Disease Prediction using Machine Learning Algorithms

Mr.Sharan L Pais, Fayiz Ahmed K, Sharanya, Shrihastha, Varshith
Senior Assistant Professor, Department of Information Science and Engineering[1]
Students Department of Information Science and Engineering[2,3,4,5]
Alva's Institute of Engineering and Technology, Mijar, Mangalore, Karnataka, India

Abstract: *The development and exploitation of several prominent Data mining techniques in numerous real-world application areas (e.g. Industry, Healthcare and Bio science) has led to the utilization of such techniques in machine learning environments, in order to extract useful pieces of information of the specified data in healthcare communities, biomedical fields etc. The accurate analysis of medical database benefits in early disease prediction, patient care and community services. The techniques of machine learning have been successfully employed in assorted applications including Disease prediction. The aim of developing classifier system using machine learning algorithms is to immensely help to solve the health-related issues by assisting the physicians to predict and diagnose diseases at an early stage. A Sample data of 4920 patients' records diagnosed with 41 diseases was selected for analysis. A dependent variable was composed of 41 diseases. 95 of 132 independent variables (symptoms) closely related to diseases were selected and optimized. This research work carried out demonstrates the disease prediction system developed using Machine learning algorithms such as the Decision Tree classifier, Random forest classifier, and Naïve Bayes classifier. The paper presents the comparative study of the results of the above algorithms used.*

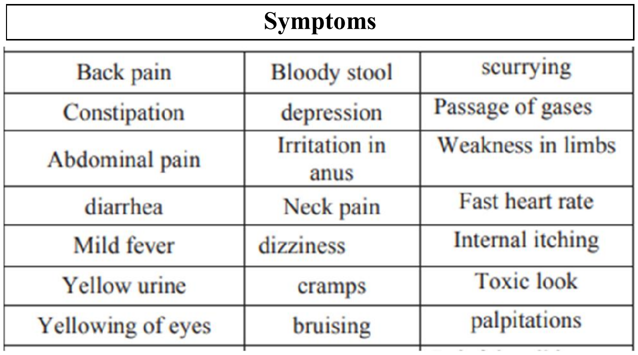
Keywords: Disease Prediction.

I. INTRODUCTION

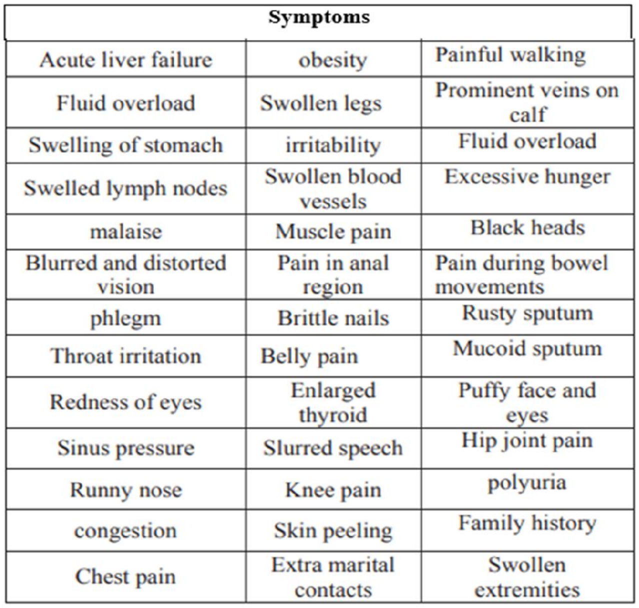
It has always been difficult to find a new medicine. A new medicine is researched and developed over a long period of time. The total number of candidate molecules in the foundation phase of drug discovery for any disease was estimated to be between 1060 and 10200 [1]. The reason for this is; It takes a long time to find the right compounds for making a new medicine. In the past, the medical industry did not have facilities that utilized machine learning strategies to investigate potential medicines. Since the advent of artificial intelligence (AI), the field of computer applications has seen significant growth. The idea of artificial intelligence is nothing more than a computerized simulation of human intelligence. The process of machine learning, which entails gathering information, developing rules for extracting it, demonstrating approximate or definite inferences, and verifying, is the foundation for the development of artificial intelligence. The precision of machine learning algorithms is the foundation of artificial intelligence's success. The availability of a substantial training dataset is primarily what determines a machine learning algorithm's accuracy. We now have a lot of data to train a system with. The integration of AI into the drug development process has evolved to a greater extent. AI is now playing a significant role in this analysis and development of drug discovery. Based on the requirements, pharmaceutical companies, AI-focused research and development institutions, and medical professionals can collaborate to investigate the new medicine. Numerous earlier works can be found in the literature for drug recommendations. A recurrent neural network (RNN) was proposed by Yasonik et al. [2] to generate molecules for drug discovery. Using transfer learning, the network was fine-tuned by investigating the generated molecules. The authors of the manuscript [3,4] describe how artificial intelligence can be used to investigate medicine. A system of evidence-based assessment is described in [5,6]. Machine learning algorithms will be used to make medical recommendations in the future .The system collects a lot of data based on the information patients provide. Utilizing these data systems enables training and the recommendation of medications. In [7], a few researchers have fostered a patient eating routine suggestion framework utilizing AI approach. For various medical diagnoses, various machine learning approaches have been developed. [8,9] provides a description of these strategies. Leung and others 10] have talked about how biologists, data scientists, and medical researchers working together on the development of genomic medicine can benefit from machine learning techniques. There are a number of papers in [11-13] that discuss the applications of machine learning techniques to medical imaging. For the purpose of predicting diseases, some authors used a deep learning approach to explore information from medical imaging data [14–16].In this work, we try to figure out which medicine is best for a disease that our system recommends. The goal of our work is to develop a machine learning-based system that can suggest medications based on symptoms. As we have seen, there are numerous diseases that, if their symptoms are similar, can be treated with the same medications. In addition, it is able to locate the chemical composition that is closest to what is needed to create the novel drug for any new diseases. The initial list of known diseases and their symptoms has been prepared. The medicines and their components are then examined in relation to the aforementioned conditions. This side effect based sickness expectation strategies might help specialists to endorsed the medication with more precision.

