International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 7, July 2020 ||

Feature Selection Using Genetic Algorithm for the Prediction of Fibrosis Stages in Hepatitis C Patients

Krishnendu K B[1] , Deepa S S [2]

M Tech Student, Dept. of CSE, Government Engineering College, Idukki, Kerala, India[1]

Associate Professor, Dept. of CSE, Government Engineering College, Idukki, Kerala, India [2]

ABSTRACT: The interpretation of liver fibrosis stages is mainly classified into four different stage which are F1-F2 (mild to moderate fibrosis) and F3-F4 (advanced fibrosis). With the help of machine learning approaches each fibrosis stage can be predicted. In this study, feature selection is performed using Genetic Algorithm (GA). With the help of different algorithms such as Decision Tree (DT), Logistic Regression (LR), Support Vector Machine (SVM), Naive Bayes (NB) and Random Forest (RF) the accuracy of these algorithms for the classification of fibrosis stages are also compared. Using GA, a set of 14 features are selected as relevant features. After the feature selection, prediction of Fibrosis is performed and then patients with fibrosis are classified into different stages (from F1-F4). This method obtained an accuracy of85.04%.

KEYWORDS*:* Liver Fibrosis, Genetic Algorithm, Feature selection, Correlation coefficient, Machine Learning Algorithms

I.INTRODUCTION

The purpose of ML is to understand the structure ofdata and fit that data into models that can be used for prediction, classification, etc. Although machine learning is a n area within computer science, it differs from traditional computational approaches [1]. Recently, 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. Using these algorithms liver fibrosis stages can be predicted.

Like prediction, feature selection is also important. Feature selection is done after data preprocessing. With the use of different features election methods such as wrapper approaches optimal features can be selected from the dataset. This method helps to increase the accuracy of the implemented model.

According to METAVIR score fibrosis stages ranges from F0 to F4 where F0 is considered as no fibrosis, F1 to F2is 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 Elastography (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. RELATED WORKS

Mahmoud El Hefnawi et al. [2], ANN and decision tree models are used for the prediction of advanced liver fibrosis in hepatitis c patients. Correlation based feature selection is used. DT gives better results. Somaya Hashem et al. [3] a mathematical model is proposed for the prediction of risk for liver fibrosis based on noninvasive methods. In this study, correlation-based feature selection is used and multi-linear regression is implemented for prediction.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 7, July 2020 ||

In Heba Ayeldeen et al. [4], a decision tree algorithm is used to predict the liver fibrosis stage in HCV patients. The features are selected based on the P-value of each variable. Dr. S. Vijayarani et al. [5], predict liver disease using Support Vector Machine (SVM) and Naive Bayes. SVM and NB are compared based on the performance factor: classification accuracy and execution time. By comparing the classification accuracy, SVM is better than NB. In the case of execution time, SVM takes more time than NB.

Manickam Ramasamy et al. [6] has implemented Decision Stump, Hoeffding Tree, J48, Logistic Model Tree (LMT), Random Forest (RF), REP (Reduced Error Pruning) Tree and Random Tree is used for the classification and comparison of Hepatitis dataset. By comparing the accuracy RF is better than any other algorithms used. Somaya Hashem et al. [7], developed a classification model using Alternating Decision Tree (ADT)for the prediction of advanced liver fibrosis which obtained an accuracy of 84.8%. The feature selection method used is Pearson correlation.

Somaya Hashem et al. [8], different machine learning algorithms such as alternating decision tree (ADT), genetic algorithm (GA), particle swarm optimization (PSO), and Multilinear regression (MReg), liver fibrosis stages are predicted and compared. The filter method is used as a feature selection method. The filter method based on the Pearson correlation coefficient is used. The highest accuracy is obtained for ADT.

Thirunavukkarasu K. et al. [9], has implemented three machine learning algorithms such as Logistic Regression, Support Vector Machine, and K-nearest neighbors for the prediction of liver diseases. In this work, feature selection is performed by finding the relationship between the attributes and the predictor variable been seen by using a pivot table. Based on the findings, attributes been selected. Accuracy of three algorithms is compared and also concluded that Logistic Regression is appropriate for predicting liver disease.

