IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

PREDICTION OF DISEASES USING SUPERVISED LEARNING

Ashish Kumar
Assistant Professor
Krishna Engineering College
Ghaziabad, India

Priya Ghansela
UG Student
Krishna Engineering College
Ghaziabad, India

Purnima Soni
UG Student
Krishna Engineering College
Ghaziabad, India

Chirag Goswami
UG Student
Krishna Engineering College
Ghaziabad, India

Parasmani Sharma
UG Student
Krishna Engineering College
Ghaziabad, India

ABSTRACT

In healthcare industry and in various applications, data mining plays a basic role to forecast various diseases. For detection of various diseases patient need to go through various tests. But mining techniques can lower the number of trails. The downgraded tests play a key role in team behaviour. This requires proper paper studies of mining methods which is used for forecasting different diseases. In this paper Naïve Bayes and Decision Tree method are described to envision various diseases. Facilities must be inventive so patient examination and cure can be done. Various machine learning algorithms can be beneficial to the medical purpose, the humans practice large and all-purpose medical datasets to analyse them for clinical insights. This can be used by physicians in various medical field which leads to satisfaction to large number of people when well accomplished. We are basically trying to contrivance functionalities of ML in a single system in a medical field. Instead of having diagnosis, diseases are predicted and executed using ML then the medical field would be smarter and more advanced. Some cases can happen when basic investigation of a disease is not in reach. Hence disease forecast can be accurately carried out.

Keywords: Machine Learning, Random Forest, Naïve Bayes, Support Vector Machine, Decision Tree

1. INTRODUCTION

ML deals with more effective methods to decisive diseases which contain well compiled and large compiled databases. ML is very useful in the field of healthcare research. Owing to the environment and living habits of the people in the specific areas, the efficacy of a disease forecast can be decreased owing to more differentiation in a different regional disease. There are therefore more rivals:

1) How is misplaced data gathered?

2) How can the geographical existence of the diseases be determined?

3) How to defeat living habit and climate problems?

To beat the problems of partial and missing data, we will combine unorganised as well as organised data to precisely forecast disease. So, we proposed Naive Bayesian [1] algorithm for both types of data so we can get the precise outcome.

Machine Learning mainly focuses in the medical field and patient care to give a better outcome. Is also used to identify different diseases and analyse it effectively. Predictive forecasting using effective and various ML methods which helps to forecast the disease more accurately and to treat patients. The medical care generates large amount of medical information which could be used to extract and indicate various diseases which can occur in future while using health data and treatment history. This unknown information in medical care data will latterly be used for making decisions in patients' health. This area needs to be developed in the field of medical care using see-through data. Another such technique in the field of healthcare is for machine learning algorithms. Medical care needs to be dynamic so enhanced decisions for patient analysis and can be done. In medical field, humans process large Machine learning in healthcare aids the humans to process huge and complex medical dataset then explore them to clinical insights. This can then be used by physicians in the medical field. Therefore, ML enhance the patient's contentment. The Decision tree [2], Naïve Bayes [3], support vector [1] machine and Random Forest methods are used in predicting diseases using health data and patient handling history.

2. LITERATURE REVIEW

The paper aims to study the latest cardiac research using various mining methods analyse numerous mining methods and to figure out that these techniques give accurate and efficient results [4].

To have a consistent model to predict the heart disease, the paper proposes rule-based approach for evaluating the precision to apply rules to distinct results of help decision trees, vector machine and systemic regression on the Cleveland Heart Disease database [5].

The primary goal is to build an Intelligent Framework using the methodology of data mining simulation, namely Naive Bayes. In this person, it is introduced as a cloud-based program that responds to the predefined queries [6].

Dr. Shirin Glader's research on developing machine-learning models to forecast the trajectory of various diseases will continue. Using machine learning, she will go over developing a model, assessing its efficiency and answering or posing different disease related questions. Her talk should explore the machine learning principle, since it is implemented using R [7].

They are developing a latest multidisplinary disease risk forecast, which depends on the convolutionary neural network, it uses organised and unorganised hospital facts. Neither of recent research cantered, to the best of our knowledge, on all forms of medical data analytics. Prediction accuracy of our proposed algorithm exceeds 94.8 per cent relative to many traditional prediction algorithms [8].

This uses machine learning algorithms in the proposed method for successful estimation of the frequency of thyroid disease in populations that are recurrent with illness. This is playing with the modified models from gathered real-life data from hospitals [9].

This article introduces a new CNN-based Quick Spectral classification algorithm that allows several hyperspectral images to create a composite image and then trains the model only once on the composite image. The model will predict each picture independently after testing [10].

