed is displayed in Fig. 2. In this algorithm the data records are generally required in another form like vectors or numbers which represent mathematical values. Since it is in mathematical form, different distance-based functions can be used to get the segregation between the records. The distance between the data points
and the test data is computed and then the probability of the data points being closer to the test data is determined. The Manhattan, Makowski, or Euclidean distances are some possible distance functions.

The equations for these distances are as follows:

$$\text{Euclidean distance} = \sqrt{\sum_{i=1}^{k} (xi - yi)^2} \quad (3)$$

$$\text{Manhattan distance} = \sum_{i=1}^{k} |xi - yi| \quad (4)$$

$$\text{Minkowski distance} = \left(\sum_{i=1}^{k} (|xi - yi|)^q\right)^{1/q} \quad (5)$$

where x and y are the coordinates, k is the number of points and q is an integer having value between 1 and 2.

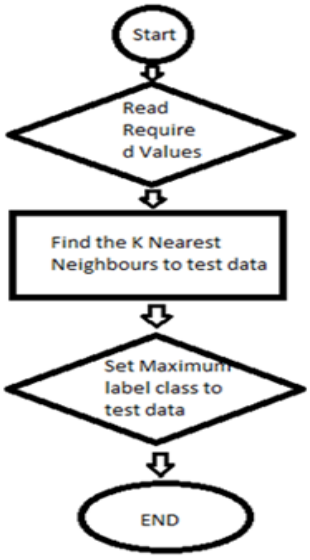

Fig. 2. KNN Architecture

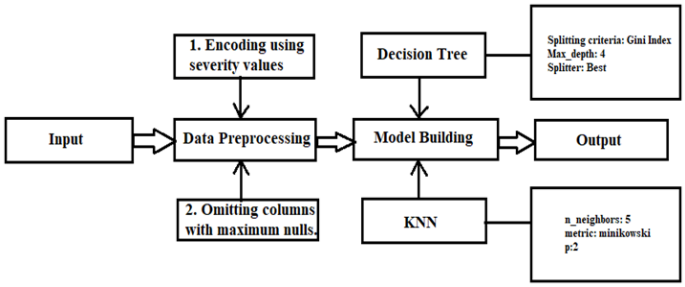

Fig. 3. Proposed System

4. About the Dataset

This section talks about the dataset which we have used for our analysis. There are four CSV files in the dataset: dataset. csv, symptom_Description.csv, symptom_precaution.csv, and symptom_severity.csv. The disease's names and associated symptoms are included in the dataset.csv file. The collection contains information about 41 distinct diseases. The dataset has about 4.9K rows and 17 symptoms for each condition, with the first 4 symptoms columns being virtually completely filled. The rest of the columns are nearly empty. The diseases and a brief explanation of each are included in the description file. It provides a basic overview of the diseases. The diseases and their precautions are listed in the precaution file. Each disease has 4 precautions included in the file. The symptoms and their corresponding severity are listed in the severity file. The weight of each symptom depicts how severe that symptom could be pertaining to that disease. In the last CSV files, the Symptom Description, gives us a description of the given symptom.

5. Experimental Analysis

In the world, nowadays there is a huge amount of data everywhere. So, it is important to better understand and to predict the upcoming features that can be extracted from the data more accurately and in an efficient manner. Data Analytics is the technique of analysing the data in order to draw future predictions and conclusions about the past events from it. There are various techniques of data analytics and nowadays various kinds of software are also available which may help in analysing the data. There are also various steps of the data analytics process. For doing proper data analytics a data needs to be collected and must be first understood what all the attributes it has. Then it must be cleaned and organized in a proper manner, so it makes a proper analysis on it. Cleaning is a very important part of the data analytics process as it helps in proper structuring of the data. For cleaning purposes, the missing values in the dataset must first be handled properly and then they must be filled or removed based on their requirement.

