BAUSTJ
BAUST Journal

Potential Disease Detection Using Naive Bayes and Random Forest Approach

Dr. Mohammed Sowket Ali[1*], Junayed Mahmud[2], S. M. Fahim Shahriar[3], Shameha Rahmatullah[4] and Abu Saleh Musa Miah[5]

[*1,2,3,4,5]Department of CSE, Bangladesh Army University of Science and Technology, Saidpur, Bangladesh

emails: [*1]*sowket@baust.edu.bd; [2]junayed.mahmud27@gmail.com; [3]fahimshahriar3951@gmail.com; and [5]abusalehcse.ru@gmail.com*

ARTICLE INFO

Article History:

Received: 25 March 2022
Revised: 29 June 2022
Accepted: 01 July 2022
Published online: 31 July 2022

Keywords:

Disease Prediction
Machine Learning
Classification Algorithms
Naive Bayes
Random Forest

ABSTRACT

Due to the growing amount of knowledge and information in the medical and healthcare industries, data mining is a crucial and effective method for predicting disease from medical datasets. In this paper, we confirmed that it is possible to predict potential diseases in an individual's victimization using Machine Learning algorithms. Based on the patient's symptoms, our method successfully predicts deadly diseases including Diabetes, COVID-19, AIDS, and others. We applied machine learning methods like Naive Bayes, Decision Trees, Random Forests, SVM, and Logistic Regression to disease predictions. After analyzing all of these algorithms, we found that Naive Bayes and Random Forest gave us the more accurate disease prediction. The medical datasets were collected from the Kaggle website. From the datasets, 80 percent of data are used for training and 20 percent are used for testing data. In order to validate our model, we used English and Bangla languages-based user interface to collect the disease symptoms. Into our GUI user can give five symptoms, and through classifiers the final required results will be obtain. The study's main contribution was to achieve high calculation accuracy for early prediction of diseases.

1. INTRODUCTION

Inspire of the increasing potential for medical care worldwide due to the increasing burden of non-infection disease and aging population, medical care access is unacceptably decreased worldwide Tomohiro Morita et al. Fewer patient are going to receive medicine based on their disease, for cardiovascular disease, probability of patient taking secondary prevention at least one medicine was 54.9% for the countries of upper-middle-income, 19.8% for low-income countries, and 30.7% for the countries of middle-income. The present world countries, the ratio of doctor and population are not enough for medical service especially lower income and middle come countries. Bangladesh is a lower income country, the doctor-nurse and doctor-population ratio is 2.5:1 and 1:12690 respectively, which is worst statistic in the world Semigran HL et al. In addition, for such resource-limited in the field of health care of the countries, Researchers have been thinking a different way to support the limitation. In this manner, the apparatus to assist patients with self-diagnosis and self-triage is direly expected to alleviate the deficiency of medical care resources. This type of thought and research to self-diagnose and medical diagnose is not the new invention of the 4th industrial revolution but before the era of an invention of internet and smart mobile phone. Find out disease, treatment tips, and track down the connected medication analysts were utilized health encyclopedia in the home library stock were the exclusive source of reliable information. Free access of knowledge and Internet have changed everything, like Jeff Arnold founded one of the top healthcare websites with 75 million visitors namely WebMd. Since that time, many researchers have been trying to develop such kind medical websites to meet the needs of patients: The people are concerned about their health and

explore to find the answers in website via online and the patient who do not access to medical services. For 10 years, capacity of individuals to get information and information has definitely changed. For example, the first largest wiki dedicated to healthcare and medicine lunched in 2006 namely WikiDoc on the internet with an initial focus on cardiovascular disease by Marios Poulos et al.

Machine learning is a branch of artificial intelligence that utilizes an assortment of statistical, probabilistic, and optimization techniques procedures that permits PCs to "learn" from past models and to recognize hard-to-discern patterns from enormous, loud, or complex informational indexes. That's why machine learning can be used in disease detection based on the symptoms given by the user. In our work, we are willing to use machine learning to develop a symptom checker. To bring out the best and most accurate result, algorithms will be used, which will be decided later based on our studies.

There are several symptom checkers available on the web. Symptom checkers are acquiring notoriety throughout the world. In the United States alone, symptom checkers were used over 100 million times a year ago by individuals; they are basic and simple to use for everybody from adolescents to grown-ups of retirement age, with numerous accessible by means of cell phone applications [19]. Some of the most popular are WebMD, Mayo Clinic, iTriage, DocResponse, etc.

