SURVEY ON MACHINE LEARNING ALGORITHMS FOR LIVER DISEASE DIAGNOSIS AND PREDICTION

K USHA RANI[1]

ASSOCIATE PROFESSOR, DEPARTMENT OF CSE, BHOJ REDDY ENGINEERING COLLEGE FOR WOMEN, VINAY NAGAR, HYDERABAD-59

HARSHITA SHARMA[2]

UG SCHOLAR, DEPARTMENT OF CSE, BHOJ REDDY ENGINEERING COLLEGE FOR WOMEN, VINAY NAGAR, HYDERABAD-59

MANASA PONNURU[3]

UG SCHOLAR, DEPARTMENT OF CSE, BHOJ REDDY ENGINEERING COLLEGE FOR WOMEN, VINAY NAGAR, HYDERABAD-59

MANISHMA GANAPURAPU[4]

UG SCHOLAR, DEPARTMENT OF CSE, BHOJ REDDY ENGINEERING COLLEGE FOR WOMEN, VINAY NAGAR, HYDERABAD-59

ABSTRACT

Machine learning plays a vital role in health care industry. It is very important in Computer Aided Diagnosis. Computer Aided Diagnosis is a quickly developing dynamic region of research in medicinal industry. The current specialists in machine learning guarantee the en-hanced precision of discernment and analysis of diseases. The computers are empowered to think by creating knowledge by learning. This procedure enables the computers to self-learn individually without being explicitly programed by the programmer .There are nu-merous sorts of Machine Learning Techniques and which are utilized to classify the data sets. They are Supervised, Unsupervised and Semi-Supervised, Reinforcement, deep learning algorithms. The principle point of this paper is to give comparative analysis of super-vised learning algorithms in medicinal area and few of the techniques utilized as a part of liver disease prediction.

Keywords: Liver Disease; Medical Data Mining; Supervised Learning; Machine Learning Techniques.

1. INTRODUCTION

The liver is a vast, substantial organ in the human body. Weighing around 3 pounds. The liver contains two vast segments, called the right and the left projections. The gallbladder sits under the liver, alongside parts of the pancreas and digestive organs. The liver and these organs work together to process, ingest, and process suste-nance. The liver's main role is to filter the harmful substances in the blood originating from the digestive system, before passing it to the rest of the body [4]. Liver damage is the one of the top deadliest disease in the world. The main causes of liver damage are Fatty liver, Liver Fibrosis, Cirrhosis, hepatitis and infections [5]. Fig 1. Shows the stages of liver damage, in the first stage healthy liver will become fatty liver due to accumulation of cholesterol and triglycerides, after few months to years fatty liver will becomes liver fibrosis, later it leads to final stage of liver damage known as cirrhosis.

Figure 1: Stages of Liver Damage

In the early stages of the liver disease, it is very difficult to identi-fy even though liver tissue has been damage moderately, it origins many medical experts repeatedly fail to diagnose the disease. This can distort to wrong medicine and treatment, so early detection is very important and necessary to save the patient [6].

1.1. LEARNING

Machine Learning [11][12] is the domain of Artificial Intelligence which is concerned with building adaptive computer system that are able to improve their competence and or efficiency through learning from input data or from their own problem solving expe-rience. Artificial Intelligence is the study of how to make comput-ers do things which at the moment, people do better. Two dimen-sions for learning is competence and efficiency.

a) Supervised Learning

The name itself says that there is supervision for training patterns about their labels. Supervised learning is nothing but classification. Training set consists of patterns along with their associated labels, but test set consist of only patterns without labels. By using any classification techniques such as NNC, ID3, SVM, ADT etc. We can build the classifier by providing training set as input. Test set as input to the classifier which is built, generate the labels for unlabeled patterns [6].

