https://erjm.journals.ekb.eg
ISSN: 1110-1180

DOI: ---------

ERJ
Engineering Research Journal
Faculty of Engineering
Menoufia University

Framework of Predicting the Acute Hepatitis C Outcomes Using Data Mining Techniques

Ahmed A. Abdel Qader[1*], Arabi E. Keshk[1], Sanaa M. Kamal[2], Khalid A. Elbahnasy[3]

[1] Department of Computer Science, Faculty of Computers and Information, Menoufia University,
[2]Department of Gastroenterology and Hepatology, Faculty of Medicine, Ain Shams University,
[3]Department of Computer Science, Faculty of Computers and Information, Ain Shams University,
**(Corresponding author Email: ahmadhamid92@outlook.com)*

ABSTRACT
Hepatitis C is a common disease that attacks the human liver. The hepatitis C infection could evolve into chronic hepatitis in almost 80% of cases. The acute stage of the C virus presents a turning point in the development of hepatitis C. Due to the lack of guidelines, physicians are not able to decide on whether to pursue clinical procedures or not. Furthermore, no one knows if it had healed off-hand, or it will need treatment. In this paper, a prediction model had been created to predict the acute hepatitis c outcomes based on data mining methods using clinical data. The dataset was collected from different centers in Egypt and Europe in the text format. The model depends on a framework that consists of six main phases. The phases are problem understanding, data realizing, preprocessing, modeling, appraisal (evaluation), and Visualization. Decision tree technique is the used data mining method that can produce a decision tree (prediction model). This study introduce a developed application based on a knowledge base. The knowledge base used the rules of prediction model as an input for the developed application. Then, the outcomes were predicted to be an output from the application. The experimental results showed that the hepatitis c virus core antigen is a reliable method for monitoring disease cases. The core antigen is a reliable monitoring tool for treatment decision making. Also, the averages of the four models that had been obtained are 92.3% of sensitivity, 88.91% of specificity and 90.12 % of accuracy.

Keywords: *Acute Hepatitis C Virus, Acute Hepatitis C Outcomes, Prediction Model, Data Mining Techniques, Decision Tree*

1. Introduction

The hepatitis C virus (HCV) infection is the cause of C disease. For that, Hepatitis C infection is the main reason for chronic liver disease worldwide as it affects 2.8% of the world population, corresponding to > 185 million infections worldwide. The highest rate of HCV infection is present in Egypt. HCV is considered a socio-economic epidemic allover Egypt [1]. HCV is prominently transmitted through unsafe injections, and poorly sterilized medical equipment. Chronic hepatitis is the result of acute HCV infection in 80% of cases. The liver cirrhosis is the result of chronic hepatitis C that evolves to liver cancer in the end stage. There are different treatments and each treatment has a specific genotype can be responded.

There may be no symptoms in the acute phase. HCV-RNA can be detected in the serum in almost all patients within 1–2 week(s) of exposure. HCV-RNA early-stage evaluations are more reliable than the anti-HCV test since at 2-6 months (window period) Seroconversion or at a later time in certain risk groups is detected. Spontaneous clearance is still possible at the first 6 months of the acute phase. HCV achieves spontaneous healing in 20-50% of patients while other cases mostly evolve to the chronic stage of the disease. 30% of the overall reported cases of HCV in Egypt are in the acute stage of the ailment.

In some cases, the infection will disappear automatically, and it will develop into a chronic

This paper was presented in the 1st Inter. Conf. in Technological University Education and its Role in Industry, Energy and Environmental Conservation (Cairo 15-16 May, 2022) Upon arrangement with Delta Technical University.

disease in others. Hepatitis C is regularly symptomless. Therefore, it is hard to detect and diagnose. The physicians can't make several challenging decisions because there is no approved procedure on whether to pursue clinical treatment or not. Moreover, academic research needed more efforts for controlling acute hepatitis C to establish standardized treatment guidelines using data from published clinical studies and focus on unresolved issues. Spontaneous viral extermination requires Cellular immune acknowledgment as a vital factor. Late relapse may occur after the first HCV-RNA, so follow-up (regular)tests of RNA levels are significant. Therefore, HCV-RNA should be monitored for a period of at least six months with 2–3 consecutive tests and subsequent testing if Core Antigen elevation is observed [2]. However, No one knows if it will resolve spontaneously or it will need a clincal treatment.

