International Journal of Engineering Sciences & Research Technology

(A Peer Reviewed Online Journal)
Impact Factor: 5.164

IJESRT

RESEARCHERID
THOMSON REUTERS

[Tomy *et al.*, 9(11): November, 2020]
IC™ Value: 3.00

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A REVIEW ON MACHINE LEARNING TECHNIQUES FOR ADVANCED HEALTH CARE SYSTEMS

Tessy Tomy[*1], Ashok K[2] & Monika Gupta[3]

[*1,2&3]Department of Electronics and Communication, New Horizon College of Engineering, Bengaluru, Karnataka, India

DOI: https://doi.org/10.29121/ijesrt.v9.i11.2020.1

ABSTRACT

Artificial intelligence is the technology that lets a machine mimic the thinking ability of a human being. Machine learning is the subset of AI, that makes this machine exhibit human behavior by making it learn from the known data, without the need of explicitly programming it. The health care sector has adopted this technology, for the development of medical procedures, maintaining huge patient's records, assist physicians in the prediction, detection, and treatment of diseases and many more. In this paper, a comparative study of six supervised machine learning algorithms namely Logistic Regression(LR),support vector machine(SVM),Decision Tree(DT).Random Forest(RF),k-nearest neighbor(k-NN),Naive Bayes (NB) are made for the classification and prediction of diseases. Result shows out of compared supervised learning algorithms here, logistic regression is performing best with an accuracy of 81.4% and the least performing is k-NN with just an accuracy of 69.01% in the classification and prediction of diseases.

KEYWORDS: Artificial intelligence · health care · Performance · Machine learning · Supervised Learning.

1. INTRODUCTION

The existence of AI has been known for a long time but it was in late 1950's its true possibilities were explored [1], thereon evolving rapidly and is now well knitted into our lives. The ability of a human being to obtain knowledge and apply it suitably on integrating with the cognitive skills is called intelligence. Now, when this intelligence is made to be exhibited by a machine, its called Artificial intelligence [2,3].

AI already has its deep roots laid in the fields of image processing, robotics [4,5], data mining, text analysis and so on and is now revolutionizing the domain of health care with remarkably performing algorithms. These algorithms can predict, understand learn and can act on the data available. The exciting promise of AI in health care is also due to the availability of sufficient medical data and various developed analytical techniques to work on it [6]. In the ongoing pandemic of COVID-19 also, this technology is essentially made use for the detection, diagnosis and for the patient monitoring [7,8].

Implementation of AI in health and medicine is achieved by several techniques of which machine learning, robotics and deep learning has its major share of contribution towards the diagnosis, assistance and collection of health records respectively [9-13].We are mainly focusing on the machine learning algorithms and their role in health care domain. Section 2 describes the importance of data in machine learning and its organization. Section 3 deals with the broad categories of machine learning algorithms. Section 4 focuses on the six different algorithms of supervised machine learning. Section 5 discusses the performance measuring metrics of the algorithms resulting into a comparative analysis of the algorithm.

2. MACHINE LEARNING:DATA AND ITS IMPORTANCE

Machine learning is a form of AI that lets machine to mimic the thinking ability of humans by exploring, identifying, and learning from data. The algorithm builds the machine learning models to train the machine on the data it is fed with. As the algorithm adaptively improve their performance as the number of samples available for learning increases, it is important to keep a check on the reliability, accuracy and volume of data [14,19]. The data available exists in either structured or unstructured form.

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

Structured Data
Structured data are the information which are stored and displayed in a well- organized manner and is also easily accessible. This might include attributes such as the patients' gender, age, height and results obtained from previous medical examination and history of any other ailments [20].

The quality and consistency of structured data needs to be maintained, with several standards which would determine the way the data is recorded and stored. This may help in the data comparison, interpretation and even for display within the electronic medical records (EMR) of a patient. Suppose, if an EMR reads a data with two parts as" height 65" meaning height is 65 inches, with no proper standards the variable name may be interpreted incorrectly, depending upon where and who is entering the data. Therefore, the labeled data needs to be carefully chosen [21]

Unstructured Data
Unstructured Data lacks the data organization, thereby making it less identifiable. It may include the data like "patient-doctor conversation", diagnosis images etc. [20]. The study shows that the health care data distribution is 20% structured and 80% unstructured [15]. This calls for the need of various tools to process this unstructured information to make it usable. This task of processing sometimes is a challenge [17-20]. The use of unstructured data can achieve a new dimension in health care data analytic. The analysis of medical images might require a robust computational environment. This needs to be explored as medical images form a significant part of patient's disease history. This may lead to an early detection and treatment helping to deliver service on time to the needy.