Machine learning is not new in the field of detecting diseases. It has been widely used in some specific areas. Artificial Neural Networks (ANNs) and Decision Trees (DTs) have been used in cancer detection and diagnosis for nearly 20 years describe by Md. Aminul Islam and Nusrat Jahan. Today machine learning strategies are being utilized in a wide scope of uses going from recognizing and characterizing tumours through X-ray and CRT images to the classification of malignancies from proteomic and genomic (microarray) examines by your MD a digital health tech company.

The Healthcare industry has become enormous business. The healthcare industry creates a lot of information day by day which can be utilized to take out data for predicting disease that can happen to a patient in the future by using the treatment history and health data. This hidden information will be subsequently utilized for instinctive dynamic for the patient's wellbeing. Additionally, this zone needs improvement by utilizing enlightening information in medical services. The significant test is the means by which to extricate the data from this information on the grounds that the sum is tremendous, so some information mining and AI methods can be utilized. Likewise, the normal result and extent of this venture is that assuming the infection can be anticipated, early treatment can be given to the patients who can diminish the danger of life and save the existence of patients.

This is where a symptom checker can come to great help. Symptom checker applications are innovation programs that utilization nitty-gritty calculations to permit clients to enter side effects and get indicative data. The potential for these applications to be used across the medical services climate is boundless, as an ever-increasing number of individuals look to the web to self-analyse their diseases to minimize their expenses and medical services availability high.

To design a prediction system for medical data classification, an early disease prediction by using Naïve Bayes and Random Forest algorithm. At first, we need the dataset and need to be implemented as per our project. GUI needs to be defined to make our project user friendly. The prediction task uses the main algorithms “Naïve Bayes” and “Random Forest” to show our desired result. The full code needs to be written with Python 3. It boosts the dataset classification and prediction. Python 3 makes a lightweight GUI and also supports Bangla UTF-8 characteristics for the Bengali language-based GUI. We use Jupyter Notebook and PyCharm as the main IDE to visual analysis. Scikit Learn model building, TKinter as GUI, and GitHub as visual control.

2. DATASET

Dataset for this project was collected from Kaggle (Neelima, 2019). The data are collected from hospitals which were structured and unstructured. It also partitioning into training and testing data sets using machine learning algorithm. This both training and testing data set are in CSV file system which containing diseases and symptoms. Which are utilized to prepare the model. In our program Read_csv() function is utilized to store the information in the data frame, named df. Utilizing replace () function, prognosis column that is the different diseases, it is replaced by the numbers from 0 to n-1, where n is the number of different diseases present in .csv record.

There are columns containing diseases, their symptoms, precautions to be taken, and their weights. This dataset can be easily cleaned by using file handling in any language. The user only needs to understand how rows and columns are arranged.

1	itching	skin_rash	nodal_ski	continuou	shivering	chills	joint_pain	stomach_	acidity	ulcers_on
2	1	1	1	0	0	0	0	0	0	0
3	0	1	1	0	0	0	0	0	0	0
4	1	0	1	0	0	0	0	0	0	0
5	1	1	0	0	0	0	0	0	0	0
6	1	1	1	0	0	0	0	0	0	0
7	0	1	1	0	0	0	0	0	0	0
8	1	0	1	0	0	0	0	0	0	0
9	1	1	0	0	0	0	0	0	0	0
10	1	1	1	0	0	0	0	0	0	0
11	1	1	1	0	0	0	0	0	0	0
12	0	0	0	1	1	1	0	0	0	0

Figure 2.1: Training Data Set (training.csv)

The type of dataset and problem is a classic supervised binary classification. Given the number of elements all with characteristics, we want to build a machine learning model to identify user behaviour. To solve the problem, we have to analyse the data, do any required transformation and normalization, apply a machine learning algorithm, train a model, check the performance of

pus_filled	blackhead	scurring	skin_peel	silver_like	small_der	inflammat	blister	red_sore_	yellow_cr	prognosis
0	0	0	0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	0	0	0	Allergy

Figure 2.2: Testing Data Set (Testing.csv)

the trained and iterate with other algorithms until we find the most performing for our type dataset. We visualized a table with the first row of the dataset to better understanding the data format. The visualization of data is a significant step in the data analysis. With this graphical visualization, we have a better understanding of the values of the features.