Data mining techniques are used in developed tools or database software in which any resercher can use, they are also used to process, identify clinical information and build a data model. Consequently, they mean that the structure and clinical information form must be processed by the method that gives the data mining techniques the proper ground to work efficiently according to the created model [3].

Thus,That study used data mining methods depend on a clinical data. The clinical data had been gathered from different centers in Egypt and Europe.The research aims to design a model that predicts acute hepatitis c outcomes.

The following sections of this paper will discuss the related work in Section 2. Section 3 is to describe clinical data and the framework phases. In Section 4, the experimental results had been discussed. Our conclusion and the future work is included in Section 5.

2. Related work

The Prediction of Acute Hepatitis C outcomes had been developed by mining Patients data, There are a lot of researches have been developed to build a model from clinical information.

In [4], the authors have developed a classification model by combining the clinical information and the serum biomarkers, the classification model predicts advanced methods of liver fibrosis. Samples that have been collected from 39,567 HCV patients were separated into two individual sets randomly. Using an alternating decision tree algorithm, two models were obtained. They used six parameters in model one and four in model two, and both of them are related to FIB-4 characteristics, then replace alanine aminotransferase with alpha-fetoprotein. They evaluated the performance of the proposed models using the receiver operating curve and sensitivity results. The results are 0.78 ROC and 86.2% for negative predictive value with 84.8% accuracy that is more accurate from FIB-4. Because of the high accuracy that came with applying the Decision Tree (DT) technique, accessing liver biopsy was reduced.

In [5], the authors have used machine learning approaches for predicting the response of interferon-based therapy from clinical data. The ANN and DT techniques had used. Patients' data of the genotype-4 of HCV had been collected from the hospital of Cairo University, these cases had treated by merging the therapy PEG-IFN-α and RBV for 48 weeks. The dataset had split by using 150 patients for the training set and 50 for testing and validation. The accuracy results are 0.76 for ANN and 0.80 for DT.

In paper [6], A framework had designed to estimate the patients' HCV response to processing from clinical information by investigating DT and ANN data mining algorithms. For that, 200 Egyptian cases of hepatitis C infection (genotype4) had covered in this research. HCV patients have been treated and observed in Cairo University Hospital for approximately two years. In the evaluation phase, using different Data Mining Techniques (DMT) built and evaluated random features in the final models. Besides, 50 cases had selected randomly to undergo testing in each iteration. The results showed that the ANN accuracy is 78%, and it is 80%for DT, while AC accuracy is 92%.

In [7], the authors developed a model to predict the treatment response of HCV by using a regression tree (CART), Statistical analysis, and classification. They developed a tool that receives data to select optimal split variables in the exploration phase, then creates a decision tree in the data mining phase. Then, in the classification phase, they classified data into homogeneous subgroups with related outcomes using the same values for analysis process of logistic regression and the CART analysis to build a set of 269.

In [8], They made a statistical analysis using the R language packages. Thus, they used the CART package to build an SVR and DT to classify patients. Classification factors had defined using a recursive partitioning algorithm to cut down the pruning of cost-complexity. They used logistic regression to find the relation between individual clinical factors and SVR. Data from 840 cases infected with chronic hepatitis C genotype 1b. SVR had been diagnosed in 465 patients, yet 375 where either non-responders or relapsed. SVR achieved a rate of 55.4%. The SVR

rate between the 48 is 55.3%, and 72-week is 56.4 % treatment groups with P=0.81 though the NR for 72 weeks was more limited in cases.

