International Journal of Scientific Research in Science, Engineering and Technology (www.ijsrset.com)
© 2020 IJSRSET | Volume 7 | Issue 1 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
DOI : https://doi.org/10.32628/IJSRSET207138

A Survey on Predicting Advanced Liver Fibrosis Using Different Machine Learning Algorithms

Krishnendu K B[1], Deepa S S[2]
[1]M Tech Scholar, Department of Computer Science and Engineering, GEC Idukki, Kerala, India
[2]Associate Professor, Department of Computer Science and Engineering, GEC Idukki, Kerala, India

ABSTRACT

Machine learning (ML) is a subsection of AI. The goal of ML is to understand the structure of data and fit that data into models that can be used for prediction, classification etc. Although machine learning is an area within computer science, it differs from traditional computational approaches. In recent years, different machine learning algorithms are used for disease prediction. Algorithms like Decision Tree (DT), Support Vector Machine (SVM), Particle Swarm Optimization (PSO), Multi- Linear Regression, Random Forest, Genetic Algorithm (GA), Artificial Neural Network (ANN), Naive Bayes, etc. are used for classification. Using these algorithms liver fibrosis stages can be predicted. This paper discusses different machine learning algorithms for the prediction of liver fibrosis stage and the performance analysis of these algorithms in various studies.

Keywords : Liver Fibrosis, Hepatitis C, Machine Learning

I. INTRODUCTION

Hepatitis C is an infection that may lead to serious liver damage. It is caused by the Hepatitis C virus (HCV) . HCV affects people in different ways and has several stages. Liver fibrosis occurs when the healthy tissue of the liver becomes scarred and therefore cannot work as well. Fibrosis is the first stage of liver scarring. If the liver becomes more scarred, it is known as liver cirrhosis.

According to METAVIR score fibrosis stages ranges from F0 to F4 where F0 is considered as no fibrosis, F1 to F2 is categorized as mild to moderate fibrosis and F3 to F4 is categorized as advanced fibrosis. F4 is the last stage(Liver cirrhosis) . Liver fibrosis is the first stage and the last stage is cirrhosis. For diagnosis and staging liver fibrosis, liver biopsy was considered as a gold standard. But it is very expensive and risky. To overcome this drawback non-invasive methods are used. Noninvasive methods help patients by reducing the pain that the patient exposed in the biopsy process.

Some noninvasive tests based on indexes derived from serum markers, such as FIB-4 score and AST-to-platelet ratio index (APRI) . Imaging techniques such as Transient Elastog- raphy (TE) uses ultrasound and vibratory waves for estimating the extent of liver fibrosis. Hepatitis is a liver disease which affects majority of the population in all age group. Diagnosing hepatitis is a major challenge for many hospitals and public health care services all over the world.

II. MACHINE LEARNING

Machine learning is an application of Artificial Intelligence (AI) and it mainly focuses on improving the learning process of computers based on their

experience. A good quality data is provided to learn and train machines by building learning models using data and different algorithms. The obtained model is used for prediction.

Different approaches used for learning can be supervised, unsupervised or semi supervised. In supervised learning, a model is designed from the inputs and their desired outputs.

A.Classification

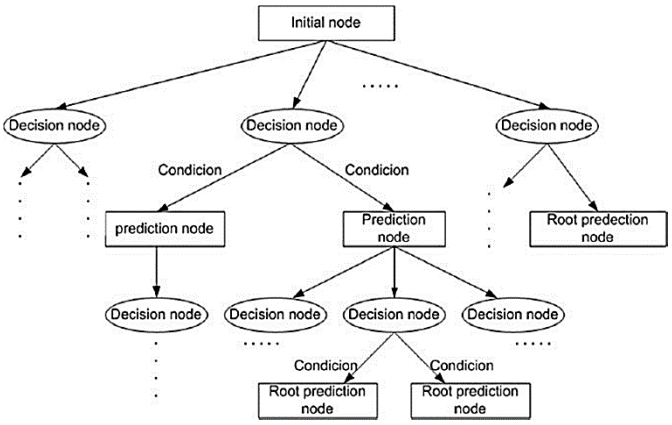
Classification is a supervised learning method in machine learning. In supervised learning, the computer program learns from the data input given to it. New observations are classified based on this learning. Some applications of classification problems are speech recognition, document classification, bio-metric identification etc.

