Response Prediction for Chronic HCV Genotype 4 Patients to DAAs

Mohammed A. Farahat
Faculty of Computers and Information
Helwan University
Cairo, Egypt

A. Abdo
Faculty of Computers and Information
Helwan University
Cairo, Egypt

Samar K. Kassim
Faculty of Medicine
Ain Shams University
Cairo, Egypt

Khaled A. Bahnasy
Faculty of Computers and Information
Ain Shams University
Cairo, Egypt

Sanaa M. Kamal
Faculty of Medicine
Ain Shams University
Cairo, Egypt

Ahmed Sharaf Eldin
Faculty of Computers and Information
Helwan University
Cairo, Egypt

***Abstract*—Hepatitis C virus (HCV) is a major cause of chronic liver disease, end stage liver disease and liver cancer in Egypt. Genotype 4 is the prevalent genotype in Egypt and has recently spread to Southern Europe particularly France, Italy, Greece and Spain. Recently, new direct acting antivirals (DAAs) have caused a revolution in HCV therapy with response rates approaching 100%. Despite the diversity of DAAs, treatment of chronic hepatitis C genotype 4 has not yet been optimized. The aim of this study is to build a framework to predict the response of chronic HCV genotype 4 patients to various DAAs by applying Data Mining Techniques (DMT) on clinical information. The framework consists of three phases which are data preprocessing phase to prepare the data before applying the DMT; DM phase to apply DMT, evaluation phase to evaluate the performance and accuracy of the built prediction model using a data mining evaluation technique. The experimental results showed that the model obtained acceptable results.**

Keywords—HCV; DMT; Decision Tree; DAAs; Prediction Model

I. INTRODUCTION

Hepatitis C virus (HCV) is a major cause of liver disease worldwide. The WHO estimated that more than 170 million persons are infected by this virus [1]. The HCV infection is transmitted parentally through injections which are unsafe; inadequate sterilization of medical equipment in some health-care places; transfusion of unscreened blood or blood products, sexual transmission or vertical transmission from mother to infant [1]. HCV causes an acute infection which evolves to chronic hepatitis in 80% of cases [3]. Some patients with chronic hepatitis C infection develop liver cirrhosis, end stage liver disease or liver cancer [4, 5]. There are six genotypes of hepatitis C which may respond differently to the treatment. Approximately, 350000 to 500000 people die each year because of hepatitis C-related liver diseases [6].

Egypt has the highest incidence and prevalence of HCV infection worldwide [7]. HCV represents a huge public health and socio-economic problem in Egypt. The prevalence of HCV in Egypt is 12% of the population (about 11 million Egyptians) [8]. The prevalent genotype in Egypt is genotype 4. Recently, HCV genotype 4 started to spread to Europe particularly France, Italy, Greece and Spain now. Thus, HCV genotype 4 became a growing problem in other areas of the world.

The treatment of HCV has progressed from interferon monotherapy to interferon and ribavirin (RBV) combination therapy then pegylated interferon (PEG-INF) and ribavirin therapy. However, interferon based therapies were associated with multiple adverse events in addition to limited response rates especially in genotypes 1 and 4. Recently, new direct acting antivirals (DAAs) resulted in very high success rates exceeding 90% with minimal adverse events. DAAs are either given as interferon free regimen or administered with pegylated interferon. DAAs represent a breakthrough in HCV treatments since the response rate exceeds 90% Thus, DAAs represent a breakthrough in eradication of HCV. However. DAAs have not been adequately evaluated in chronic HCV genotype 4.

Data mining is not all about the used tools or database software. Data mining itself depends on building a suitable data model and structure which can be used to process, identify, and build the needed medical and clinical information. In spite of the source data form and structure, structuring and organizing the clinical information in a format which allows the data mining techniques to run in as efficient a model as possible.

Therefore, this study is designed to build an application to predict the response of chronic HCV genotype 4 patients to DAAs by applying Data Mining Techniques (DMT) on clinical information.

The remainder of this paper is structured as follows. First, we discussed the related work in Section II. This is followed by a description of the clinical data and the phases of our framework in Section III. The experimental results are discussed in Section IV. We conclude our paper in Section V and give an outlook to the future work.

[1] http://www.who.int/mediacentre/factsheets/fs164/en/

II. Related Work

Many researches have been developed to predict patients' response to treatment of HCV from clinical information using different data mining techniques.