A.K.M Sazzadur Rahman et al. [10], has implemented six algorithms such as Logistic Regression, K-Nearest Neighbors, Decision Tree, Support Vector Machine, Naive Bayes, and Random Forest for accuracy comparison. Correlation based feature selection is used. This work evaluates the performance of different Machine Learning algorithms to reduce the high cost of chronic liver disease diagnosis by prediction.

Muktevi Srivenkatesh[11],has evaluated five classifiers that are Naive Bayes, logistic regression, support vector machines, Random Forest, and K-Nearest Neighbor. The objective of this work is to pick the most efficient algorithm by comparing accuracy.

III. FEATURE SELECTION

The main idea of feature selection is to remove irrelevant features before learning algorithm. Mainly there are three categories of feature selection methods: filter method, wrap- per method, and hybrid method. By removing the irrelevant features using these methods, the classification accuracy can beimproved.

A. Filter Method

The filter approach is faster than any other method in data analysis. One example of this method is correlation. Correlation can be defined as a measure of how strongly one variable depends on another. If two features are highly correlated then we can predict one from another. The disadvantage of the filter method is that it does not give a guaranty of optimal feature subset generation.

1) Pearson correlation coefficient: This is a measure frequently used in machine learning. The strength of the linear relationship between two variables is summarized between the values -1 and1.

- Positive correlation: High correlation between variables when the correlation coefficient is a positive value.
- Zero correlation: No correlation between variables when the correlation coefficient is zero.
- Negative correlation: When the correlation coefficient is a negative value then it means that values of one variable decrease as the values of another increase.

The formula for calculating Pearson correlation coefficient is:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 7, July 2020 ||

B. Wrapper Method

In order to get better results from a few features wrapper methods are used. This method gives the optimal subset of feature by iteratively. Some examples of the wrapper methods are Genetic algorithm, forward feature selection, backward feature elimination etc.

1) Genetic Algorithm: is a search-based optimization algorithm based on the principles of Genetics and natural selection. This algorithm is used to find optimal solutions to problems in research and machine learning. In this algorithm, optimization means to finding the values of input in such a way that we get thebestoutputvalue.Figure1showstheflowchartofGA.

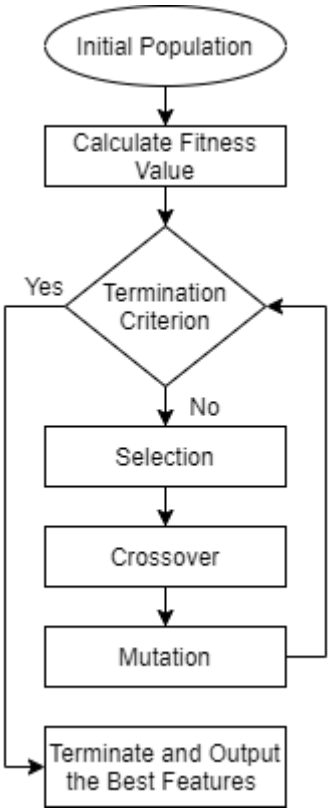

Fig. 1. Flowchart of GA

The main steps involved in this algorithm are:

- Initial population: Features in the dataset are selected randomly to create the initial population. This is the initial step. If a feature is included then its value will be 1, else 0.
- Fitness value: For the feature selection fitness value can be calculated using any of the machine learning algorithms as their prediction accuracy. Fitness value is calculated at each iteration to eliminate the set of features with low accuracy. This process stops when there is no change inaccuracy.
- Selection: Individuals with high accuracy is considered as better and selected for the next step.
- Crossover: The fittest individuals are selected to pairtheir genes to create newoff-springs.
- Mutation: Some genes of new off-springs are changed. i.e., changing 1 to 0 or 0 to 1. Thus, new individuals are produced.