Different types of classification algorithms in machine learn- ing are:

- Decision Tree (DT)
- Support Vector Machine (SVM)
- Random Forest
- Neural Networks
- Naive Bayes
- Logistic Regression
- Genetic Algorithm
- Particle Swarm Optimization

1) Decision Tree: A decision tree is a modeling technique used for classification, clustering, and prediction activities. The decision tree uses a divide and conquer to split the problem search space into subsets. Figure 1 [10] shows an example for a decision tree. The nodes are the features and the arcs represents the splitting of features. A tree is constructed for classification. The tree built is applied to the dataset. Some of the decision tree-based algorithms are CART, ID3, C4.5 etc.

ID3 uses information theory to build a decision tree. In ID3, the feature with high information gain is selected as the root node. Each arc contains the split of that attribute. The improved version of the ID3 algorithm is C4.5. C4.5 can handle missing data and continuous data.



2) Support Vector Machine: Support Vector Machine (SVM) is a classifier that performs classification tasks by constructing hyper planes in a multidimensional space that separates cases of different class labels. SVM supports both regression and classification tasks and can handle multiple continuous and categorical variables.

For categorical variables a dummy variable is created with case values as either 0 or 1. Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. SVM can be used for classification and feature selection.

3) Random Forest: From the name itself, it is clear that random forest is a collection of trees (forest) . It is also known as a random decision forest. It generates multiple trees for classification. It creates a large number of individual decision trees. The mean prediction of individual trees is the output of random forest.

4) Neural Network: A neural network is a circuit of neu- rons, or an artificial neural network, composed of nodes. A neural network is network, for solving artificial intelligence (AI) problems. The connections of the neuron are defined as weights.

NN contains mainly three layers: an input layer, one or more hidden layers and an output layer. The input

values are given to the input layer. The hidden layers performs the prediction and the output is obtained in the output layer.

These networks may be used for predictive modeling and applications where they can be trained via a data set. Self- learning resulting from experience can occur within networks, which can acquire conclusions from a complex and unrelated set of information.

5) Naive Bayes: Naive Bayes is a classification algorithm that is a probabilistic algorithm that takes advantage of proba- bility theory. It is a family of algorithms that works based on Bayes theorem and consider every attribute is independent of each other.

6) Logistic Regression: Logistic regression is a type of a supervised machine learning algorithm. It makes a prediction that has a binary outcome from the past data. Logistic regres- sion usually returns the result in a very short time. Hence, it is used as a benchmarking model.

7) Genetic Algorithm: It is an example of evolutionary computing method [11]. It is a search based optimization technique based on principles of genetics and natural selection. Optimization means finding the value of inputs in such a way that we get the best output values. GA find optimal or near optimal solutions to difficult problem which otherwise would take a lifetime to solve.

8) Particle Swarm Optimization: It is also an iterative optimization technique like GA. A group of birds is termed as swarm.The idea behind this approach is, finding the food particle is to follow the birds which are nearest to the food particle. This behavior of birds is simulated in the computation environment. Thus an optimized result will be obtained using this algorithm.

III. LITERATURE REVIEW

This literature review focus on various machine learning algorithms used for the prediction of liver fibrosis and to compare their performance.

Mahmoud El Hefnawi et al. [1], prediction of advanced liver fibrosis in hepatitis c patients using ANN and decision tree. ANN gives better results using the features: Albumin, AFP, Viral load, fibrosis score, and HAI and decision tree gives acceptable results from only 3 features: ALT, Fibrosis score and HAI. The CART classifier is used for constructing a decision tree. This study concludes than the decision tree gives a better result than ANN. Neural networks needs more features than decision tree.

