Symptoms based disease diagnosis and treatment recommendation

Sanchit Singhal
sanchit.singhal@outlook.com

Shubham Rana
rana.shubham92@gmail.com

Vibhor Khanduri
vibhorkhanduri.10@gmail.com

***Abstract*— Health is the most important aspect of a person's life. One needs to maintain their health if they wish to be happy. It is said that only a healthy body can have a healthy mind which has a positive impact on an individual's performance. Research suggests that a healthy person is more likely to lead a successful life than someone who is unhealthy or not conscious about their health. But even healthy individuals tend to fall sick due to various reasons. The most trusted and reliable means by which a sick person gets their health check-ups, diagnosis of disease and recommendation for its treatment all over the world are hospitals. Numerous treatments are available for various diseases, but no single person can possibly know about all the diseases and its various treatments. Therefore, the issue is that there's no place where we can get the detail of the diseases and their treatments. What if there's a place where one can find their health issues by just entering their symptoms and get recommendations for treatment as well? This will enable people to find out the cause of their problem and get a solution without having to visit a doctor. This research proposes an idea of developing a system which can eliminate this tedious process of visiting hospitals and making appointments with a doctor to get treatment. It intends to apply the various concepts of NLP and machine learning for building a chatbot application. People will be able to interact with the chatbot application like how they do with other human beings and through a series of queries, the application will try predicting the disease and recommend treatment for it. The proposed system will be immensely useful to people as they will be able to conduct their daily health check-ups, it will make people aware of the status of their health and motivate people to take proper measures to be healthy absolutely free of cost.**

Keywords—healthcare; disease prediction; chatbot; treament recommendation

I. INTRODUCTION

A prosperous society is where all its people are healthy and happy. Thus, it is imperative to keep yourself healthy if one wishes to lead a happy life [1]. It is said that only a healthy body can have a sound mind which has a positive effect on the performance and personality of individuals. But these days, individuals are less conscious of their wellbeing. In their hectic life, they neglect to care about the most important aspect of their life and are less mindful of their wellbeing. According to a recent news by TOI, we observe that individuals do not care about their health at all and do not have the patience to go through the process of meeting a doctor in hospitals i.e. reaching the hospital, making an appointment, waiting for hours etc [2]. This fast paced and hectic life does not have any place for health. People give the example of their busy lives as an excuse for neglecting their health which later turns into a major problem for them.

In today's times, we see that social media apps like WhatsApp, messenger, Facebook etc. have become a major part of everyone's life. Everything, from talking to family and friends to conducting business is now carried out on these applications [3]. People prefer texting than talking on call or meeting someone in person. They find it comfortable and easy.

In the proposed framework, a medical chatbot is being built which will communicate and motivate users to discuss about health issues they're facing and on the basis of symptoms provided by them, the medical chatbot will return diagnosis. The chatbot will be able to figure out symptoms from the interaction it had with the user. Using the symptoms provided, the medical chatbot will predict disease and recommend the treatment required [4]. Therefore, the chatbot will be able to accurately diagnose the patient with the help of simple analysis of symptoms and an interactive approach executed with help of NLP.

Medical chatbots have a very positive effect on the wellbeing of people in a society. It has greater accuracy and leaves no scope for human errors [5]. People nowadays tend to be extremely addicted to internet but are negligent about their health status. They neglect the need of treatment for small problems which eventually become a cause of worry in the future. The suggested system is the solution for all these problems. The idea is focused on building a chatbot which will be cost free and available all day long anywhere in the world. The fact that this system is cost free and is accessible from wherever the user wants, be it in there working environment, will push people to use it. It liberates people from the tedious task which is involved in consulting specialized doctor.

The execution of such a system can spread more awareness among individuals with respect to their health and the need to take measures for a healthy body. With this system, there'll be a decline in no. of individuals neglecting their health because of tiring processes of hospital appointment [6]. People will be able to communicate with the medical chatbot in a similar manner like how they do with other individuals without having to leave their work. The system ensures there'll be no disturbance in their work and is extremely user friendly. So, this can be a way to help individuals be cautious regarding

their health status with the use of chatbot and therefore helps individuals to maintain a healthy life, thus playing a major part in healthcare of the society [7].