In this paper they have developed machine learning algorithms to accurately forecast infectious disease epidemics in populations with repeated diseases. We are proposing a new multi-modal disease forecasting method for convolutionary neural networks, using unstructured and structured medical data [11].

The paper describes about the advanced machine learning methods called averaged single-dependence estimators with assumption of resolving the problems of prediction, with the result obtained by DNA microarray gene expression, if a given cancer would revert during a limited period of time, generally 5 years. The statistical difficulty, we use an entropy-based approach to gene selection to pick specific analytical genes that are precisely responsible for forecasting recurrence [12].

This paper addresses the key difficulties of assessing and distinguishing patients and classifying them with TB with HIV and those with HIV without TB disease traces. This article is a short overview of the methods used to identify the co infected patients with HIV / TB [13].

The last described method is especially interesting, because it is a component of a trend that evolves into personalized, prescient medication. In conducting this report, we performed a comprehensive review of the distinctive styles of machine learning methods being used, the forms of knowledge being organized, and the application of these strategies in development modelling and visualization. In comparison, multiple transmitted reports tend to be low of an acceptable standard of acceptance or checking [14].

This paper proposes using well-known ML methods such as Decision Tree, Logistic Regression, Support Vector Machines (SVM), K Nearest Neighbour (KNN) and ensemble arrangement to mechanically design an intellectual diagnostic assistant for forecasting various allergic diseases all over Turkey [15].

In the given method, 40 digital images are gathered using AOCD unit and MIT unit database respectively. The multi-SVM is arrangement that is supervised which deals with methods that are used to classify images for classification processing. The diagnostic method includes two phase phases such as preparation and research, the preparation data set's function values are matched with each type's test data set [16].

Machine learning is a developing area of data science that deals with multiple method that help machines to learn from occurrences. The main goal of the project is to create a program that can achieve early diabetes forecasting for patients with greater precision by integrating the effects of numerous methods in machine-learning. The goal is to forecast diabetes with the help of three various supervised learning algorithms: ANN, SVM and Logistic regression. It also deals with the earliest detection of diabetic disease [17].

It needs tons of testing in modern medical diagnosis which may hinder the detection of illness. Therefore, the strategies of mining will allow medical experience to render the diagnosis judgment utilizing a machine-aided judgment that support a program.Comprehensive study on different mining methods used to predict disease that are described in the article[18].

The aim of this challenge is to develop a system that takes a cancer medicinal dataset as a knowledge and then conducts data set analysis to produce findings that lets therapeutic researchers understand the state of the disease. The paper focuses on the data that obtained from pre-treatment stage in order to fix the coarse details and arrange their previous details of the table, and prediction cancer in initial stages which is provided as feedback to K-Nearest Neighbour detection method [19].

The paper evaluates different ML method to find the efficiency of such methods during their initial dengue detection. Mining is detection method for illnesses such as, dengue. To evaluate and equate the findings Weka toolbox is used [20].

The article, deals with some strategies of picture segmentation that identify acne lesions and strategies of ML, which differentiate specific acne lesions. Our findings revealed that between k-means, texture study, HSV model segmentation methods, two k-means clusters better than others with a reliability of approximately 70% [21].

An article uses wide dataset of Maharashtra uses two common mining allotment techniques ANN and SVM for forecasting malaria. Data is recognized for all 35 districts of Maharashtra, from 2011 to 2014. As a result, it is noted that SVM is more précised than ANN.SVM takes only 15-20 days to forecast the outbreak. Hence, the accuracy can be increased by taking more training into consideration. At country level the system can be implemented [22].

Using large-scale computing methods to mine unstructured medical data. They applied a multimodal neural convolutionary network to both structured and unstructured classification results. We obtained higher precision than

the current approaches, and the pace at which the algorithm was reached was therefore quicker in comparison than some. Latent factor analysis was used to renovate incomplete data [23].

In this paper they have developed machine learning algorithms to accurately forecast infectious disease epidemics in populations with repeated diseases. We are proposing a new multi-modal disease forecasting method for convolutionary neural networks, using unstructured and structured medical data [24].

3. PROPOSED METHOD

In this proposed we can get the large volume of a healthcare big data, then the data considered as training data. Decision Tree method forecast diseases and sub disease. Map reduced methods are carried out to enhance operational competence. It minimizes query improvement time. Accuracy is bettered using different machine learning methods. The considered system starts with the idea that was net accomplished by the antecedent. It arranges specific ratios for definite client pattern out his conditions. Hence, making our application to all at economical cost.

