Infectious Disease

Prediction of Hepatitis disease using ensemble learning methods

MOHAMMAD MAHDI MAJZOOBI[1,2], SEPIDEH NAMDAR[1], ROYA NAJAFI-VOSOUGH[3], ALI ABBAS HAJILOOI[4], HOSSEIN MAHJUB[5*]

[1] Department of Infectious Diseases, Hamadan University of Medical Sciences, Hamadan, Iran; [2] Brucellosis Research Center, Hamadan University of Medical Sciences, Hamadan, Iran; [3] Department of Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran; [4] Hamadan University of Medical Sciences, Hamadan, Iran; [5] Research Center for Health Sciences, Department of Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran

Keywords

Hepatitis B virus • Hepatitis C virus • Ensemble learning • Data analysis

Summary

Objective. *Hepatitis is one of the chronic diseases that can lead to liver cirrhosis and hepatocellular carcinoma, which cause deaths around the world. Hence, early diagnosis is needed to control, treat, and reduce the effects of this disease. This study's main goal was to compare the performance of traditional and ensemble learning methods for predicting hepatitis B virus (HBV), and hepatitis C virus (HCV). Also, important variables related to HBV and HCV were identified.*

Methods. *This case-control study was conducted in Hamadan Province, in the west of Iran, between 2014 to 2019. It included 534 subjects (267 cases and 267 controls). The bagging, random forest, AdaBoost, and logistic regression were used for predicting HBV and HCV. These methods' performance was evaluated using accuracy.*

Results. *According to the results, the accuracy of bagging, random forest, Adaboost, and logistic regression were 0.65 ± 0.03, 0.66 ± 0.03, 0.62 ± 0.04, and 0.64 ± 0.03, respectively, with random forest showing the best performance for predicting HBV. This method showed that ALT was the most important variable for predicting HBV. The the accuracy of random forest was 0.77±0.03 for predicting HCV. Also, the random forest showed that the order of variable importance has belonged to AST, ALT, and age for predicting HCV.*

Conclusion. *This study showed that random forest performed better than other methods for predicting HBV and HCV.*

Introduction

Hepatitis is one of the dangerous diseases that result from viral infections [1, 2]. This virus attacks the liver leading to its inflammation. Inflammation may lead to the death of the liver cells and affect the functionality of the liver [3]. Five main types of hepatitis have been identified, namely hepatitis A, B, C, D, and E viruses [4]. The most common types of these are hepatitis A virus, hepatitis B virus (HBV), and hepatitis C virus (HCV). Among these, HBV and HCV will cause chronic hepatitis, liver cirrhosis, and hepatocellular carcinoma [2, 5, 6].

It is estimated that 257 and 71 million people around the world are currently infected with HBV and HCV, respectively [7, 8]. The prevalence of HBV depends on the geographic area, and its overall prevalence was estimated at 3.6% in the world [9]. The HCV global prevalence in adults is 2.5% [10]. Furthermore, the incidence of HCV was estimated between 0.5% and 2.8% in various studies [11, 12]. According to previous studies, African and Asian countries have the highest prevalence of HBV and HCV [13, 14]. In Iran, the prevalence of HBV and HCV was about 2.2 and 0.5% in the general population, respectively [15, 16].

Prediction of chronic diseases plays an important role in health informatics. Hepatitis is one of the chronic diseases that can lead to liver cirrhosis and hepatocellular carcinoma, which cause deaths around the world. Therefore, early diagnosis is needed to control, treat, and reduce the effects of this disease [17-19]. In the past few years, machine learning methods have been widely employed for predicting chronic diseases [2, 5, 13, 17, 20-23]. Among them, ensemble learning methods such as bagging, random forest, and AdaBoost are powerful methods. These methods can achieve better learning performance by combining several weak learners [24-26]. Despite the several studies that have used ensemble learning methods for the prediction of hepatitis disease, only a few of them have been performed in Iran [27, 28]. Furthermore, hepatitis is a public health concern, especially in developing countries such as Iran [29, 30]. Hence, this study's main goal was to compare the performance of three ensemble learning methods (including bagging, random forest, Adaboost) and logistics regression as traditional methods for predicting hepatitis diseases (HBV and HCV). Important variables related to HBV and HCV were also identified.

