Disease Diagnosis System Using Machine Learning

ABSTRACT

The efficient use of data mining in virtual sectors such as e-commerce, and commerce has led to its use in other industries. The medical environment is still rich but weaker in technical analysis field. There is a lot of information that can occur within medical systems. Using powerful analytics tools to identify the hidden relationships with the current data trends. Disease is a term that provides a large number of conditions connected to the heath care. These medical conditions describe unexpected health conditions that directly control all the organs of the body. Medical data mining methods such as corporate management mines, classification, integration is used to analyze various types of common physical problems. Separation is an important problem in data mining. Many popular clips make decision trees to produce category models. Data classification is based on the ID3 decision tree algorithm that leads to accuracy, data are estimated to use entropy verification methods based on cross-sectional and segmentation and results are compared. The database used for machine learning is divided into 3 parts - training, testing, and finally validation. This approach uses a training set to train a model and define its appropriate parameters. A test set is required to test a professional model and its standard performance. It is estimated that 70% of people in India can catch common illnesses such as viruses, flu, coughs, colds etc. every two months. Because most people do not realize that common allergies can be symptoms of something very serious, 25% of people suddenly die from ignoring the first normal symptoms. Therefore, identifying or predicting the disease early using machine learning (ML) is very important to avoid any unwanted injuries.

Keywords: Disease prediction; naïve bayes; machine learning; data processing; data splitting, cross-folding.

1. INTRODUCTION

Machine learning is a study of specific algorithms and mathematical models to perform a given task without explicit use commands, relying on patterns and problem instead. Machine learning algorithms create a mathematical model based on given data, known as "training data", to make predictions without explicit speculation is scheduled to perform this function. The main objective of machine learning is to generate patterns in the given dataset and practice. Patterns based on patterns are often difficult to answer business questions, find one analyzes styles and helps solve problems. There are many machine learning apps on the market. The top categories are:

- Banking
- Financial Market Analysis
- Medical diagnosis
- Indigenous Language Processing
- Emotional Analysis
- Recommendation Programs
- Time Prediction etc.

Disease Prediction using Machine Learning is a system which predicts the symptoms based on the information provided to the system. It also predicts the disease of the patient or the user based on the information or the symptoms he/she enter into the system and provides the accurate results based on that information. If the patient is not much serious and the user just wants to know the type of disease, he/she has been through. It is a system which provides the user the tips and tricks to maintain the health system of the user and it provides a way to find out the disease using this prediction.

Now a day's health industry plays major role in curing the diseases of the patients so this is also some kind of help for the health industry to tell the user and also it is useful for the user in case he/she doesn't want to go to the hospital or any other clinics, so just by entering the symptoms and all other useful information the user can get to know the disease he/she is suffering from and the health industry can also get benefit from this system by just asking the symptoms from the user and entering in the system and in just few seconds they can tell the exact and up to some extent the accurate diseases. Now a day's doctors are adopting many recent methodologies and upgraded technologies for identification and diagnosing not only common disease, but also many other diseases other than common. The successful treatment is always credited by the accurate/correct diagnosis. Doctors may sometimes fail to take correct/accurate decisions while diagnosing a patient, therefore disease prediction systems using machine learning approach to support in such cases to get accurate diagnosis. The project disease prediction using machine learning is developed to overcome general disease in earlier stages as we all know in competitive environment of economic development the mankind has involved so much that he/she is not concerned about health according to research there are 40% peoples how ignores about general disease which leads to harmful disease later. The main reason of ignorance is laziness to consult a doctor. The peoples have involved themselves so much that they have no time to take an appointment and consult the doctor which later results into fatal disease. According to research there are 70% peoples in India suffers from general disease and 25%of peoples face death due to early ignorance the main motive to develop this project is that a user can sit at their convenient place and have a check-up of their health the UI is designed in such a simple way that everyone can easily operate on it and can have a check-up.

