Frontiers in Health Informatics www.ijmi.ir
2023; 12: 137 Open Access

# Effective factors in diagnosing the degree of hepatitis C using machine learning

Mohammadjavad Sayadi[1], Vijayakumar Varadarajan[2,3,4], Elahe Gozali[5,6], Malihe Sadeghi[7]*

[1]Department of Computer Engineering, Technical and Vocational University (TVU), Tehran, Iran
[2]School of Computer Science and Engineering, The University of New South Wales, Sydney, Australia
[3]Dean International, Ajeenkya D Y Patil University, Pune, India
[4]School of Business and Management, Geneva, Switzerland
[5]Health and Biomedical Informatics Research Center, Urmia University of Medical Sciences, Urmia, Iran
[6]Department of Health Information Technology, School of Allied Medical Sciences, Urmia University of Medical Sciences, Urmia, Iran
[7]Department of Health Information Technology, Sorkheh School of Allied Medical Sciences, Semnan University of Medical Sciences, Semnan, Iran

**Article Info**

***Article type:***
*Research*

***Article History:***
*Received: 2023-03-04*
*Accepted: 2023-04-10*
*Published: 2023-04-16*

*** Corresponding author:***
*Malihe Sadeghi*

*Department of Health Information Technology, Sorkheh School of Allied Medical Sciences, Semnan University of Medical Sciences, Semnan, Iran*

*Email: sadeghiii.m@gmail.com*

***Keywords:***
*Hepatitis C*
*Machine Learning*
*HCV*

## ABSTRACT

Introduction: Hepatitis C virus (HCV) is a major public health threat, which can be treated if diagnosed early, but unfortunately, many people with chronic diseases are not diagnosed until the final stages. Machine learning and its techniques can be very helpful in diagnosis. This study examines the factors affecting hepatitis C diagnosis using machine learning.

Material and Methods: A total of 27 features were used with a dataset containing 1385 records of patients with different grades of HCV. The dataset was clean and preprocessed to ensure accuracy and consistency. To reduce the dimension of the dataset and determine the effective features three feature selection, Pearson Correlation, ANOVA, and Random Forest, were applied. Among all the algorithms, KNN, random forests, and Deep Neural Networks were selected to be utilized, and then their evaluation metrics, such as Accuracy and Recall. To create prediction models, fifteen features were selected for the mentioned machine learning algorithms.

Results: Performance evaluation of these models based on accuracy showed that Deep Learning with Accuracy = 92.067 had the highest performance. KNN and Random Forest had almost the same performance after Deep Learning. This performance was achieved on dataset containing features that were selected by ANOVA feature selection.

Conclusion: Machine learning has been very effective in solving many challenges in the field of health. This study showed that using data-mining algorithms also can be useful for HCV diagnosing. The proposed model in this study can help physicians diagnose the degree of HCV at an affordable and with high accuracy.

Cite this paper as:
Sayadi M, Varadarajan V, Gozali E, Sadeghi M. Effective factors in diagnosing the degree of hepatitis C using machine learning. Front Health Inform. 2023; 12: 137. DOI: 10.30699/fhi.v12i0.440

## INTRODUCTION

Hepatitis C is a type of liver disease caused by the C virus (HCV) that can be deadly if undetected [1]. This virus is transmitted through blood, and the infection enters small amounts of blood and spreads in different ways [2]. HCV can cause or provoke chronic hepatitis and acute hepatitis, ranging from a mild illness lasting a few weeks to severe illness and death [2]. This disease can progress slowly and even cause cancer [1]. Around 71 million people worldwide are affected by this disease [3]. The risk of this disease can be reduced substantially through prompt diagnosis [4].

Artificial intelligence has entered health care in recent years, and its use in health has attracted much more attention than any other industry [5]. Artificial intelligence is applied in medicine to use automated diagnostic processes and to monitor people who need health care [6-8].

The increased use of artificial intelligence in

Copyright© 2023, Published by Frontiers in Health Informatics. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International (CC BY) License (http://creativecommons.org/).

healthcare processes will allow care providers to automate a significant amount of their work processes, freeing up more time for medical professionals to perform other tasks that cannot be automated [5, 9]. One of the most popular forms of artificial intelligence is machine learning which, as one of the subsets of artificial intelligence, has many uses in diagnosing and predicting the occurrence of various types of diseases [10]. Machine learning algorithms collect data and use it to create models to take intelligent actions. In recent years, many studies have been conducted by authors in the field of health who have used machine learning algorithms to build prediction models based on clinical records [6]. In the field of hepatitis, several studies have been conducted using machine learning to diagnose hepatitis and its degree [11-13].

