International Journal of Research in Circuits, Devices and Systems

E-ISSN: 2708-454X
P-ISSN: 2708-4531
IJRCDS 2022; 3(1): 77-83

www.circuitsjournal.com
Received: 12-11-2020
Accepted: 20-12-2020

Nagaraj S
School of Computing Science and Engineering, Vellore Institute of Technology, Chennai, Tamil Nadu, India

Correspondence
Nagaraj S
School of Computing Science and Engineering, Vellore Institute of Technology, Chennai, Tamil Nadu, India

The role of machine learning algorithms for diagnosing disease

Nagaraj S

Abstract
Nowadays, machine learning algorithms are playing very essential role in the medical sector, mainly for diagnosing disease from the medical database. Various companies are using these methods for the early detection of diseases and improve medical diagnostics. The main motivation or aim of this paper is to give an overall overview of the machine learning algorithms that are used for the determining and prediction of various diseases such as Naïve Bayes, logistic regression, support vector machine, K-nearest neighbour, K-means clustering, decision tree, and random forest.

Keywords: Disease diagnostics, machine learning, classification algorithm

1. Introduction
Diagnosis is the way of determination of disease, some signs are non-particular, and thus the major demanding task is the disease diagnosis pointing out disease is the important thing for the treatment of any kind of diseases. Machine learning is the field that can able to forecast the disease diagnosis based on the prior training data. Many of the scientists across the global have created various methods of machine learning to diagnosis numerous diseases. Machine learning offers the possibility for machines to learn without being specifically programmed. Implementation of model by algorithm of machine learning can predict a primitive stage of disease diagnosis. The diagnosis and proper treatment are the best method to keep down the death rate by any disease, hence many of the scientists are Implemented new technical models for prediction of disease by machine learning Algorithm [1, 2].
Machine learning models learn from patterns in given training examples without explicit instructions and then use influence to develop useful predictions. The major health problems diseases like HIV, Blood Cancer, Breast Cancer, diabetes and heart syndrome impacts persons health greatly and even may lead to death. Due to greater improvement or progress of Artificial intelligence and machine learning, a handful of classifiers and clustering algorithms such as support vector machine (SVM), Decision Tree, Naive Bayes, k-nearest, Random forest are some of the algorithms than can give solution to this circumstance [3].
In medical applications, the classification 0f machine learning techniques is done because it related to problems in day-to-day life. Classification algorithms initially use the training data to build its model and then obtained model is subjected to test data to obtain proprietary predictions. These techniques can drastically minimize faults that occur in diagnosis and can able to obtain results in shorter time period.

2. Machine learning types
Machine learning which is the branch of an artificial intelligence that allows machine to think as same as human being and make a decision on their possess with intervention of human being. Machine learning is the process of automatically building machines to learn without being programmed. The main goal or aim of machine learning is to make a computer program that can able to access the training data and make it use for its learning process. There are many types of machine learning being exist as illustrated in Fig 1.

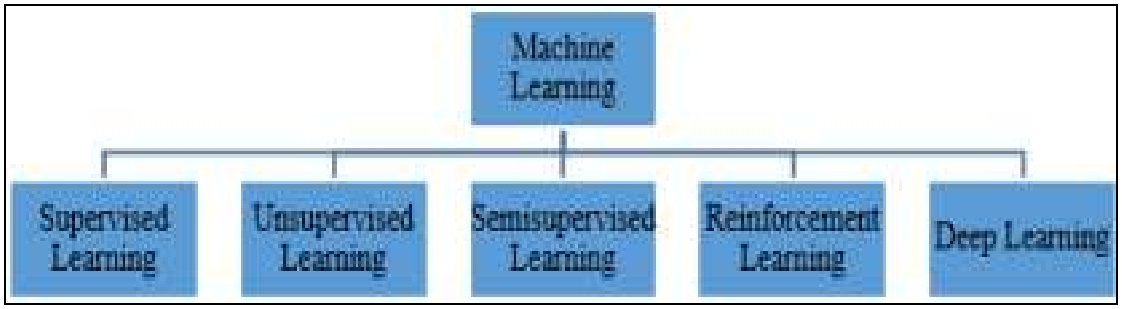

Fig 1: Types of machine learning

This paragraph would brief introduce each type of machine learning. The supervised machine learning means the algorithms learns to predict data from input data, and this type has both input and output data. In the case of unsupervised machine learning the algorithms has only the input data and build its model from only input data. And the semi-supervised machine learning type has dependency on both the types (supervised and unsupervised) [4]. They use both labelled and unlabelled it means some input have label data other do not have label data. In the case of reinforcement machine learning the system make attempt to learn through its interaction with the environment and reward the desired the action and punish undesired ones. The wide range of these machine learning application utilised in medical field such as disease diagnosis. The deep learning which is the subset of machine learning that consists number of layers that holding multiple stages of perception, such that every layer gets access the information from the previous layer and result would be passed to next layer.