1.1 Advantages

- Diagnosis improves the efficiency of treatments and avoiding complications for the infected patient in long term.
- Undiagnosed patients will transmit the disease to society. Early diagnosis will stop an outbreak and helps to prevent it from spreading.
- Misuse of antibiotics contribute to antibiotic resistance. Diagnostic tests will determine when antibiotics are a correct treatment or not.
- Disease prediction system provides a web platform predicting / detecting the infected disease based on the number of symptoms exist in the body.
- The user can identify the various existing symptoms in the patients and can find the probabilistic figures of diseases.
- This will also be a feasible option for those

who have mild symptoms and do not feel the need to get immediate medical help.

1.2 Challenges

- Collection of datasets to train the model to meet the accurate prediction of the diseases.
- Validation of the results has to be done by the well qualified doctors in their respective fields.
- Prediction systems have to be developed to make it interactive for the users.

N. Kumar et.al. [1] presented the prediction system for various diseases based on clinical data collected therefore automated disease prediction is easier. Heart diseases prediction system using data mining algorithms is presented in [2-4]. U. Ojha et.al. [5] proposed breast cancer prediction system using data mining Technique. Some disease prediction models are proposed using machine learning algorithm [6-8], convolution NN [9], Support vector machine [10] and Detection of Hepatitis (A, B, C and E) Viruses using Naive bays and nearest neighbor classifier [11-12]. This paper is organized as follows. Section I gives an introduction on disease detection system using machine learning. Section II proposed problem definition. Section III presented the experimental results. Section IV presented the conclusion followed by future scope.

2. PROBLEM DEFINITION

One of the major problems facing both developed and underdeveloped countries is the difficulty of treating sick people. Despite the lack of medical technology in various hospitals, most of those countries use the richness of their resources to meet this challenge and yet they cannot meet the pressure to provide quality medical services to its people. It has become increasingly difficult to find a lasting solution to the issue of traditional medical diagnostics that is characterized by inaccuracy and ambiguity.

Many researchers have proposed many algorithms and technologies based on Genetic Algorithms, Fuzzy Logic and Artificial Neural Network to ensure the correctness and precision of the drug industry which are the part of Artificial Intelligence. It is estimated that 70% of people in India can catch common illnesses such as viruses, flu, coughs, colds etc. every two months. Because most people do not realize that common allergies can be symptoms of something very serious, 25% of people suddenly die from ignoring the first normal symptoms. This can be a unsafe circumstances for the people and sometimes it is very scary. Therefore, identifying and detecting / predicting the diseases early are very important to avoid any unwanted injuries. Existing programs are programs dedicated to a selected disease or are in the research phase of algorithms where they involve common diseases. Figure 1 shows the basic structure of the system which includes the data processing and data flow required to produce the result.

3. EXPERIMENTAL RESULTS

The experimentation carried out on:

i. **Front end:** Bootstrap, CSS, HTML JavaScript, Jquery
ii. **Back end:** Python framework (Django)
iii. **Database:** PostgreSQL
iv. **Tools:** PgMyadmin, Orange

3.1 Data Pre-Processing

Project based on machine learning, preprocessing of data is the initial or basic step. The next implementation step is complex in nature and involves collecting, selecting, preprocessing, and applying various transformation on data. This process can be divided into some phases for easier implementation.

3.2 Data Representation/Visulization

A more information/data represented in graphical form is easier to analyze and understand i.e. to create templates, slides, charts and diagrams.

3.3 Data Cleaning

Cleaning of data means removing unwanted data (Noises) and inconsistencies present in the available dataset i.e. redundancy/duplication in an available data. Using imputation techniques, data scientist can substituting the missing values with mean values i.e. fill in missing data.