Somaya Hashem et al. [2] propose a mathematical model to predict the level of risk for liver fibrosis based on noninva- sive methods. Noninvasive methods aim is to provide useful information to help patients to reduce the use of liver biopsy. In this study, multi-linear regression is used for prediction.

The dataset used contains the features: age, gender, body mass index (BMI), grade of fibrosis and the activity, albumin, total bilirubin, indirect bilirubin, Alanine aminotransferase (ALT), as part ate aminotransferase (AST), Alfa-fetoprotein (AFP), Alkaline Phosphatase (ALP), gamma-glutamyl trans- ferase (GGT), International Normalized Ratio (INR), quantity of HCV RNA, White Blood Cells (WBC) count, Hemoglobin (Hb), platelet, creatinine, serology finding, Glucose, Postpran- dial glucose test (PC%), and HDL-cholesterol.

TABLE I. EXECUTION TIME TAKEN FOR CLASSIFICATION OF LIVER DISEASE

Algorithm	Execution time in ms
SVM	3210
Naive Bayes	1670

TABLE II. ACCURACY OF DIFFERENT CLASSIFIERS

Classifier	Accuracy (%)
Decision stump	83.75
Hoeffding tree	78.75
J48	86.25
Logistic model tree	85.00
Random forest	87.50
REP	80.00
Random tree	78.75

TABLE III
FEATURES OF TWO MODELS IN DECISION TREE

Model	Features used
Model 1	Age, Body Mass Index (BMI), Alpha-fetoprotein (AFP), Aspartate Aminotransferase (AST), Platelet Count, Albumin
Model 2	Age, AFP, Platelet Count, AST

TABLE IV
ADVANTAGES AND DISADVANTAGES OF DECISION TREE ALGORITHMS

	Advantages	Disadvantages
CART	Handles missing values automatically	Poor modeling in a linear structure
ID3	Easy to understand	Can suffer from over fitting
C4.5	Memory efficient than ID3	High training samples are needed
J48	Omits the missing values, decision trees pruning, continuous attribute value ranges, derivation of rules	Run complexity of algorithm depends on the depth of the tree

The correlation and P value are calculated to select the features. The features with P<0.01 are selected. Using the selected features, a model is created by the multi-linear regression. Prediction of fibrosis is performed using the model. In Heba Ayeldeen et al. [3], the machine learning model based on a decision tree is used to predict the liver fibrosis stage in HCV patients. The accuracy obtained in this work is 93.7%. There are nine significant biomarkers which are liver function tests and other molecular tests such as AST, ALT, ALB, T.BIL, D.BIL, GGT, HA, -macroglobulin and ApoA2.

The P-value for each variable is calculated and features selected for decision making are HA, GGT and 2-MC. It is found that HA level increases with increased fibrosis level and is efficient in predicting different grades of fibrosis.

Dr. S. Vijayarani et al. [4], predict liver disease using machine learning techniques such as Support Vector Machine (SVM) and Naive Bayes. Indian Liver Patient Dataset (ILPD) from the UCI Repository is used here. From the experiments, SVM is more accurate but takes more time for execution. Thus, we can say that SVM is not the best algorithm for the prediction of liver diseases. Table 1 shows the execution time taken for SVM and Nave Bayes for the prediction of liver dataset.

Manickam Ramasamy et al. [5] Decision Stump, Hoeffding Tree, J48, Logistic Model Tree (LMT), Random Forest, REP (Reduced Error Pruning) Tree and Random Tree is used on the Hepatitis dataset for the classification and comparison.

From table 2 it is clear that random forest classifier is more accurate than any other classifiers and there is only a slight difference between random forest and J48. This study concludes that random forest is the best classifier for the prediction of liver diseases.

Somaya Hashem et al. [6], developed a classification model for the prediction of advanced liver fibrosis. Alternating De- cision Tree (ADT) is used for prediction and obtained an accuracy of 84.8%. Using the dataset two models are designed.

TABLE V
COMPARATIVE ANALYSIS OF DECISION TREE ALGORITHMS

	Procedure	Pruning
CART	Constructs binary decision tree	Post pruning based on cost complexity measure
ID3	Top down decision tree construction	Pre pruning using a single pass algorithm
C4.5	Top down decision tree construction	Pre pruning using a single pass algorithm
J48	Top down decision tree construction	Two pruning methods: sub tree replacement and sub tree raising

Table 3 shows the features selected for each model.