In [2], Mohammed M.Eissa et al. used Rough Granular Neural Network model and Artificial Neural Network (ANN) for Making Treatment Decisions of Hepatitis C. This data was collected from clinical trials of a newly developed medication for HCV. It consists of 119 cases; each of which is described by 28 attributes: 23 numerical and 5 categorical attributes, the intention of the dataset is to forecast the presence or absence of the hepatitis virus related to the proposed medication. The rough set technique had been used to discover the dependency among the attributes, and to reduce the attributes and their values before the original information, remove redundant information, reduce the dimension of the information space, provide a simpler neural network training sets, and then construct and train the neural network. The experimental results show that the proposed hybrid model can acquire the advantages from the two data-mining methods (Rune Space (RS) and ANN). In addition, the integration of the Rough Sets and ANN together can produce a positive effect, enhancing model performance.

In [6] E. M. F. El Houby, et. al. applied knowledge discovery technique to predict HCV patients' response to treatment, which is a combined therapy Peg-IFN and RBV, according to a set of features. The proposed framework consists of two phases which are pre-processing and data mining. A database of 200 Egyptian cases was constructed from patients with hepatitis C virus genotype 4, who treated with combined therapy Peg-IFN and RBV for two years. This data was collected at Cairo University hospital. For each patient a record composed 12 features was registered, in addition to response feature. Associative Classification (AC) has been used to predict response to treatment in patients. AC technique has been used to generate a set of Class Association Rules (CARs). The most suitable CARs are selected to build a classifier which predicts patient's response to treatment from the selected features. The accuracy of the algorithm is high reach up to 90%.

In [7] M. ElHefnawi, et. al. made a prediction of response to Interferon-based therapy in Egyptian patients with Chronic Hepatitis C using machine-learning approaches. They used ANN and DT techniques. The study included 200 Hepatitis C patients with genotype 4 at Cairo University Hospital who were treated with combined therapy PEG-IFN-α and RBV for 48 weeks. The data was divided into 150 cases for training and 50 for validation; the maximum accuracy for ANN and DT were 0.76 and 0.80 respectively.

In [8] M. M. Eissa, et. al. introduced a Hybrid Rough Genetic Model to classify the effects of a new medication for HCV treatment through Hybrid Rough Genetic Model which has been used to predict response to new medications for HCV treatment in patients with hepatitis C virus (HCV). This data was collected from clinical trials of a newly developed medication for HCV [32, 33]. It consists of 119 cases; each of which is described by 28 attributes: 23 numerical and 5 categorical attributes. During the experiment HCV Dataset was divided into training set and test set with splitting factor 0.25. The proposed model included 4 phases (data preprocessing, data reduction, rule generation and classifications of HCV data). the proposed hybrid model can acquire the advantages from the two data-mining methods (Rough Sets and Genetic Algorithms) and therefore, produce superlative results .The Integrating Rough Sets and Genetic Algorithms together can produce a positive effect, enhancing model performance.

In [9] a framework has been developed to compare different data mining techniques' performance in predicting patients' response to treatment of HCV from clinical information. Three data mining techniques which are: (ANN, AC and DT). Data from 200 Egyptian patients with hepatitis C virus genotype 4, who were treated with combined therapy IFN plus RBV for 2 years, was collected at Cairo University Hospital. In evaluation phase, all the models built using different DM techniques for various candidate features subsets have been evaluated using test dataset of 50 cases which have been selected randomly in each iteration. This dataset is independent of the model building dataset (i.e., training dataset). According to evaluation results, the highest performance model can be selected. The best accuracy for the AC is 92% while for ANN it is 78% and it is 80% for DT.

In [10], Lin E, et. al. have used two families of classification algorithms, including Multilayer Feed Forward Neural Network (MFNN) and logistic regression as a basis for comparisons. An MFNN is one type of ANN models where connections between the units do not form a directed cycle. These classifiers were performed using the Waikato Environment for Knowledge Analysis (WEKA) software. There were 523 participants, including 350 Sustained virological response (SVRs) and 173 Non-viral response (NRs). They further converted the clinical diagnostic data into numerical forms, that is, 1 for "SVR" and 0 for "NR", respectively. To measure the performance of prediction models, they defined the accuracy as the proportion of true predicted participants of all tested participants. In addition, they used the receiver operating characteristic (ROC) methodology and calculated the area under the ROC curve (AUC).