Previous studies have not had the same approach in reporting the results of presenting the hepatitis C disease degree diagnosis model using machine learning algorithms, but what has been determined from their review is that the evaluation indicators of the models implemented in these studies degree [11-13] such as model accuracy, it is not high. The set of features used in them does not cover all the parameters needed in diagnosis. The present study has tried to include all the parameters affecting the diagnosis of the degree of hepatitis C disease in the data set and use different methods to select the features so that the best and most effective features are used for diagnosis in modeling. Finally, a model should be used to diagnose the degree of the disease that has high accuracy and reliability.

## MATERIAL AND METHODS

This study utilizes a retrospective observational design which has been conducted to find the effective factors for diagnosing the degree of Hepatitis C Virus (HCV).

### Dataset

In this study, a dataset was used which contains 27 features and 1385 records of patients with different grades of HCV, which can be categorized into three categories: (1) Demographic, (2) Symptom, (3) Liver Function Tests (LFTs) and (4) Blood Panels before the start of treatment (Table 1).

**Table 1: Dataset and features description**

| Feature name | Feature Category | Description | Data type |
|---|---|---|---|
| Age | Demographic | Age of patient | Numeric |
| Gender | Demographic | Gender of patient (1: Male, 2: Female) | Numeric (bool) |
| BMI | Symptom | Body Mass Index of patient | Numeric |
| Fever | Symptom | Presence of Fever (1: Absence, 2: Presence) | Numeric (bool) |
| Nausea/Vomiting | Symptom | Presence of Nausea/Vomiting (1: Absence, 2: Presence) | Numeric (bool) |
| Headache | Symptom | Presence of Headache (1: Absence, 2: Presence) | Numeric (bool) |
| Diarrhea | Symptom | Presence of Diarrhea (1: Absence, 2: Presence) | Numeric (bool) |
| Fatigue & generalized bone ache | Symptom | Presence of Fatigue & generalized bone ache (1: Absence, 2: Presence) | Numeric (bool) |
| Jaundice | Symptom | Presence of Jaundice (1: Absence, 2: Presence) | Numeric (bool) |
| Epigastric pain | Symptom | Presence of Epigastric pain (1: Absence, 2: Presence) | Numeric (bool) |
| WBC | Liver Function Tests | White Blood Cell count | Numeric |
| RBC | Liver Function Tests | Red Blood Cell count | Numeric |
| HGB | Liver Function Tests | Haemoglobin | Numeric |
| Plat | Liver Function Tests | Platelet count | Numeric |
| AST 1 | Blood Panels | Aspartate Aminotransferase Enzyme Level at one week | Numeric |
| ALT 1 | Blood Panels | Alanine Aminotransferase Enzyme Level at one week | Numeric |
| ALT4 | Blood Panels | Alanine Aminotransferase Enzyme Level at four weeks | Numeric |
| ALT 12 | Blood Panels | Alanine Aminotransferase Enzyme Level at 12 weeks | Numeric |
| ALT 24 | Blood Panels | Alanine Aminotransferase Enzyme Level at 24 weeks | Numeric |
| ALT4 36 | Blood Panels | Alanine Aminotransferase Enzyme Level at 36 weeks | Numeric |
| ALT 48 | Blood Panels | Alanine Aminotransferase Enzyme Level at 48 weeks | Numeric |
| ALT after 24 w | Blood Panels | Alanine Aminotransferase Enzyme Level after 24 weeks | Numeric |
| RNA Base | Blood Panels | RNA at the start of treatment | Numeric |
| RNA 4 | Blood Panels | RNA at four weeks | Numeric |
| RNA 12 | Blood Panels | RNA at 12 weeks | Numeric |
| RNA EOT | Blood Panels | RNA at the End of Treatment | Numeric |
| RNA EF | Blood Panels | RNA Elongation Factor | Numeric |
| Baseline histological Grading | Target Column | Baseline histological Grading (0-13) | Numeric |

### Data Preprocessing

The dataset was cleaned and preprocessed to ensure accuracy and consistency. Missing values were handled using appropriate techniques (imputation or exclusion). Categorical variables were encoded into numerical representations. Data normalization or standardization was applied to ensure consistency in scale.