3. CATEGORIES OF MACHINE LEARNING ALGORITHM

As the machine learn to improvise itself from the training data without the need of being explicitly programmed, they may be broadly categorized into four major types of learning [21-23]

Supervised Learning
Under this Supervised learning, the algorithms are trained with labelled set of data and then applied on unlabeled data set to sort them into class having common traits. They are widely used for the predictive analysis of diseases [19].

Unsupervised Learning
Unsupervised produces untrained algorithm, which acquires the feature learning from the unlabeled data set itself and uses this learning on new data set to identify the class of data.

Semi-Supervised learning
When the labeled data set available for learning is not sufficient to build the model, this algorithm also uses the unlabeled data to have sufficient data for training. Therefore, it exhibits the mix behavior of supervised as well as unsupervised learning.

Reinforcement learning
The algorithm learns to take the best possible classification action on interacting with the potentially complex environment, after several trial-error attempts and result evaluation [25]. Fig.1 collectively describes the defining features of above-mentioned learning types.

4. SUPERVISED LEARNING ALGORITHMS

As discussed, these algorithms are mainly used in the domain of health care for the disease prediction and diagnosis, where the model is trained first using the labelled data i.e., data with proper feature definition, and then applied on new data for the expected outcome [19]. All the instances of data are represented with attributes having a value. Supervised models predict values or labels of new instances with their given attributes [29]. Fig.2 explains it with an example. This has again two variants i.e., the model generating a discrete predicted value will be called the classification and the one generating a continuous predicted value is called regression [19].

The Six most used supervised ML algorithms are explained here for their way of classifying of data.

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

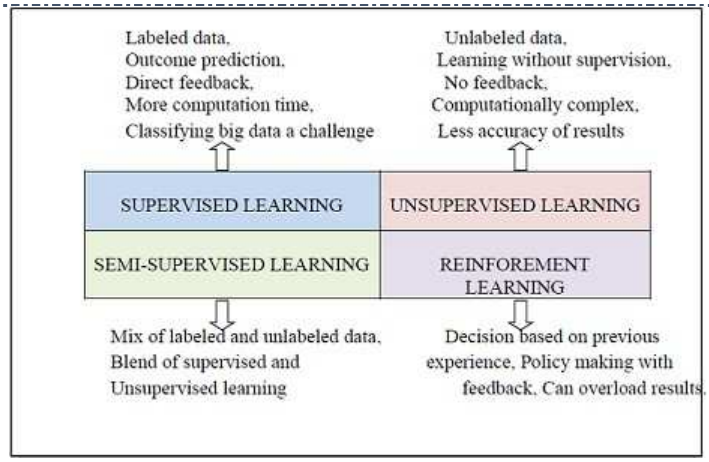

Fig. 1. Categories of machine learning

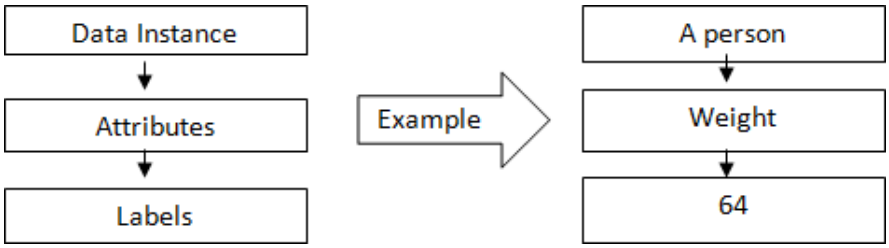

Fig. 2. Example illustrating supervised machine learning

Logistic Regression (LR)
Logistic Regression is a common binary classifier used to categorize an instance into a direct '0' or '1' class or in health care terms, into a class of 'diseased' or 'not diseased' class. Extension of this algorithm called the multi nominal logistic regression is used to evaluate a relation between one dependent and one or more independent variable resulting into a probable value between '0' or '1', thereby raising a need to set a threshold to bring the outcome value nearer to the well-defined classes [26]