II. OVERVIEW

The dataset we have considered consists of 132 symptoms, the combination or permutations of which leads to 41 diseases. Based on the 4920 records of patients, we aim to develop a prediction model that takes in the symptoms from the user and predicts the disease he is more likely to have. The considered symptoms are:

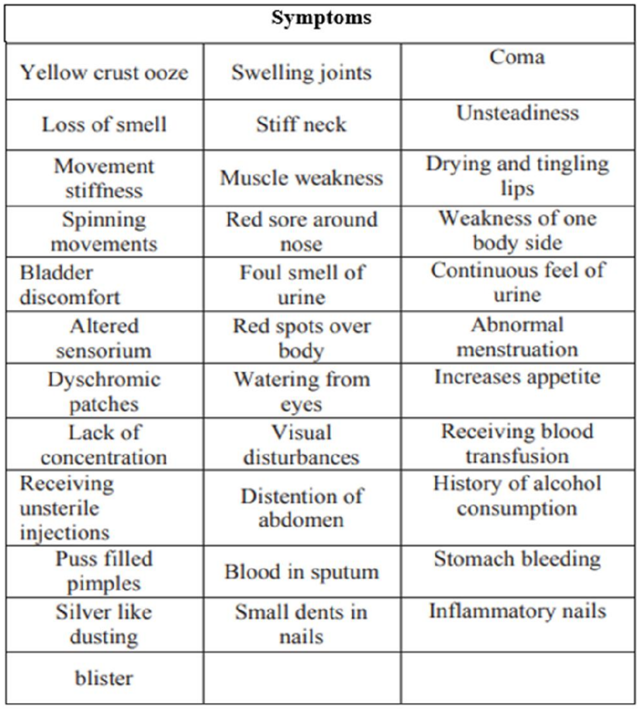


Symptoms		
Back pain	Bloody stool	scurrying
Constipation	depression	Passage of gases
Abdominal pain	Irritation in anus	Weakness in limbs
diarrhea	Neck pain	Fast heart rate
Mild fever	dizziness	Internal itching
Yellow urine	cramps	Toxic look
Yellowing of eyes	bruising	palpitations