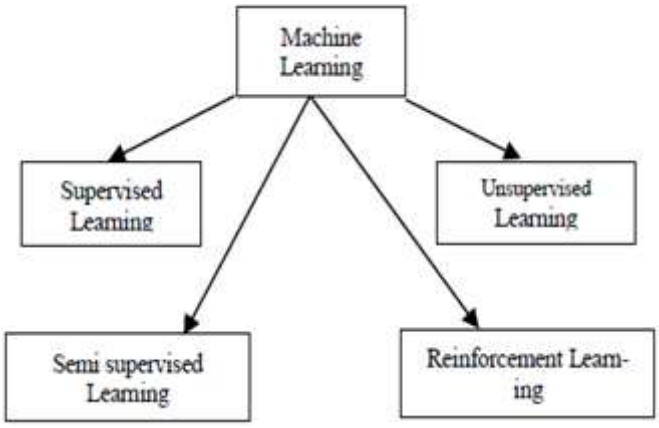

Figure 2: Categories of Machine Learning

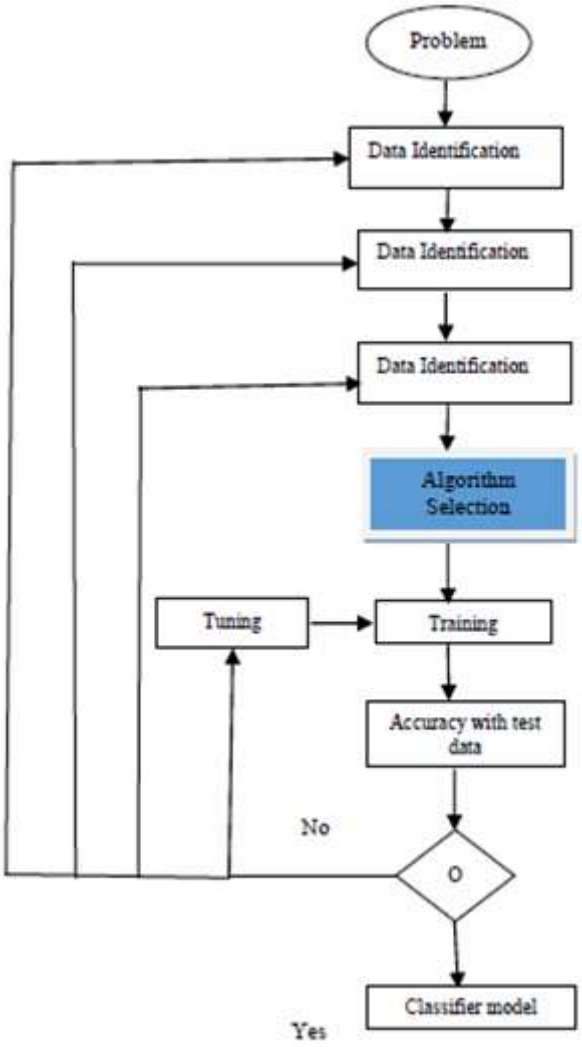

Figure 3: Supervised Learning Process

1) The first step is collecting the data set. Datasets can be col-lected from UCI repository where some benchmark datasets are available the important features in the dataset are identi-fied by using some simplest techniques such as before or from experts given information. The datasets contains noisy data and feature values, therefore preprocessing techniques are essential.

2) The second step is the data preparation and data prepro-cessing. Instance selection reduces sample size and enables a data mining algorithm with very large datasets to function and work effectively. Feature subset selection is the process of identifying and removing irrelevant and redundant fea-tures as possible (yu& Liu, 2004) [6].

3) Algorithm selection: The algorithm selection is very critical step in the process of supervised machine learning. The classifier evaluator is based on accuracy. (The percentage of current prediction divided by the total number of predic-tions).

4) There are at least three techniques which are used to calcu-late classifier accuracy. Such as 2/3 and 1/3 rule, cross vali-dation and leave one out cross validation [7].