C. Embedded Method

This is a technique which perform feature selection as part of the model construction process. This approach tends to be between filters and wrappers in terms of computational complexity.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 7, July 2020 ||

VI. PROPOSED APPROACH

The proposed system contains four main modules. The process starts with an initial dataset that contains 29 features. The dataset contains the values of patients with and without fibrosis. The dataset also contains 4 class values, which are the different stages of fibrosis. Figure 2 shows the system design of the proposed method.

Data preprocessing is an initial step in Machine Learning as the quality of data and the useful information that can be derived from it directly affects the ability of the model to learn; therefore, it is extremely important to preprocess the data before feeding it into the model. Redundant data are removed and, outliers and missing values are replaced in this module.

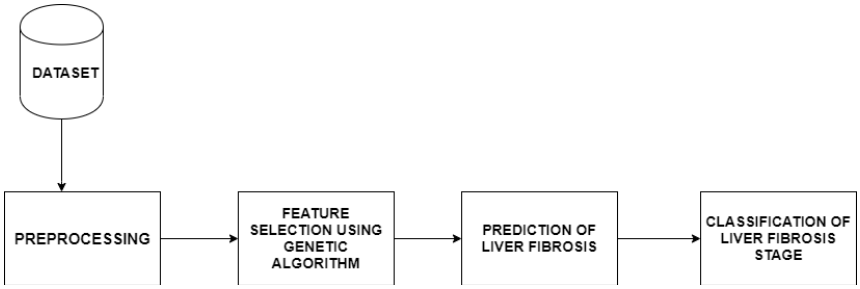

Fig. 2. System Design

Feature Selection is one of the main concepts in machine learning which hugely impacts the performance of the model. Feature Selection is important when the number of features is very large. This process reduces training time and evaluation time. GA is the algorithm used for feature selection. The fitness value for this process is calculated with the help of machine learning algorithms. The features selected by the GA will be used for the prediction of fibrosis.

For classification decision tree is used. Decision Trees are algorithms for predictive modeling in machine learning. After feature selection a set of features from a new dataset with the same number of rows. This dataset is used for the predictionof fibrosis.Thepatientswithoutfibrosishavestage0andpatients with fibrosis are considered to be a value 1. The dataset is split into two: a training set and a testing set. For the prediction, a decision tree classifier is used. The classification phase gives the fibrosis stage of patients with fibrosis. It classifies the fibrosisstagefromF1toF4.

V. RESULTANALYSIS

Feature selection was used in model constructions to eliminate the redundant features for model simplicity in order to make them easier to interpret by users. In this study, we made a comparison between feature selection using correlation and Genetic Algorithm (GA), with the help of different machine learning, approaches on the prediction of liver fibrosis in Chronic Hepatitis C patients. This study helps to find the best feature selection method and best machine learning algorithm for the prediction offibrosis.

There are two different approaches to the feature selection process. The first one is to make an independent assessment, based on the general characteristics of data. Methods be- longing to this approach are called filter methods (Pearson correlation coefficient) because the feature set is filtered out before model construction.

The second approachis to use a machine learning algorithm (GA) to evaluate different subsets of features and finally select the one with the best performance on classification accuracy. The latter algorithm will be used in the end to build a predictive model. Methods in this category are called wrapper method because the arising algorithm wraps the whole feature selection process.

In the first method, after preprocessing correlation coefficient is calculated for each feature. Features with correlation greater than .009 is selected for creating the machine learning model. from the 28 features in dataset, 21 features are selected using correlation. Then five different machine learning algorithms are used to find the best algorithm that can be used for the prediction of Liver Fibrosis. The dataset with 29 features is given to the algorithms. From the figure 3, Decision Tree has the highest accuracy(81.67%) among the other four algorithms. By comparing the accuracy, Decision Tree and Random Forest have the highest accuracies.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 7, July 2020 ||