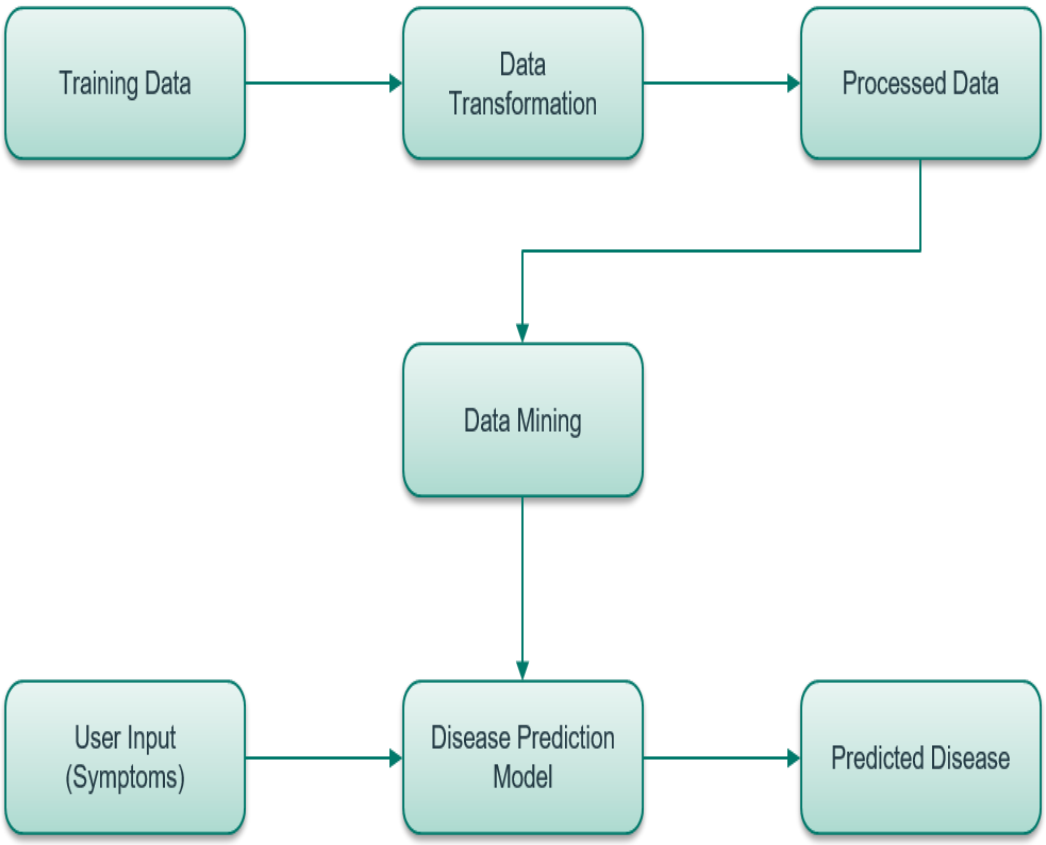


Fig. 1. System architecture

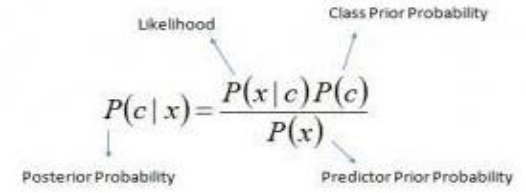
3.4 Data Gathering / Collection

Data collection has been done from the internet to identify the disease here the real symptoms of the disease is collected i.e., no dummy values are entered. Table I shows the dataset of disease and symptoms. Disease data based on symptoms are compiled from available health related internet media and kaggle.com website. This .csv file contains 5000 rows of record of the patients with their symptoms (132 types of different symptoms) and their corresponding disease (40 class of general disease). Table II shows the sample record of the patients and their symptoms.

3.5 Algorithm Implemented

In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Naive Bayes model is easy to understand and with some efforts it can be implemented successfully. This classifier is useful for very large database. Along with simplicity, Naive Bayes performs better as compared to other available classification methods.