In [11], Masayuki Kurosaki, et. al. used classification and regression tree (CART) and Statistical analysis to build a predictive model of response to the treatment in HCV. The software automatically explore the data to search for optimal split variables, builds a decision tree structure and finally classifies all subjects into particular subgroups that are homogeneous with respect to the outcome of interest. The CART analysis was carried out on the model building set of 269 patients using the same variables as logistic regression analysis.

In [12], Kazuaki Chayama et al. made a statistical analysis using the R software package (http://www.r-project.org). CART analysis was used to generate a decision tree by classifying patients by SVR, based on a recursive partitioning algorithm with minimal cost-complexity pruning to identify optimal classification factors. The association between SVR and individual clinical factors was assessed using logistic

regression. Data was collected from 840 genotype 1b chronic hepatitis C patients. In this study 465 patients achieved an SVR, whereas 375 patients were either non-responders or relapsers, yielding an overall SVR rate of 55.4%. The rate of SVR did not differ significantly between the 48- and 72-week treatment groups (55.3 vs. 56.4%, respectively; P = 0.81), but the NR rate was significantly lower in patients who were treated for 72 weeks.

In [13], Naglaa Zayed et al. made a study on retrospective data belonging to 3719 adult patients with chronic HCV infection of both sexes who were diagnosed by anti-HCV antibodies. Data cleansing was applied for detecting, correcting or removing corrupt or inaccurate record from database in addition to the removing of typographical errors or validating and correcting values against a known list of entities. High quality data was characterized by accuracy, integrity, completeness, validity, consistency, uniformity and unique-ness. Weka implementation of C4.5 (WEKA J48) decision-tree learning algorithm was applied using 19 clinical, bio-chemical, virologic and histologic pre-treatment attributes form the data of 3719 Egyptian patients with chronic HCV. The universality of the decision-tree model was validated using both internal and external validation to confirm the reproducibility of the results. They applied Statistical and Multivariable logistic regression analysis on the data. They concluded that the Prediction of treatment outcome in chronic HCV patients genotype-4 (HCV-G4) has been an important debate since even with the application of combination therapy for 48 weeks only around 50% of patients will respond.

In this research we are going to build an application to predict the response of chronic HCV genotype 4 patients to DAAs as it has not been adequately evaluated in chronic HCV genotype 4.

III. Material and Methods

A. Clinical Data

Data from 420 patients infected with hepatitis C virus genotype 4 from different centers in Europe and Egypt are analyzed. Patients were treated with four different regimens of DAAs with or without pegylated interferon. For each patient there is a record that contains 13 main features which are age, treatment regimen group, gender, body mass index (BMI), white blood cells (WBC), hemoglobin, platelets, baseline PCR, baseline core antigen, aspartate aminotransferase (AST), alanine aminotransferase (ALT), histologic grading and histologic staging. All patients were evaluated at baseline and at different time points during treatment and follow up. The treatment endpoint was sustained virologic response defines as undetectable HCV RNA 12 weeks following termination of therapy.

B. Prediction Model

In this research, a framework has been built to predict the response of Chronic HCV genotype four patients to DAAs by applying Data Mining Techniques (DMT) on clinical information, and then extract the result of DMT to be a Knowledge Base for our application to perform the prediction process. Fig. 1 shows the proposed framework which consists of (1) Data preprocessing phase to prepare the data before applying the DMT; (2) DM phase to apply DMT, (3) evaluation phase to evaluate the performance and accuracy of the built model using a data mining evaluation technique.