3. Machine learning algorithms

The various machine learning algorithms are used in the identification of disease (disease diagnosis). some of them are discussed below.

A. K Nearest Neighbour Algorithm (KNN)

K Nearest Neighbour model (KNN) is the simplest model and one among widely used machine learning process for pattern recognition, classification of problems and regression. KNN access neighbours among the data using Euclidean distance points of data. This algorithm is utilised for regression and classification of problems. The value of k where k is constant) would identify all the same existing features cases with the new case and surrounded all cases to find new case for similar category [5].

Hence value of k is most significant and have to choose carefully because it may lead to system overfitting if it is very small. There are some limitations such as poor performance when there is extensive training data set. The general idea that represents K Nearest Neighbour algorithm for classification of new data point when value of k is equal to 9. The classification of new point is (0.6,0.45) as shown with the "X". the two possible classes presented in the large dotted circle as shown in Fig 2 which has three object of triangle class and the five circle class objects. Utilising the Euclidean distance, the algorithm going to classify the "X" which is the new point to the circle.

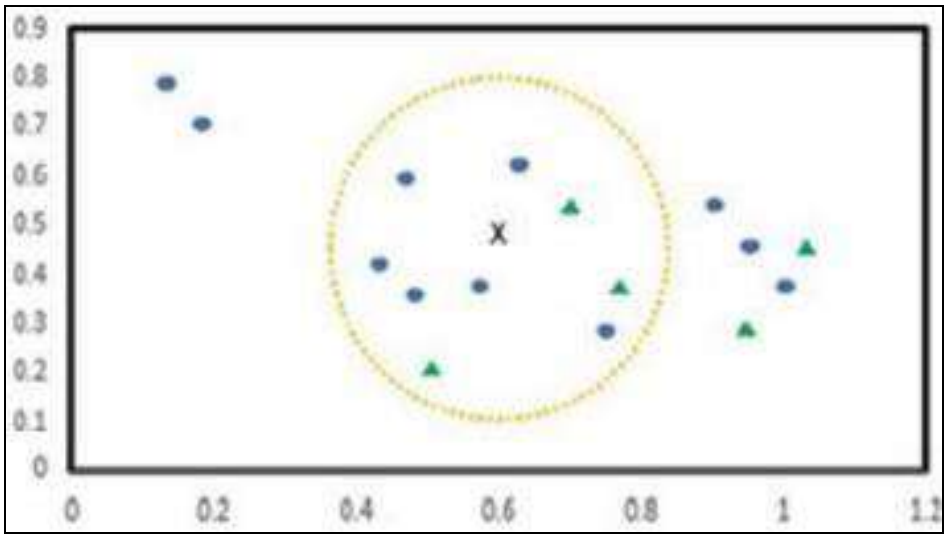

Fig 2: Large dotted circle

B. K-Means Clustering Algorithm

K-Means Clustering which is an unsupervised machine learning algorithm usually used for the clustering process based on nearest neighbour. The based on the similarity between them the data can clustered into K cluster where K is an integer Number and the value of K mush be known for the algorithm to operate. The K-Means is one of the widely used algorithm for clustering, and it has ability of detecting new data of right cluster according the majority of the distance. The choice of centroids of k cluster is randomly performed initially, then points to be allocated to its closer centroids and recomputed centroids for the newly formed assembled group. K-Means are sensitive particularly to outliers and noise, as few are influenced by centroids. The

one benefit of K-Means is the, this method can be easily implemented and its interpret and effective in computational terms. The drawback of this K-Means is the estimation or determining the value of k is difficult when the clusters are globular, efficiency suffers. The Fig 3 represent the K-Means algorithm graphically. In the first step two sets of objects are present. Then centroids of both sets will be decided. The clusters that generate the various dataset clusters are formed again according to the centroid. Until the best clusters are reached this process will be repeated.