## Feature Selection

Feature selection techniques were employed to identify the most relevant factors. Univariate and multivariate statistical analysis methods were used, and feature selection criteria were based on statistical significance, correlation analysis, or domain knowledge. This study used Pearson Correlation, ANOVA, and Random Forest as the main feature selection algorithms.

## Modeling

Various machine learning algorithms were applied to build predictive models. Among all the algorithms, KNN, random forests, and Deep Neural Network were selected to be utilized and then their evaluation metrics, such as Accuracy, Recall and Specificity were compared.

Accuracy=(TP+TN)/(TP+TN+FP+FN) (Eq. 1)

Recall= (2*TP)/(TP+FN) (Eq. 2)

Specificity=(2*TN)/(TN+FP) (Eq. 3)

In all the modeling algorithms, Holdout was used to split up the dataset into "Train" and "Test" sets. The models were trained on the training set and evaluated on the testing set.

# RESULTS

This section will present the results of feature selection methods and machine learning models. Python programming language version 3.8.1 was used for all the implementations.

Table 2 shows the features selected by Pearson Correlation, ANOVA, and Random Forest.

To create prediction models, fifteen features were selected for the mentioned machine learning algorithms. Performance evaluation of these models based on accuracy showed that Deep Learning with Accuracy = 92.067 had the highest performance. KNN and Random Forest had almost the same performance. Table 3 shows all the evaluation metrics for these machine-learning algorithms.

ANOVA method selected 11 features as the best features to predict the target column. According to Table 4, Deep Learning and Random Forest achieved the best accuracy of 93.51 and 93.029, respectively, which are the highest performance. Table 4 shows all the evaluation metrics for these machine learning algorithms.

**Table 2: Selected features by three feature selection methods**

| Pearson Correlation | ANOVA | Random Forest |
|---|---|---|
| Age | Age | Age |
| BMI | Gender | BMI |
| Fever | Fever | Nausea or Vomiting |
| Nausea or Vomiting | Nausea or Vomiting | Plat |
| Diarrhea | Diarrhea | AST1 |
| Fatigue and generalized | WBC | ALT 1 |
| Bone ache | HGB | ALT 4 |
| WBC | ALT 4 | ALT 24 |
| HGB | ALT 24 | RNA 4 |
| Plat | RNA 4 | ALT 36 |
| AST 1 | RNA 12 | RNA EOT |
| ALT4 | | RNA Base |
| ALT 48 | | |
| RNA 4 | | |
| RNA 12 | | |
| RNA EOT | | |

**Table 3: Machine learning models performance for Pearson Correlation**

| Method | Accuracy | Specificity | Recall (Sensitivity) |
|---|---|---|---|
| Deep Learning | 92.067 | 92.098 | 91.768 |
| KNN | 90.144 | 90.089 | 82.682 |
| Random Forest | 90.385 | 90.324 | 83.993 |

**Table 4: Machine learning models performance for ANOVA**

| Method | Accuracy | Specificity | Recall (Sensitivity) |
|---|---|---|---|
| Deep Learning | 93.51 | 93.442 | 93.354 |
| KNN | 89.283 | 88.464 | 86.67 |
| Random Forest | 93.029 | 92.989 | 92.746 |

**Table 5: Machine learning models performance for Random Forest**

| Method | Accuracy | Specificity | Recall (Sensitivity) |
|---|---|---|---|
| Deep Learning | 92.548 | 91.323 | 92.44 |
| KNN | 91.686 | 89.949 | 91.368 |
| Random Forest | 88.983 | 89.492 | 88.386 |

Twelve Features were selected by Random Forest and then mentioned machine learning algorithms were utilized on these features. For this feature selection, Deep Learning had the best performance with an accuracy of 92.548, and then KNN achieved second place with an accuracy of 91.686. Table 5 shows the evaluation metrics for these machine learning algorithms.

Fig 1 shows the comparison of evaluation metrics for the best model of each selected feature method.