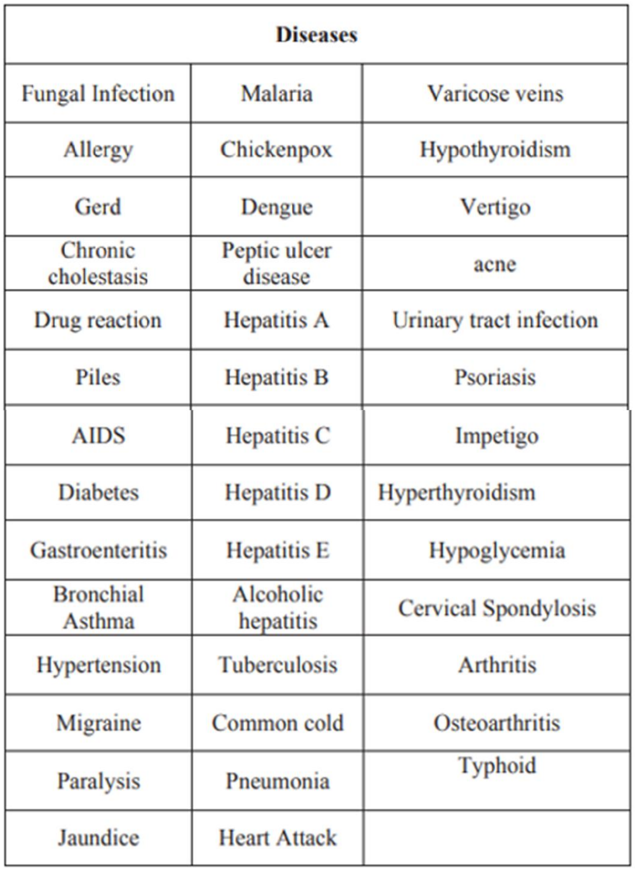
Symptoms		
Acute liver failure	obesity	Painful walking
Fluid overload	Swollen legs	Prominent veins on calf
Swelling of stomach	irritability	Fluid overload
Swelled lymph nodes	Swollen blood vessels	Excessive hunger
malaise	Muscle pain	Black heads
Blurred and distorted vision	Pain in anal region	Pain during bowel movements
phlegm	Brittle nails	Rusty sputum
Throat irritation	Belly pain	Mucoid sputum
Redness of eyes	Enlarged thyroid	Puffy face and eyes
Sinus pressure	Slurred speech	Hip joint pain
Runny nose	Knee pain	polyuria
congestion	Skin peeling	Family history
Chest pain	Extra marital contacts	Swollen extremities

DOI: 10.48175/IJARSCT-7825



Symptoms		
Yellow crust ooze	Swelling joints	Coma
Loss of smell	Stiff neck	Unsteadiness
Movement stiffness	Muscle weakness	Drying and tingling lips
Spinning movements	Red sore around nose	Weakness of one body side
Bladder discomfort	Foul smell of urine	Continuous feel of urine
Altered sensorium	Red spots over body	Abnormal menstruation
Dyschromic patches	Watering from eyes	Increases appetite
Lack of concentration	Visual disturbances	Receiving blood transfusion
Receiving unsterile injections	Distention of abdomen	History of alcohol consumption
Puss filled pimples	Blood in sputum	Stomach bleeding
Silver like dusting	Small dents in nails	Inflammatory nails
blister		

The diseases considered are:

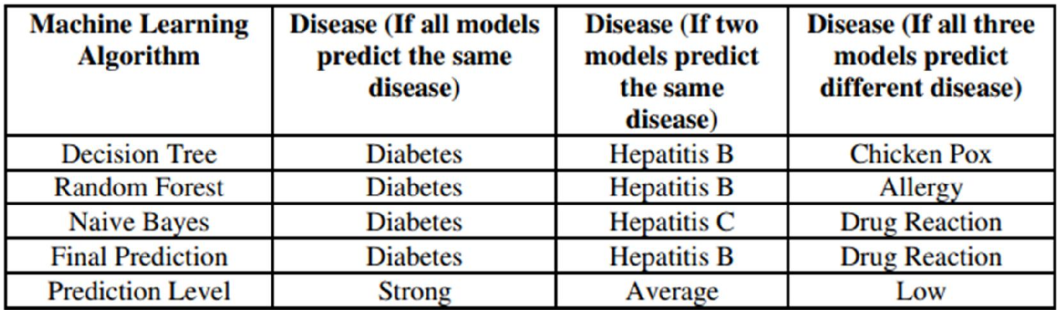


Diseases		
Fungal Infection	Malaria	Varicose veins
Allergy	Chickenpox	Hypothyroidism
Gerd	Dengue	Vertigo
Chronic cholestasis	Peptic ulcer disease	acne
Drug reaction	Hepatitis A	Urinary tract infection
Piles	Hepatitis B	Psoriasis
AIDS	Hepatitis C	Impetigo
Diabetes	Hepatitis D	Hyperthyroidism
Gastroenteritis	Hepatitis E	Hypoglycemia
Bronchial Asthma	Alcoholic hepatitis	Cervical Spondylosis
Hypertension	Tuberculosis	Arthritis
Migraine	Common cold	Osteoarthritis
Paralysis	Pneumonia	Typhoid
Jaundice	Heart Attack	

DOI: 10.48175/IJARSCT-7825

III. METHODOLOGY

The Decision Tree Classifier [19], Random Forest Classifier [22], and Naive Bayes Classifier [23] are the three data mining algorithms that are utilized in the implementation of the disease prediction system. First, we used each of the three classifiers to train our disease prediction system separately, and then we looked at the results. because accurate diagnosis and prediction of a disease are crucial to a patient's successful treatment. Consequently, we have assigned distinct prediction levels based on the predictions made by multiple classifiers. Similarly, a single classifier can predict a distinct disease while two classifiers can predict a particular disease. As a result, we have deemed the level of prediction to be high if the same disease is predicted by all of the classifiers. On the other hand, we take a disease's average level of prediction into account if it is predicted by two classifiers but not by just one. If all classifiers correctly predict different diseases, the Naive Bayes classifier is used to make the final prediction. Because the Naive-based classifier provides greater accuracy and avoids the issue of overfitting. Some diseases' prediction levels are described in detail in Table 1. The block diagram of the proposed architecture is depicted in Figure 1. The description and working of the algorithms are given below.



Machine Learning Algorithm	Disease (If all models predict the same disease)	Disease (If two models predict the same disease)	Disease (If all three models predict different disease)
Decision Tree	Diabetes	Hepatitis B	Chicken Pox
Random Forest	Diabetes	Hepatitis B	Allergy
Naive Bayes	Diabetes	Hepatitis C	Drug Reaction
Final Prediction	Diabetes	Hepatitis B	Drug Reaction
Prediction Level	Strong	Average	Low

Table 1: Final prediction and level of prediction of diseases

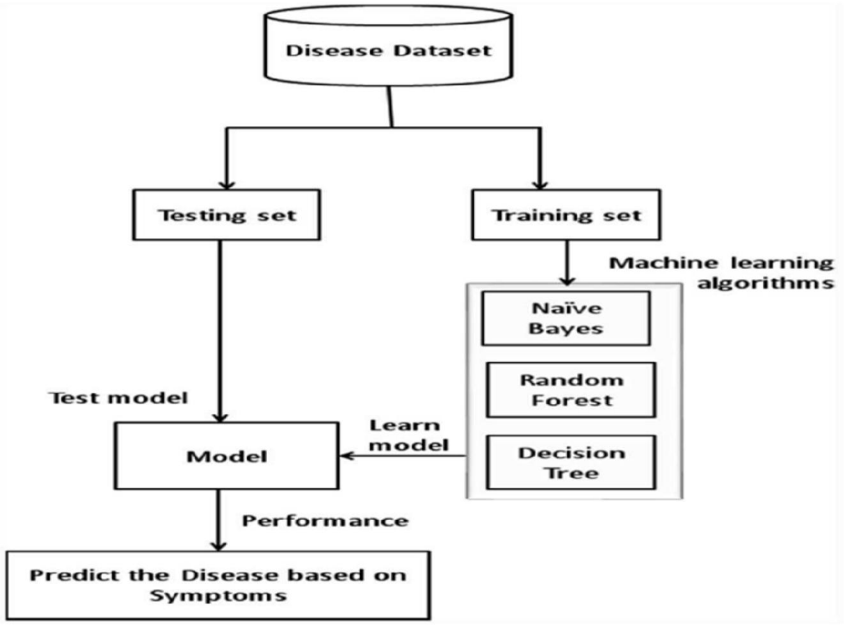


Figure 1: Block diagram of prediction model

3.1 Decision Tree Classifier

It is a decision tree-framed classification model. Every node in this tree specifies a test for the attribute, and each branch that comes from that node resembles one of the attribute's promising values. It divides the dataset into smaller and smaller subsets by learning a series of overt "if-then" rules on feature values (symptoms in our case), which allows it to predict