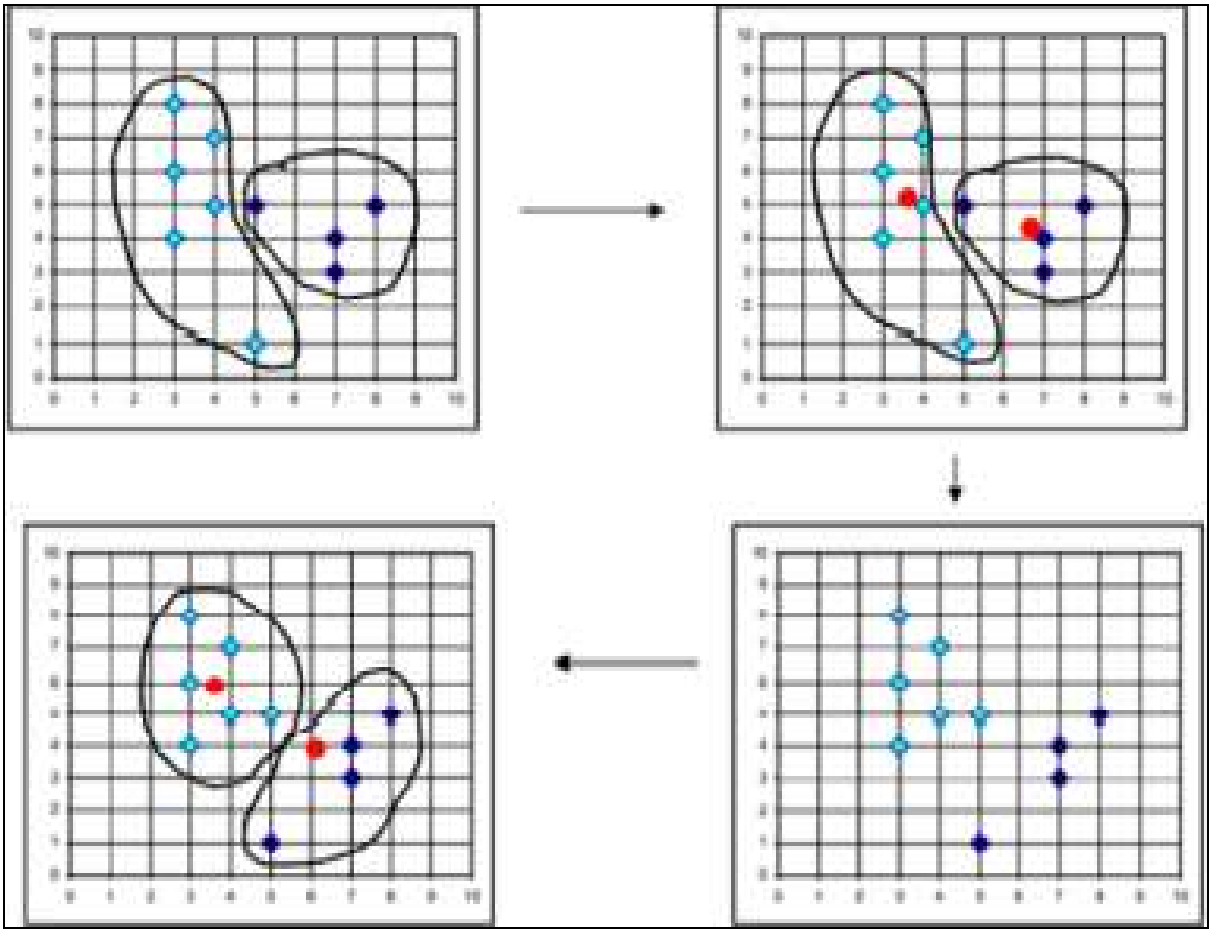

Fig 3: Represent the K-Means algorithm graphically

C. Support Vector Machine

The Support Vector Machine SVM has appear to be dominant for distinct classification problems. It strives to find the best hyperplane within classes by identifying the number of points on the class descriptors' edge. The distance between classes is called as the margin. The greater correctness for the classification can be achieved when there is elevated margin [6, 7]. The data points on the border are called support vectors. SVM is utilised for both classification of problems and regression. This proceed works well for solving a problem in the form of a linear and nonlinear dataset. SVM algorithm utilises several kernel types such as linear radial basis function (RBF), Polynomial, and Sigmoid for a prediction model. SVM works on high-dimensional space for features and selects the finest hyperplane for classification of data points into two classes. It is well organised for smaller and larger datasets that cannot be operated. Fig 4 illustrates a best example of the SVM algorithm utilising hyperplane for diagnosis of diabetes data.