Support Vector Machine (SVM)
SVM is another potential classifier algorithm, which can work evidently even with small set of attributes values and performs the classification without making any hypothesis on the data distribution and their inter-dependency [27]. Classification is done by constructing a hyper plane with maximum marginal distance that separates the data into two distinct classes, resulting in reduced error of classification [24]

Decision Tree (DT)
Decision Tree algorithm works on the principle of nodes and branches where every node corresponds to a attribute in a set or group that needs to be classified and every branch refers to a value, the node can act on. Traversing through the nodes of the tree to test for its condition for classifying the input sample, helps in collecting more information about the class [29]

NaïveBytes (NB)
The architecture of this algorithm, adapts the conditional probability concepts [23] which states that the probability of an 'happening' depends upon the fore- hand knowledge of the condition in connection with the event. This algorithm takes in fact that no attribute in a class has inter-dependency on one another, therefore changing the value of one feature does not directly affect the other in the group [27]

K-nearest neighbor(k-NN)
Although this algorithm can be used in the classification as well as regression problems, study shows its more effectiveness in the classification problem. Here the new data is classified into the well-defined classes with its

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

predicted values from the closely matched neighboring value in the training set. The value of 'k' signifies, the number of neighbors, the new data would consider performing the classification [29]. This prediction scheme may also help in the disease diagnosis, by grouping the diseases showing similar symptoms.

Random Forest (RF)

These can be visualized as the collection of several decision trees. As decision- trees are sensitive with respect to their training data, it might incur classification error, even for small change in input data, especially when the trees are deeply grown. In RF, the set of inputs from the data to be classified, is sent through various DT's and each DT gives its classified outcome after operating on few attributes from the input data. Then RF classifies it by considering either the most number of votes obtained for an outcome of DT or by taking average of trees under the RF [24].

Accuracy

It is the rate of classification i.e., how accurate is the algorithm in predicting the diseases.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{TP+TN}{TOTAL\ NUMBER\ OF\ SAMPLES} \quad \text{(i)}$$

Precision

This metric contributes towards the measurement of accuracy of an algorithm
High precision means the instance labeled as positive is actually positive.

Recall/sensitivity

This metric defines the sensitivity or completeness of the algorithm.

$$\text{Recall/sensitivity} = \frac{TP}{TP+FN} \quad \text{(iii)}$$

Correct classification is indicated by a high recall/sensitivity value.

F-measure

This helps to have a measurement that represents both of precision and recall. It is the weighted average of both of them.

$$\text{F-measure} = \frac{2*Precision*Recall}{Precision+Recall} \quad \text{(iv)}$$

Specificity

It is the metric that measures the true negative values [42]

$$\text{Specificity} = \frac{TN}{TN+FP} \quad \text{(v)}$$

Table 1. Performance Comparison of Algorithms (in percentage)

Algorithm	Logistic Regression (LR)	Support Vector Ma-chine (SVM)	Decision Tree (DT)	Naive Bytes (NB)	k-nearest neighbor (k-NN)	Random Forest (RF)
Accuracy	81.4	73.61	75.9	72.44	69.01	79.81
Precision	85.5	67.4	78.10	70.00	71.6	83.45
Recall/sensitivity	82.6	77.7	70.79	68.31	71.34	81.45
F-measure	82.8	69.1	75.71	67.5	71.8	82.3
specificity	69.6	58.6	74.75	67.65	65.9	77.8

On comparing the Logistic Regression(LR), Support Vector Ma-chine(SVM), Decision Tree(DT), NaiveBytes(NB), k-nearest neighbor(k-NN), Random Forest(RF) for its performance in the classification and prediction of diseases such as liver, heart, diabetes and all other diseases in general, logistic regression algorithm showed the best

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

performance with an accuracy of 81.4 % and a k-NN gave the poorest performance of 69.01 %. All the metrics here are calculated by considering the average of all individuals researches already made [31-41].