If the error rate is very high, we must return to previous stage of supervised learning process. A variety of factors must be exam-ined and rectified by repeating the process.

b) Unsupervised learning

In contrast to supervised learning, unsupervised learning is noth-ing but clustering, where patterns are unlabeled[11][12]. By applying these unsupervised algorithms, researchers hope to discover unknown, but useful class of item [8]. Unsupervised learning is particularly known as clustering. Cluster-ing is ever present and a wealth of clustering has been developed to solve different problems in different specific fields. However, there is no clustering algorithm that can be universally used to solve all problems. "It has been very difficult to develop a unified framework for reasoning about it (clustering) at a technical level and profoundly diverse approaches to clustering" [9]. According to AK Jain [8] clustering methods are classified into five catego-ries partitioning, hierarchical, Density based, Grid based, Model based methods.

c) Semi Supervised Learning

Semi supervised learning is another type of machine learning.it falls between supervised Learning(training data with label) and unsupervised learning(training data without label).these algo-rithms perform very well when we have less number of labeled data and large amount of unlabeled data [10].

d) Reinforcement Learning

Reinforcement learning is a another type of Machine Learning algorithms which permits software agents and machines to auto-matically define the ideal behavior within the specific context, in order to maximize the performance. Simple reward feedback is required for the agent to learn its behavior. This process is known as reinforcement signal [11].

2. Background

Machine learning algorithms are very helpful in providing vital statistics, real-time data, and advanced analytics in terms of the patient's disease, lab test results, blood pressure, family history, clinical trial data, and more to doctors. Now a days Machine learn-ing algorithms are very useful for extracting and examining the medical data in order to build certain prediction models to rise the accuracy of diagnosis in any specific disease. However, only few works in machine learning investigate liver disorders, although this disease is aggressively increasing and becoming one of the most fatal diseases in some countries [12].

- Omar S. Soliman, Eman Abo Elhamd used two algorithms, one is Particle Swarm Optimization algorithm and another algorithm is Least Squares Support Vector Machine (LS-SVM) to propose a hybrid classification model for HCV diagnosis. Authors used Principle Component Analysis algorithm for extraction of feature vectors. Modified-PSO Algorithm is used to search for the optimal values of LS-SVM parameters. The proposed model was imple-mented and evaluated on the target HCV data set from UCI re-pository databases. From the experimental results the proposed system obtained highest accuracy than the other systems [13].

- Moloud Abdar, Mariam Zomorodi-Moghadam, Resul Das, I-Hsien Ting, proposed computer aided Diagnostic method by using novel tree based algorithms, which are C5.0 algorithm and Chi-square Automatic Interaction Detector (CHAID) algorithm for liver disease prediction.in this proposed method authors used C5.0 algorithm via Boosting technique to achieve the highest accuracy as well as the production of rules on liver disease dataset [14].

- Sadiyah Noor Novita Alfisahrin, Teddy Mantoro [15] were ap-plied three techniques, which are Decision Tree, Naive Bayes, and NBTree algorithms for diagnosis of liver disease. They have im-plemented classification model and obtained highest accuracy by using NB Tree algorithm. Authors concluded that the Naive Bayes algorithm gives the fastest computation time followed by Decision Tree and NB Tree algorithm and also proved that number of clas-sification rule of NB Tree algorithm is simpler than the number of classification rule produced by Decision Tree algorithm.

- Bendi Venkata Ramana, proposed five classification algorithms for liver disease diagnosis. Authors applied Naive Bayes classifica-tion, C 4.5 Decision Tree, Back Propagation, K-Nearest Neighbor and Support Vector Machine algorithms on two different type of liver Data sets, which are BUPA liver disorder data and India Liver Patient Data (ILPD).in this paper the above algorithms are considered for evaluating their classification performance in terms of Accuracy, Pre-cision, Sensitivity and Specificity in classifying liver patient's dataset.Finally they obtained highest accuracy by using K-Nearest Neighbor and Support Vector Machine algorithms [16].

- Sumedh Sontakke, Jay Lohokare, Reshul Dani, proposed two methods in order to classify the chronic liver disease, one method is a symptomatic approach to diagnosis, and second one involves a genetic approach to the diagnosis. Proposed approach is the applica-tion of Artificial Neural Networks and Multi-Layer Perceptron to Micro-Array Analysis. They used these two methods to improve the efficiency of two algorithms Back Propagation and Support Vector Machine (SVM) to classify the liver disease. Authors achieved highest accuracy by using Back-Propagation algorithm [17].