3. METHODOLOGY

Our application for disease prediction that take patients medical data and predicts whether a patient has a disease or not using a machine learning algorithm based on following architecture.

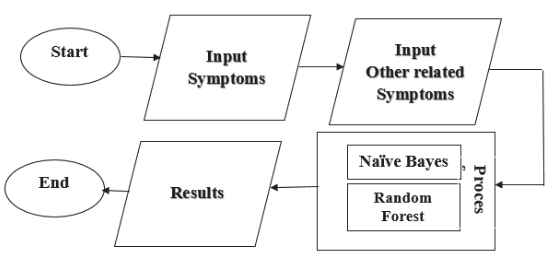

Figure 3.1: System Architecture

A. *Naïve Bayes*

The Naïve Bayes (NB) is a classification technique based on Bayes' theorem describe by Lindley DV. The probability of an event based on the prior knowledge of conditions associate with that event can be described by Naïve Bayes theorem. This classifier assumes a selected feature in a class that is not directly associate with any other feature, although features for that class could have interdependence among themselves,

$$P(c|x) = \frac{P(X|C)P(C)}{P(x)} \quad (1)$$

where,

P(c|x) = Posterior Probabilty,

P(x|c) = Likelihood,

P(c) = Class Prior Probability, and

P(x) = Predictor Prior Probability.

B. *Random Forest*

The Random Forest (RF) is an ensemble classifier and consisting of many DTs same as the way a forest is a collection of many trees describe by Breiman L., Random Forest and Mach Learn. Random Forest is based on the concept of collaborative learning, which is a process of combining multiple classifiers to solve a complex problem and to enhance the performance of the model.

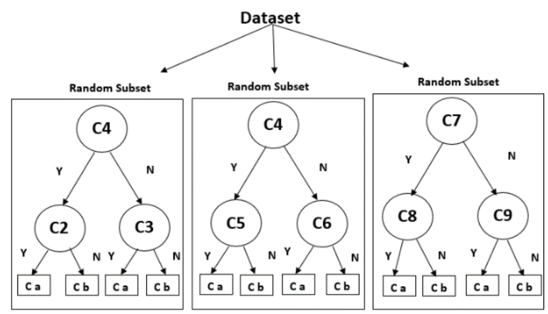

Figure 3.2: Random Forest

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given datasets and takes the average to improve the predictive accuracy of that datasets". Instead of relying on one decision tree, the random forest takes the prediction from each tree and is constructed on the majority votes of predictions and it predicts the ultimate output.

4. DESCRIPTION OF DIFFERENT MODULES

A. *Scikit Learn*

Scikit Learn is the most useful library for machine learning in Python. We use it in our project to build

machine learning models. The sklearn library contains a lot of effective tools for machine learning and statistical modelling, including classification, regression, clustering, and dimensionality reduction, Scikit-Learn: https://www.analyticsvidhya.com/blog/2015/01/scikit-learn-python-mac hine-learning-tool/.

B. Tkinter

Tkinter is the standard Python interface to the Tk GUI toolkit. Both Tk and Tkinter are available on most Unix platforms, as well as on Windows systems, https://docs.python.org/3/library/tkinter- .html.

5. PARAMETERS EXTRACTIO FROM THE SELECTED ALGORITHMS

We use the pre-train database (Training.csv), which is utilized to prepare the model. train_test_split() for splitting test model data and train model data. It helps to fit every model correctly. Here 80% of data is for training and 20% for testing. Hopefully, we got a good accuracy score. Therefore, every model can predict correctly.

As our dataset is human-made, so it is tough to find the accuracy of our algorithms'. So, we analyse our algorithms' by giving them the same type of symptoms and analyzing them by their predicting results. After comparing these algorithms, we conclude to the decision that Naïve Bayes and Random Forest gave us more perfect prediction than any other algorithms.

6. APPLICATION USER INTERFACE

After train the model, now we are ready to run the application. The interface is very user friendly for every people. Our target users are mostly Bangladeshi and also English-speaking foreigners. Keep in mind; we made two different user interfaces in our application. One is in English for foreigners, and the other is Bangla for Bangladeshi people.