In [9], the authors made a model predict the outcome of chronic HCV genotype-4 (HCV-G4) patients' treatment. HCV samples had collected from adult patients. The DT algorithm had applied to 3719 patients using Weka implementation of C4.5 (weka J48). For results validation, both internal and external validation was used, and applied statistical and Multivariable logistic regression analysis on the cases. The results made it clear that only 50% of patients with chronic HCV will respond to the 48 weeks of combination therapy.

In [10], longitudinal research had presented to determine the cost-effectiveness and accuracy results of hepatitis C core antigen assay in observing the acknowledgment to pegylated interferon (PEG-IFN) and ribavirin therapy for chronic HCV genotype 4 (G4). The results showed that Hepatitis C core Ag is an efficient way of monitoring patients with rapid turnaround time and less expensive compared to other tools. 410 patients with chronic hepatitis C genotype 4 met inclusion criteria and were enrolled in the study. The sustained virological response rate (SVR) was 66.34%.

In [11], They used differnet classification techniques to classfiy the patients suspected of being infected with HCV, the uses techniques the Support Vector Machine (SVM), Gaussian Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), Logistic regression (LR), and K-nearest neighbors 1 (KNN) algorithm . They used data set that had been provided bu university of california and Irvine Machine Learning Repository .The results showed that the area under the curve (AUC) was 0.921 for LR, 963 for KNN, 0.953for DT, 0.972 for SVM, 0.896 for Gaussian NB, and 0.998for RF models .

3. DATA and methods

3.1. Clinical Data

A global collection of data from 300 HCV patients were analyzed. The dataset was collected from different centers in Egypt and Europe. For each case, there are 59 main features that are group, gender, age, occupation, social class, smoking, marital status, transmission, fever, nausea/vomiting , headache, diarrhea, malaise, jaundice, dark urine, epic pain , hip pain, liv span , WBC base (x1000) , WBC (x1000) 4 to 48 weeks, RBC base (M), RBC (M) 4 to 48 Weeks, HB base, , HB 4 to 48 Weeks, Base Plateletsx1000, Plateletsx1000 4 to 48 Weeks, Bil Base, Bil 2 to 48 Weeks, ALT Base, ALT 2 to 48 weeks, ALTEF, Genotype, HCV AB(baseline),RNA base ,RNA 2 to 72 weeks , IL 28 B CC, IL 28 B CT, IL 28 B TT, Baseline core Ag, Week 4 to 72 weeks core Ag, Category. All cases had estimated at baseline and various periods through treatment and carry-out tests.

3.2. Model-To-Predict

This research has presented a unique model to predict the outcomes of acute hepatitis C. DMT is mainly applied clinical data. Also, this study introduce a developed application based on a knowledge base. The knowledge base used the results of DMT as an input for the developed application. Then, the outcomes were predicted to be an output from the application.

Figure. 1 shows the paper's framework depends on and includes six main phases which are business understanding phase to identify the central objectives of our project by identifying the variables that need to be predicted, and data understanding phase to collect and understand the data features before applying the preparation phase. Moreover, the data preparation phase to put the data in a suitable form, so the DMT can work properly in the modeling phase. Besides, the evaluation phase aims to evaluating the model performance. In the evaluation method, the accuracy result of the developed model and another statistical results had been obtained by using the data mining process. Visualization phase aims to communicate with the results had been obtained.