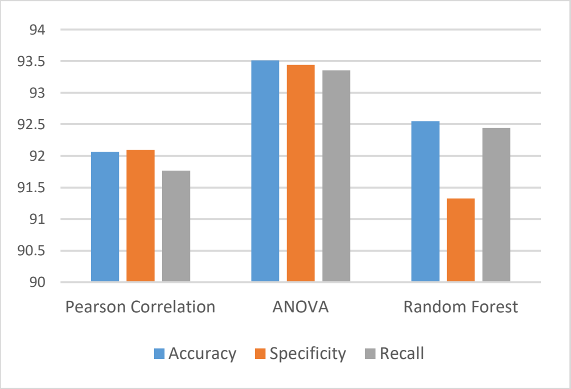

**Fig 1: Comparing the best model for each feature selection method**

## DISCUSSION

Data mining and machine learning have added new approaches to the field of medical diagnosis. Many studies have been done in this field and many systems have been implemented. The diagnosis of HCV degree is one of the challenges addressed in this study, and an attempt was made to determine the factors affecting its diagnosis using machine learning tools. To conduct this study, three feature selection methods, Pearson Correlation, ANOVA, and Random Forest, were selected and applied to the dataset. Then, each feature set was used in Deep Learning, Random Forest, and KNN algorithms for modeling. According to the modeling results, Deep Learning was the best algorithm for modeling this dataset for all the feature selection methods.

Comparing the evaluation metrics of the best models on these three feature selection methods showed that ANOVA was the best feature selection method (Fig 1), and consequently, the most effective factors on disease diagnosis were Age, Gender, Fever, Nausea or Vomiting, Diarrhea, WBC, HGB, ALT 4, ALT 24, RNA 4, RNA 12.

Although the comparison of different studies is meaningless due to the difference in data sets, researchers' approach, and modeling conditions, it helps us to know our position in this field. For this purpose, some cases of past studies are reviewed to compare their methods and results with this study and to understand the results of this study better. In a study using machine learning, Butt et al. [12] presented a model to detect the degree of hepatitis C, and in this way, they used various non-invasive serum biochemical markers and clinical data of patients to identify and predict the stage of hepatitis C disease. In this study, an intelligent hepatitis C stage detection system (IHSDS) was presented to predict the stage of hepatitis C in humans using machine learning tools. In this study, 29 features were used, of which 19 were selected for the research, and the accuracy of the proposed system was reported as 98.89% during training and 94.44% during validation.

In another study, Tsvetkov et al. [13] implemented a model for detecting the stage of liver fibrosis in patients with chronic viral hepatitis C using machine learning. In this study, 1240 patient records were examined with chronic viral Hepatitis C. There were 28 primary variables, among which 9 parameters were selected as the most important predictive parameters to determine the probability of liver fibrosis. This study mentions neither the method of feature selection nor the type of algorithm that performed best and led to the creation of an optimized model. Finally, the implemented model's accuracy, sensitivity and specificity were reported as 80.56%, 66.67 and 94.44 respectively. This study has designed a model based on machine learning for diagnosing the Stage of liver fibrosis in patients. Machine learning models were developed and tested using the mentioned dataset. Nine usual prognostic factors were selected as the essential predictor factors. They finally reported the highest accuracy of 80.56% for their model.

Barakat et al. [11] used machine learning algorithms such as Random Forest to develop a model to predict the stage of liver fibrosis in children infected with the hepatitis C virus. In their model, the Random Forest algorithm achieved the highest accuracy of 90.3%.

In all the mentioned studies, the feature selection method was not reported, and based on the different predictor features, one machine learning algorithm was used for modeling. For the evaluation of the implemented model, the only reported metric is Butt et al. [12]. the study was Precision. Tsvetkov et al. [13] reported Accuracy, Sensitivity, and Specificity as evaluation metrics, and Barkat et al. [11] reported accuracy.

As mentioned in the research method and results of the present study, three feature selection methods were used, and after comparing these three methods, eleven features selected by the ANOVA method as the better feature selection method were used for the next step. Then three machine learning algorithms were implemented for modeling, and their results were compared. As the final result, the model created by the Deep Learning algorithm was selected as the best model. Also, three evaluation metrics of accuracy, sensitivity, and specificity were calculated and reported. In addition to the differences reported in the research method compared to the previous studies, the final model evaluation metrics results in the present study were better than previous studies.

## CONCLUSION

Since identifying the degree of Hepatitis C is very important to determine the treatment plan, doctors need to perform various tests to diagnose the degree of hepatitis C, and patients must undergo multiple tests and examinations, which has a very high cost and time for patients.