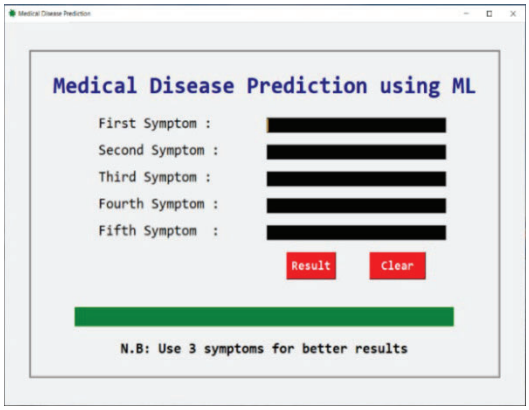

Figure 6.1: User Interface (English)

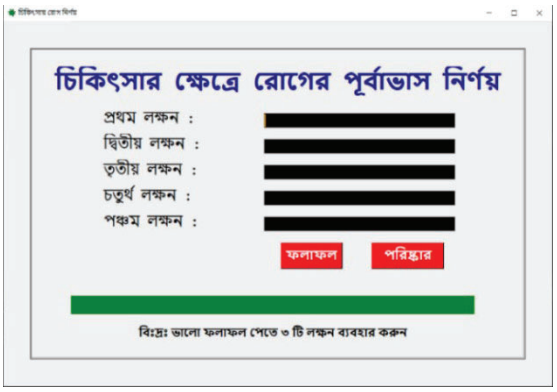

Figure 6.2: User Interface (Bangla)

7. RESULT ANALYSIS

We analysis five supervised learning algorithms in our project, which are Decision Tree, Naive Bayes, Random Forest, SVM and Logistic Regression. Our both English and Bangla Interface work properly. Here user can give input maximum five symptoms and through classifiers the final required results will be obtain.

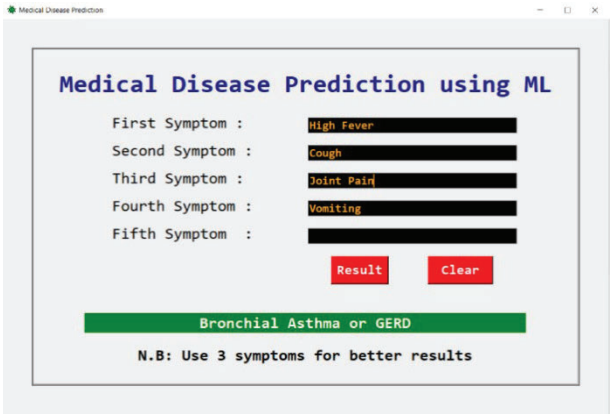

Figure 7.1: Figures of Predicted Result in English

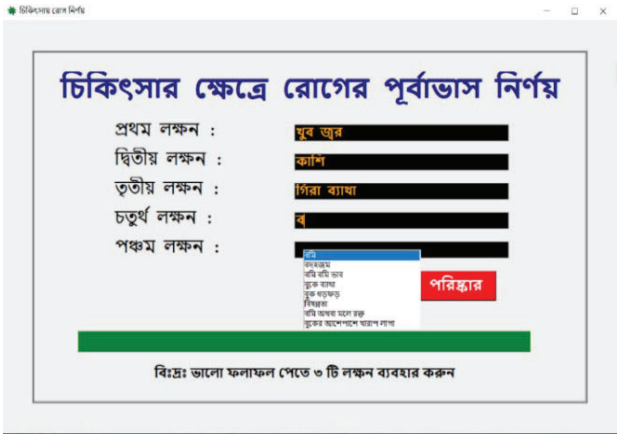

Figure 7.2: Figures of Predicted Result in Bangla

A. Prediction Analysis

We took five tests to analyze our prediction task and denoted them as:

Table 1 Prediction Analysis

Test 1	High Fever Joint Pain Breathlessness **Decision Tree:** ["Heart attack"] **NOT OK** High Fever Joint Pain Breathlessness **Naïve Bayes:** ["Bronchial Asthma"] **OK** High Fever Joint Pain Breathlessness **Random Forest:** ["Bronchial Asthma"] **OK** High Fever Joint Pain Breathlessness **Logistic Regression:** ["Bronchial Asthma"] **OK** High Fever Joint Pain Breathlessness **SVM**: ["AIDS"] **NOT OK**
Test 2	Stiff Neck Headache Pain Behind The Eyes **Decision Tree:** ['Arthritis'] **NOT OK** Stiff Neck Headache Pain Behind The Eyes **Naive Bayes:** ['Migraine'] **OK** Stiff Neck Headache Pain Behind The Eyes **Random Forest:** ['Dengue'] **OK** Stiff Neck Headache Pain Behind The Eyes **Logistic Regression:** ['Paralysis (brain hemorrhage)'] **NOT OK** Stiff Neck Headache Pain Behind The Eyes **SVM:** ['Paralysis (brain hemorrhage)'] **NOT OK**
Test 3	Excessive Hunger Blurred And Distorted Vision Irregular Sugar Level **Decision Tree:** ['Varicose veins'] **NOT OK** Excessive Hunger Blurred And Distorted Vision Irregular Sugar Level **Naive Bayes:** ['Migraine'] **OK** Excessive Hunger Blurred And Distorted Vision Irregular Sugar Level **Random Forest:** ['Diabetes '] **OK** Excessive Hunger Blurred And Distorted Vision Irregular Sugar Level **Logistic Regression:** ['Diabetes '] **OK** Excessive Hunger Blurred And Distorted Vision Irregular Sugar Level **SVM:** ['Urinary tract infection'] **NOT OK**
Test 4	Coma Stomach Bleeding Yellowish Skin **Decision Tree:** ['Hepatitis E'] **OK** Coma Stomach Bleeding Yellowish Skin **Naive Bayes:** ['Hepatitis C'] **OK** Coma Stomach Bleeding Yellowish Skin **Random Forest:** ['Hepatitis E'] **OK** Coma Stomach Bleeding Yellowish Skin **Logistic Regression:** ['Hepatitis E'] **OK** Coma Stomach Bleeding Yellowish Skin **SVM:** ['Urinary tract infection'] **NOT OK**
Test 5	Passage Of Gases Loss Of Appetite High Fever **Decision Tree:** ['AIDS'] **NOT OK** Passage Of Gases Loss Of Appetite High Fever **Naive Bayes:** ['Peptic ulcer'] **OK** Passage Of Gases Loss Of Appetite High Fever **Random Forest:** ['Peptic ulcer'] **OK** Passage Of Gases Loss Of Appetite High Fever **Logistic Regression:** ['Peptic ulcer'] **OK** Passage Of Gases Loss Of Appetite High Fever SVM: ['AIDS'] **NOT OK**

NOT OK = Wrong Prediction, OK= Desired Prediction.

Here we analysis our algorithms to choose which algorithms are suitable for our project. We basically analysis our algorithms' manually. Because, we use a pre-train dataset so it is tough to find the accuracy from that algorithm. So, we do manually our analysis. We use same symptoms in different algorithms and as a result we see that Naïve Bayes and Random Forest gives us the best results among all the algorithms. By analyzing our algorithms', we choose Naïve Bayes and Random Forest.

Table 2 Test Result Summary

NO.	Algorithms	Result	
		OK	NOT OK
1	Naive Bayes	5	0
3	Random Forest	5	0
4	Logistic Regression	4	1
5	SVM	0	5

In Table 2 – "Test Summary" we can see that Naïve Bayes and Random Forest gives us perfect results rather than others. Logistic Regression also gives us a good result. But, despite being good algorithm Decision tree and SVM can not achieve the good results. In table-2 we saw Naïve Bayes predicted 5 tests correctly,

Decision Tree show 1 correct and 4 wrong predictions. On the other hand, Random Forest did perform same as Naïve Bayes. Logistic Regression did 4 correct prediction and 1 wrong prediction. SVM didn't show any correct prediction. So, from the table-2 we can easily choose best algorithm for our project.

8. CONCLUSIONS

In our application, we use the **"pre-train model"** medical dataset, which comprises separate training and testing data sets, for disease prediction. Our application can help a doctor as an assistant and gives a primary overview of the disease condition and perhaps even save both the patient and the doctor time. In this paper, Naïve Bayes, Decision Tree, Random Forest, SVM, and Logistic Regression machine algorithms are used to predict a small set of relations between attributes in the databases to build a trustworthy classifier. The main contribution of the study to attain high calculation accuracy for early prediction of diseases with a regular speed. Our application can be used by the user of worldwide in International English language as well as regional Bangla language users. This approach could help advance medical research, patient care, and social services.