The paper structure is as follows: section II describes about the existing work in this area, section III is about proposed application of research work, section IV is architecture of the work done and section V describes the experimental results followed by conclusion and future scopes in section VI.

II. LITERATURE REVIEW

In this section, we will discuss about the various conclusion derived from the papers read as part of the background study.

F. G. Woldemichael and S. Menaria, [8], focuses on the Medical conclusion learning design through the assembled information on diabetes and make smart therapeutic decision emotionally support networks to support doctors. The main objective of the assessment is collect Intelligent Diabetes Disease Prediction System which provides examination of diabetes ailment using the database of patients of diabetes. The framework proposed the utilization of algos such as Bayesian and KNN (K-Nearest Neighbor) to be applied on the database of patients of diabetes and analyzing them by considering the attribute of diabetes which are not similar for predicting the diabetes illness.

Miss Swati Y. Dugane and Prof. Karuna G. Bagd [9] discusses about a system which tells the users about the disease he/she may be suffering from on the basis of the inputs entered by them in that system. It will also let the user know about the concerned specialist the user should get treatment from for that particular disease (e.g. Endocrinologist for diabetes, gastroenterologist for intestine disorder etc). The system will ask the user to elaborate on the symptom by asking specific questions to come to final and precise conclusion.

S. Palaniappan and R. Awang [10] discusses about a model of Intelligent Heart Disease Prediction System (IHDPS) with data mining techniques, namely, Decision Trees, Naive Bayes and Neural Network. Preliminary results indicate that every technique have a unique advantage in achieving the goal of the predefined mining goal. IHDPS has the ability to answer difficult "what if"; queries which conventional decision support system cannot. With the use of medical profiles such as age, sex, bp and blood sugar it has the ability to predict the chances of patients suffering heart diseases. It empowers a lot of knowledge, e.g. pattern, relationship between the different medical factors linked with heart diseases, to be establish.

S. H. Koppad and A. Kumar, [11], the author talks about his research analysis with Data mining systems of big data, using Decision tree for enhancing the performance for COPD diagnosis in patients. This centralize clinical repository of data consists of patients personal details and is referred with the unique aadhaar number which assists in knowing the treatment received by every patient in different hospital and about the doctors who treated them. The preliminary results show a very high precision and efficiency in the diagnosis of COPD in patients by the system proposed.

Deepthi, Y., Kalyan, K.P., Vyas, M., Radhika, K., Babu, D.K. and Rao, N.K[12] discusses about how machine learning can be used in predicting diseases based on symptoms. The various algorithms of machine learning like Naive Bayes, Decision Tree and Random Forest are used on given dataset to predict the disease. Python has been used for the purpose of implementation. This research shows which algorithm is the best on the basis of their accuracy. This accuracy of the algorithm is found by how the dataset given has performed.

BIRNBAUM [13] presented a high level introduction on data mining which further relate to the surveillance of the healthcare data. Traditional statistic and data mining are compared, algorithm and designs of data mining, and advantage of automated data system are explained.

P. Soucy and G. W. Mineau [14], proposed to use KNN algorithm for the purpose of text categorization. It suggests the use of a features selection method which searches and find the related feature for learning task at the moment with the help of features interaction (which is based on word interdependencies). This permits to extensively diminish the amount of features selected from where to learn, turning our KNN algorithm appropriate in places where both the volumes of document and also the size of vocabulary are very high, as with World Wide Web. Thus, KNN algorithm being proposed turns very efficient for classification of text document in that sense (in the sense how predictable and interpretable they are), as is shown. The simplicity (with respect its execution and fine-tuning) become its most valuable asset for the purpose of in-the-field application.

D. Madhu, C. J. N. Jain, E. Sebastain, S. Shaji and A. Ajayakumar [15] suggests an idea about creating a system that uses artificial intelligence for prediction of disease on the basis of the symptom and suggest the treatments available for the disease. This system would also be able be show the compositions of the different medicines and how to use them. This will help people to take correct treatment. Thus, people will be aware about their health status and can take correct protection.

S. Mohan, C. Thirumalai and G. Srivastava [16] suggests a method whose aim is finding features by using the various machine learning technique which will result in improving the precision of cardiovascular disease. The model is presented with various combinations of feature and different techniques of classification known.