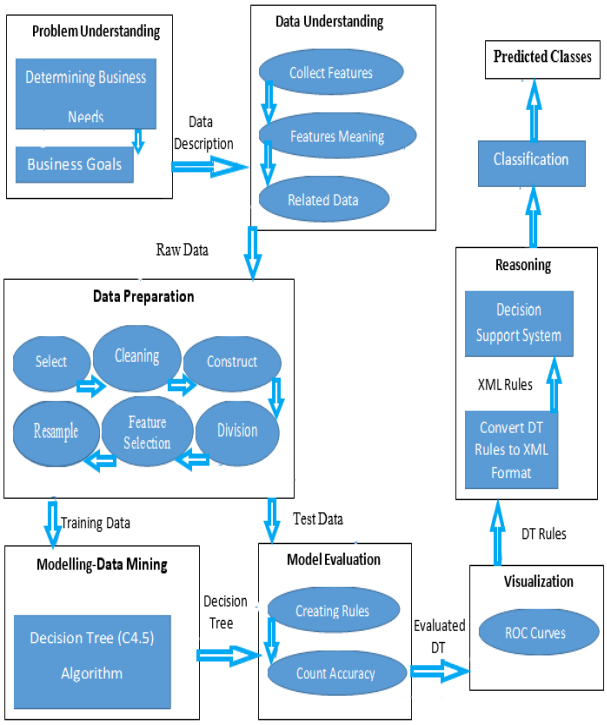

Figure. 1. Prediction framework of the acute hepatitis c outcomes

3.2.1. Business Understanding Phase

In this phase, the needs and goals must be identified to solve the problem.

The business needs phase defines the problem of acute hepatitis c outcomes, the paper aims to predict if the patients will be healed spontaneously or not, the prediction process depends on clinical data for cases with acute hepatitis c infection. The dataset contains two instances of a class label named Spontaneous Recovery and Chronic.

The business goals phase is to minimize the burden cost of treatment and determine if the patients need a treatment or not.

3.2.2. Data Understanding Phase

This phase aims to collect and know the meanings of the features and related data.

a) The features collection step is to gather the features that describe the state of patients through a medical test such as RNA, core antigens, with their follow-up test, etc.
b) Acknowledging the meaning of different Features helps in understanding the results, the raw data, and create the rules that helps us in the prediction process and the decision making.
c) Related data is to collect the relevant information wisely to solve the problem.

3.2.3. Data preparation Phase

This phase aims to select, clean, construct, divide, select features, and resample the data to be suitable for the model and for applying DMT. A select phase is to select manually the main features and ignore the features that contain the same values and the records numbering. Data cleaning is the phase that works to eliminate the empty values and their records. The construction phase is to organize and sort the data if needed to be readable and ready for the division phase. Furthermore, clinical data will be divided according to treatment groups which are (G1, G2, G3, and G4). A subset of data has been selected through the feature selection phase to be relevant to DMT. Both the DMT and feature selection method is independent.

In the study, fourteen features had extracted for G1, fifteen during G2, ten for G3, and sixteen for G4. The class label (Category) is considered as the result of patient status which is yes (chronic and needs treatment) or no (spontaneous recovery and does not need treatment). Resampling is to divide the data into two sets where the first set is 90% of the whole data and the second set is 10% of the whole data using unsupervised resample filter in Weka software for each treatment group.

3.2.4. Data Mining Phase

The C4.5 decision-tree learning algorithm in weka software (Weka J48) had been used for the four training data sets. The output of that learning algorithm is decision trees where C4.5 uses the ID3 algorithm. That algorithm depends on some premises that had explained in the next subsections.

3.2.4.1. The Tree structure

Algorithm's main pillars that had been used,

- If whole states produce a similar class, the tree converts to a leaf. Then this leaf returns with the class name.
- A test run on every attribute calculates their potential information and calculates the information gain.
- Find the most relevant property to the branch.

There are a lot of advantages of using C4.5 decision tree learning algorithm such as it's widely used and it's very easy to understand for technical or non-technical people and that feature is very important for this study, so It is not necessary to have any statistical knowledge. Also, it's not affected by some issues like the anomalies and outliers in data comapring to most of the machine learning algorithms. On the other side there are distavantages of using decission tree learning algorithm like over-fitting problem, overfitting is a problem that makes the genereated tree isn't generalized and create over complex tree, but This problem can be solved by setting constraints on model parameters and pruning the generated tree.

