Research Article

Machine Learning Algorithms for Detection: A Survey and Classification

[1]Deepa N. Reddy, [2]Priyanka R, [3]Sanjana S, [4]Santrupti. M. Bagali, [5]Sara Sadiya

[1]Department of Electronics and Communication, BMS Institute of Technology & Management, Bengaluru, India.
Email: reddydeepa2680@gmail.com
[2]Department of Electronics and Communication, BMS Institute of Technology & Management, Bengaluru, India.
Email: priyankaraju9480@gmail.com
[3]Department of Electronics and Communication, BMS Institute of Technology & Management, Bengaluru, India.
Email: sanjanadathathri123@gmail.com
[4]Department of Electronics and Communication, BMS Institute of Technology & Management, Bengaluru, India.
Email: santruptibagali@gmail.com
[5]Department of Electronics and Communication, BMS Institute of Technology & Management, Bengaluru, India.
Email: mnsara2288@gmail.com

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 28 April 2021

Abstract: There is an enormous amount of data being dealt with by the medical field on a daily basis. Using a conventional method for handling data can affect the accuracy of the results. Early recognition of the disease is crucial for the analysis of patient medicines and specialists. The objective of this paper is to provide a comprehensive review of the techniques used in disease detection. Machine learning algorithms can be used to find out facts in medical research, particularly disease prediction. Machine learning algorithms such as Support vector machine [SVM], Decision trees, Bayes classifiers, K-Nearest Neighbours [KNN] Ensemble classifier techniques, etc. are used to determine different ailments. The use of machine learning algorithms can lead to fast and high accuracy prediction of diseases. This research paper analyses how machine learning techniques and algorithms are used to predict different diseases and their types. This paper provides an extensive survey of the machine learning techniques used for the prediction of chronic kidney disease, liver disease, haematological diseases, Alzheimer's disease, and urinary tract infections.

Index Terms: Machine Learning Algorithm, Disease, Neural Networks, Decision tree.

1. Introduction

Machine learning is a subset of artificial intelligence that contains algorithms or methods, for naturally creating models from data. Machine Learning is the technique of making computers learn and act like human beings by feeding data sets and information without being specifically programmed[1][2]. A machine learning system learns from experience, unlike a machine that performs a task by following clear-cut rules[3][4]. A rule-based system will perform a task the same way every time, whereas the performance of a machine learning system can be made better through training and testing by exposing the algorithm to more data. The Machine Learning model can be broadly classified into three categories.

2. Machine Learning:Classification

2.1. Supervised Learning Model

Supervised learning as the name suggests works under supervision, that is the machine predicts after being trained by the data that is labelled. Data for which the target answer is already known is called a labelled data. [5] The labelled data is fed to the machine which analyses and learns the relation of these images with its labels, based on its features. Now when a new image is fed to the machine without any label, with the aid of only the past data set, the machine is able to predict accurately and give the output. Example algorithms include: The Back Propagation Neural Network and Logistic Regression[6][7].

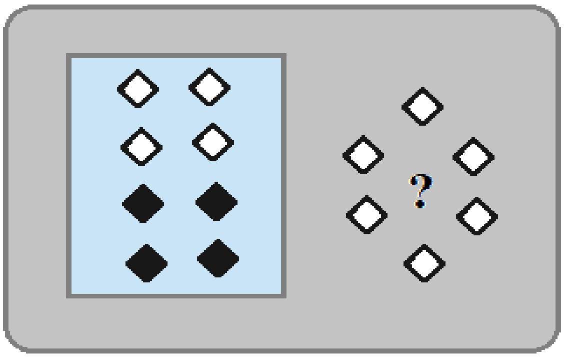

Fig.1. Supervised Learning Algorithm

2.2. Unsupervised Learning Model

Unsupervised machine learning Model lacks supervision, by not training the machine nor allowing it to act on the data which is not labeled. Therefore, machine tires to link the patterns and give the response. The machine recognizes the patterns from given set of data and clusters them based on their similarities, patterns, etc. [8] Unsupervised learning can be additionally grouped into association and clustering. Example problems include association role learning and dimensionality reduction clustering. Example algorithms are: the Apriori algorithm and K-Means [9].Unsupervised Learning models are extensively used over real life data sets[10].