A. K. Mishra, P.K. Keserwani, S.G. Samaddar, H.B. Lamichaney, & A.K. Mishra, [17] discuss about the various machine learning algorithms like GBM, XGBoost etc which will be used for calculating the performance of each algorithm on a pre-selected database. Comparative analysis have been carried out for comparing the different results. Thus, for maximizing the probable output, a blend of algorithm has likewise been tested as some ensemble model and the best output of products of combination is used for building the decision support system for prediction in healthcare. The framework designed gives an idea of efficiency of system.

III. PROPOSED APPLICATION

The project on Symptoms based disease prediction and treatment recommendation system aims on the comparative

study of various algorithms of machine learning dealing with prediction and analysis of data. Through this project we deduce the accuracy of the models and compare their performance for predicting the disease based on the symptoms provided by the user. In this project we have implemented three machine learning predictive algorithms namely:

1. Naïve Bayes
2. Decision tree
3. Random forest

Using these three algorithms we will predict the disease from the symptoms and test the accuracy of the model on various test data.

From the results of the accuracy we will also discover the best suited algorithm for the creation of predictive model to predict the disease from symptoms accurately. The objective of our project is to predict disease on the basis of symptoms and recommend the treatment.

The user will interact with the chatbot and enter the symptoms. Chatbot is the intermediate between the user and the model. The input is taken by the model through chatbot and then the input is processed to natural language which is understandable by the system. After conversion, the data taken by the user, different classification models are applied to map the data with symptoms and using these models, prediction is made about the disease. After the classification of disease, association model is applied to get the treatments of the disease. Once the disease prediction and treatment recommendation is made then these are converted from nlp to language that is understandable by the user and the output is displayed to the user using chatbot.

The methodology to be adapted consist of:

1) Research on Topic and setting a workflow:

Firstly, we would be exploring the topic in depth and gathering information and study in depth the use cases which will help in understanding the topic. According to the research setting a proper flow for working.

2) Collection of data:

Obtaining a proper data set for our use which would be used for testing and training of the model.

3) Building the model:

Project can be divided into 3 major parts

1)Chatbot

2)Symptoms-to-disease mapping

3)Disease-to-Treatment mapping

4) Prediction of the result:

The output can be obtained after the processing and mapping of input with relevant data.

IV ARCHITECTURE SYSTEM

In the proposed system the user inputs the symptoms and models predicts the disease using various models. In this project we have used different algorithms to check the accuracy and the performance for the prediction. The system architecture diagram comprises of various stages and steps that are involved in the development of the model and predicting the result.

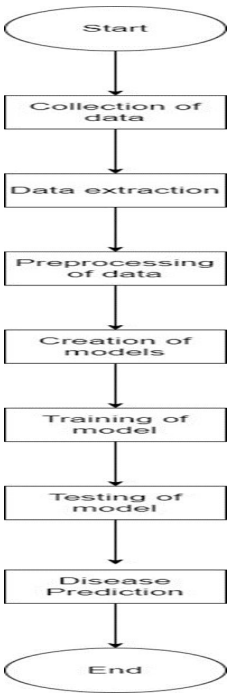

Figure 2 Flowchart of architecture system

The Fig 2 depicts the flow of processing of data and obtaining the accuracy of various models. In the project we have worked on obtaining the data and we processed it and extracted various significant variables from it which are needed to obtain the results using the predictive models and giving accurate results.

A) PROJECT STRUCTURE

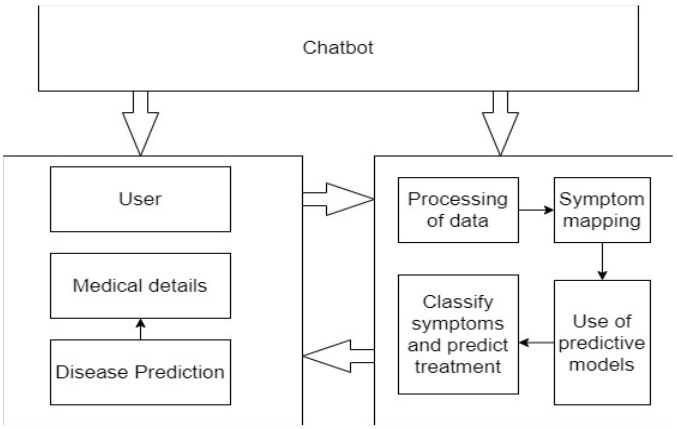

Figure 3 Project model

The above figure displays the structure the project we are working on. The system consists of a chatbot which takes input from the user. The user inputs the symptoms through the chatbot and obtains the prediction of its disease and the basic treatment required for the same.