DOI: 10.48175/IJARSCT-7825

our objective (i.e., disease prediction). Decision nodes and leaf nodes are the two parts of the decision tree classifier.

- Decision node: If a node is further splits into sub-nodes, then this node is called as the decision node. In this presented work, all the symptoms (features) are considered as decision nodes.
- Leaf node: The nodes from which there is no subordinate nodes coming off are considered as leaf nodes. In other words, they don't further split the data anymore. At the level of leaf nodes, we achieve the classification level. Leaf node represents the classification which is the decision of a class. In this work the diseases are correspond to the leaf nodes.

3.2 Random forest classifier

Random forest is a popular machine learning algorithm that gives excellent results most of the time. It is pretty easy to use for the classification purpose. The drawback of using decision tree algorithm is that it suffers from the overfitting problems. Basically, Random forest classifier crafts a set of decision trees from an arbitrarily chosen subset of the training set. Finally, it collects the outcomes from different decision trees to decide the final prediction. It is a kind of ensemble learning based meta estimator that ensembles a many decision tree classifiers on various sub-samples of the data.

3.3 Naive bayes classifier

Naive Bayes classifier is a supervised learning approach. It uses the Bayes theorem concept for solving the classification problems. It is mostly appropriate to use in those classification problem which have a high-dimensional dataset. It is one of the simplest and effective classification algorithms which can be used in the rapid development of the machine learning models with quick predictions. The basic Naive Bayes concept is that each feature contributes independently and equally towards obtaining the results. One more specialty of this algorithm is; it needs very less computational power.

$$P(D|S) = \frac{P(S|D)P(D)}{P(S)}$$

According to our work, we have mapped the formulation parameters of the Bays theorem. In the above formula, D denotes class (Disease) and S denotes Features (Symptoms).

Key Terms

- P(D) -: Prior probability is the proportion of Disease in the considered data set.
- P(D|S)-: Posterior probability
- P(S|D)-: Likelihood is the probability of classification a disease in presence of some other symptoms.
- P(S)-: Predictor Prior Probability is the proportion of symptoms in the dataset.

Example

Considering medical record available in Table 2, we estimate the Naïve Bayes results for the set of symptoms. If a particular symptom is present, then we have indicated it by 1. Similarly, if that symptom is not found then it is marked as 0. The disease prediction has been explained by the following example. Let us consider, for a disease the symptoms details are as follows:

- High fever= Present (denoted by value'1')
- Shivering=Present (denoted by value '1')
- features considered = 4
- Rash=Absent (denoted by value '0')
- Joint Pain=Present (denoted by value '1')
- Symptoms=High fever, Rash, Shivering, Joint Pain.
- Classes = Dengue, Zika Virus

DOI: 10.48175/IJARSCT-7825

High Fever	Rash	Shivering	Joint Pain	Disease
1	0	1	0	Dengue
1	1	0	0	Zika Virus
0	1	0	1	Zika Virus
0	0	1	1	Dengue
1	0	1	1	Dengue
1	1	0	1	Dengue
1	1	1	1	Zika Virus
0	1	1	1	Zika Virus
1	1	1	0	Zika Virus
0	1	1	0	Dengue
1	0	0	1	Dengue
0	1	0	0	Dengue

Table 2: Sample medical record

As per the dataset we have considered:

1. Likelihood = P (Feature=symptoms |Class=Dengue, Zika Virus)
2. Marginal Likelihood= P(Features=symptoms)
3. Prior Likelihood= P(Class)

The prediction is thus made by comparing the posterior probabilities for each class (i.e. For each disease) after observing the input symptoms.

'S1' for 'High fever',
'S2' for 'Rash',
'S3' for 'Shivering',
'S4' for 'Joint Pain' and 'D' for 'Diseases(class)'.