Model 1 is created using 6 features and model 2 with 4 features. With 4 features model 2 acquired the highest accuracy for the prediction of fibrosis stage. Using these models, the stage of fibrosis can be identified. If the output value of the decision tree is greater than or equal to zero then, the patient is having advanced fibrosis of stage F3 or F4 else the stage is F0 to F2 which is moderate or mild fibrosis.

The four features used in model 2 are independent variables with high P-value and accepted correlation. The use of alpha- fetoprotein AFP as a feature for predicting advanced fibrosis in addition to using ADT improves the results compared to those of the FIB-4 algorithm which uses ALT instead.

S. Nagaparameshwara Chary et al. [7], this paper compares various decision tree algorithms for classification and to find their performance analysis. The decision tree classifiers such as ID3, CART, C4.5, and J48 are used for this comparative study. This study explains about these algorithms. J48 is a Java implementation and optimized version of C4.5 where C4.5 is the successor of ID3.

TABLE VI. COMPARISON OF DIFFERENT MODELS

Algorithm	Accuracy (%)
Logistic Regression	72
SVM	71
K- Nearest Neighbor	97.47
Decision Tree	66.14
Random Forest	87.25
Neural Network	86.32
Ensembled Method	71.53

TABLE VII. ACCURACY OF DIFFERENT ALGORITHMS FOR PREDICTION

Algorithm	Accuracy (%)
ADT	84.4
GA	69.6
MReg	69.1
PSO	66.4

The table 4 shows the findings of a comparative study from the survey. Table 5 shows different pruning techniques and procedures used in decision tree algorithms. These four methods uses Greedy approach. The different measures used by htese methods are Gini Diversity Index, Entropy and eliminated for getting better results.

The features such as age, aspartate aminotransferase (AST), platelet count, and albumin were found to be independent predictors of fibrosis, with P-value < 0.0001 and accepted correlation (r >0.1) with fibrosis. Therefore, these variables have been used for the prediction of advanced fibrosis in these models. The dataset contains 22,690 training data and 16,877 testing data.

By comparing the accuracy of the algorithms with these features, a decision tree is more accurate (84.4%) than the other algorithms such as GA (69.6%) , MReg (69.1%) and PSO (66.64%) .

V.V. Ramalingam et al. [8], is a comparative study on the classification algorithms such as logistic regression, SVM, K- nearest neighbor, DT, Random Forest, Neural Network and En- sembled method. These algorithms use the datasets discussed here are liver disease-related to hepatitis and hepatocellular carcinoma.

Most of the algorithms get different accuracy if they change the dataset. For the ensembled method of the liver disease dataset, this framework achieved an accuracy of 71.53% on the Indian liver disease patient dataset and 67.54% accuracy on the Bupa liver disease dataset. The accuracy of different algorithms used are shown in Table 5.

In Priyanga Chandrasekar et al. [9], an entropy-based dis- cretization method is used in J48 for improving the clas- sification accuracy . The numeric attributes are discretized and the performance of this approach with the J48 classifier is tested and compared with the performance of the J48 classifier without discretization. The prediction accuracy of J48 with discretization is high when compared with the J48 without discretization. The discretized J48 model improves construction time.

Somaya Hashem et al. [10], using different machine learning algorithms such as alternating decision tree (ADT), genetic algorithm (GA), particle swarm optimization (PSO) and multi- linear regression (MReg), liver fibrosis stages are predicted and compared (Table 6). In the preprocessing step, filter method is used as a feature selection method. Filter method based on Pearson correlation coefficient is

used. Redundant features are implementation and optimized version of C4.5 where C4.5 is the successor of ID3.

IV. CONCLUSION

In this survey, different machine learning algorithms which are used for the prediction of liver fibrosis where compared. By observing the accuracy of each algorithm, it is found that accuracy may depend on the dataset chosen. In most cases, the decision tree and random forest gives better results than any other classifiers. C4.5 is the classifier used to generate a decision tree.