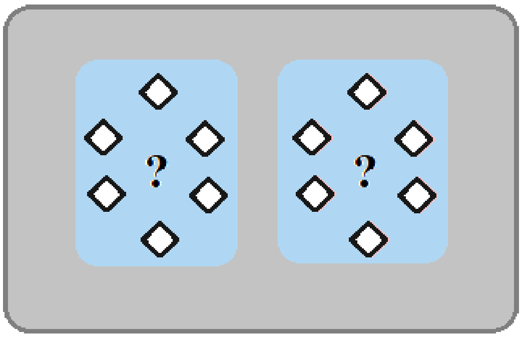

Fig.2. Unsupervised Learning Algorithm

2.3. Semi-supervised Learning Model

The amalgamation of supervised learning and unsupervised learning is referred to as Semi Supervised learning. This learning has a combination of labeled and unlabeled data set. Manually labeling all the data set is not practically possible, but can label some portions of the data and use that portion to train our model. The labeled data can be used as a training set for our model. We use our model to predict on the unlabeled part of the data set and label them. This operation of labelling the unlabeled data in tandem with the output that was forecasted by our neural network is called pseudo labelling. After designating the unlabeled data, then we train our model with the full data set [11][12].

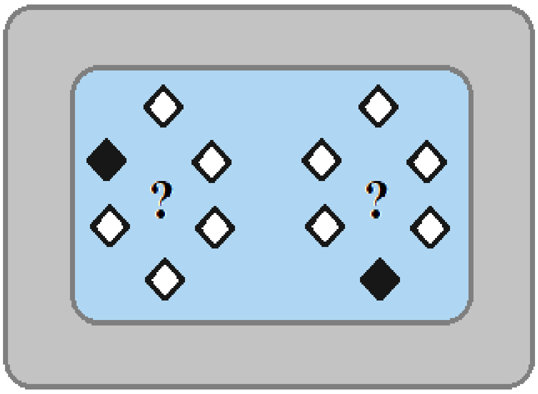

Fig.3. Semi-supervised Learning Algorithm

Table 1. Summary of the Classification Models [13][14]

TYPES OF LEARNING		
Supervised Machine learning	**Unsupervised Machine learning**	**Semi-Supervised Machine learning**
Models/Method 1. Linear Regression 2. Support Vector Machine (SVM) 3. Nearest Neighbour 4. Random Forest 5. Naïve Bayes 6. Decision Trees	Models/Method 1.Heirarchical clustering 2.Principal Component Analysis 3.Independent Component Analysis 4.K Nearest Neighbours	Models/Method 1.Continuity Assumption 2.Manifold Assumption 3.Cluster Assumption

3. Machine Learning Techniques for Disease Predictionbased Self-adaptation with Reusability

3.1. Liver Disease

Liver disease (LD) is one of those uncertain diseases which are difficult to diagnose even though the symptoms are seen at an early stage. The reason it is difficult to diagnose is that the symptoms are not prominent in the early stage as the liver is capable of functioning at a partially damaged state. Early diagnosis can be life-saving. Even though the diagnosis of this disease at an early stage is a challenge to the medical industry, it can still be detected. Diagnoses at an early stage can increase the patient’s life span substantially.

The most effective ML algorithms Back Propagation, Support Vector Machine (SVM), Naïve Bayes, Random Forest, and K- Nearest Neighbors (KNN) are used in the study and analysis of this disease [15].

The outcome of the analysis is as follows: Back-propagation algorithms give the accuracy of 73.2%[15], SVM gives an accuracy of 71% [15], Random forest provides an accuracy of 74%[16], KNN gives an accuracy of 62%[16], Naïve Bayes gives an accuracy of 95.1% [17].