Since machine learning has been very effective in solving many challenges in the field of health, this study was conducted to help doctors in diagnosing the degree of hepatitis C. In this study, three feature selection techniques were first used to select the most useful features for HVC analyzing. Then for each selected feature set, three machine-learning algorithms were applied, and a diagnosing model for detecting Hepatitis C was reported. This study showed that using data-mining algorithms can be helpful to for HCV diagnosing. The proposed model in this study can help physicians diagnose the degree of HCV at an affordable and with high accuracy.

## AUTHOR'S CONTRIBUTION

All authors contributed to the literature review, design, data collection and analysis, drafting the manuscript, read and approved the final manuscript.

## CONFLICTS OF INTEREST

The authors declare no conflicts of interest regarding the publication of this study.

## FINANCIAL DISCLOSURE

No financial interests related to the material of this manuscript have been declared.

## REFERENCES

1. Abrantes J, Torres DS, Brandão–Mello CE. The many difficulties and subtleties in the cognitive assessment of chronic hepatitis C infection. Int J Hepatol. 2020; 2020: 9675235. PMID: 32257447 DOI: 10.1155/2020/9675235 [PubMed]

2. Pietschmann T, Brown RJ. Hepatitis C virus. Trends Microbiol. 2019; 27(4): 379-80. PMID: 30709707 DOI: 10.1016/j.tim.2019.01.001 [PubMed]

3. Abu-Freha N, Mathew Jacob B, Elhoashla A, Afawi Z, Abu-Hammad T, Elsana F, et al. Chronic hepatitis C: Diagnosis and treatment made easy. Eur J Gen Pract. 2022; 28(1): 102-8. PMID: 35579223 DOI: 10.1080/13814788.2022.2056161 [PubMed]

4. Dermont M, Sullivan R, Sibal B, Foster G, Mandal S. Hepatitis C diagnosis and management: a primary care and public health partnership approach. Br J Gen Pract. 2022; 72(715): 89-92. PMID: 35091416 DOI: 10.3399/bjgp22X718529 [PubMed]

5. Ahuja AS. The impact of artificial intelligence in medicine on the future role of the physician. PeerJ. 2019; 7: e7702. PMID: 31592346 DOI: 10.7717/peerj.7702 [PubMed]

6. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med. 2019; 25(1): 30-6. PMID: 30617336 DOI: 10.1038/s41591-018-0307-0 [PubMed]

7. Safdari R, Kazemi Arpanahi H, Langarizadeh M, Ghazisaeidi M, Dargahi H, Zendenhdel K. Design a fuzzy rule-based expert system to aid earlier diagnosis of gastric cancer. Acta Inform Med. 2018; 26(1): 19-23. PMID: 29719308 DOI: 10.5455/aim.2018.26.19-23 [PubMed]

8. Samad-Soltani T, Ghanei M, Langarizadeh M. Development of a fuzzy decision support system to determine the severity of obstructive pulmonary in chemical injured victims. Acta Inform Med. 2015; 23(3): 138-41. PMID: 26236078 DOI: 10.5455/aim.2015.23.138-141 [PubMed]

9. Karami M, Fatehi M, Torabi M, Langarizadeh M, Rahimi A, Safdari R. Enhance hospital performance from intellectual capital to business intelligence. Radiol Manage. 2013; 35(6): 30-5. PMID: 24475528 [PubMed]

10. Kaur I, Behl T, Aleya L, Rahman H, Kumar A, Arora S, et al. Artificial intelligence as a fundamental tool in management of infectious diseases and its current implementation in COVID-19 pandemic. Environ Sci Pollut Res Int. 2021; 28(30): 40515-32. PMID: 34036497 DOI: 10.1007/s11356-021-13823-8 [PubMed]

11. Barakat NH, Barakat SH, Ahmed N. Prediction and staging of hepatic fibrosis in children with hepatitis c virus: A machine learning approach. Healthc Inform Res. 2019; 25(3): 173-81. PMID: 31406609 DOI: 10.4258/hir.2019.25.3.173 [PubMed]

12. Butt MB, Alfayad M, Saqib S, Khan M, Ahmad M, Khan MA, et al. Diagnosing the stage of hepatitis C using machine learning. J Healthc Eng. 2021; 2021: 8062410. PMID: 35028114 DOI: 10.1155/2021/8062410 [PubMed]

13. Tsvetkov V, Tokin I, Lioznov D. Machine learning model for diagnosing the stage of liver fibrosis in patients with chronic viral hepatitis C. Preprints.org. 2021; 2021: 2021020488.