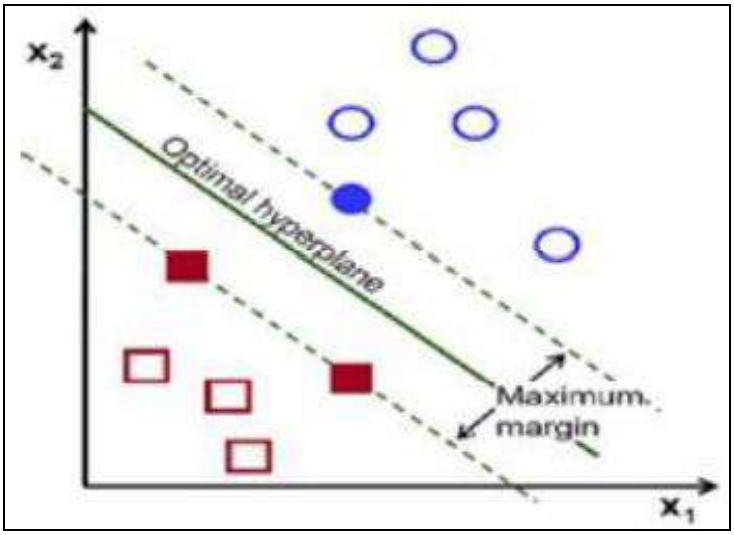

Fig 4: Illustrate SVM algorithm utilising hyperplane for diagnosis of diabetes data.

D. Naive Bayes algorithm

Naïve Bayes (NB) is a statistical technique based on classification algorithm. Naïve Bayes is standard excellent algorithm in a machine learning application due to its simplicity intelligibility in allowing all features to contribute equivalently to the final conclusion. Computational efficiency is equivalent to this simplicity, making the NB approach exciting and appropriate for distinct fields. The NB classification's key part is prior, posterior, and class conditional probability. This way has various benefits, such as easy for huge datasets and very useful. It could be used for binary and multiclass problems of classification. A tiny amount of training data is essential and can be utilised for both discrete and steady data. The application of this algorithm can be utilised for filtrate spam emails and classification of documents [9].

E. Decision Tree Algorithm

The Decision Tree (DT) is a supervised machine learning algorithm which is useful to solve regression and classification issues by continuously dividing data depending on a particular variable. The data is divided into the nodes, and the tree's leaf illustrate the concluding decisions. The motivation or purpose of the decision tree is to build a model that can be used to forecast the variable which is targeted by learning straightforward decision rules obtained from training data. The tree structure is created using the training data in a training process. The leaf nodes have the name of the class, and a decision node is a non-leaf node. The decision tree control categorical and numerical data. The nonlinear relationship within arguments does not affect the efficiency of the tree. No pre-processing of the data is needed. When the tree is frequently built, the probability of overfitting can happen [10, 11]. Fig 5 illustrates a simple decision tree that contain of the root node, one child node, and three leaf's nodes. One application of decision tree is utilised widely in the medical field. For example, using DT to find breast cancer, the leaf nodes of the tree are split into two groups (Benign or Malignant). To conclude whether the tumour is benign or malignant based on the Clump Thickness (CT), protocols will be defined within the selected data set characteristics. Fig 5 represents an example of using the DT algorithm to predict a breast cancer.