The pre-defined dataset is imported, and this dataset will be used for training the model. When the dataset is imported, then only the symptoms which are empty are replaced or imputed with NA values. Then the data needs to be cleaned to get rid of the NA values in the dataset. Percentage of NA values in each column are calculated in order to see how much data in each column are empty. As, the attributes which are almost empty are mostly of no use, so these columns are dropped. We are dropping the columns which have a percentage of NA values almost equal to 90% and more. Therefore, in the dataset we chose to drop off the attributes from Symptom_13 to Symptom_17. This is done purely based on prediction or purely based on the assumption that these columns might not affect the training of the model that was made as these columns have a greater number of missing values. The number of unique symptoms and several unique diseases are also calculated to get an idea of the amount of data which will be worked upon. Then the most occurring symptom and the least occurring symptom in the entire dataset is found out. From the dataset, we see that Fatigue is the most occurring symptom in the dataset. So, tiredness or fatigue happens in most diseases. All diseases weaken the body, so people feel some kind of tiredness in case of any symptom.

Symptom percentage per disease is used to check which unique symptoms in the dataset occur most frequently among all the diseases present there.

Data Visualization which can be treated as a major part of data analytics are carried out to better visualize the data. Data Visualization comes handy in situations where there is a large amount of data and in order to better understand it, we need to look at it, the maxima, or the minima in it. Data Visualization provides a very handy tool for the developers or the data analysts to get a better understanding of the data all at once. It also provides a better way of knowing about the dataset, the outliers in it, the past and the current trends, and the patterns in it. Various advantages and disadvantages are also there in plotting the data or its visualization. In the dataset, data visualization techniques were applied in order to get a clear understanding of it. By using bar plot, it is seen that headache and swollen legs are the most common or the most frequent symptoms occurring in the dataset. So, we can say using the bar plot, that headache is the most common symptom in all diseases across the dataset. All the diseases that people had, weight loss and swollen legs were also common as their count was maximum in case of symptoms. Also, the severity of the symptoms whether one symptom is more severe or less severe are plotted. After plotting, it is seen from the dataset that fluid overload, chest pain and coma have the most level of severity among all the symptoms in the dataset. Chest pain is the most severe symptom as it may be the cause of an underlying heart attack which may harm a person silently.

6. Performance Metrics

In order to evaluate the performance and effectiveness of the machine learning algorithms deployed, the following four performance metrics were used to evaluate the performance of the two algorithms used.

1. Accuracy: Accuracy of a model gives the total correct classification made by the model. It is the ratio of the total number of correct classifications to the total classifications made.

Accuracy: (TP+TN)/(TP+TN+FP+FN) (6)

2. Precision: Precision is used to measure the proportion of correctly classified positive instances. It is the ratio of the correctly classified positive instances to the sum of total positive classified instances.

Precision: TP/(TP+FP) (7)

3. Recall: Recall also known as sensitivity is the ratio of the total correctly classified positive instances to the total number of actual positive instances

Recall: TP/(TP+FN) (8)

4. F1 Score: F1 combines precision and recall of a model and is calculated as the harmonic mean of the two.

F1Score: TP/(TP+(1/2)*(FN+FP)) (9)

where,

TP is True Positive.

FP is False Positive.

TN is True Negative.

FN is False Negative.

Observation and Results

This section mentions the observation and results obtained. The project above shows that when doing data analytics and machine learning, we see that data analysis is easily done by getting the frequency of symptoms from which we can see the most and least occurring symptoms and plot it on the graph. By running the machine learning code, we get the predictive disease as well as the accuracy score. From the predictive output, we can import the disease description and precautions from the other datasets.

The result can be inferred that on inputting disease's unique symptoms, the disease can be accurately predicted while inputting symptoms that are general and occurring in multiple diseases, the accuracy of disease prediction drops down. Thus, the dataset needs more disease-unique symptoms to increase the accuracy score of the ml code.

Most symptoms related to the disease are only important because they increase the predictive accuracy score, most occurring symptoms occur for most of the diseases and reduce the predictive accuracy score.