3.2.4.2. Information gain

The method of counting gain depends on the entropy, and entropy calculates the sample homogeneity. If the entropy equals zero, it means that the sample is totally homogeneous. On the other side if the example had equally split, it produces an entropy of one. Where pi is the proportion of instances for Set of features S to i-th class:

Entropy $E(S) = -\sum_{i=1}^{n} p_i . log_2 p_i$ (1)

But the entropy only measures the quality of a set of samples. The average weight of all results was estimated next to the split process had done using their size:

Weighted Average

$$I(S,A) = \sum_i \frac{|S_i|}{|S|} . E(S_i) \quad (2)$$

And finally, Gain defined by: Information Gain

$$G(S,A) = E(S) - I(S,A) \quad (3)$$

The property that maximizes the difference had picked, or that decreases the un-orderedness most.
The maximization of information gain and the minimizing of average entropy is our aim, where E(S) is constant for all attributes A [3].

3.2.4.3. Pruning

Pruning make the tree more general and minimize the erros of classification, that may be caused by selecting a particular training set. It also helps us to avoid the outliers of the result. An ill-defined subset of instances was seleted from all data sets which differ from its neighbored instances. The pruning process comes next to the production of the tree for the training instances and make the tree more genereal.

3.2.4.4. Results

C4.5 is one of the embeded algorithms in weka software. It takes the training sets as inputs and produces the predicted classes as outputs. The predicted classes are about the outcomes of acute hepatitis C for the four groups on the Core Antigen result, e.g. chronic or spontaneous recovery for patients. Figure. 2 displays the DT output by using WEKA J48 on the G2 data set.

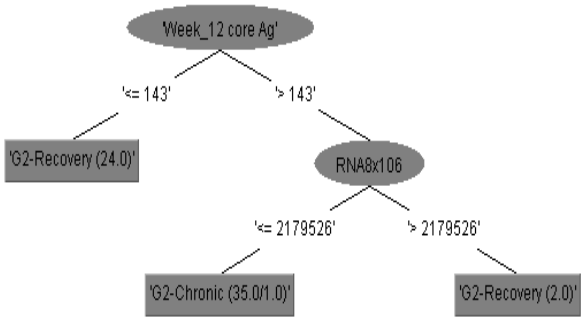

Figure 2- Decision Tree of Group 2 created by WEKA J48 (C4.5) algorithm

3.2.5. Evaluation Phase

The hold-out validation method has been performed on the four decision tree models by using test data sets. The cross-validation can be performed by the holdout method. Both of the training set and test set is the result of the whole data set after the splitting process [3]. The function approximator utilizes the training instances to match the predicted values produced by a testing set function, which has never perceived these output earlier. The mean absolute test set error of the resulting errors is calculated to evaluate the model. This method takes less time than other methods.

There are a lot of statisitcis that had been computed and obtained like Area Under Curve(AUC) and Acuuracy and Senesitivity and Specifity, lets consider that TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative.

Sensitivity (True positive Rate) TPR = TP / (TP + FN) (4)

Specificity (True Negative Rate) TNR = TN / (FP + TN) (5)

Accuracy ACC = (TP + TN) / (TP + TN+FP+FN) (6)

3.2.6. Reasoning

This study used a C# programming language to develop a DSS application. That application can help healthcare providers in the prediction process, and it can be viewed as an expert tool.

The resulting model of the DT method is the knowledge base of this tool. The application input is XML rules which are produced by converting the model of a decision tree on WEKA.

4. Experimental Results

The current section presents the results that have been obtained from the proposed framework by using DMT. A global collection of data from 300 HCV patients were analyzed. The model performance had been evaluated by extensive and conducted experimental studies. The clinical data had been divided into four groups according to the treatment group. Each group was selected by the feature selection algorithm, fourteen features were selected for G1, fifteen for G2, ten for G3, and sixteen for G4. 90% of data were used as a training set, and the other 10% were used as a test set.