3.2. Chronic Kidney Disease

Kidney is one of the vital organs in human body. They perform major functions like excretion, filtration of blood and osmoregulation. In other words, we can say that it helps in removing all the unnecessary and toxic material from the body. Every year in India, around a million cases of renal failure are diagnosed. Chronic kidney disease (CKD) is also called as renal failure. CKD is a slow and dangerous disease that leads to periodic loss of kidney function with the duration of time. It will develop to permanent kidney failure. The following Symptoms are seen if it’s not diagnosed and cured at early stage: week bones, Blood Pressure, anemia, decreased immune response, electrolytes accumulation, poor nutrition health and nerve damage, and built-up wastes in blood and body. The symptoms of CKD develop slowly and are unpredictable and aren’t confined and specific to only CKD. Sometimes the symptoms are not even observed with some patients. Machine learning can be used here in order to predict whether or not the person has CKD. This is done by using the old records of CKD affected patients in order to train the predicting model of machine learning algorithm. The stage of chronic kidney disease can be analyzed using the technique called Glomerular Filtration Rate (GFR). This is also used to find the level of kidney function. It majorly uses the patient’s blood creatinine for its calculation.

By using data of CKD patients with machine learning algorithms such as Support Vector Machine, Decision Tree, Random Forest a model that provides maximum accuracy for predicting CKD can be built.

From the outcome of the algorithms, it is noticed that, SVM provides an accuracy of 96.75%[18], decision tree algorithm gives the accuracy of 91.75% [18], Random Forest gives 99.16% [19].

3.3. Haematology Diagnosis

Blood affects human life in many different ways. It circulates throughout the body like a postman and visits all organs. Blood should reflect the growth in change. Different parameter values in the blood analysis tests can be used to detect this change.

The pathological conditions that affect blood producing organs or the blood are termed as haematological diseases(HD). This group includes a variety of blood cancers, different types of anemia such as severe aplastic anemia (SAA), thalassemia, iron deficiency anemia, sickle cell disease, and hemorrhagic conditions. These also include idiopathic thrombocytopenic purpura, congenital neutropenia etc.

Several parameters such as age, genders, symptoms and other health conditions are considered by the doctors to choose the specific test for detecting that disease. For successful treatment of these diseases, quick and accurate medical diagnosis is very important. In the survey, using laboratory blood test results and machine learning algorithms two models were built to predict the hematological diseases. The first model used a dataset containing different blood test parameters to detect the diseases and in the second model a reduced data set was used whose parameters were measured once the patient is admitted [20].

When the top five most likely haematological diseases were considered, the accuracy of the first and second model were 88% and 86% respectively. For determining the most likely disease the accuracy was 59% and 57% respectively. On conducting a clinical test, the accuracies of both the predictive models were very much similar to the results of haematology specialists. This was the first study which concluded that we can detect a haematological disease successfully with blood test samples alone, by using machine learning algorithms [20]. Usually, the data set used in haematological surveys contained a set of 50-60 parameters, a few of them were WBC (white blood cell concentration), PLT (platelet concentration), MCV (mean corpuscular volume), HCT (haematocrit), LYMPH % (lymphocyte concentration), NEUT% (percentage of neutrophils), MONO% (percentage of monocytes), IMI# (immature WBC concentration), PLT-X (average platelet concentration) [21]. From these studies we get to know that the results of the blood test contain a lot more information that is generally not recognized by the physicians. This remarkable result opens up exceptional possibilities in the field of medical diagnosis [22].

Among all the machine learning algorithms the ones used in this study include used:

Support Vector Machine (Linear and RBF)- the scikit implementation of SVC (Support Vector Classifier) is used; Naïve Bayesian Classifier-the experiment above used the scikit lean implementation, Gaussian NB; Decision Tree; Random Forest. On comparing different machine learning algorithms that were applied to a large number of data sets, it was observed that overall Random Forest was the best algorithm to be used.

When the results of different classifiers were examined, the accuracies of these classifiers ranged between 71.2% and 98.16%. The accuracies were found to be NaiveBayes-81.60%, Bayesian network-92.86%, Multilayer Perception-91.8-%, Decision Tree-97.00%, SVM-71.20%, Random Forests-97.12% [22].