Bayes theorem computes posterior probability P(c |x) from prior probability class / predictor i.e. P(c), P(x) and P(x |c). This equation is represented as:



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Where,

P(c|x) = posterior probability of class (c, target) given predictor (x, attributes).
P(c) = prior probability of class.
P(x|c) = likelihood which is the probability of predictor given class.
P(x) = prior probability of predictor

Table 1. Datasets: Disease and symptoms

Sr. No.	Disease	Symptoms
1	Malaria	Chills, Vomiting, high fever, sweating, …
2	Allergy	Continuous sneezing, shivering, chills, …
3	Arthritis	Muscle weakness, stiff neck, …
4	Typhoid	Chills, vomiting, fatigue, high fever, …
5	Gastroenteritis	Vomiting, sunken eyes, dehydration, …

Naïve Bayes Classifier algorithm is using in our proposed system. This classifier works on the principle of Bayes Theorem

$$P(Y|X1, \dots \dots \dots, Xn) = \frac{P(Y)P(X1, \dots \dots, Xn|Y)}{P(X1, \dots \dots Xn)}$$

Where,

Y is the class variable
X1, X2, …,Xn ---are the dependent features

From above equation we can rewrite as:

$$P(Disease|symptom_1, symptom_2, \dots \dots ., symptom_n) = \frac{P(Disease)P(symptom_1, \dots \dots .., symptom_n|Disease)}{P(symptom_1, symptom_2, \dots \dots \dots \dots .. Symptom_n)}$$

Using the naive independence assumption:

P (symptom-1,.symptom-N|Disease) = P (Symptom-I | Disease)

Where

I = 1, 2, , N

The value P (Symptom-I | Disease) of can be calculated by using multinomial Naïve Bayes which is given by:

$$P(symptomi\ |Disease) = \frac{N_{yi} + \alpha}{N_y + \alpha n}$$

Where,

Nyi is the same disease frequency
Ny is the total disease symptoms
n is the total number of symptoms
α is Laplace and always 1

Smoothing the value of P (Disease) can be calculated by using Laplace Law of Succession which is given by:

$$P(Disease) = \frac{N(Disease) + 1}{N + 2}$$

Where,

N (Disease) means same disease frequency
N represents total number of disease count