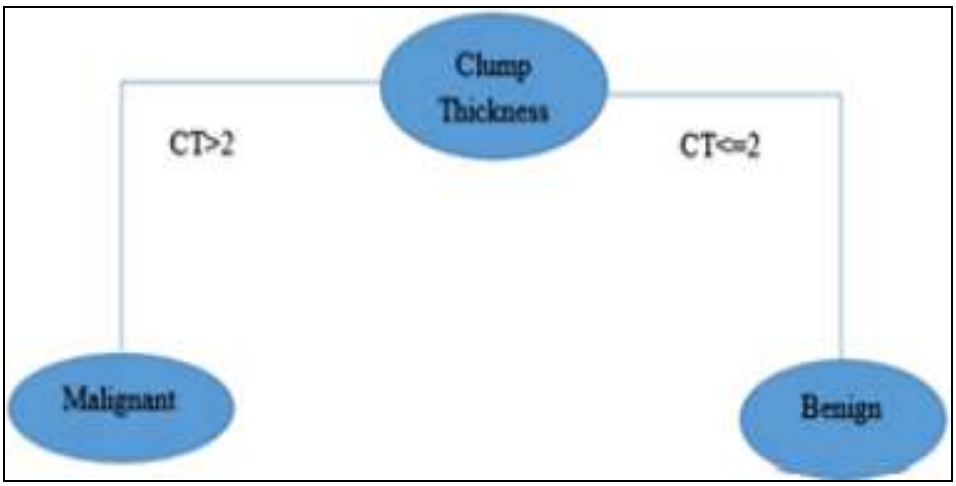

Fig 5: Represents an example of using the DT algorithm to predict a breast cancer.

F. Logistic Regression

Logistic regression is a supervised machine learning algorithm which is used for problem solving in the form of binary classification. Logistic regression is a mathematical model and used with logistic function to model binary classification, and for logistic regression, there are many

more complex extensions logistic regression is a regression model that identify that a specified data item or entry is probably to belong to a identified class using the regression model. Logistic regression uses a sigmoid function to model the data, as shown in Fig 6. Logistic regression has several principal or key points, like application simplicity, computational effectiveness, training-based effectiveness, ease of regularization [12, 13]. For input features, no scaling is required. Even so, the potential to resolve a nonlinear problem and susceptible to overfitting.

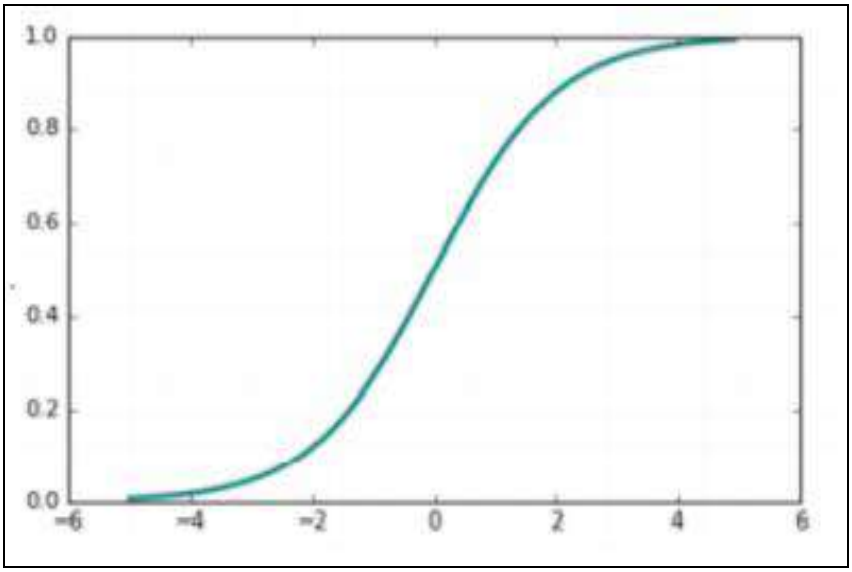

Fig 6: Logistic Function

G. Deep Learning

Deep learning is a subset of machine learning which involves a number of nonlinear transformations. Deep learning has numerous algorithms that can be able to learn/ to view input data from many layers of processing with complex structures. There are several DL architectures, such as recurrent neural networks (RNN), deep autoencoders, and convolutional neural networks (CNN) [14, 15]. These algorithms are used in distinct fields such as natural language processing, speech recognition, and medicals field. RNN is one of the deep learning algorithms which has internal memory to store the latest information usually, memory units in RNN architecture have the links to themselves, that moves information from the execution in the before. RNN changes the nature of the current forward process to adjust to the context of current input. The idea of the region of interest in MRI medical images generally distributed among many adjacent slices which results in having resemblance in succeeding slices [16, 17, 18, 19].

Note that researchers can found some other essential research points towards machine learning and deep learning in [19-30]. The articles [19-30] provide a detail description, issues and importance of such learning techniques in different application (in detail).