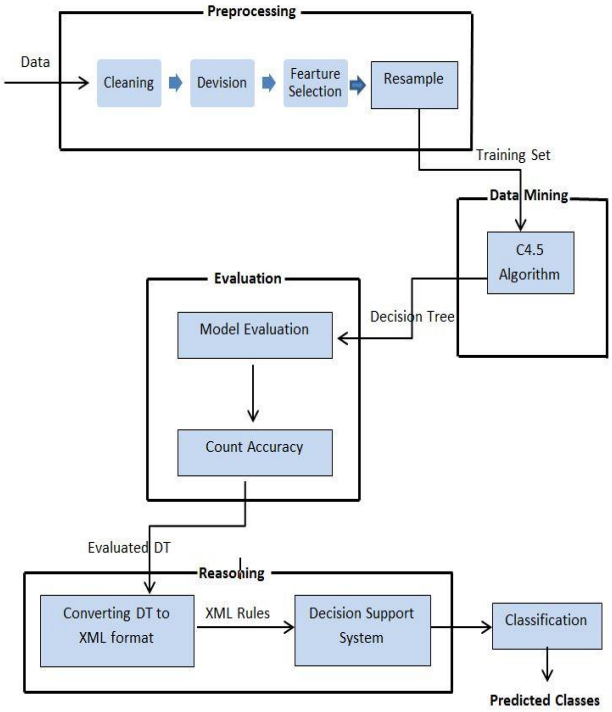


Fig. 1. The Framework of predicting the response of HCV genotype 4 to DAAs

1) Data Preprocessing Phase

In this phase, a series of steps were applied to clean, divide, select the most suitable features for the model from the patient data for applying DMT and resample the data sets into training and test sets.

Data cleaning phase is to clean the data and remove the records that contains empty values. Then the clinical data has been divided into four groups according to the regimen of DAAs treatments which are labeled as (TR1, TR2, TR3 and TR4).

Feature Selection phase is to select a subset of features relevant to the target DMT from all the features of the data set. In the filtering approach; the feature selection algorithm is independent of the DMT which applied to the selected features.

In this research eight features of 13 were selected for TR1, seven for TR2, five for TR3 and nine for TR4.The class label (Result) is considered as the PCR of the 48thweek (24 weeks following termination of therapy)

For Resampling; each treatment group has been divided into: 90% of the data as a training set and 10% as a testing set using unsupervised resample filter in Weka.

2) Data Mining Phase

Weka implementation of C4.5 (Weka J48) decision-tree learning algorithm was applied the four training data sets.

C4.5; is based on the ID3 algorithm and tries to find simple (or small) decision trees (DT's). Some premises on which this algorithm is based will be presented in the following sections.

a) Construction

Some premises guide this algorithm, such as:

- If all cases are of the same class, the tree is a leaf and the leaf is returned labeled with this class;
- For each attribute, calculate the potential information provided by a test on the attribute (based on the probabilities of each case having a particular value for the attribute). Also calculate the gain in information that would result from a test on the attribute (based on the probabilities of each case with a particular value for the attribute being of a particular class);
- Depending on the current selection criterion, find the best attribute to branch on.

b) Counting gain

This process uses the "Entropy", i.e. a measure of the disorder of the data. The Entropy of $\vec{y}$ is calculated by

$$\text{Entropy}(\vec{y}) = -\sum_{1} \frac{|yj|}{|\vec{y}|}\log\frac{|yj|}{|\vec{y}|} \quad (1)$$

Iterating over all possible values of $\vec{y}$. The conditionalEntropy is

$$\text{Entropy}(j|\vec{y}) = \frac{|yj|}{|\vec{y}|}\log\frac{|yj|}{|\vec{y}|} \quad (2)$$

and finally, we define Gain by

$$\text{Gain}(\vec{y}, j) = \text{Entropy}(\vec{y} - \text{Entropy}(j|\vec{y})) \quad (3)$$

The aim is to maximize the Gain, dividing by overall entropy due to split argument $\vec{y}$ by value j [15].

c) Pruning

Pruning is a significant step to the result because of the outliers. All data sets include a little subset of instances which are not well-defined, and vary from the other ones in its neighborhood. After the whole creation processes of the tree, which classify all the training set instances, it is pruned. This is to minimize classification errors which can be occurred because of specialization in the training set; we do this to make the tree more general.

d) Results

To show concrete examples of the C4.5 algorithm application, WEKA software tool has been used on training sets. The resulting classes are about the effect of the four treatments on the PCR result, e.g. TR1_Yes or TR1_No. Fig. 2 shows the resulting DT, using C4.5 implementation from WEKA on TR3 data set (as the smallest tree).