Table 1: Comprehensive View of Machine Learning Algorithms for Liver Disease Diagnosis and Prediction

Title	Machine Learning Algorithms	Accuracy	Data Set	Conclusion
Performance analysis of classification algorithms on early detection of Liver disease –ELSEVIER-2016[14]	C5.0 algorithm CHAID algorithm	93.75 65	UCI repository	Boosting techniques in C5.0 algorithm leads to increased accuracy and speed in creating rules
Data Mining Techniques For Optimization of Liver Disease Classification IEEE-2017[15]	Decision tree Naive Bayes NB Tree Naive Bayes classification	66.14 56.14 67.01 95.07	UCI repository	Obtained highest accuracy by using NB Tree algorithm
A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis IJDMS-2011[16]	C 4.5 Decision Tree, Back Propagation, K-Nearest Neighbor SVM	96.27 96.93 97.47 97.07	BUPA Liver Disorders datasets, (UCI) Machine Learning Repository[20]	KNN, Back propagation are giving better results with all the feature set combinations
Diagnosis of Liver Diseases using Machine Learning IEEE-2017[17]	SVM Back Propagation	71 73.2	(UCI) Machine Learning Repository	Back propagation are giving highest accuracy
Prediction of Liver Fibrosis stages by Machine Learning model: A Decision Tree Approach* IEEE-2015[18]	Decision Tree	93.7	Department of Biochemistry and Molecular Biology of Kasralainy Hospital of Cairo University	Achieved Highest accuracy rate by using Decision Tree algorithm
Comparison of Machine earning approaches for Prediction of advanced liver Fibrosis in Chronic Hepatitis C Patients IEEE-2016 [19]	PSO GA MReg ADT	66.4 69.6 69.1 84.4	Egyptian National Committee for control of viral Hepatitis Data Base ,NTP HCV Patients data(39,567)	ADT Model achieved Highest Accuracy rate

- Heba Ayeldeen, Olfat Shaker, Ghada Ayeldeen, Khaled M. Anwar, proposed machine learning technique model based on decision tree classifier to forecast individuals' liver fibrosis .they identified in-formative biomarkers provided from different laboratory tests through containing the highly statistical data mining techniques to predict the hepatic fibrosis in Egyptian patients infected with hepatitis C virus. achieved accuracy is 93.7% by using decision tree classifier.

- Somaya Hashem, Gamal Esmat, Wafaa Elakel, Shahira Habashy, Safaa Abdel Raouf, Mohamed Elhefnawi, Mohamed El-Adawy, Mahmoud ElHefnawi, developed Particle swarm optimization, decision tree, multi-linear regression and genetic algorithm models in order to predict the advanced fibrosis stage for chronic HCV patients. Authors concluded that Particle Swarm Optimization model attained highest correlation coefficient 0.28 with presence of advanced fibrosis and AD tree model achieved the highest accuracy

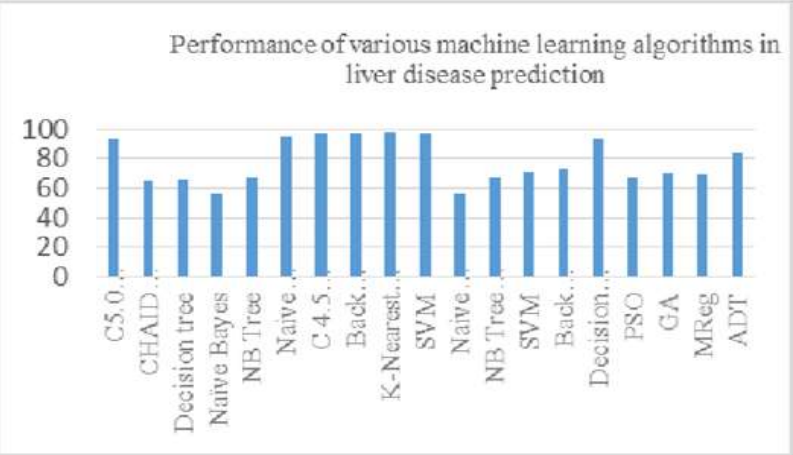

Figure 4: Performance of Various Supervised Learning Algorithms in Liver Disease Prediction

3. CONCLUSION

This paper provides an idea of recent machine learning algorithms available for detection and diagnosis of liver disease. From the study it can be clearly observed that different supervised learning algorithms K-Nearest Neighbour and Support Vector Machine provide enhanced accuracy on detection of liver diseases.