3.4. Alzheimer's Disease

Alzheimer's Disease (AD) is a neurodegenerative disorder. The cause of this is uncertain and it is mostly seen in aged people. It is one of the most common cause that leads to dementia. Selective memory impairment is seen as the early symptom of this. As of now, only treatments to amend some of the symptoms are available and no cure is available. The most commonly used model and their analysis and accuracy is as follows:

The most frequently applied algorithm is Support Vector Machine (SVM) and the second most frequently used is the Naïve Bayes algorithm. From the results it is observed that the generalized linear model surpasses the other classifiers with an accuracy of 88.24% during the test period. Accuracies seen for Naive Bayes algorithms and deep learning are 74.65% and 78.32% respectively. Accuracies for KNN (K- Nearest Neighbors) and Decision Tree are 43.26 and 74.22 respectively[23].

96 out of the total of 1454 showed the instruments psychometric properties. 89 papers explain the paper and pencil test. From the studies it is observed that, the Montreal Cognitive Assessment showed an effective screening test for memory clinic testing [24].

3.5. Urinary Tract Infections

A collective term that describes any infection or abnormality affecting all or any part of the urinary tract namely kidneys, urinary bladder, urethra and the ureters is known as UTI or the Urinary Tract Infection. The urinary system of human beings can be split into two sections. The upper tract consists of the kidneys and the ureteric (ureter). The lower tract consists of the vesica urinaria (urinary bladder) and the urethra.

In the local primary care, Cystitis (commonly known as Urinary Tract Infection or UTI) is considered as one of the most common bacterial infections, affecting approximately 150 million people every year worldwide [25]. It is seen in older adults much more than the younger ones. It is very crucial to detect Urinary Tract Infections in early stages especially in older adults as delayed treatments might lead to further complications and can be catastrophic. An experiment was conducted whose aim was to validate, train and compare different models to detect Urinary Tract Infection's using a validation dataset.

Different models were developed in this experiment to predict the Urinary Tract Infection using seven machine learning algorithms namely- Support Vector Machine (SVC), Elastic Net, Logistic Regression, Extreme Gradient Boosting, Random Forest, Neural Network and Adaptive Boosting. The machine algorithm- Logistic Regression, which is regularly used in the field of medicine was named as a side-by-side comparison of baselines. The other algorithms were chosen for relative ease in implementation, resiliency to overfitting, their ability to model non-linear associations and since they were widely accepted by the machine learning community.

The accuracies of different models were found to be: Random Forests-87.4%, Adaboost-85.6%, Support Vector Machine-86.3%, Elastic Net-86.4%, Logistic Regression-86.4%, Neural Networks-86.3% and XG Boost-87.5% [26].

4. Discussion and Conclusion

Machine learning is the science of making computers gain an understanding of and behave like humans, providing data and information without being specifically programmed. It can be classified into Supervised learning model, Semi-Supervised learning model and Unsupervised learning model. In this paper, a detailed review of machine learning classification is discussed. Machine learning uses various algorithm or models to perform a specific task. The primary use of these algorithms is in the medical prediction field. The focus is on the use and amalgamation of different algorithms for predicting different types of diseases using machine learning. Various research works with some effective techniques done by different people is studied.

In this paper, five such diseases of high priority are considered and different machine learning models are used for its prediction. Models like Back Propagation, Support vector machine, Naive bayes, Random forest, and K-Nearest Neighbors are used for the study and analysis of Liver disease, among these, Naive Bayes gave the highest accuracy of 95.1%. Chronic Kidney Disease is the second disease and models like Decision tree, support vector machine, random forest are used for analysis. The maximum accuracy of 99.16% is observed by the Random forest model. The analysis and prediction of Hematology included models like support vector machine, Naive Bayesian Classifier, Decision tree, Random Forest and the highest accuracy of 97.12 was obtained by Random Forest making it the most efficient model for the prediction of this disease. The study of Alzheimer’s Disease included models such as Support vector machine, Naive Bayes, Random Forest. Reasonable accuracy of 74.65% is obtained for Naive Bayes. Different models were developed for UTI prediction employing seven machine learning algorithms namely- Neural Network, Random Forest, Support Vector Machine, Logistic Regression, Extreme Gradient Boosting, Adaptive Boosting and Elastic Net. Accuracy of 90.4% is obtained by XG Boost which proved to be the highest among all other models. From this study a wide overview of the relative performances of different variants of supervised machine learning for disease prediction is provided. The information provided on relative performances by this study can be used by researches in selecting appropriate machine learning algorithms for their studies.