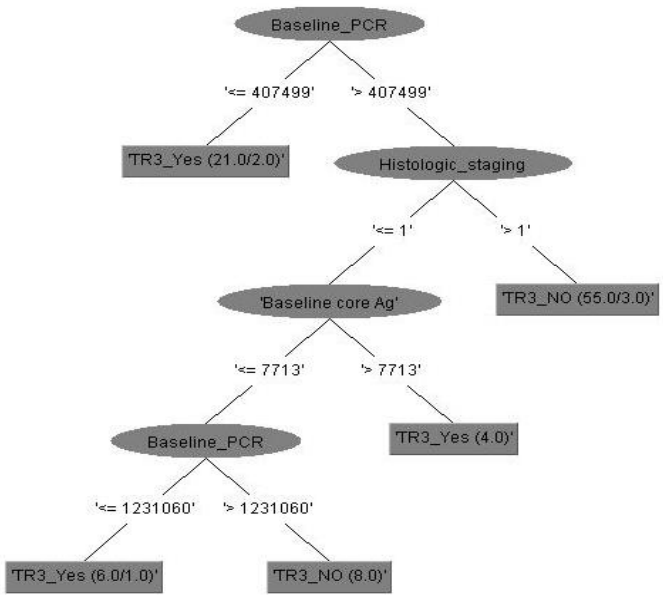


Fig. 2. DT of TR3, Created by the C4.5 algorithm

3) Evaluation Phase

The universality of the four DT models was validated using the test data sets by the hold-out validation method.

The holdout method is considered as the simplest type of cross validation. The data set is divided into two different sets, which are known as the training set and the testing set [14]. The function approximator uses the training set only to fit a function which is used to predict the output values for the testing set data which has never seen these output values before. Then the errors it makes are gathered as before to give the mean absolute test set error, which will be used to evaluate the model. This method is usually preferable to the other methods and takes less time to compute.

C. Reasoning

In the reasoning phase an application has been developed with C# programming language to perform the prediction operation. It can be considered as an expert system.

The knowledge base of this application is the model of decision tree algorithm. It is applied on Weka which delivers rules which has been converted into XML rules format to be used as an input to our DSS application.

IV. Experimental Results

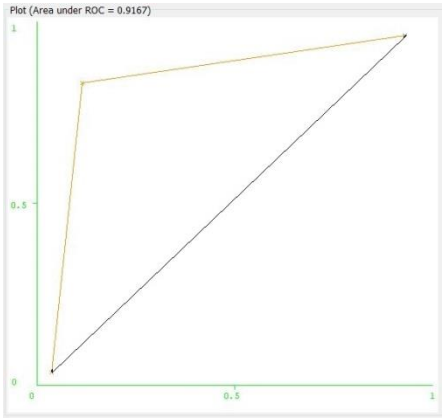
This section shows an empirical performance evaluation of the proposed framework using the applied DM techniques. Data from 420 patients infected with hepatitis C virus genotype 4 from different centers in Europe and Egypt were

used. Extensive experimental studies had been conducted in order to evaluate the model performance. The clinical data has been divided into four groups according to the regimen of DAAs treatments. Feature selection algorithm has been applied on each group. Eight features of 13 were selected for TR1, seven for TR2, five for TR3 and nine for TR4.A subset of 10% of the data had been selected to test the model and the 90% used to build the classifier.

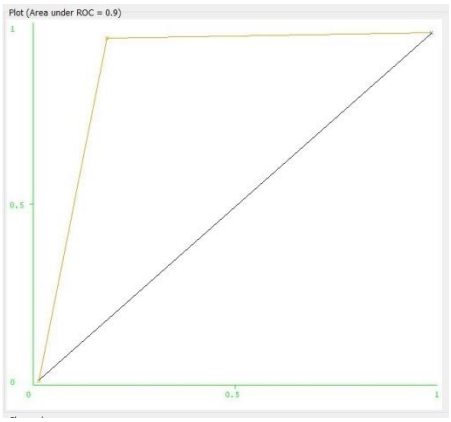
After applying the model; a large scale of statistical information were obtained. These performance measures had been used to evaluate the model as shown in Table 1. This table shows the performance for each treatment group. The DT of the 1sttreatment had 83.3% sensitivity, 100% specificity and 90.9% accuracy. The DT of the 2nd treatment had 80% sensitivity, 100% specificity and 90% accuracy. The DT of the 3rd treatment had 100% sensitivity, 71.4% specificity and 81.8% accuracy. Finally the DT of the 4thtreatment had 57.1% sensitivity, 100% specificity and 75% accuracy. The averages of the four decision trees are 80% of sensitivity, 93% of specificity and 84% of accuracy.