The System consist of two layers

1)Front-end (Chatbot)

Parameter	Naïve Bayes	Decision Tree	Random Forest
Deterministic/ Non-deterministic	Non-deterministic	Deterministic	Non-deterministic
Effectiveness on	Small data	Large data	Great with high dimension data
Speed	fast on large data	Faster than Naïve bayes	Slower than decision tree
Accuracy	Require large data for accurate result	Highly accurate	Highly accurate on large dataset

2)Back-end (Machine learning models)
Front-end (Chatbot)
The chatbot is the first layer of the system that takes input from the user and is the interface between the user and the models. Basically, the chatbot takes input from the user containing the symptoms and other required details and after the prediction it returns the prediction of disease and the treatment to the user.
Back-end (Machine learning models)
The back-end or the second layer of the system comprise of the machine learning models and the pre processing of the data for obtaining the results. The model takes input from the chatbot and process the input to natural language that is understandable by the system. After converting the text to natural language, the data is mapped with the symptoms using various models and a prediction is made based on the symptoms. After predicting the disease, the output of disease and the treatment is converted from nlp to language that is understandable by the user and the output is sent to the chatbot and is displayed to the user.

B) ALGORITHM USED

In our project, we have used classification algorithms to predict disease from symptoms. The classification problems we have used to predict disease are: -
1)Naive Bayes: -
It is a machine learning algorithm which is used to deal with classification problems. It is one of the fastest algorithms among the classification algorithms. It is based on the concept of the Bayes theorem [18]. In this technique, conditional probability is used to classify the data. Therefore, the Naive Bayes algorithm deduces the probability of record being in a particular class, based on the values of the attributes. It is based on the assumption that all the attributes are independent of each other. That is why the continous variables need to be converted into discrete variables [19].
Naive Bayes algorithm can be classified into 3 different categories:
- Gaussian Naive Bayes
- Bernoulli Naive Bayes
- Multinomial Naive Bayes

The selection of the type of Naive Bayes algorithm based on the data we are dealing with.
2)Decision Tree: -
Decision tree is a tree structure that is based on the principle of conditions. It is an efficient algorithm which is used mainly for predictive analysis. A Decision Tree has mainly the attributes- internal nodes, branches and the terminal node [20].

Internal node can be attributed as a test, branches as the conclusion of the test and leaf node as the class label. This algorithm is preferred in case of supervised models. It can be used for classification as well as regression problems.

There are two types of Decision trees namely-
- Categorical Variable Decision Tree
- Continuous Variable Decision Tree

3)Random Forest: -
Random Forest is a machine learning algorithm that produces great result most of the time without even hyper parameter tuning. It can be used for both classification and regression task [21]. It is also a supervised learning algorithm. Random forest adds additional randomness to the model, while growing the trees. In place of searching for the most significant characteristic while splitting a node, it searches for the best characteristic among a random subset of characteristics. As a result, this model provides wide diversity. The hyper parameters like n-estimators, max_features, min_sample_leaf can be used to increase the predictive power of model whereas the hyper parameters like n_jobs , random_state , oob_score can be used to increase the model's speed. This algorithm is a great choice for anyone who needs to develop a model quickly. It is a fast, simple and flexible algorithm [22].

Apart from the classification algorithms, we will be using association theorems like apriori theorem for recommending treatment of the disease.

C) COMPARATIVE STUDY OF ALGORITHMS
The table below displays the comparison of various algorithms that we have used in building the models. Based on certain factor the table allows us to compare the algorithms and makes us understand pros and cons of each based on size of dataset.

Table 1 Comparison table

Comparisons:
1)Decision tree vs Random Forest:
- Random Forest model will be less inclined to overfitting than decision tree, and gives a more summed up arrangement.
- Random Forest is stronger and more precise than decision trees.

2) Naive Bayes vs Decision tree:
- Decision tree pruning may disregard some critical qualities in preparing information, which can lead the precision for a toss.
- Decision tree is a discriminative model, whereas Naive bayes is a generative model.