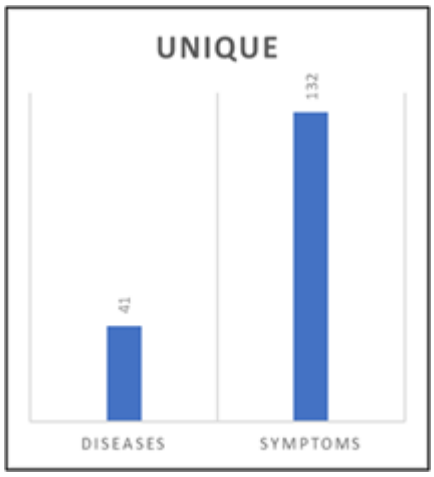

Fig. 4. Unique Symptoms and Diseases

From the dataset taken in this project, we observe that the data consists of 41 unique diseases while having 132 unique symptoms as shown in Fig 2 and Fig 3 respectively. The most occurring symptom in the whole dataset is fatigue while the least occurring symptom is the foul smell of urine.

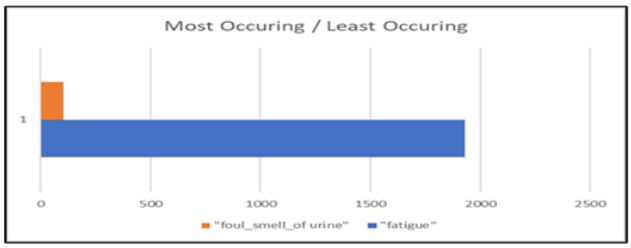

Fig. 5. Most and Least Occurring Symptoms

Table 1. Classification Report Decision Tree.

	Precision	Recall	F1-score	Support
accuracy			0.94	984
macro avg	0.96	0.95	0.93	984
weighted avg	0.96	0.94	0.93	984

From the machine learning output of the model, we can see that on inputting multiple symptoms which are occurring in multiple diseases, the predicted output may or may not be correct since misleading symptoms can lead to the wrong predicted output. The probability of correct prediction is given by the accuracy score of the machine learning algorithms. For certain diseases it's been displayed in Table 2.

Table 2. Report of metric values for certain diseases

	Vertigo	AIDS	Acne	Alcoholic Hepatitis	Allergy
precision	1.00	1.00	1.00	1.00	1.000000
recall	1.00	1.00	1.00	1.00	0.909091
f1-score	1.00	1.00	1.00	1.00	0.952381

support	17.00	13.0	21.0 0	16.00	22.00 000

Table 3. Classification Report KNN

	accuracy	macro-avg	weighted-avg
precision	0.995935	0.995935	0.99629
recall	0.995935	0.996807	0.995935
f1-score	0.995935	0.996218	0.995949
support	0.995935	738.0	738.00

It can be said in case of machine learning models that the prediction of each disease is different and based on classification certain analyses were made. Disease which had less symptoms had a good accuracy of being accurately predicted. Certain diseases had more unique symptoms when compared with others which helped the model in training them better. Certain diseases that valued symptoms and their uniqueness were also different which made it difficult for models to predict them with great accuracy. In case of switching the position or order of symptoms no change in prediction is noticed. In case of severity values, disease with a unique set of such values were predicted with great ease. Values which were repeated in case of severity could also be predicted with great metric values. The reason is the model has learned the dataset well. From the classification report in Table 1 and Table 3, the difference in results when pre-processing techniques were different can be seen.

7. Conclusion

This project helps us in determining the disease of a patient just by analysing the symptoms the patient has with some symptoms. It also helped us in analysing the data so as to predict what algorithms are giving more accurate results. It also helps in categorizing the symptoms and helps in analysing and modifying the dataset as per the needs of the model and algorithms to predict. In the future, we can add a user interface where the user can add all the symptoms manually or by voice recognition so that the model can predict the disease at ease just by telling the symptoms. The model can also be trained in such a way that it can also show the other related symptoms pertaining to the most favourable disease.

References

[1] M. Akhil jabbar, B.L. Deekshatulu, Priti Chandra, Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm, Procedia Technology, Volume 10, 2013, Pages 85-94, ISSN 2212-0173, https://doi.org/10.1016/j.protcy.2013.12.340.

[2] D. Dahiwade, G. Patle and E. Meshram, "Designing Disease Prediction Model Using Machine Learning Approach," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019, pp. 1211-1215, doi: 10.1109/ICCMC.2019.8819782.