Statistical information was obtained after applying the model. Table 1 shows the resulting performance measures of the model evaluation for each treatment group. The decision tree of 1st group showed sensitivity by, specificity by 91.18 % and accuracy by 90.16 %. The decision tree of the 2nd group showed sensitivity by 100 %, specificity by 89.47 % and accuracy by 93.44 %. The decision tree of the 3rd group showed sensitivity by 88 %, specificity by 86.1 % and accuracy by 86.89 %. Finally the decision tree of the 4th group showed sensitivity by 92.31%, specificity by 88.89 % and accuracy by 90 %. The averages of the four decision

trees showed sensitivity by 92.3%, 88.91% specificity by 88.91% and accuracy by 90.12 %.

A comparison had been made for the four models using their Receiver Operating Characteristics (ROC) curves and specificity and Sensitivity values at the optimal cutoff points in Figure. 7, a comparison between the accuracies for the four models in figure. 8 and a comparison between AUC (Area Under Curve) for the models in figure.9.

Table 1. The statistical results of the four models decision tree after testing

DT Number	Size of Tree	Number of Leaves	TP	TN	Positive Predictive Value %	Negative Predictive Value %	Sensitivity %	Specificity %	AUC %	Accuracy %
G1-DT	5	3	24	31	88.9	91.2	88.9	91.2	92.1	90.2
G2-DT	5	3	2	34	85.2	100	100	89.47	94.12	93.4
G3-DT	5	3	22	31	81.5	91.2	88	86.1	85.29	86.9
G4-DT	5	3	12	24	80	96	92.3	88.9	88.4	90
Average			20	30	83.9	94.6	92.3	88.9	89.96	90.1%