TABLE I. Performance of the Decision Trees of the 4 DAAs Combination After Testing

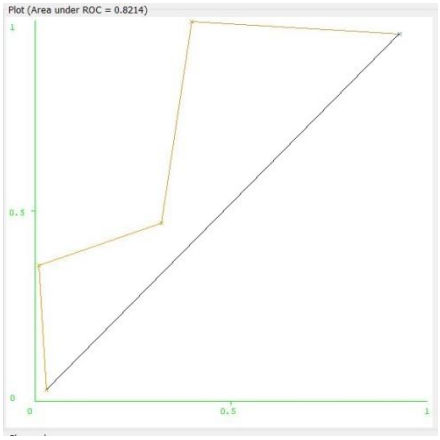
DT number	Size of the tree	Number of leaves	TP	TN	Positive Predictive value %	Negative Predictive value %	Sensitivity	Specificity	AUC%	Accuracy%
TR1_DT	25	13	5	5	100%	83.3%	83.3%	100%	91.67%	90.9%
TR2_DT	31	16	4	5	100%	83.3%	80%	100%	90%	90%
TR3_DT	9	5	4	5	66.7%	100%	100%	71.4%	82.1%	81.8%
TR4_DT	25	13	4	5	100%	62.5%	57.1%	100%	80%	75%
Average			4	5	92%	82%	80%	93%	86%	84%



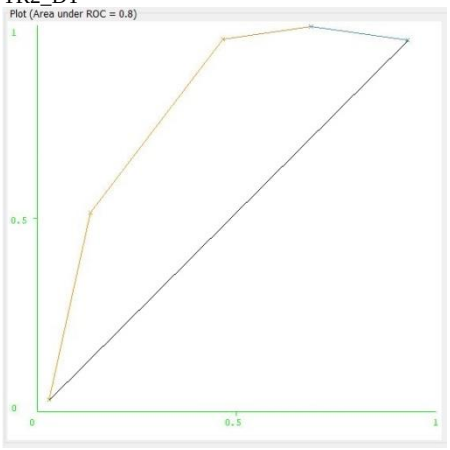
TR1_DT



TR2_DT



TR3_DT



TR4_DT

Fig. 3. The ROC curves for the 4 models with their sensitivity and specificity

Fig. 3 shows a Comparison between the Receiver Operating Characteristic (ROC) curves for the four models and their sensitivity and specificity values at the optimal cutoff points.

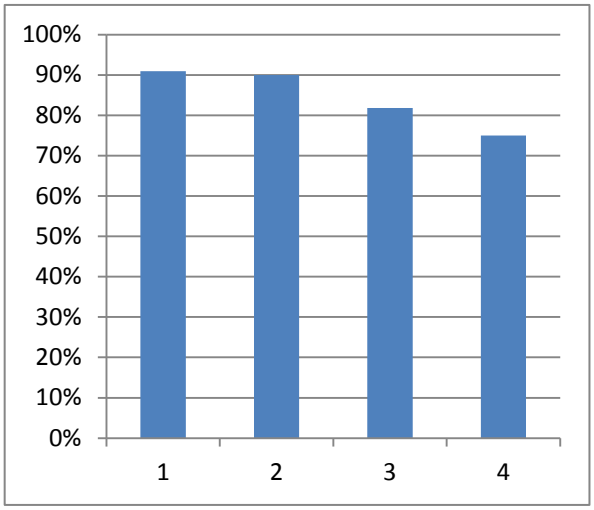


Fig. 4. Comparison of accuracy for the 4 models

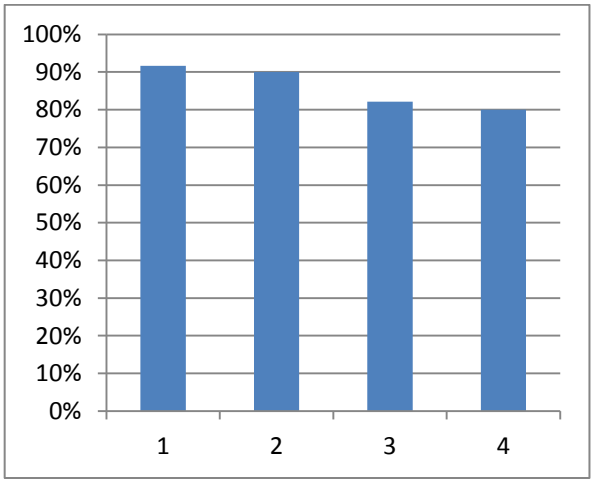


Fig. 5. Comparison of AUC for the 4 models

Fig. 4 shows a comparison between the four models regarding their accuracy. While Fig. 5 shows the comparison of Area Under Curve (AUC) for the model.