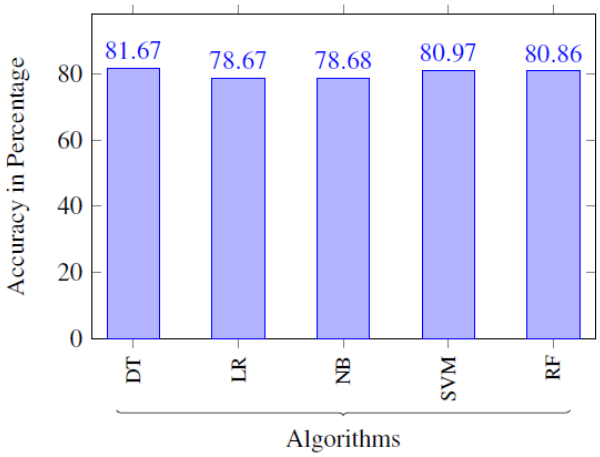

Fig. 3. Accuracy (Correlation)

The proposed method is to find the features using GA. One of the main steps in GA is to find the fitness value in each iteration. Here, to find the fitness value, machine learning algorithms are used. After preprocessing, feature selection using GA is performed. Figure 4 shows the highest accuracy of the final features by different algorithms. By comparing the accuracy, DT obtained an accuracy of 85.04%. Thus, DT is used for the next process. We also used other algorithms for the feature selection using GA to find fitness value. Figure 4 shows the accuracy of different algorithms when used for calculating the fitness value(Accuracy).

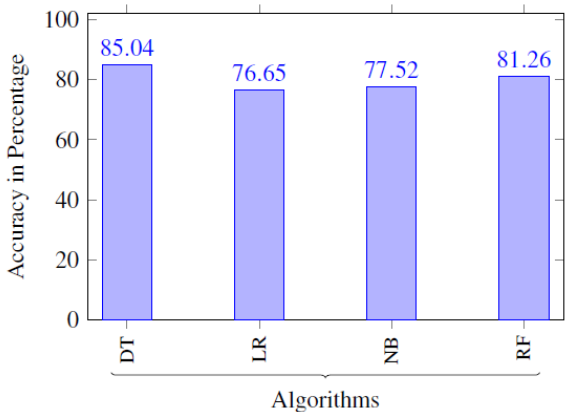

Fig. 4. Fitness value of each algorithms during feature selection

The five algorithms such as DT, LR, NB, SVM and RF are used. Each algorithm is used to calculate features fitness value. The accuracy of algorithms given are the accuracies of features selected. When DT is used for calculating fitness value, 14 features are selected which gives the accuracy 85.04%. LR selects only 2 features and gives an accuracy 76.65%. NB selects 19 features to produce an accuracy of 77.52%. When SVM is used for feature selection it takes more time than others to generate features. So, it is excluded from generating feature selection. RF selects 4 features with an accuracy of 81.26%. By comparing the algorithms DT is more accurate than other algorithms.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 7, July 2020 ||

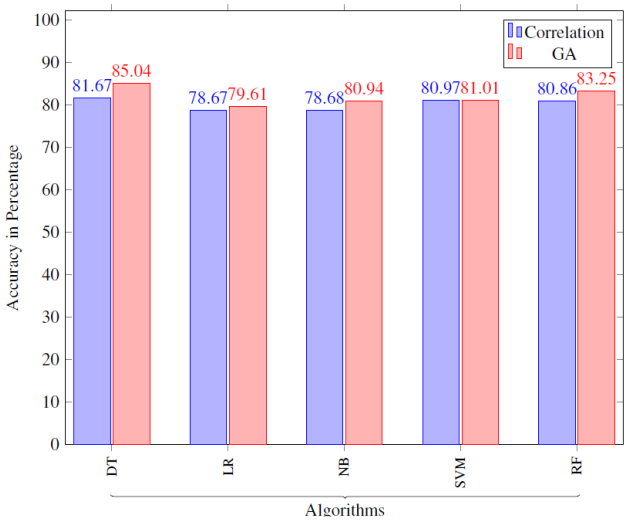

Fig. 5. Accuracy Comparison

Figure 5 is based on the accuracy comparison of features selected using correlation and GA. Blue colour shows the accuracy obtained from the features selected from correlation and red colour shows the accuracies obtained from the features selected using GA. From the figure it is clear that feature selection using GA can improve the prediction accuracy and more relevant features can be selected than the previous method which uses correlation for feature selection.