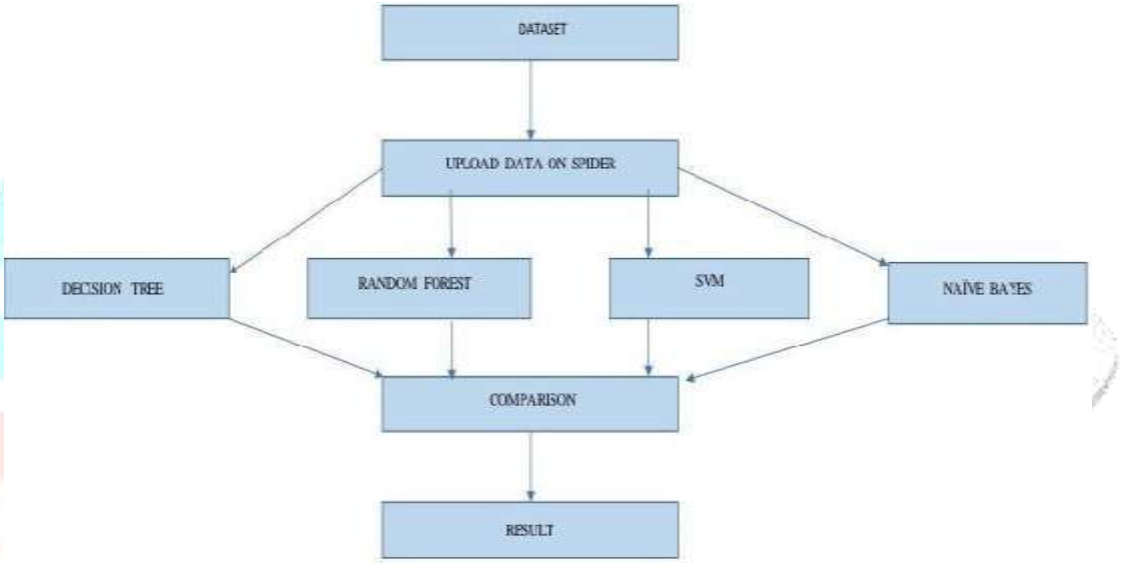

Fig (a). Flowchart of workflow

4. PRE-PROCESSING STEPS

Step 1: Import libraries

Step 2: Import the dataset required:

Involve gathering of medical information artifacts from several sources like hospitals, discharge slips of patients and from UCI repository

Step 3: Taking consideration of absent data in dataset:

It will remove all the unnecessary data and extract important features from data.

Step 4: Encoding categorized data

Step 5: Categorized the dataset into training and testing set:

model will be trained on the dataset of diseases to do the prediction accurately and produce Accuracy.

Step 6: Feature Scaling

5. IMPLEMENTATION

Machine learning predicts diseases but it can't forecast sub types of disease which are caused due the incidence of any one disease. It is not able to predict all likely condition of the individuals. It deals with only organized data. The standing administrations organize a combination of learning algorithms which thoughtfully are in envisioning diseases. Therefore, limitations are spotted with the predominant system. Firstly, Prevailing the organizations which are dearer to some of the people who could be paid to these calculation systems.

In addition, the guess method is both casual and infinite such that a computer can forecast a positive disease but cannot presume the sub-class of the disease that will decide the truth of a single bug.

Implementation Methods

5.1 Decision Tree

These models generally are used for data mining and to study statistics and to bring the tree as well as its guidelines that are need to do predictions. The forecasting can be done on the absolute values, when cases are located in the group's Decision tree[2] is basically a division which is used to build tree structure in which each node is child node, which is representing the functionality of the objective attribute or division of the instances, or decision node, conforming a few test which can be approved out on a solo characteristic-value, having one barrier with sub-tree for individually thinkable result on test .It is used to classify a case that starting from base of tree ,affecting it to attain child node, which can afford the group of the illustration.

5.2. Random Forest

It is a supervised method which deals with arrangement and regression. Mainly, it is considered for arrangement problems. As we are aware that a forest is collection of trees which meant vigorous forest. Alike random forest [1] method generate decision trees on statistics trials and then becomes the forecast after a piece of them and in conclusion it chooses the best solution for voting. A cooperative approach is improved than an individual decision tree as it reduces the over fitting by averaging the conclusion.

5.3 SVM (Support Vector Machine)

It is supervised algorithm which scans the data which can be used for arrangement and regression process. On the basis of the training examples, each noted as associated to more than one group. It enhances the model that allot it to another model and creating a non-probabilistic binary line division. SVM [1] illustrates the points in space, plotted on the distinct groups that are separated by a clear gap. New illustrations are then plotted using the similar spaces and projected to a class created on the gap on which they will lie.