3)Random Forest vs Naïve Bayes:

• Random Forest works well with both categorical and continuous variables
• Naive Bayes requires a small amount of training data to estimate the test data. So, the training period is less.

Disadvantages of algorithms

1. The main limitation of Naive Bayes is the assumption of independent predictor features.
2. Smoothing turns out to be a over-head.
3. For a Decision tree sometimes calculation can go far more complex compared to other algorithms.
4. A little change in the information can cause a huge change in the structure of the decision tree causing unsteadiness
5. Random Forest require much more time to train as compared to decision tree
6. Random forest is complex as Random Forest creates a lot of trees as compared to decision tree and combines their outputs.

V. RESULT

We implemented the model on the disease dataset and predicted the from the various symptoms given by the user. Also we recommended the generic treatment and precautions that needs to be observed in the predicted disease

We excecuted the model on the disease dataset and predicted the from the various symptoms given by the user. Also, we recommended the generic treatment and precautions that needs to be observed in the predicted disease. We also hyper tuned the algorithm using various parameters. We have used decision tree classifier because of its accuracy and better performance on large data set as compared to rest of two algorithms.

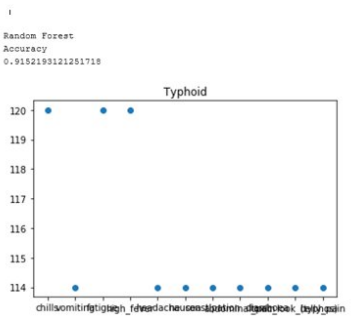

Figure 3 Random forest accuracy

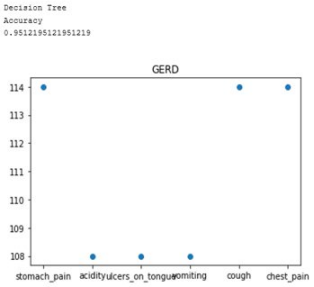

Figure 4 Decision tree accuracy

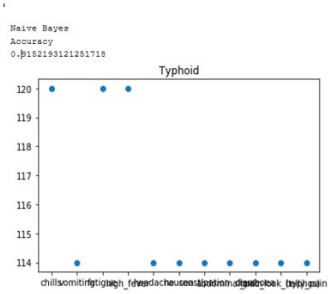

Figure 5 Naive Bayes accuracy

Above figures 3,4 and 5 shows the comparative study of algorithms based on the testing of models (Decision tree, random forest and naïve bayes) in same dataset and using same set of inputs under similar conditions. These models give different output, decision tree provides better accuracy among all three algorithms (95.1%) whereas naïve bayes and random forest give similar prediction and accuracy of almost 91.2% .Using these results from our experimentation we came to a conclusion and decided to implement our model using decision tree algorithm along with various parameters (hyper parameter tuning) and increase the accuracy further.

Figure 6 Result of hyper-tuned algorithm

Below, figure 6 displays the result of our implication of model using parameterized and optimized decision tree model. Through this we have achieved an accuracy of 97-98% based on various factors and set of inputs. The model takes various inputs from user and give prediction of the disease and what precautions one must take for the disease. It also displays the accuracy of the model for that particular prediction.

VI. CONCLUSION AND FUTURE SCOPE

Till now we have worked on the models of machine learning and implemented the predictive models on our dataset to check the accuracy of algorithms and see how precise and fast they are. We also worked on the large data set and used more predictive and associative algorithms to fine tune the model and obtain more accurate prediction from the model. We also created a interface for user and predict the treatment for disease based on the symptoms inputted by the user. In future we would be using stacking to emzemble the models and enhance the accuracy of model. Also we would improve the UI for the user and make it more friendly and easy to use . Vision of the project is to accurately predict the disease and treatment for the same, so that the user can get consultation at its ease and minimal physical contact with anyone. In this project we implemented various algorithms that we used to predict the disease from the symptoms and based on the accuracy we compared the pros and the cons of the models and how useful they are for us to use in our use case. Depending on the speed, handling of data, and accuracy we can wisely choose which algorithm we should use and what is the limitation for each of the algorithm. In future we will test and implement some more algorithms and develop a suitable model for the prediction. At the end of this project we feel we would be able to provide medical consultation to each and every person even if they are far away from the medical facility at a low cost and ease of their home.