REFERENCES

[1] Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. BMJ. 2015;351:h3480. Doi: 10.1136/bmj.h3480.
[2] Tomohiro Morita, Abidur Rahman, Takanori Hasegawa, Akihiko Ozaki, Tetsuya Tanimoto, The Potential Possibility of Symptom Checker, International journal of health policy and management. http://ijhpm.comInt J Health Policy Manag 2017, 6(10), 615–616.
[3] Marios Poulos. Knowledge-based system for prognosis of specific types of cancer using Elman neural network. December 20, 2012. Doi: 10.5430/air.v2n2p62.
[4] Cruz JA, Wishart DS. Applications of Machine Learning in Cancer Prediction and Prognosis. Cancer Informatics. January 2006. doi:10.1177/117693510600200030.
[5] Tomohiro Morita1, Abidur Rahman, Takanori Hasegawa, Akihiko Ozaki, Tetsuya Tanimoto. The Potential Possibility of Symptom Checker. 2017. Doi:10.15171 /ijhpm.2017.41.
[6] Decision Tree: Dataaspirant." How Decision Tree Algorithm works". April 21, 2017. https://dataaspirant.com/how-decision-tree-algorithm-works/.
[7] Quinlan JR. Induction of decision trees. Mach Learn. 1986;1(1):81–106. Doi: 10.1007/bf00116251.
[8] Naïve Bayes: Lindley DV. "Fiducial distributions and Bayes' theorem". J Royal Stat Soc. Series B (Methodological). 1958; 1:102–7. Doi: 10.1111/j.2517-6161.1958.tb00 278.x.
[9] Random Forest:Breiman L. Random forests. Mach Learn. 2001; Doi:10.14511/jlm.2329.5430.
[10] Support Vector Machine: Joachims T. Making large-scale SVM learning practical. SFB 475: Komplexitätsreduktion Multivariaten Datenstrukturen, Univ. Dortmund, Dortmund, Tech. Rep. 1998.[Doi:10.1107/s0108768107031758/bs5044sup1.cif].
[11] Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied logistic regression. Wiley; 2013. https://books.google.com.bd/books?hl=en&lr=&id=64JYAwAAQBAJ&oi=fnd&pg=PR13&dq=%5B7%5D+Hosmer+Jr+DW,+Lemeshow+S,+Sturdivant+RX.+Applied+logistic+regression.+Wiley%3B+2013.&ots=DskSaW5qlP&sig=fMgVa0lxSBsZK-2CnA5UHckICfY&redir_esc=y#v=onepage &q&f=false.
[12] Database 2019: https://www.kaggle.com/neelima98/disease-prediction-using-machine-learning?fbclid=IwAR0rYD UCCcGWa8-YAXSkCAvOMJYdFNHjYOg0UjO3 7t7i2JNmm1BglbB9mOU&select=Training.csv.
[13] Pandas: https://pandas.pydata.org/.
[14] Scikit-Learn:https://www.analyticsvidhya.com/blog/2015/01/scikit-learn-python-machine-learning-tool/.
[15] Tkinter: https://docs.python.org/3/library/tkinter.html.
[16] Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. BMJ. 2015;351:h3480. Doi: 10.1136/bmj.h3480.
[17] Joseph A. Cruz, David S. Wishart. Applications of Machine Learning in Cancer Prediction and Prognosis. January 1, 2006. Doi: 10.1177/117693510600200030.
[18] Reports: "Symptom Checkers Will See You Now!" https://aboutdigitalhealth.com/2019/04/15/ai-will-see-you-now.
[19] Marios Poulos. Knowledge-based system for prognosis of specific types of cancer using Elman neural network. December 20, 2012. Doi: 10.5430/air.v2n2p62.
[20] Md. Aminul Islam, Nusrat Jahan. Prediction of Onset Diabetes using Machine Learning Techniques. December 2017. Doi: 10.5120/ijca2017916020.
[21] Geeksforgeeks. 2019, https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/.
[22] Dishashree Gupta. June 1, 2017, https:// www. analyticsvid hya.com/blog/2017/06/transfer-learning-the-art-of-fine-tu ning-a-pre-trained-model/.
[23] https://www.beckershospitalreview.com/healthcare-inform ation-technology/8-things-to-know-about-online-sympto m-checker-applications.html
[24] Joseph A. Cruz, David S. Wishart. Applications of Machine Learning in Cancer Prediction and Prognosis. 2006. Doi: 10.1177/117693510600200030.
[25] https://en.wikipedia.org/wiki/Your.MD#~:text= Free%20BasicsPerformance,the%20European%20laws %20on%20privacy.