5.4 Naive-Bayes:

It is a transparent method used for creating divisions: imitation that allows class documents to problematic cases, which indicates some detailed value, where the division documents are taken out from limited set. There's no method to train divisions, but a group of methods which are based on a mutual opinion. The divisions Naive Bayes [11] find that the importance of a single function is independent of the importance of any additional function defined by the division variable. We consider an apple that can be red, round and with diameter 10 cm. It considers each feature to share it independently to the possibility that this is an apple, irrespective of any probable connections on the basis of colour, roundness, and diameter trait.

Accuracy: Accuracy is one metric for estimating arrangement models. Casually it is the fraction of forecasts our model got right. it is the total correction prediction to the total number input samples.

Kappa Coefficient: It is used to control only those instances that may have been correctly classified by chance. = (total precision – random precision) / (1-random precision)

Confusion Matrix: This table is frequently used to determine the administration of an arrangement representation for a set of test data which us having known value.

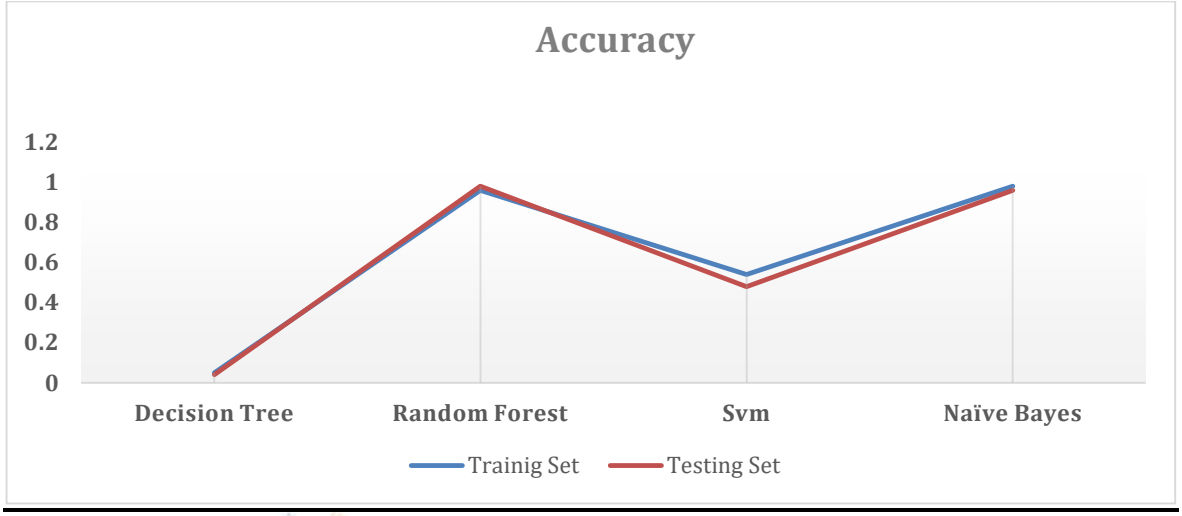

Fig. (b) Accuracy Chart

6. RESULTS

Table 1. Accuracy based on various algorithms

ALGORITHMS	Training Set Accuracy	Testing Set Accuracy
Decision Tree	0.46	0.55
Random Forest	0.95	0.95
SVM	0.48	0.54
Naive Bayes	0.93	0.93

7. CONCLUSION

Different types of methods are used to summarise the existence of various diseases. Governing the routine of each method and apply it to a particular area where it is needed. Use more significant trait assortment process to recover the accuracy and execution of method.

In conclusion, only a bordering progress is attained by the formation of analytical model for numerous illnesses of patients, hence it is concluded that we need combinational and complex representations to raise the efficiency for predicting the initial arrival of diseases. So, we will feed large amount of data to the database and it will get improve with the time. There are multiple methods to enhance the efficiency in addition of scalability of the system. With time restriction, subsequent analysis can be done for future reference. We will be making the use of different discretion techniques, and various electing methods and decision tree forms i.e. information gain in addition of gain ratio. We are willing to discover other rules like clustering algorithms, association rule and logistic regression.

8. REFERENCES

[1] Animesh Hazra, Subrata Kumar Mandal, Amit Gupta, Arkomita Mukherjee and Asmita Mukherjee, "Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review", Advances in Computational Sciences and Technology ISSN 0973-6107 Volume 10pp. 2137-2159,2017.

[2] R. Vijaya Kumar Reddy, K. Prudvi Raju, M. Jogendra Kumar, CH. Sujatha and P. Ravi Prakash, "Prediction of Heart Disease Using Decision Tree Approach", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 3, 2016.