4. Analysis of Different Machine Learning Algorithms in a specific disease diagnosis

4.1 Heart Disease

SVM provide highest accuracy of 94.60% in 2012 as shown in Fig 7. In various application fields, SVM present best performance result. Attribute or features used by Parthiban and Srivatsa in 2012 are exactly responded by SVM. In 2015, Otoom *et al*. used SVM variant called SMO. It also uses FS technique to find good features. SVM replay to these features and provide the accuracy of 85.1% but it is relatively low as in 2012. Training and testing set of both data sets are distinct, and types of are different.

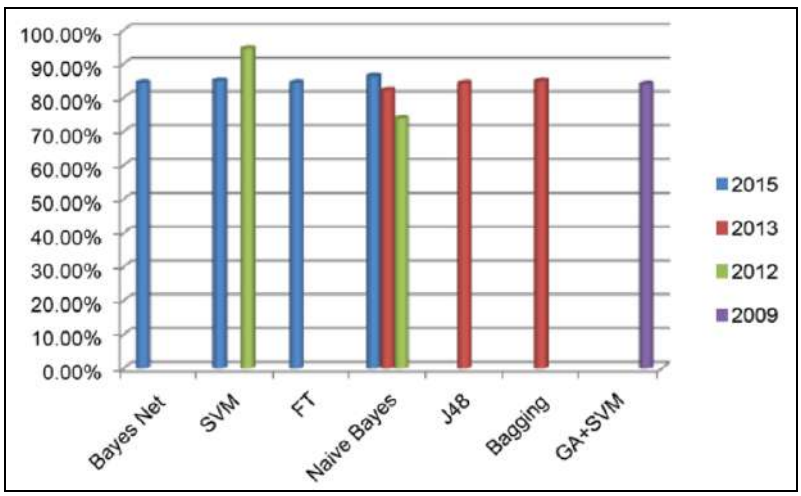

Fig 7: Machine learning algorithm's accuracy to detect heart disease.

Advantages and Disadvantages of SVM

Advantages: it will build proper classifiers and fewer over fitting, robust to noise.

Disadvantages: It is a binary classifier. For the classification of multi-class, it can use pair wise classification. And its Computational cost is very high, so it runs slow.

4.2 Diabetes Disease

Naive Bayes related system is useful for diagnosis of Diabetes. Naive Bayes offers provide maximum accuracy of 95% in 2012. The output show that this system can do good detection with less error and also this technique is essential to diagnose diabetes. But in 2015, accuracy given by Naive Bayes is less. It like 79.5652% or 79.57% accuracy. This given model for prediction of Diabetes disease want more training data for creation and testing. Fig.8 shows the Accuracy graph of Algorithms for the diagnosis of Diabetes disease according to time.

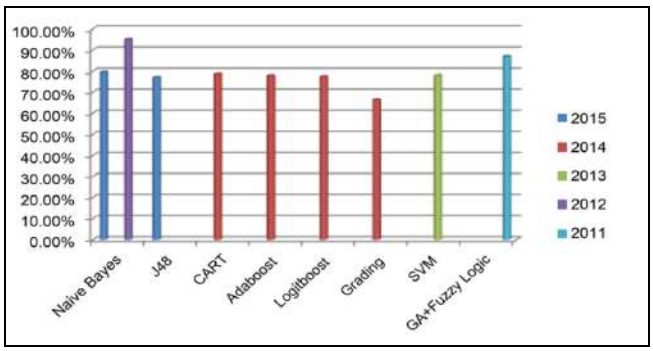

Fig 8: Accuracy of machine learning algorithms to detect diabetes disease.

Advantages and Disadvantages of Naive Bayes

Advantages: It intensify the classification performance by removing the unwanted features. And it has good performance. It usually takes less computational time.

Disadvantages: This algorithm wants huge amount of data to obtain best results. It is bad as they store entire the training examples.

Less over fitting wants best computational effort. Sample Size should be large as possible. And it is time taking. Engineering Judgment does not implement the relations between input and output variables so that the model acts like a black box.

5. Conclusion

Machine learning has come out with the medical field for providing tools and analyzing the data related to diseases. Hence machine learning algorithms play an important role in arriving the early prediction of diseases. This paper gives a review of a distinct machine learning algorithms for finding diseases, and standard datasets have been used in numerous diseases such as liver, chronic kidney, breast cancer, heart syndrome, brain tumours, and several other diseases. A result found by researchers has been tabulated to predict the diseases by ML algorithms. After comparing various papers for distinct models that predicted diseases, it shown that various algorithms have good accuracy for detection SVM, K-nearest neighbours, random forest, and the decision tree. Even, the precision of the similar algorithm can differ from one dataset to other because several essential factors affect the model's correctness and performance, like datasets, feature selection, and features. One more important point found in this review is that the model's accuracy and performance can be increased by using a different algorithm to build one ensemble model.