Table 2. Comprehensive Survey on the machine Learning Techniques for Disease Detection

MODEL/DISEASE	LIVER	CHRONIC KIDNEY	ALZHEIMER’S	HAEMATOLOGY DIAGNOSIS	UTI
SVM	71%[15]	96.75%[18]	-	71.2% [15]	86.3% [20]
RANDOM FOREST	74%[16]	99.16%[19]	-	97.12% [15]	87.4% [20]
DECISION TREE/ XGBOOST	-	91.75%[18]	74.22%[23]	97% [15]	87.5% [20]
BACK PROPAGATION	73.2%[15]	-	-	-	-
KNN	62%[16]	-	43.26%[23]	-	-
NAÏVE BAYES	95.1%[17]	-	74.65%[23]	81.6% [15]	-
LINEAR REGRESSION	-	-	88.24%[23]	-	-
DEEP LEARNING	-	-	78.32%[23]	-	-
ADABOOST	-	-	-	-	85.6% [20]
LOGISTIC REGRESSION	-	-	-	-	86.4% [20]
NEURAL NETWORK/ MULTILAYER PERCEPTRON	-	-	-	91.8% [15]	86.3% [20]
BAYESIAN NETWORK	-	-	-	92.86% [15]	-

References

1. Nithya, B., and V. Ilango. "Predictive analytics in health care using machine learning tools and techniques." In 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 492-499. IEEE, 2017.
2. Sharmila, S. Leoni, C. Dharuman, and Perumal Venkatesan. "Disease classification using machine learning algorithms-a comparative study." International Journal of Pure and Applied Mathematics 114, no. 6 (2017): 1-10
3. Das, Kajaree, and Rabi Narayan Behera. "A survey on machine learning: concept, algorithms and applications." International Journal of Innovative Research in Computer and Communication Engineering 5, no. 2 (2017): 1301-1309.
4. Kaur, Sunpreet, and Sonika Jindal. "A survey on machine learning algorithms." Int J Innovative Res Adv Eng (IJIRAE) 3, no. 11 (2016): 2349-2763
5. Narayanan, Uma, Athira Unnikrishnan, Varghese Paul, and Shelbi Joseph. "A survey on various supervised classification algorithms." In 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), pp. 2118-2124. IEEE, 2017
6. Dahiwade, Dhiraj, Gajanan Patle, and Ektaa Meshram. "Designing disease prediction model using machine learning approach." In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), pp. 1211-1215. IEEE, 2019
7. Osisanwo, F. Y., J. E. T. Akinsola, O. Awodele, J. O. Hinmikaiye, O. Olakanmi, and J. Akinjobi. "Supervised machine learning algorithms: classification and comparison." International Journal of Computer Trends and Technology (IJCTT) 48, no. 3 (2017): 128-138.
8. Jamgade, Akash C., and S. D. Zade. "Disease prediction using machine learning." International Research Journal of Engineering and Technology 6, no. 5 (2019): 6937-6938.
9. Lim, Sunghoon, Conrad S. Tucker, and Soundar Kumara. "An unsupervised machine learning model for discovering latent infectious diseases using social media data." Journal of biomedical informatics 66 (2017): 82-94.
10. Sathya, Ramadass, and Annamma Abraham. "Comparison of supervised and unsupervised learning algorithms for pattern classification." International Journal of Advanced Research in Artificial Intelligence 2, no. 2 (2013): 34-38.
11. Reddy, Y. C. A. P., P. Viswanath, and B. Eswara Reddy. "Semi-supervised learning: A brief review." Int J Eng Technol 7, no. 1.8 (2018): 81.
12. Obulesu, O., M. Mahendra, and M. ThrilokReddy. "Machine learning techniques and tools: A survey." In 2018 International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 605-611. IEEE, 2018.