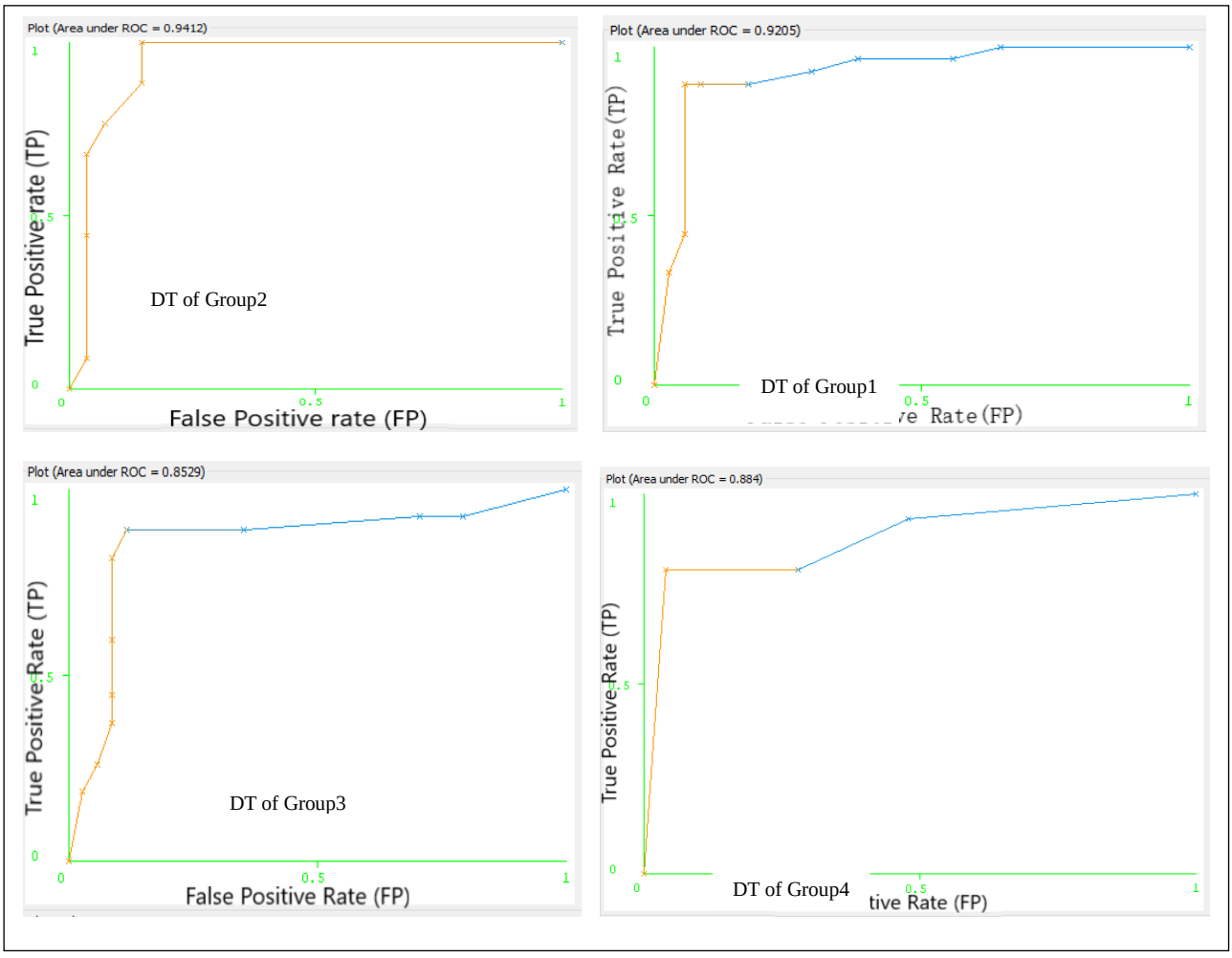

Figure 7. Comparison of the four models according to their ROC Curves

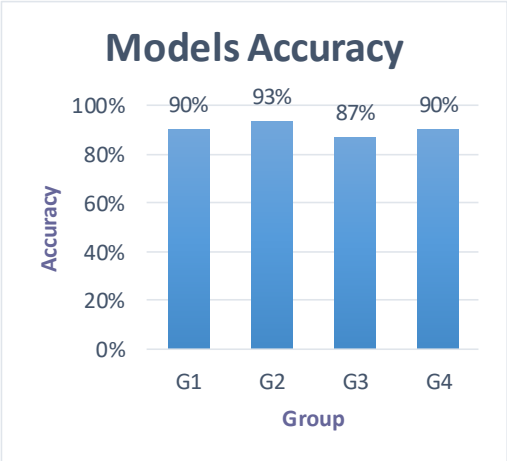

Figure. 8 comparing the accuracy for each model of the four models

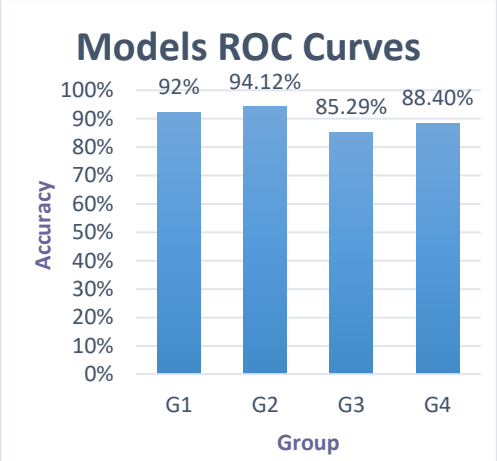

Figure. 9 comparing AUC for each model of the four models

5. CONCLUSIONS

In this study, a system had been developed to predict the acute hepatitis c outcomes. This can be approached when data mining methods are applied on clinical information. A global collection of data from 300 HCV patients were analyzed. Four groups had been extracted after the division process of clinical data according to the treatment group. We have selected the features for each group by using the feature selection algorithm and then we have extracted the result of DTs after the prediction was performed by using decision tree.