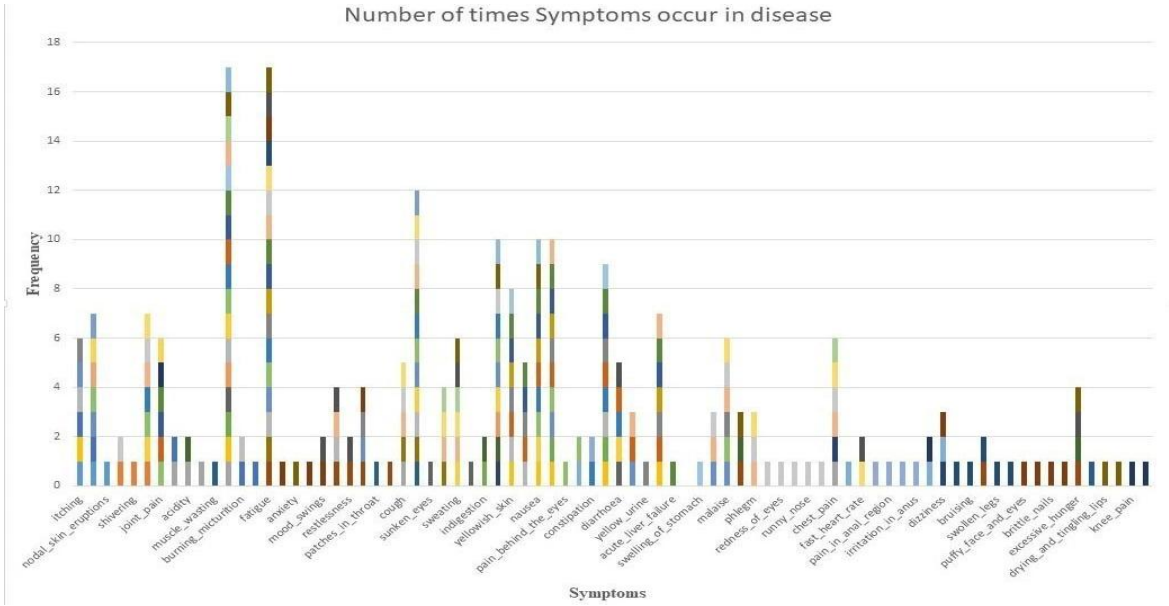


Fig. 2 (a). Occurrence of symptoms

Table 2(a). Testing dataset

palpitation	*nful_walk*	*filled_pim*	*blackhead*	*scurring*	*kin_peelin*	*r_like_du*	*_dents_in*	*mmatory_*	*blister*	*re_around*	*w_crust_*	*prognosis*
0	0	0	0	0	0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	0	0	0	0	0	GERD
0	0	0	0	0	0	0	0	0	0	0	0	Chronic cholestasis
0	0	0	0	0	0	0	0	0	0	0	0	Drug Reaction
0	0	0	0	0	0	0	0	0	0	0	0	Peptic ulcer disease
0	0	0	0	0	0	0	0	0	0	0	0	AIDS
0	0	0	0	0	0	0	0	0	0	0	0	Diabetes
0	0	0	0	0	0	0	0	0	0	0	0	Gastroenteritis
0	0	0	0	0	0	0	0	0	0	0	0	Bronchial Asthma
0	0	0	0	0	0	0	0	0	0	0	0	Hypertension
0	0	0	0	0	0	0	0	0	0	0	0	Migraine
0	0	0	0	0	0	0	0	0	0	0	0	Cervical spondylosis
0	0	0	0	0	0	0	0	0	0	0	0	Paralysis (brain hem
0	0	0	0	0	0	0	0	0	0	0	0	Jaundice
0	0	0	0	0	0	0	0	0	0	0	0	Malaria
0	0	0	0	0	0	0	0	0	0	0	0	Chicken pox
0	0	0	0	0	0	0	0	0	0	0	0	Dengue
0	0	0	0	0	0	0	0	0	0	0	0	Typhoid
0	0	0	0	0	0	0	0	0	0	0	0	H- A
0	0	0	0	0	0	0	0	0	1	0	0	H- B
0	0	0	1	0	0	0	0	0	0	0	0	H- C
0	0	0	0	0	1	0	0	0	0	0	0	H- D
0	0	0	0	0	0	0	0	0	0	0	1	H- E
1	0	0	0	0	0	0	0	0	0	0	0	Alcoholic hepatitis

Table 2(B). Testing dataset

itching	*skin rash*	*_skin_eru*	*nuous_sne*	*shivering*	*chills*	*joint_pain*	*omach_pa*	*acidity*	*rs_on_ton*	*scle_wast*	*vomiting*
1	1	1	0	0	0	0	0	0	0	0	0
0	0	0	1	1	1	0	0	0	0	0	0
0	0	0	0	0	0	0	1	1	1	0	1
1	0	0	0	0	0	0	0	0	0	0	1
1	1	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1
1	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	1	0	0	0	0	0	1
1	1	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	1	1	0	0	0	0	1
0	0	0	0	0	1	0	0	0	0	0	1
0	0	0	0	0	0	1	0	0	0	0	1
1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	1