V. Conclusion and Future Work

In this research, a framework has been built to predict the response of Chronic HCV genotype 4 patients to DAAs by applying Data Mining Techniques (DMT) on clinical information. Data from 420 patients infected with hepatitis C virus genotype 4 from different centers in Europe and Egypt has been analyzed. Patients were treated with four different regimens of DAAs with or without pegylated interferon. The clinical data has been divided into four groups according to the regimen of DAAs treatments. Feature selection algorithm has been applied on each group. Decision Tree has been applied for the prediction, after that extraction of the result of DTs was performed. This constructed a Knowledge Base for our application to perform the prediction operation. The experimental results showed that the four groups give acceptable results. The best accuracy was 90.9% for the 1st group.

In the future, more data sets will be used to train other classifiers and to try more experiments. Also other techniques will be applied and more than one technique will be combined to reach as high accuracy as possible.

References

[1] S.M. Kamal, I.A. Nasser, “Hepatitis C genotype 4: What we know and what we don't yet know,” Hepatology, vol. 47, no. 4, pp. 1371-83, Apr 2008.

[2] Mohammed M.Eissa et al., "Rough - Granular Neural Network Model for Making Treatment Decisions of Hepatitis C," the 9th International Conference on INFOrmatics and Systems (INFOS2014), December, 2014.

[3] Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques 3rd edition, 2011.

[4] A. A. Freitas, “A survey of evolutionary algorithms for data mining and knowledge discovery,” in Advances in Evolutionary Computation, pp. 819–845, Springer, Berlin, Germany, 2001.

[5] M. F. Enas, “Analysis of associative classification for prediction of HCV response to treatment,” International Journal of Computer Applications, vol. 63, no. 15, pp. 38–44, 2013.

[6] E. M. F. El Houby and M. S. Hassan, “Using associative classification for treatment response prediction,” Journal of Applied Sciences Research, vol. 8, no. 10, pp. 5089–5095, 2012.

[7] M. ElHefnawi, M. Abdalla, S. Ahmed et al., “Accurate prediction of response to Interferon-based therapy in Egyptian patients with Chronic Hepatitis C using machine-learning approaches, ”in Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 803–810,August, 2012.

[8] M. M. Eissa, M. Elmogy, M. Hashem, and F. A. Badria,“Hybrid rough genetic algorithm model for making treatment decisions of hepatitis C,” in Proceedings of the 2nd Conference of Engineering and Technology and International (ICET ’14), German University in Egypt, Cairo, Egypt, 2014.

[9] Enas M. F. El Houby, "A Framework for Prediction of Response to HCV Therapy Using Different Data Mining Techniques," Hindawi Publishing Corporation, 2014.

[10] Lin E, Hwang Y, Wang SC, "Pharmacogenomics of drug efficacy in the interferon treatment of chronic hepatitis C using classification algorithms," Advances and Applications in Bioinformatics and Chemistry journal, 2010.

[11] Masayuki Kurosaki, et al, "A predictive model of response to peginterferon ribavirin in chronic hepatitis C using classification and regression tree analysis," Hepatology Research, 2010.

[12] Kazuaki Chayama et al., "Factors predictive of sustained virological response following 72 weeks of combination therapy for genotype 1b hepatitis C," J Gastroenterol (2011) 46:545–555, 2011.

[13] Naglaa Zayed et al., "The assessment of data mining for the prediction of therapeutic outcome in 3719 Egyptian patients with chronic hepatitis C," clinics and Research in Hepatology and Gastroenterology, 2012.

[14] Vijay Kumar Mago and Nitin Bhatia, Cross-Disciplinary Applications of Artificial Intelligence and Pattern Recognition: Advancing Technologies, 2011

[15] A. Sivasankari, S. Sudarvizhi and S. Radhika Amirtha Bai, “Comparative study of different clustering and decision tree for data mining algorithm,” International Journal of Computer Science and Information Technology Research, Vol. 2, Issue 3, pp. 221-232, 2014.S