From the noninvasive methods, machine learning techniques can predict the fibrosis stage of patients. This is very useful for patients to avoid liver biopsy. Due to the best results obtained from the decision tree, it is considered as a suitable algorithm for predicting liver disease such as fibrosis.

V. REFERENCES

[1]. Mahmoud ElHefnawi, Mahmoud Abdalla, Safaa Ahmed, Wafaa Elakel, Gamal Esmat, Maissa Elraziky, Shaima Khamis and Marwa Hassan. Accurate Prediction of Advanced Liver Fibrosis Using the Decision Tree Learning Algorithm in Chronic Hepatitis C Egyptian Patients, Gastroenterology Research and Practice, Volume 2016.

[2]. Somaya Hashem, Shahira Habashy, Wafaa El Akel, Safaa Abdel Raouf, Gamal Esmat, Mohamed El Adawy and Mahmoud El Hefnawi. A Simple multi linear regression model for predicting fibrosis scores in chronic Egyptian hepatitis C virus patients. International Journal of Bio Technology and Research (IJBTR). 2014 Jun; Vol. 4, Issue 3.

[3]. Heba Ayeldeen, Olfat Shaker, Ghada Ayeldeen, Khaled M. Anwar, Prediction of Liver Fibrosis stages by Machine Learning model: A Decision Tree Approach, IEEE Third world conference, 2015.

[4]. Dr. S. Vijayarani, Mr. S.Dhayanand, Liver Disease Prediction using SVM and Nave Bayes Algorithms, International Journal of Science, Engineering and Technology Research (IJSETR) Volume 4, Issue 4, April 2015.

[5]. Manickam Ramasamy, Shanthi Selvaraj, Dr. M. Mayilvaganan, An Em- pirical Analysis of Decision Tree Algorithms: Modeling Hepatitis Data, 2015 IEEE International Conference on Engineering and Technology (ICETECH), 20th March 2015, Coimbatore, TN, India.

[6]. Somaya Hashem, Gamal Esmat, Wafaa Elakel, Shahira Habashy, Safaa Abdel Raouf, Mohamed Elhefnawi, Mohamed El-Adawy, Mahmoud ElHefnawi, Accurate Prediction of Advanced Liver Fibrosis Using the Decision Tree Learning Algorithm in Chronic Hepatitis C Egyptian Patients, Hindawi Publishing Corporation Gastroenterology Research and Practice Volume 2016.

[7]. S. Nagaparameshwara Chary, Dr. B.Rama A Survey on Comparative Analysis of Decision Tree Algorithms in Data Mining , International Conference on Innovative Applications in Engineering and Information Technology(ICIAEIT-2017).

[8]. V.V. Ramalingam, A.Pandian, R. Ragavendran Machine Learning Tech- niques on Liver Disease -A Survey, International Journal of Engineering & Technology, 2018.

[9]. Priyanga Chandrasekar, Kai Qian, Hossain Shahriar and Prabir Bhat- tacharya, Improving the Prediction Accuracy of Decision Tree Mining with Data Preprocessing, 2017 IEEE 41st Annual Computer Software and Applications Conference.

[10]. Somaya Hashem, Gamal Esmat, Wafaa Elakel , Shahira Habashy , Safaa Abdel Raouf, Mohamed Elhefnawi , Mohamed El Adawy , Mahmoud ElHefnawi. Comparison of Machine Learning Approaches for Prediction of Advanced Liver

Fibrosis in Chronic Hepatitis C Patients, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2017.

[11]. Margaret H. Dunham, Dataminning: Intoductory and advanced topics.

Cite this article as :

Krishnendu K B, Deepa S S, "A Survey on Predicting Advanced Liver Fibrosis Using Different Machine Learning Algorithms", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 7 Issue 1, pp. 177-183, January-February 2020. Available at doi : https://doi.org/10.32628/IJSRSET207138
Journal URL : http://ijsrset.com/IJSRSET207138