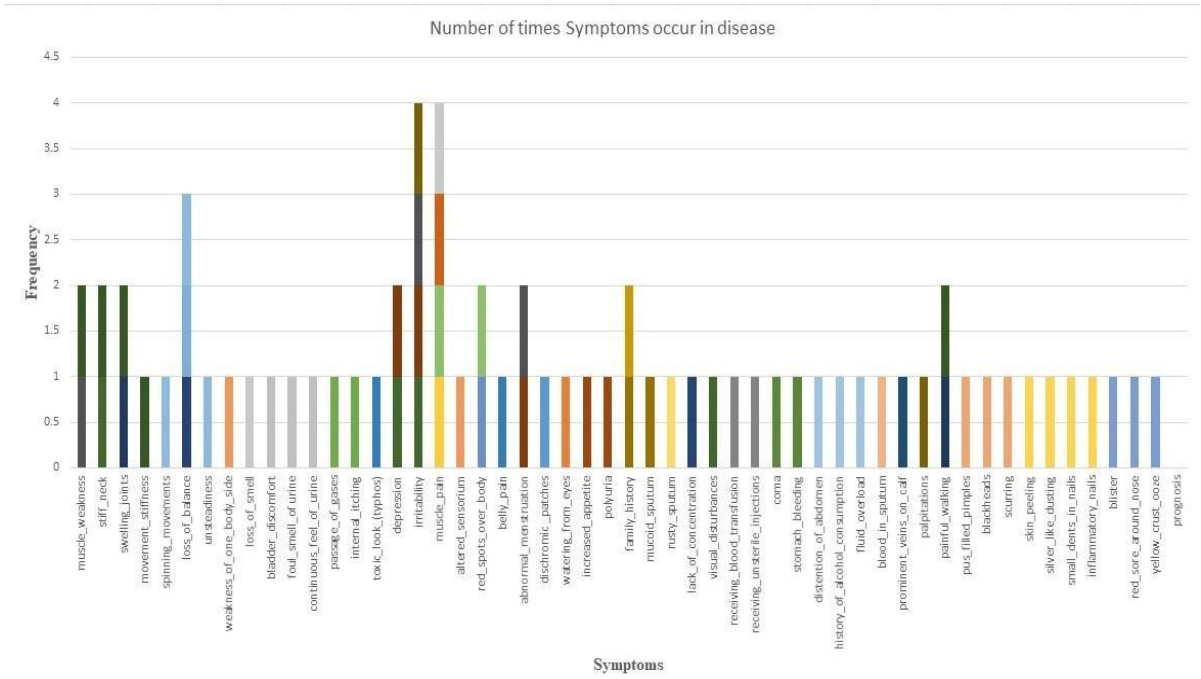


Fig. 2(b). Occurrence of symptoms

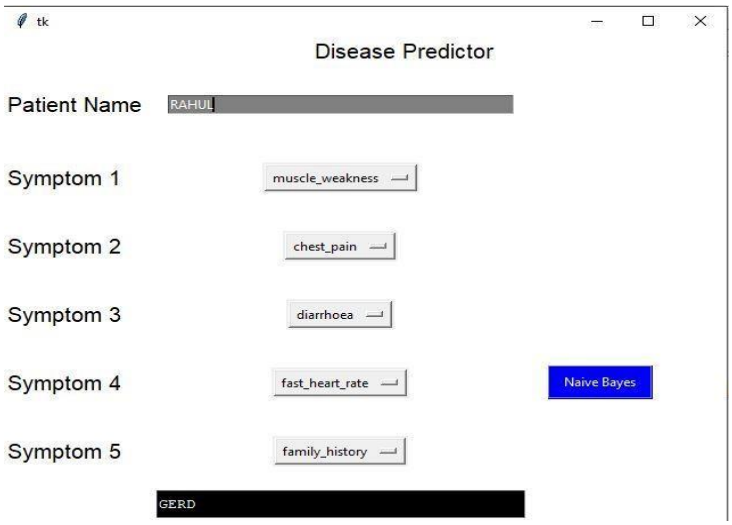


Fig. 3. System interface

The above two tables show some rows of Testing dataset wherein column represents name of Symptoms and the rows of datasets contains names of various diseases. The data is in binary format so “1” in row of disease represents the occurrence of its corresponding symptom in that disease and “0” represents non-occurrence.

Fig. 2 shows frequency of symptoms occurring in various diseases where X-axis of graph shows name of the various symptoms and Y-axis of the graph contains frequency of symptoms. Also, by analyzing the graph it can be seen that “patches in throat" and “sweating" these symptoms shows highest frequency which means in 17 different diseases these common symptoms are observed in the dataset.

The figure 3 shows the user interface of the system where patient/user has to enter his/her name and also has to provide at least 5 symptoms which will be taken as a input by the system and these input will be given to trained

model which will produce name of the predicted disease as a output.