REFERENCES

[1] R. B. Mathew, S. Varghese, S. E. Joy and S. S. Alex, "Chatbot for Disease Prediction and Treatment Recommendation using Machine Learning," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), pp. 851-856, 2019, doi: 10.1109/ICOEI.2019.8862707.

[2] Sethi, Ramandeep & Thumar, Aniket & Jain, Vaibhav & Chavan, Sachin, “Disease Prediction Application based on Symptoms,” International Journal of Scientific Research in Computer Science, Engineering and Information Technology, pp: 641-646, 2019.

[3] M. Ferdous, J. Debnath and N. R. Chakraborty, "Machine Learning Algorithms in Healthcare: A Literature Survey," 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp: 1-6, 2020.

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

[5] M. M. Rahman, R. Amin, M. N. Khan Liton and N. Hossain, "Disha: An Implementation of Machine Learning Based Bangla Healthcare Chatbot," 2019 22nd International Conference on Computer and Information Technology (ICCIT), pp: 1-6, 2019.

[6] Saini, Akanksha, A. J. Meitei, and Jitenkumar Singh, "Machine Learning in Healthcare: A Review." 2021 Available at SSRN 3834096.

[7] Dharwadkar, Rashmi, and Neeta A. Deshpande. "A medical chatbot," International Journal of Computer Trends Technology, Vol. 60, No.1, 2018

[8] F. G. Woldemichael and S. Menaria, "Prediction of Diabetes Using Data Mining Techniques," 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI), pp: 414-418, 2018.

[9] http://ijarcet.org/wp-content/uploads/IJARCET-VOL-6-ISSUE-9-1339-1343.pdf

[10] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," 2008 IEEE/ACS International Conference on Computer Systems and Applications, pp. 108-115, 2008, doi: 10.1109/AICCSA.2008.4493524.

[11] S. H. Koppad and A. Kumar, "Application of big data analytics in healthcare system to predict COPD," 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT), pp: 1-5, 2016.

[12] Deepthi, Y., Kalyan, K.P., Vyas, M., Radhika, K., Babu, D.K. and Rao, N.K., “Disease Prediction Based on Symptoms Using Machine Learning,” In Energy Systems, Drives and Automations, pp: 561-569, 2020, Springer, Singapore.

[13] BIRNBAUM, EDITED BY DAVID, "Application of data mining techniques to healthcare data," 2004 Infection control and hospital epidemiology, 2004.
[14] P. Soucy and G. W. Mineau, "A simple KNN algorithm for text categorization," Proceedings 2001 IEEE International Conference on Data Mining, pp: 647-648, 2001.
[15] D. Madhu, C. J. N. Jain, E. Sebastain, S. Shaji and A. Ajayakumar, "A novel approach for medical assistance using trained chatbot," 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT), pp: 243-246, 2017.
[16] S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp: 81542-81554, 2019.
[17] A. K. Mishra, P.K. Keserwani, S.G. Samaddar, H.B. Lamichaney, & A.K. Mishra, "A decision support system in healthcare prediction", In Advanced Computational and Communication Paradigms, pp: 156-167, Springer, Singapore, 2018.
[18] Divya, S., Indumathi, V., Ishwarya, S., Priyasankari, M. and Devi, S.K., "A self-diagnosis medical chatbot using artificial intelligence," Journal of Web Development and Web Designing, Vol. 3, No. 1, pp.1-7, 2018.
[19] Rish, Irina, "An empirical study of the naive Bayes classifier," 2001 IJCAI 2001 workshop on empirical methods in artificial intelligence. Vol. 3. No. 22. 2001.
[20] P. H. Swain and H. Hauska, "The decision tree classifier: Design and potential," in IEEE Transactions on Geoscience Electronics, Vol. 15, No. 3, pp: 142-147, 1997.
[21] Khalilia, M., Chakraborty, S. and Popescu, M., "Predicting disease risks from highly imbalanced data using random forest," BMC medical informatics and decision making, Vol. 11, No. 1, pp:1-13, 2011.
[22] R. Devika, S. V. Avilala and V. Subramaniyaswamy, "Comparative Study of Classifier for Chronic Kidney Disease prediction using Naive Bayes, KNN and Random Forest," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), pp: 679-684, 2019.