VI. CONCLUSION

This study is a comparison between different machine learning approaches for the prediction of liver fibrosis stages in Chronic Hepatitis C patients. GA is used for feature selection and is compared with feature selection using correlation. By analyzing the accuracies, we proved that feature selection using GA is more efficient than the previous methods. The proposed method obtained an accuracy of 85.04%.

We also proved that DT is the most suitable algorithm for the prediction of Fibrosis. Using GA form the 28 features in the dataset, it selects only 14 features as the most relevant features for the prediction. But when using correlation, it selected 21 features. Also concluded that we could predict fibrosis stage from F1 to F4 for HCV patients using different machine learning approaches with high accuracy. Hence the proposed model could be used as an acceptable, safe, and low-cost alternating for the prediction of fibrosis stages rather than risky alternative tools such as liver biopsy in chronic hepatitis C virus patients.

VII. FUTURE WORK

The proposed future work is:

- To work with a different dataset which contains different set of features and a greater number of data. By combining the features new subset can be obtained for better results.
- To combine other wrapper approaches with GA for per- forming feature selection and to generate classification model.

REFERENCES

[1] 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),Volume7Issue1,pp.177-183,January-February2020.

[2] MahmoudElHefnawi,MahmoudAbdalla,SafaaAhmed,WafaaElakel, 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.

[3] Somaya Hashem, Shahira Habashy, Wafaa El Akel, Safaa Abdel Raouf, Gamal Esmat, Mohamed El Adawy and

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 7, July 2020 ||

Mahmoud El Hefnawi. A Simple multi linear regression model for predicting fibrosis scores in chronic Egyptian hepatitis C virus patients. International Journal of Bio TechnologyandResearch(IJBTR).2014Jun;Vol.4,Issue3.

[4] Heba Ayeldeen, Olfat Shaker, Ghada Ayeldeen, Khaled M. Anwar, Prediction of Liver Fibrosis stages by Machine Learning model: A DecisionTreeApproach,IEEEThirdworldconference,2015.

[5] 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, April2015.

[6] 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),20thMarch2015,Coimbatore,TN,India.

[7] 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.

[8] 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/ACMTransactions on Computational Biology and Bioinformatics, 2017.

[9] Thirunavukkarasu K., Ajay S. Singh, Md Irfan and Abhishek Chowdhury," Prediction of Liver Disease using Classification Algorithms",2018 4th International Conference on Computing Communication andAutomation (ICCCA).

[10] A.K.M Sazzadur Rahman, F. M. Javed Mehedi Shamrat, Zarrin Tasnim,Joy Roy, Syed Akhter Hossain," A Comparative Study on LiverDisease Prediction Using Supervised Machine Learning Algorithms",International Journal of Scientific & Technology Research, Volume 8,Issue 11, November 2019.

[11] Muktevi Srivenkatesh," Performance Evolution of Different MachineLearning Algorithms for Prediction of Liver Disease", InternationalJournal of Innovative Technology and Exploring Engineering (IJITEE),Volume-9 Issue-2, December 2019.

[12] S. Nagaparameshwara Chary, Dr. B.Rama A Survey on ComparativeAnalysis of Decision Tree Algorithms in Data Mining , InternationalConference on Innovative Applications in Engineering and InformationTechnology(ICIAEIT-2017).

[13] V.V. Ramalingam, A.Pandian, R. Ragavendran Machine Learning Techniqueson Liver Disease -A Survey, International Journal of Engineering& Technology, 2018.

[14] Priyanga Chandrasekar, Kai Qian, Hossain Shahriar and Prabir Bhattacharya,Improving the Prediction Accuracy of Decision Tree Miningwith Data Preprocessing, 2017 IEEE 41st Annual Computer Softwareand Applications Conference.

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