An application had been developed that depends on the extracted results that were obtained as a Knowledge Base to predict the patient state.

The experimental results showed that hepatitis c virus core antigen does a reliable process to monitor the cases of the disease and is a cost-effective monitoring tool for treatment decision making. Also, the model obtained acceptable results. 93.4% is the most accurate rate we achieved that is much better than others in the related work that had been illustrated in a previous section.

Finally, the business goals had been achieved as it's demonstrated in the business understanding phase. This study produced a prediction model that includes Core Antigen, the core antigen can minimize the cost since it is less expensive than other methods.

In future work, larger datasets and various techniques will utilize for attaining high accuracy.

6. References

[1] S. M. Kamal and I. A. Nasser, "Hepatitis C genotype 4: What we know and what we don't yet know," *Hepatology*, vol. 47, pp. 1371--1383, 2008.

[2] T. Santantonio, J. Wiegand and J. T. Gerlach, "Acute hepatitis C: current status and remaining challenges," *Journal of hepatology*, vol. 49, pp. 625--633, 2008.

[3] A. A. Freitas, "A survey of evolutionary algorithms for data mining and knowledge discovery," in *Advances in evolutionary computing*, Springer, pp. 819--845, 2003.

[4] S. Hashem, G. Esmat, W. Elakel, S. Habashy, S. Abdel Raouf, S. Darweesh, M. Soliman, M. Elhefnawi, M. El-Adawy and M. ElHefnawi, "Accurate prediction of advanced liver fibrosis using the decision tree learning algorithm in chronic hepatitis C Egyptian patients," *Gastroenterology research and practice*, vol. 2016, p. 7, 2016.

[5] M. ElHefnawi, M. Abdalla, S. Ahmed, W. Elakel, G. Esmat, M. Elraziky, S. Khamis and M. Hassan, "Accurate prediction of

response to Interferon-based therapy in Egyptian patients with Chronic Hepatitis C using machine-learning approaches," in *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 771--778, 2021.

[6] E. M. El Houby, "A framework for prediction of response to HCV therapy using different data mining techniques," *Advances in bioinformatics*, vol. 2014, p. 7, 2014.

[7] M. Kurosaki, K. Matsunaga, I. Hirayama, T. Tanaka, M. Sato, Y. Yasui, N. Tamaki, T. Hosokawa, K. Ueda and K. Tsuchiya, "A predictive model of response to peginterferon ribavirin in chronic hepatitis C using classification and regression tree analysis," *Hepatology research*, vol. 40, pp. 251--260, 2010.

[8] K. Chayama, C. N. Hayes, K. Yoshioka, H. Moriwaki, T. Okanoue, S. Sakisaka, T. Takehara, M. Oketani, J. Toyota and N. Izumi, "Factors predictive of sustained virological response following 72 weeks of combination therapy for genotype 1b hepatitis C," *Journal of gastroenterology*, vol. 46, pp. 545--555, 2011.

[9] N. Zayed, A. B. Awad, W. El-Akel, W. Doss, T. Awad, A. Radwan and M. Mabrouk, "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*, vol. 37, pp. 254--261, 2013.

[10] S. Kamal, S. Kassim, E. El Gohary, A. Fouad, L. Nabegh, T. Hafez, K. Bahnasy, H. Hassan and D. Ghoraba, "The accuracy and cost-effectiveness of hepatitis C core antigen assay in the monitoring of anti-viral therapy in patients with chronic hepatitis C genotype 4," *Alimentary pharmacology \& therapeutics*, vol. 42, pp. 307--318, 2015.

[11] Reza Safdari, Amir Deghatipour, Marsa Gholamzadeh, Keivan Maghooli, " Applying data mining techniques to classify patients with suspected hepatitis C virus infection," Elsevier-*Intelligent Medicine*, vol. 1, ppI. S2667-1026(22)00002-X, 2022.