[3] N. Deepika and K. Chandra Shekar, "Association rule for classification of Heart Attack Patients", International Journal of Advanced Engineering Science and Technologies, Vol. 11, No. 2, pp. 253 – 257, 2011.

[4] Animesh Hazra, Arkomita Mukherjee, Amit Gupta, Asmita Mukherjee, "Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review", Research Gate Publications, pp.2137-2159,2017

[5] T. Mythili, Dev Mukherji, Nikita Padaila and Abhiram Naidu, "A Disease Prediction Model using SVM-Decision Trees- Logistic Regression (SDL)", International Journal of Computer Applications, vol. 68, 2013

[6] Shadab Adam Pattekari and Asma Parveen, "Prediction System for Disease using Naive Bayes", International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624. Vol 3, Issue 3, pp 290-294, 2012.

[7] Chen, Yixue Hao, Kai Hwang, Fellow, IEEE, Lu Wang, and Lin Wang, Disease Prediction by Machine Learning over Big Data from Healthcare Communities, 2169-3536 (c) IEEE, 2017.

[8] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications, vol. 17, no. 8, pp. 0975-8887, 2011.

[9] Dhomse Kanchan B. "Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis", MET's Institute of Engineering, IEEE, INSPEC Number: 16980418, 2016.

[10] Oubida Alaoui Mdaghri, Mourad El Yadari, Abdelillah Ben youssef, Ab-dellah El Kenz Faculty of Science Rabat Morocco, Rabat, Study and analysis of Data Mining for Healthcare IEEE ,2016.

[11] Chen, Y. Ma, J. Song, C. Lai, B. Hu, Smart Clothing: Connecting Human with Clouds and Big Data for Sustainable Health Monitoring, ACM/Springer Mobile Networks and Applications Vol. 21, No. 5, pp.825C845, 2016.

[12] Shoon Lei Win, Zaw Htike, Faridah Yusof, Ibrahim A. Noor batch," cancer recurrence prediction using machine learning", International Journal of Computational Science and Information Technology, Vol.2, No.2, 2014.

[13] Ashwini D.V and Dr. Seema S, "Machine Learning Approach to Detect Tuberculosis in patients with or without HIV co- infection", International Journal of Computational Science and Information Technologies, Vol 6(3), 2015.

[14] Mehta Banu H, "Liver Disease Prediction using Machine-Learning Algorithms", International Journal of Engineering and Advanced Technology (IJEAT), Volume-8 Issue-6, 2019.

[15] Sevinç İlhan Omurca, Ekin Ekinci, Bengisu Çakmak, Selin Gizem Özkan, "Using Machine Learning Approaches for Prediction of the Types of Asthmatic Allergy across the Turkey ", data science and applications, Vol. 2, No.2, 2019

[16] S. Reena Parvin, O.A. Mohamed Jafar, "Prediction of Skin Diseases using Data Mining Techniques", International Journal of Advance Research in Computer and Communication Engineering, Vol.6, Issue 7, 2017.

[17] Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques", Int. Journal of Engineering Research and Application, Vol. 8, Issue 1, (Part -II) 2018.

[18] Ahelam Tikotikar, Mallikarjun Kodabagi, "A survey on technique for prediction of disease in medical data", International Conference on Smart Technology for Smart Nation, 2017.

[19] Dr.B. Santhosh Kuma1, T. Daniya, Dr. Ajayan, "Breast Cancer Prediction Using Machine Learning Algorithms", International Journal of Advanced Science and Technology, Vol. 29, No. 03, 2020.
[20] N.Rajathi, S.Kanagaraj, R.Brahmanambika and K.Manjubarkavi, "Early Detection of Dengue Using Machine Learning Algorithms", International Journal of Pure and Applied Mathematics, Vol. 118, No.18, 2018.
[21] Nasim Alamdari, Kouhyar Tavakolian, Minhal Alhashim, MD FAAD, and Reza Fazel-Rezai, "Detection and Classification of Acne Lesions in Acne Patients: A Mobile Application", 2016.
[22] Vijeta Sharma, Ajai Kumar, Lakshmi Panat, Dr. Ganesh Karajkhede, Anuradha lele , "Malaria Outbreak Prediction Model Using Machine Learning", International Journal Of Advanced Research in Computer Engineering And Technology, Vol.4, Issue 12, 2015.
[23] Hlaudi Daniel, Masethe Mosima and Anna Masethe Prediction of "Heart Disease using Classification Algorithms" [Journal] // World Congress on Engineering and Computer Science (WCECS) - Vol. II, 2014.
[24] Richard Osuala and Ognjen Arandjelovic University of St Andrews, United Kingdom, Visualization of Patient Specie Disease Risk Prediction IEEE, 2017.