Firstly, the probability for Dengue is estimated (i.e. the class=Dengue with input symptoms as follows:
"High fever=Present"; "Rash=Absent"; "Shivering=Present"; "Joint Pain=Present")

Thus, the formula modifies to:

P (S=Dengue | S1=Present, S2=Absent, S3=Present, S4=Present) = P (S1=Present, S2=Absent. S3=Present, S4=Present | S=Dengue) * P(S=Dengue) = P (S1=Present | S=Dengue) * P (S2=Absent | S=Dengue) * P (S3=Present | S=Dengue) * P (S4=Present | S=Dengue) * P(S=Dengue)

=4/12 * 4/12 * 5/12 *5/12*8/12 =0.01286

Secondly, the probability for Zika Virus is estimated (i.e. the class=Zika Virus with the same input symptoms as mentioned in the above step)

P (S=Zika Virus | S1=Present, S2=Absent, S3=Present, S4=Present)= P (S1=Present, S2=Absent. S3=Present, S4=Present | S=Zika Virus) * P(S=Zika Virus) = P (S1=Present | S=Zika Virus) * P (S2=Absent | S=Zika Virus) * P (S3=Present | S=Zika Virus) * P (S4=Present | S=Zika Virus) * P(S=Zika Virus)

= 3/12 * 0 * 2/12 * 2/12 * 4/12

= 0.002348 0.0128 > 0.0023 --- > P(S=Dengue) > P(S=Zika Virus)

Thus, we can predict that the considered data point belongs to the class "Dengue", i.e. the patient with the symptoms High fever, Shivering, and Joint Pain are more likely to have Dengue than Zika Virus.

IV. IMPLEMENTATION AND RESULTS

4.1 Performance of Algorithms on Training Data

The system was trained on medical record of 4920 patients prone to 41 diseases which was due to the combination of various symptoms. We have considered 95 symptoms out of 132 symptoms to avoid overfitting. We used the K fold cross validation technique (K=5) to check the performance of all three algorithms on the dataset.

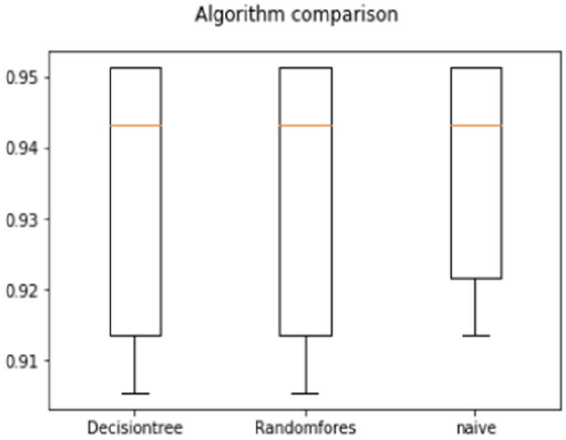
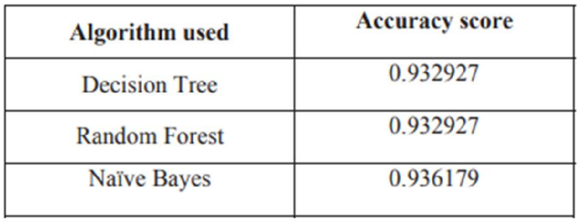


Figure 2 : Box and whisker plot of comparision of algorithms performance on training set

The above figure is a box and whisker plot showing the spread of the accuracy scores across each cross-validation fold (K=5) for each algorithm. From these results, we can infer that all the three algorithms work exceptionally well on the dataset. However, Naïve Bayes is perhaps working a little better when compared to the other two algorithms.