6. References

1. Shaik Razia P, Swathi Prathyusha N, Vamsi Krishna N, Sathya Sumana. A Review on Disease Diagnosis Using Machine Learning Techniques. International Journal of Pure and Applied Mathematics. 2017;117:16.
2. Diyar Qader Zeebaree, Habibollah Haron, Adnan Mohsin Abdulazeez, Dilovan Asaad Zebari. Trainable Model Based on New Uniform LBP Feature to Identify the Risk of the Breast Cancer. In 2019 International Conference on Advanced Science and Engineering (I.C.O.A.S.E.) (pp.). IEEE, 2019, 106-111.
3. Iswanto Iswanto, Laxmi Lydia E, Shankar K, Phong Thanh Nguyen, Wahidah Hashim, Andino. Identifying Diseases and Diagnosis using Machine Learning. International Journal of Engineering and Advanced Technology. 2019, 8.
4. Reem Alassaf A, *et al*. Preemptive Diagnosis of Chronic Kidney Disease Using Machine Learning Techniques. International Conference on innovations in Information Technology (I.T.), IEEE, 2018.
5. Gopi Battineni, Getu Gamo Sagaro, Nalini Chinatalapudi, Francesco Amenta. Applications of Machine Learning Predictive Models in the Chronic Disease Diagnosis. Journal of Personalized Medicine, 2020.
6. Diyar Qader Zeebare, Habibollah Haron, Adnan Mohsin Abdulazeez, Dilovan Asaad Zebari. Machine learning and Region Growing for Breast Cancer Segmentation. International Conference on Advanced Science and Engineering, IEEE, 2019.
7. Hossam Meshref. Cardiovascular Disease Diagnosis: A Machine Learning Interpretation Approach, International Journal of Advanced Computer Science and Applications. 2019;10:12.
8. Joel Jacob, Joseph Chakkalakal Mathew, Johns Mathew, Elizabeth Issac. Diagnosis of Liver Disease Using Machine Learning Techniques. International Research Journal of Engineering and Technology. 2018;05:04.
9. Siddhesh Iyer, Shivkumar Thevar, Priyamurgan Guruswamy, Ujwala Ravale. Heart Disease Prediction Using Machine Learning. International Research Journal of Modernization in Engineering Technology and Science. 2020;02:07.
10. Dilovan Asaad Zebari, Diyar Qader Zeebaree, Adnan Mohsin Abdulazeez, Habibollah Haron, Haza Nuzly Abdull Hamed. Improved Threshold Based and Trainable Fully Automated Segmentation for Breast Cancer Boundary and Pectoral Muscle in Mammogram Images, IEEE Access, 2020, 8.
11. Sneha Grampurohit, Chetan Sagarnal. Disease Prediction using Machine Learning Algorithms. International Conference for Emerging Technology (I.N.C.E.T.), Belgaum, India. 2020, 5-7.
12. Pahulpreet Singh Kohli, Shriya Arora. Application of Machine Learning in Disease Prediction. International Conference on Computing Communication and Automation (I.C.C.C.A.), IEEE, 2018.