13. Uddin, Shahadat, Arif Khan, Md Ekramul Hossain, and Mohammad Ali Moni. "Comparing different supervised machine learning algorithms for disease prediction." BMC medical informatics and decision making 19, no. 1 (2019): 1-16.
14. Gupta, Shubham, Vishal Bharti, and Anil Kumar. "A survey on various machine learning algorithms for disease prediction." Int. J. Recent Technol. Eng 7, no. 6c (2019): 84-87.
15. Sontakke, Sumedh, Jay Lohokare, and Reshul Dani. "Diagnosis of liver diseases using machine learning." In 2017 International Conference on Emerging Trends & Innovation in ICT (ICEI), pp. 129-133. IEEE, 2017.
16. Rahman, AKM Sazzadur, FM Javed Mehedi Shamrat, Zarrin Tasnim, Joy Roy, and Syed Akhter Hossain. "A Comparative Study On Liver Disease Prediction Using Supervised Machine Learning Algorithms." International Journal of Scientific & Technology Research 8, no. 11 (2019): 419-422.
17. El-Shafeiy, Engy A., Ali I. El-Desouky, and Sally M. Elghamrawy. "Prediction of liver diseases based on machine learning technique for big data." In International conference on advanced machine learning technologies and applications, pp. 362-374. Springer, Cham, 2018.
18. Tekale, Siddheshwar, Pranjal Shingavi, Sukanya Wandhekar, and Ankit Chatorikar. "Prediction of chronic kidney disease using machine learning algorithm." International Journal of Advanced Research in Computer and Communication Engineering 7, no. 10 (2018): 92-96..
19. Revathy, S., B. Bharathi, P. Jevanthi, and M. Ramesh. "Chronic Kidney Disease Prediction using Machine Learning Models." . International Journal of Engineering and Advanced Technology (IJEAT) 9, no. 1 (2019).
20. Gunčar, Gregor, Matjaž Kukar, Mateja Notar, Miran Brvar, Peter Černelč, Manca Notar, and Marko Notar. "An application of machine learning to haematological diagnosis." Scientific reports 8, no. 1 (2018): 1-12.

21. Pekelharing, J. M., O. Hauss, R. De Jonge, J. Lokhoff, J. Sodikromo, M. Spaans, R. Brouwer, S. De Lathouder, and R. Hinzmann. "Haematology reference intervals for established and novel parameters in healthy adults." Sysmex Journal International 20, no. 1 (2010): 1-9.
22. Alsheref, Fahad Kamal, and Wael Hassan Gomaa. "Blood diseases detection using classical machine learning algorithms." International Journal of Advanced Computer Science and Applications (IJACSA) 10, no. 7 (2019).
23. Shahbaz, Muhammad, Shahzad Ali, Aziz Guergachi, Aneeta Niazi, and Amina Umer. "Classification of Alzheimer's Disease using Machine Learning Techniques." In DATA, pp. 296-303. 2019.
24. De Roeck, Ellen Elisa, Peter Paul De Deyn, Eva Dierckx, and Sebastiaan Engelborghs. "Brief cognitive screening instruments for early detection of Alzheimer’s disease: a systematic review." Alzheimer's research & therapy 11, no. 1 (2019): 1-14.
25. Flores-Mireles, Ana L., Jennifer N. Walker, Michael Caparon, and Scott J. Hultgren. "Urinary tract infections: epidemiology, mechanisms of infection and treatment options." Nature reviews microbiology 13, no. 5 (2015): 269-284.
26. Taylor, R. Andrew, Christopher L. Moore, Kei-Hoi Cheung, and Cynthia Brandt. "Predicting urinary tract infections in the emergency department with machine learning." PloS one 13, no. 3 (2018): e0194085.