[3] S. Ambesange, R. Nadagoudar, R. Uppin, V. Patil, S. Patil and S. Patil, "Liver Diseases Prediction using KNN with Hyper Parameter Tuning Techniques," 2020 IEEE Bangalore Humanitarian Technology Conference (B-HTC), 2020, pp. 1-6, doi: 10.1109/B-HTC50970.2020.9297949.

[4] D. Rahmat, A. A. Putra, Hamrin and A. W. Setiawan, "Heart Disease Prediction Using K-Nearest Neighbor," 2021 International Conference on Electrical Engineering and Informatics (ICEEI), 2021, pp. 1-6, doi: 10.1109/ICEEI52609.2021.9611110.

[5] P. Deepika and S. Sasikala, "Enhanced Model for Prediction and Classification of Cardiovascular Disease using Decision Tree with Particle Swarm Optimization," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 1068-1072, doi: 10.1109/ICECA49313.2020.9297398.

[6] N. Rochmawati et al., "Covid Symptom Severity Using Decision Tree," 2020 Third International Conference on Vocational Education and Electrical Engineering (ICVEE), 2020, pp. 1-5, doi: 10.1109/ICVEE50212.2020.9243246.

[7] A. M. Elsayad, M. Al-Dhaifallah and A. M. Nassef, "Analysis and Diagnosis of Erythemato-Squamous Diseases Using CHAID Decision Trees," 2018 15th International Multi-Conference on Systems, Signals & Devices (SSD), 2018, pp. 252-262, doi: 10.1109/SSD.2018.8570553.

[8] M. Pak and M. Shin, "Developing disease risk prediction model based on environmental factors," The 18th IEEE International Symposium on Consumer Electronics (ISCE 2014), 2014, pp. 1-2, doi: 10.1109/ISCE.2014.6884338.

[9] P. S. Kohli and S. Arora, "Application of Machine Learning in Disease Prediction," 2018 4th International Conference on Computing Communication and Automation (ICCCA), 2018, pp. 1-4, doi: 10.1109/CCAA.2018.8777449.

[10] M. Chakarverti, S. Yadav and R. Rajan, "Classification Technique for Heart Disease Prediction in Data Mining," 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), 2019, pp. 1578-1582, doi: 10.1109/ICICICT46008.2019.8993191.

[11] S. Ambekar and R. Phalnikar, "Disease Risk Prediction by Using Convolutional Neural Network," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018, pp. 1-5, doi: 10.1109/ICCUBEA.2018.8697423.

[12] J. Archenaa1 and E. A. Mary Anita 2A, "Survey Of Big Data Analytics in Healthcare and Government", 8 May,2015 https://www.sciencedirect.com/science/article/pii/S1877050915005220

[13] Wills, Mary J. Decisions Through Data: Analytics in Healthcare. Journal of Healthcare Management: July–August 2014 - Volume 59 - Issue 4 - p 254-262

[14] Bakot, K., Ślęzak, A. The use of Big Data Analytics in healthcare. *J Big Data* 9, 3 (2022). https://doi.org/10.1186/s40537-021-00553-4

[15] Advances in Mathematics: Scientific Journal 9 (2020), no.10, 8207–8215 ISSN: 1857-8365 (printed); 1857-8438 (electronic) https://doi.org/10.37418/amsj.9.10.50 Spec. Iss. on AOAOCEP-2020.

[16] T. N. Pandey, A. K. Jagadev, S. K. Mohapatra and S. Dehuri, "Credit risk analysis using machine learning classifiers," *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, Chennai, India, 2017, pp. 1850-1854, doi: 10.1109/ICECDS.2017.8389769.

[17] Ahmed, U., Issa, G. F., Khan, M. A., Aftab, S., Khan, M. F., Said, R. A., ... & Ahmad, M. (2022). Prediction of diabetes empowered with fused machine learning. *IEEE Access*, *10*, 8529-8538.

[18] Arumugam, K., Naved, M., Shinde, P. P., Leiva-Chauca, O., Huaman-Osorio, A., & Gonzales-Yanac, T. (2023). Multiple disease prediction using Machine learning algorithms. *Materials Today: Proceedings*, *80*, 3682-3685.