13. Adnan Mohsin Abdulazeez, Baraa Wasfi Salim, Diyar Qader Zeebaree, Dana Doghramachi. Comparison of VPN Protocols at Network Layer Focusing on Wire Guard Protocol. IJIM. 2020;14:18.
14. Berina Alić, Lejla Gurbeta, Almir Badnjević. Machine Learning Techniques for Classification of Diabetes and Cardiovascular Diseases, 6th Mediterranean Conference on Embedded Computing, 2017.
15. Divya Jain, Vijendra Singh. Feature selection and classification systems for chronic disease prediction: A review, Elsevier, 2018.
16. Huseyin Polat, Homay Danaei Mehr, Aydin Cetin. Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods, published online springer, 2017.
17. Golmei Shaheamlung, Harshpreet Kaur, Mandeep Kaur. Survey on machine learning techniques for the diagnosis of liver disease, International Conference on Intelligent Engineering and Management, IEEE, 2020.
18. Diyar Qader Zeebaree, AdnanMohsin Abdulazeez, Dilovan Asaad Zebari, Habibollah Haron, Haza Nuzly Abdull Hamed. Multi-Level Fusion in Ultrasound for Cancer Detection Based on Uniform LBP Features, Computers, Materials & Continua. 2021;66(3)3363-3382.
19. Amit Kumar Tyagi, Aswathy Aghila SUG, Sreenath N. AARIN: Affordable, Accurate, Reliable and INnovative Mechanism to Protect a Medical Cyber-Physical System using Blockchain Technology IJIN. 2021;2:175-183.
20. Nair MM, Tyagi AK, Sreenath N. The Future with Industry 4.0 at the Core of Society 5.0: Open Issues, Future Opportunities and Challenges. International Conference on Computer Communication and Informatics (ICCCI). 2021, 1-7. Doi: 10.1109/ICCCI50826.2021.9402498.
21. Tyagi AK, Fernandez TF, Mishra S, Kumari S. Intelligent Automation Systems at the Core of Industry 4.0. In: Abraham A., Piuri V., Gandhi N., Siarry P., Kaklauskas A., Madureira A. (eds) Intelligent Systems Design and Applications. ISDA 2020. Advances in Intelligent Systems and Computing. Springer, Cham, 2021, 1351. https://doi.org/10.1007/978-3-030-71187-0_1
22. Goyal Deepti, Tyagi Amit. A Look at Top 35 Problems in the Computer Science Field for the Next Decade, 2020. 10.1201/9781003052098-40.
23. Amit Kumar Tyagi, Dr. Meenu Gupta, Aswathy SU, Chetanya Ved. Healthcare Solutions for Smart Era: An Useful Explanation from User's Perspective, in the Book "Recent Trends in Blockchain for Information Systems Security and Privacy, CRC Press, 2021.
24. Varsha R, Nair SM, Tyagi AK, Aswathy SU, Radha Krishnan R. The Future with Advanced Analytics: A Sequential Analysis of the Disruptive Technology's Scope. In: Abraham A., Hanne T., Castillo O., Gandhi N., Nogueira Rios T., Hong TP. (eds) Hybrid Intelligent Systems. HIS 2020. Advances in Intelligent Systems and Computing, Springer, Cham, 2021, 1375. https://doi.org/10.1007/978-3-030-73050-5_56
25. Tyagi Amit Kumar, Nair Meghna Manoj, Niladhuri Sreenath, Abraham Ajith. Security, Privacy Research issues in Various Computing Platforms: A Survey and the Road Ahead. Journal of Information Assurance & Security. 2020;15(1):1-16. 16p.
26. Ramesh Prasad Tharu. Multiple regression model fitted for job satisfaction of employees working in saving and cooperative organization. Int J Stat Appl Math 2019;4(4):43-49.
27. Madhav AVS, Tyagi AK. The World with Future Technologies (Post-COVID-19): Open Issues, Challenges, and the Road Ahead. In: Tyagi A.K., Abraham A., Kaklauskas A. (eds) Intelligent Interactive Multimedia Systems for e-Healthcare Applications. Springer, Singapore, 2022. https://doi.org/10.1007/978-981-16-6542-4_22
28. Mishra S, Tyagi AK. The Role of Machine Learning Techniques in Internet of Things-Based Cloud Applications. In: Pal S., De D., Buyya R. (eds) Artificial Intelligence-based Internet of Things Systems. Internet of Things (Technology, Communications and Computing). Springer, Cham, 2022. https://doi.org/10.1007/978-3-030-87059-1_4
29. Akshara Pramod, Harsh Sankar Naicker, Amit Kumar Tyagi. Machine Learning and Deep Learning: Open Issues and Future Research Directions for Next Ten Years. Book: Computational Analysis and Understanding of Deep Learning for Medical Care: Principles, Methods, and Applications, Wiley Scrivener, 2020.
30. Amit Kumar Tyagi, Poonam Chahal. Artificial Intelligence and Machine Learning Algorithms, Book: Challenges and Applications for Implementing Machine Learning in Computer Vision, IGI Global, 2020. Doi: 10.4018/978-1-7998-0182-5.ch008
31. Amit Kumar Tyagi, Rekha G. Challenges of Applying Deep Learning in Real-World Applications, Book: Challenges and Applications for Implementing Machine Learning in Computer Vision, IGI Global, 2020, 92-118. Doi: 10.4018/978-1-7998-0182-5.ch004.