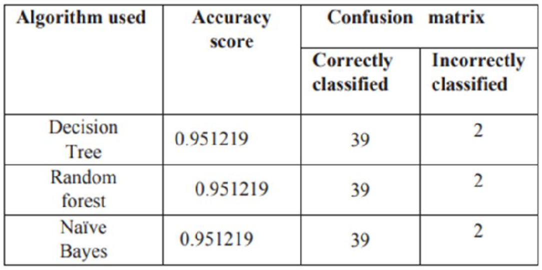
The accuracy score of each algorithm after training were:



Algorithm used	Accuracy score
Decision Tree	0.932927
Random Forest	0.932927
Naïve Bayes	0.936179

Table 3: Accuracy table

Performance of Algorithms on test data after training, the system was tested on 41 new patients records considering 95 symptoms. The accuracy score and the confusion matrix is given as by:



Algorithm used	Accuracy score	Confusion matrix	
		Correctly classified	Incorrectly classified
Decision Tree	0.951219	39	2
Random forest	0.951219	39	2
Naïve Bayes	0.951219	39	2

Table 4: Accuracy and confusion matrix

From the above table, we can infer that all the algorithms have equal accuracy score. The accuracy in terms of percentage: 95.12 percentage.

V. CONCLUSION

From the historical development of machine learning and its applications in medical sector, it can be shown that systems and methodologies have been emerged that has enabled sophisticated data analysis by simple and straightforward use of machine learning algorithms. This paper presents a comprehensive comparative study of three algorithms performance on a medical record each yielding an accuracy up to 95 percent. The performance is analyzed through confusion matrix and accuracy score. Artificial Intelligence will play even more important role in data analysis in the future due to the availability of huge data produced and stored by the modern technology.

REFERENCES

[1]. Lin, E., Lin, C.H. and Lane, H.Y., 2020. Relevant applications of generative adversarial networks in drug design and discovery: molecular de novo design, dimensionality reduction, and de novo peptide and protein design. Molecules, 25(14), p.3250.

[2]. Yasonik, J., 2020. Multiobjective de novo drug design with recurrent neural networks and nondominated sorting. Journal of Cheminformatics, 12(1), pp.1-9.

[3]. Mintz, Y. and Brodie, R., 2019. Introduction to artificial intelligence in medicine. Minimally Invasive Therapy & Allied Technologies, 28(2), pp.73-81.

[4]. Hamet, P. and Tremblay, J., 2017. Artificial intelligence in medicine. Metabolism, 69, pp.S36- S40.

[5]. Rajkomar, A., Dean, J. and Kohane, I., 2019. Machine learning in medicine. New England Journal of Medicine, 380(14), pp.1347-1358.

[6]. Sidey-Gibbons, J.A. and Sidey-Gibbons, C.J., 2019. Machine learning in medicine: a practical introduction. BMC medical research methodology, 19(1), pp.1-18.

[7]. Iwendi, C., Khan, S., Anajemba, J.H., Bashir, A.K. and Noor, F., 2020. Realizing an efficient IoMT-assisted patient diet recommendation system through machine learning model. IEEE Access, 8, pp.28462-28474.

[8]. Kononenko, I., 2001. Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in medicine, 23(1), pp.89-109.

[9]. Kononenko, I., Bratko, I. and Kukar, M., 1997. Application of machine learning to medical diagnosis. Machine Learning and Data Mining: Methods and Applications, 389, p.408.

[10]. Leung, M.K., Delong, A., Alipanahi, B. and Frey, B.J., 2015. Machine learning in genomic medicine: a review of computational problems and data sets. Proceedings of the IEEE, 104(1), pp.176-197.

[11]. Erickson, B.J., Korfiatis, P., Akkus, Z. and Kline, T.L., 2017. Machine learning for medical imaging. Radiographics, 37(2), pp.505-515.

[12]. Giger, M.L., 2018. Machine learning in medical imaging. Journal of the American College of Radiology, 15(3), pp.512-520.

[13]. Wernick, M.N., Yang, Y., Brankov, J.G., Yourganov, G. and Strother, S.C., 2010. Machine learning in medical imaging. IEEE signal processing magazine, 27(4), pp.25-38.

[14]. Suzuki, K., 2017. Overview of deep learning in medical imaging. Radiological physics and technology, 10(3), pp.257-273.

[15]. Lee, J.G., Jun, S., Cho, Y.W., Lee, H., Kim, G.B., Seo, J.B. and Kim, N., 2017. Deep learning in medical imaging: general overview. Korean journal of radiology, 18(4), p.570.

[16]. Lundervold, A.S. and Lundervold, A., 2019. An overview of deep learning in medical imaging focusing on MRI. Zeitschrift für Medizinische Physik, 29(2), pp.102-127.

DOI: 10.48175/IJARSCT-7825