5. CONCLUSION

As health care data is very sensitive, these algorithms needs to be chosen after a very careful investigation, for its different performance validating measures. Here only the broader level classification algorithms are considered. The variants of the algorithm considered here may yield better results. These algorithms can be subjected to other cross-validation techniques for its accuracy measurement. As available of unstructured data is in abundance, it could be a game changer in the health care domain. The need for proper overall standardization in maintaining digital healthcare data is an evident need

REFERENCES

[1] Jahanzeb, Shabbir., Tarique, Anwer.: Artificial Intelligence and its Role in Near Future. Journal of Latex Class Files, Vol.14, Issue 8 (2015)
[2] Avneet, Pannu.: Artificial Intelligence and its Application in Different Areas. Inter- national Journal of Engineering and Innovative Technology (IJEIT) Vol.4, Issue 10 (2015)
[3] Harjit, Singh.: Artificial Intelligence Revolution and India's AI Development: Chal- lenges and Scope. IJSRSET, Vol.3, Issue 3 (2017)
[4] Xin, Zhang., Wang, Dahu.:Application of artificial intelligence algorithms in im- age processing. Journal of Visual Communication and Image Representation, Vol.61, Pages 42-49, (2019)
[5] Tran., Bach., Xuan.: Global Evolution of Research in Artificial Intelligence in health and Medicine: A Bibliometric Study. Journal of clinical medicine vol. 8,(2019)
[6] Jiang.,Fei, Jiang.,Yong, Zhi.: Artificial intelligence in health care: past, present and future. BMJ. 2. svn. 10.1136/svn-2017-000101 (2017)
[7] Raju, Vaishyaa., Mohd,Javaid.:Artificial Intelligence (AI) applications for COVID- 19 pandemic.Elsevier, Diabetes Metabolic Syndrome: Clinical Research Reviews, Vol.14, Issue 4 (2020)
[8] Mei, X., Lee, H., Diao, K.: Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. Nat Med (2020)
[9] T, J, Wroge., Y, O¨ zkanca., C, Demiroglu.: Parkinson's Disease Diagnosis Using Machine Learning and Voice. 2018 IEEE Signal Processing in Medicine and Biology Symposium (SPMB), pp. 1-7, (2018)
[10] Tanveer., Richhariya., Bharat.,Khan.: Machine Learning Techniques for the Diag- nosis of Alzheimer's Disease: A Review. ACM Transactions on Multimedia Comput- ing, Communications and Applications. 16. 35. 10.1145/3344998 (2020)
[11] Davatzikos, C.:Machine learning in neuroimaging: Progress and challenges. Neu- roImage, 197, 652–656 (2019)
[12] Heidingsfeld., Michael.,Feuer., Ronny.,Karlovic.: A Force-controlled Human- assistive Robot for Laparoscopic Surgery. IEEE International Conference on Systems, Man and Cybernetics, 10.1109/SMC.2014.6974460 (2014)
[13] Jose, Roberto, Ayala, Solares., Francesca, Elisa, Diletta, Raimondi.: Deep learning for electronic health records: A comparative review of multiple deep neural architec- tures. Journal of Biomedical Informatics,Vol. 101 (2020)
[14] Thomas, Davenport.,Ravi, Kalakota.: The potential for artificial intelligence in health care. Future health care Journal, Vol. 6, No 2: 94–8 (2019)
[15] Kong,H, J.: Managing Unstructured Big Data in health care System. Health care informatics research, 25(1), 1–2 (2019)
[16] P,K, Sahoo., S, K,Mohapatra., S, Wu.: Analyzing health care Big Data With Pre- diction for Future health Condition. IEEE Access, vol. 4, pp. 9786-9799,(2016)
[17] Yuji, Roh., Geon, Heo., Steven, Euijong, Whang.: A Survey on Data Collection for Machine Learning A Big Data - AI Integration Perspective. arXiv:1811.03402v2 [cs.LG] (2019)
[18] Esteva, A., Robicquet, A., Ramsundar, B.: A guide to deep learning in health care.,Nat Med 25, 24–29 (2019)
[19] Sidey,Gibbons, J., Sidey,Gibbons, C.: Machine learning in medicine: a practical introduction. BMC Med Res Methodol 19, 64 (2019)