[19] Biswas, N., Ali, M. M., Rahaman, M. A., Islam, M., Mia, M. R., Azam, S., ... & Moni, M. A. (2023). Research Article Machine Learning-Based Model to Predict Heart Disease in Early Stage Employing Different Feature Selection Techniques.

[20] Ghosh, H., Tusher, M.A., Rahat, I.S., Khasim, S., Mohanty, S.N. (2023). Water Quality Assessment Through Predictive Machine Learning. In: Intelligent Computing and Networking. IC-ICN 2023. Lecture Notes in Networks and Systems, vol 699. Springer, Singapore. https://doi.org/10.1007/978-981-99-3177-4_6

[21] Rahat IS, Ghosh H, Shaik K, Khasim S, Rajaram G. Unraveling the Heterogeneity of Lower-Grade Gliomas: Deep Learning-Assisted Flair Segmentation and Genomic Analysis of Brain MR Images. EAI Endorsed Trans Perv Health Tech [Internet]. 2023 Sep. 29 [cited 2023 Oct. 2];9.https://doi.org/10.4108/eetpht.9.4016

[22] Ghosh H, Rahat IS, Shaik K, Khasim S, Yesubabu M. Potato Leaf Disease Recognition and Prediction using Convolutional Neural Networks. EAI Endorsed Scal Inf Syst [Internet]. 2023 Sep. 21 https://doi.org/10.4108/eetsis.3937

[23] Mandava, S. R. Vinta, H. Ghosh, and I. S. Rahat, "An All-Inclusive Machine Learning and Deep Learning Method for Forecasting Cardiovascular Disease in Bangladeshi Population", EAI Endorsed Trans Perv Health Tech, vol. 9, Oct. 2023.https://doi.org/10.4108/eetpht.9.4052

[24] Mandava, M.; Vinta, S. R.; Ghosh, H.; Rahat, I. S. Identification and Categorization of Yellow Rust Infection in Wheat through Deep Learning Techniques. EAI Endorsed Trans IoT 2023, 10. https://doi.org/10.4108/eetiot.4603

[25] Khasim, I. S. Rahat, H. Ghosh, K. Shaik, and S. K. Panda, "Using Deep Learning and Machine Learning: Real-Time Discernment and Diagnostics of Rice-Leaf Diseases in Bangladesh", EAI Endorsed Trans IoT, vol. 10, Dec. 2023 https://doi.org/10.4108/eetiot.4579

[26] Khasim, H. Ghosh, I. S. Rahat, K. Shaik, and M. Yesubabu, "Deciphering Microorganisms through Intelligent Image Recognition: Machine Learning and Deep Learning Approaches, Challenges, and Advancements", EAI Endorsed Trans IoT, vol. 10, Nov. 2023. https://doi.org/10.4108/eetiot.4484

[27] Mohanty, S.N.; Ghosh, H.; Rahat, I.S.; Reddy, C.V.R. Advanced Deep Learning Models for Corn Leaf Disease Classification: A Field Study in Bangladesh. Eng. Proc. 2023, 59, 69. https://doi.org/10.3390/engproc2023059069

[28] Alenezi, F.; Armghan, A.; Mohanty, S.N.; Jhaveri, R.H.; Tiwari, P. Block-Greedy and CNN Based Underwater Image Dehazing for Novel Depth Estimation and Optimal Ambient Light. Water 2021, 13, 3470. https://doi.org/10.3390/w13233470

[29] Hegde, S., & Mundada, M. R. (2021). Early prediction of chronic disease using an efficient machine learning algorithm through adaptive probabilistic divergence based feature selection approach. *International Journal of Pervasive Computing and Communications*, *17*(1), 20-36.

[30] Pandey, T.N., Mahakud, R.R., Patra, B., Giri, P.K., Dehuri, S. (2022). Performance of Machine Learning Techniques Before and After COVID-19 on Indian Foreign Exchange Rate. In: Dehuri, S., Prasad Mishra, B.S., Mallick, P.K., Cho, SB. (eds) Biologically Inspired Techniques in Many Criteria Decision Making. Smart Innovation, Systems and Technologies, vol 271. Springer, Singapore. https://doi.org/10.1007/978-981-16-8739-6_41.