4. CONCLUSION

In this paper we propose symptoms-based disease diagnosis system which diagnosis diseases based on Machine learning algorithm. Our findings from the studied literature, we acknowledged that the predictions done earlier did not use a large dataset. Manipulation over large data set, system will result in better prediction and better detection of the symptoms. Naive Bayes algorithm performs better on large data set and also outperforms in terms of prediction of the symptoms / disease . This developed system is very much useful in Health Care Industry and other Industries. The proposed system will be developed using different existing predicting algorithms for better prediction in the system. These includes:

- Increase algorithm accuracy
- To add more algorithms
- Improvising the algorithms to increase efficiency of the system and improve its working.
- To make it a complete healthcare diagnosis system package to be used in hospitals.

In future, this system can be used by people in remote areas who are not in reach of doctors currently and It can also be used in military operations, where the soldiers visit remote areas don't have access to a doctor when needed. Still, there is a scope of improvement in better predicting using various machine learning algorithms.

CONSENT

It is not applicable.

ETHICAL APPROVAL

It is not applicable.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Kumar N, Khatri S. Implementing WEKA for medical data classification and early disease prediction, 3rd International Conference on Computational Intelligence & Communication Technology (CICT), Ghaziabad, India. 2017;1-6.
DOI: 10.1109/CIACT.2017.7977277
2. Gandhi M, Singh SN. Predictions in heart disease using techniques of data mining, 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), Greater Noida, India. 2015;520-525.
DOI: 10.1109/ABLAZE.2015.7154917
3. Sultana M, Haider A, Uddin MS. Analysis of data mining techniques for heart disease prediction, 2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), Dhaka, Bangladesh. 2016;1-5.
DOI: 10.1109/CEEICT.2016.7873142
4. Sowmiya C, Sumitra P. Analytical study of heart disease diagnosis using classification techniques, 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), Srivilliputtur, India. 2017;1-5.
DOI: 10.1109/ITCOSP.2017.8303115
5. Ojha U, Goel S. A study on prediction of breast cancer recurrence using data mining techniques, 2017 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence, Noida, India. 2017;527-530.
DOI: 10.1109/CONFLUENCE.2017.7943207
6. Hendrik Anton Meijer. Systems and methods for rare disease prediction and treatment, Information and Software Technology. 2015;47(8):55.
7. Dahiwade D, Patle G, Meshram E. Designing Disease Prediction Model Using Machine Learning Approach, 2019 3rd International conference on Computing Methodologies and Communication (ICCMC), Erode, India. 2019;1211-1215.
DOI: 10.1109/ICCMC.2019.8819782
8. Kohli PS, Arora S. Application of Machine Learning in Disease Prediction, 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India. 2018;1-4.
DOI: 10.1109/CCAA.2018.8777449
9. Ambekar S, Phalnikar R. Disease risk prediction by using convolutional neural network, Fourth International Con ference on Computing Communication Control and

Automation (ICCUBEA), Pune, India. 2018; 1-5.
DOI: 10.1109/ICCUBEA.2018.8697423

10. Patil M, Lobo VB, Puranik P, Pawaskar A, Pai A, Mishra R. A proposed model for lifestyle disease prediction using support vector machine, 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Bengaluru, India. 2018;1-6.
DOI: 10.1109/ICCCNT.2018.8493897
11. Altayeva A, Zharas S, Cho YI. Medical decision-making diagnosis system integrating k-means and Naïve Bayes algorithms, 2016 16th International Conference on Control, Automation and Systems (ICCAS), Gyeongju, Korea (South). 2016;1087-1092.
DOI: 10.1109/ICCAS.2016.7832446
12. Ali Say. The impact of medication quality on patient rehospitalization rate. International Journal of Respiratory Care. 2019;15(1):12–14.
13. Ahmed N. Public health outcome framework application on obesity patients. International Journal of Respiratory Care. 2019; 15(1):15–18.
14. Trishna TI, Emon SU, Ema RR, Sajal GIH, Kundu S, Islam T. Detection of Hepatitis (A, B, C and E) Viruses Based on Random Forest, K-nearest and Naïve Bayes Classifier, 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India. 2019;1-7.
DOI:
10.1109/ICCCNT45670.2019.8944455