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

[20] M, Chen., Y, Hao., K, Hwang.: Disease Prediction by Machine Learning Over Big Data From health care Communities. IEEE Access, vol. 5, pp. 8869-8879 (2017)
[21] https://health snap.io/structured-unstructured-health -data/
[22] K, Shailaja., B, Seetharamulu., M,A,Jabbar.: Machine Learning in health care: A Review. Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, 2018, pp. 910-914 (2018)
[23] Dey, Ayon.: Machine Learning Algorithms : A Review. (2016)
[24] Uddin, S., Khan, A., Hossain, M.: Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak 19, 281 (2019).
[25] Chao,Yu., Jiming, Liu., Shamim, Nemati.: Reinforcement Learning in health care: A Survey. arXiv:1908.08796 vol.4,(2020)
[26] Boateng., Ernest, Yeboah., Abaye, Daniel.: A Review of the Logistic Regression Model with Emphasis on Medical Research. Journal of Data Analysis and Information Processing. 07. 190-207. 10.4236/jdaip.2019.74012.(2019)
[27] Yu, W., Liu, T., Valdez, R.: Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC medical informatics and decision making, 10, 16 (2010)
[28] Sureskumar., Kalaiselvi.: Naive Bayesian Classification Approach in health care Applications. (2017)
[29] Cho, G., Yim, J., Choi, Y., Ko, J.: Review of Machine Learning Algorithms for Diagnosing Mental Illness. Psychiatry investigation, 16(4), 262–269 (2019)
[30] Kendale, S., Kulkarni, P., Rosenberg, AD.: Supervised Machine-learning Predictive Analytics for Prediction of Postinduction Hypotension. Anesthesiology,129(4):675- 688 (2018)
[31] Mohammad, Pourhomayoun., Mahdi, Shakibi.: Predicting Mortality Risk in Pa- tients with COVID-19 Using Artificial Intelligence to Help Medical Decision-Making. doi:https://doi.org/10.1101/2020.03.30.20047308 (2020)
[32] Kumar G, Dinesh.,K Arumugaraj.,Kumar D, Santhosh .: Prediction of Cardio- vascular Disease Using Machine Learning Algorithms. International Conference on Current Trends towards Converging Technologies (2018)
[33] Javad, Hassannataj, Joloudari., Hamid, Saadatfar., Abdollah, Dehzangi.:Computer-aided decision-making for predicting liver disease using PSO-based optimized SVM with feature selection. Informatics in Medicine Un- locked,Vol. 17 (2019)
[34] Deepti, Sisodia., Dilip, Singh, Sisodi.:Prediction of Diabetes using Classification Algorithms.Procedia Computer Science,Vol. 132,Pages 1578-1585 (2018)
[35] M,Banu, Priya1., P, Laura, Juliet2., P,R, Tamilselvi3.: Performance Analysis of Liver Disease Prediction Using Machine Learning Algorithms. International Research Journal of Engineering and Technology (IRJET) Vol.5 Issue 1 (2018)
[36] Jakka, Aishwarya.,Jakka, Vakula.: Performance Evaluation of Machine Learning Models for Diabetes Prediction. 10.35940/ijitee.K2155.0981119 (2019)
[37] Sakr,S., Elshawi, R., Ahmed, AM.: Comparison of machine learning techniques to predict all-cause mortality using fitness data: the Henry ford exercise testing (FIT) project. BMC Med Inform Decis Mak, 17(1):174 (2017)
[38] R, Kalaiselvi., G, Santhoshni.: A Comparative Study on Predicting the Probabil- ity of Liver Disease. International Journal of Engineering Research and Technology (IJERT) Vol.08, Issue 10 (2019)
[39] S, M, M, Hasan., M, A, Mamun.:Comparative Analysis of Classification Ap- proaches for Heart Disease Prediction. International Conference on Computer, Com- munication, Chemical, Material and Electronic Engineering (IC4ME2),pp. 1-4 (2018)
[40] Bashir., Saba,Khan., Zain, Khan.:Improving Heart Disease Prediction Using Fea- ture Selection Approaches. 619-623. 10.1109/IBCAST.2019.8667106 (2019)
[41] G, T, Reddy.:An Ensemble based Machine Learning model for Diabetic Retinopa- thy Classification. International Conference on Emerging Trends in Information Tech- nology and Engineering (ic-ETITE)pp. 1-6,(2020)
[42] Alaa,Tharwat.: Classification assessment methods,Applied Computing and Infor- matics, ISSN 2210-8327,(2018)