REFERENCES

[1] H. M. Gilligan, D. M. Venesy dan F. D. Gordon, 100 Questions & Answers about Liver, Heart, and Kidney Transplantation A Lahey Clinic Guide, United States of America: Jones & Bartlett Learning, 2011.

[2] D.A. Saleh F. Shebl M. Abdel-Hamid et al. "Incidence and risk fac-tors for hepatitis C infection in a cohort of women in rural Egypt"Trans. R. Soc. Trop. Med. Hyg. vol. 102 pp. 921928 2008. https://doi.org/10.1016/j.trstmh.2008.04.011.

[3] S. B. Kotsiantis, Supervised Machine Learning: A Review of Clas-sification Techniques, Informatica (2007) 249-268 249.

[4] P. Kshirsagar, S. Akojwar, Nidhi D. Bajaj , “A hybridised neural network and optimisation algorithms for prediction and classification of neurological disorders” International Journal of Biomedical Engineering and Technology ,vol. 28,Issue 4,2018.

[5] P. Kshirsagar and S. Akojwar, "Novel approach for classification and prediction of non linear chaotic databases," 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), 2016, pp. 514-518, doi: 10.1109/ICEEOT.2016.7755667,2016

[6] Kshirsagar, P.R., Akojwar, S.G., Dhanoriya, R, “ Classification of ECG-signals using artificial neural networks”, In: Proceedings of International Conference on Intelligent Technologies and Engineering Systems, Lecture Notes in Electrical Engineering, vol. 345. Springer, Cham (2014).

[7] P Viswanath, K rajesh, C.Lavanya and Y C A Padmanabha Reddy, “A Selective Incremental Approach for Transductive Nearest Neighbour Classification. Proceedings of IEEE International Conference of RAICS-2011.pp (221-226), 2011

[8] Jain, A.K., Murty, M. N., and Flynn, P. (1999), Data clustering: A review, ACM Computing Surveys, 31(3): 264–323 https://doi.org/10.1145/331499.331504.

[9] J. Kleinberg, “An impossibility theorem for clustering,” in Proc. 2002 Conf. Advances in Neural Information Processing Systems, vol. 15, 2002, pp. 463–470.

[10] Yoshua Bengio, Olivier Delalleau, Nicolas Le Roux. In Semi-Supervised Learning (2006), pp. 193-216

[11] H. Manoharan, et al., “Examining the effect of aquaculture using sensor-based technology with machine learning algorithm”, Aquaculture Research, 51 (11) (2020), pp. 4748-4758.

[12] S. Sundaramurthy, S. C and P. Kshirsagar, "Prediction and Classification of Rheumatoid Arthritis using Ensemble Machine Learning Approaches," *2020 International Conference on Decision Aid Sciences and Application (DASA)*, 2020, pp. 17-21, doi: 10.1109/DASA51403.2020.9317253.

[13] Jose, D., Chithara, A.N., Nirmal Kumar, P. et al. Automatic Detection of Lung Cancer Nodules in Computerized Tomography Images. Natl. Acad. Sci. Lett. 40, 161–166 (2017). https://doi.org/10.1007/s40009-017-0549-2

[14] Ramesh D., Jose D., Keerthana R., Krishnaveni V. (2018) Detection of Pulmonary Nodules Using Thresholding and Fractal Analysis. In: Hemanth D., Smys S. (eds) Computational Vision and Bio Inspired Computing. Lecture Notes in Computational Vision and Biomechanics, vol 28. Springer, Cham. https://doi.org/10.1007/978-3-319-71767-8_80

[15] M Prathiba et al, Automated Melanoma Recognition in Dermoscopy Images via Very Deep Residual Networks, 2019 IOP Conf. Ser.: Mater. Sci. Eng. 561 012107