Methods

Study desig

This case-control study was conducted in Hamadan Province, Western Iran. This study was approved by the ethics committee of Hamadan University of Medical Sciences (IR.UMSHA.REC.1396.330).

https://doi.org/10.15167/2421-4248/jpmh2022.63.3.2515

Between 2014 and 2018, 267 patients with a definite diagnosis of HBV or HCV (131 HBV, 131 HCV and 5 HBV and HCV), as the case group, were referred to the hepatitis clinic and the infectious diseases clinic of Hamedan Health Center. The control group was selected from among 267 people referred to Sina Hospital and Dey Hamedan Laboratory during 2018-2019. The second author collected case and control group data using a checklist. The checklist included data related to demographic characteristics (age, sex) and the results of laboratory tests. All participants were over 15 years of age. Non-cooperation of the participants to perform further laboratory tests, ultrasound or any other follow-up was excluded from this study. Informed consent was obtained from the participants after explaining the objectives of the study.

Data Collection and preparation

The data collection tools included an information form on demographic characteristics (age, sex) and laboratory tests. Of any contributor, under sterile conditions, a 10 cc blood sample was obtained by an expert sampler. At the time of admission to the laboratory, alanine aminotransferase (ALT), aspartate aminotransferase (AST), cholesterol (CHOL), triglyceride (TG), fasting blood sugar (FBS), body mass index (BMI) were measured and recorded. Missing values of these variables were imputed with the multiple imputations method. For an exact evaluation of probable causes of elevated ALT, including viral hepatitis and fatty liver, HBsAg, HBCAb, and HCVAb as well as liver sonography were requested for all of the non-hepatitis groups with high ALT. An abnormal measure of CHOL, TG, and blood sugar is defined as above 200, 200, and 115 respectively. The aforementioned cutoffs were determined according to the recommendation of the kit's manufacturer (Roche Biotech).

Statistical Analysis

Bagging, proposed by Breiman, is one of the most popular and earliest ensemble learning methods. This method uses bootstrap resampling to create multiple training subsets from the given original training dataset. Then each training subsets is used to construct a classifier, which is also called a based learner. Eventually, all based learners aggregate into the final prediction model [24].

Random forest, also proposed by Breiman, is a tree-based ensemble learning method. In this method, each tree is grown by a bootstrap sample, which is obtained by randomly resampling from the training dataset. When building each tree, at each nodeot tree, a subset of predictors was selected randomly and among which, the best predictor is chosen for splitting. Eventually, predictions are obtained by averaging the results of all the trees. The random forest can also estimate the importance of predictors using the Gini Index, which makes the results more interpretable [25].

AdaBoost, proposed by Freund and Schapire, is one of the most popular ensemble learning methods that was used boosting algorithms. This method creates a subset of the training dataset. Then an initial classifier-based model is constructed by assigning the same weight for instances. Each boosting iteration assigns weight to the training instances so that the next learner concentrates on reweighted instances that were misclassified previously. Eventually, the final model is a weighted sum of all the classifier-based models [26].

Logistic regression as a traditional method was performed to assess the effect of prognostic factors on HBV and HCV. This method can determine the direction of association of variables on outcome. The results are also easy to interpret [31].

The cross-validation method has been used for evaluating the performance of three ensemble learning methods and logistic regression, in which the dataset was randomly divided into training (70%) and test (30%) sets. Then, the discrimination ability of each methods was assessed by accuracy. This procedure was repeated 100 times and the average values of accuracy were computed.

Software packages

The statistical analyses were performed using R Version 3.6.3 [32], with the following packages "adabag", "COREelearn" and "randomForest".

Results

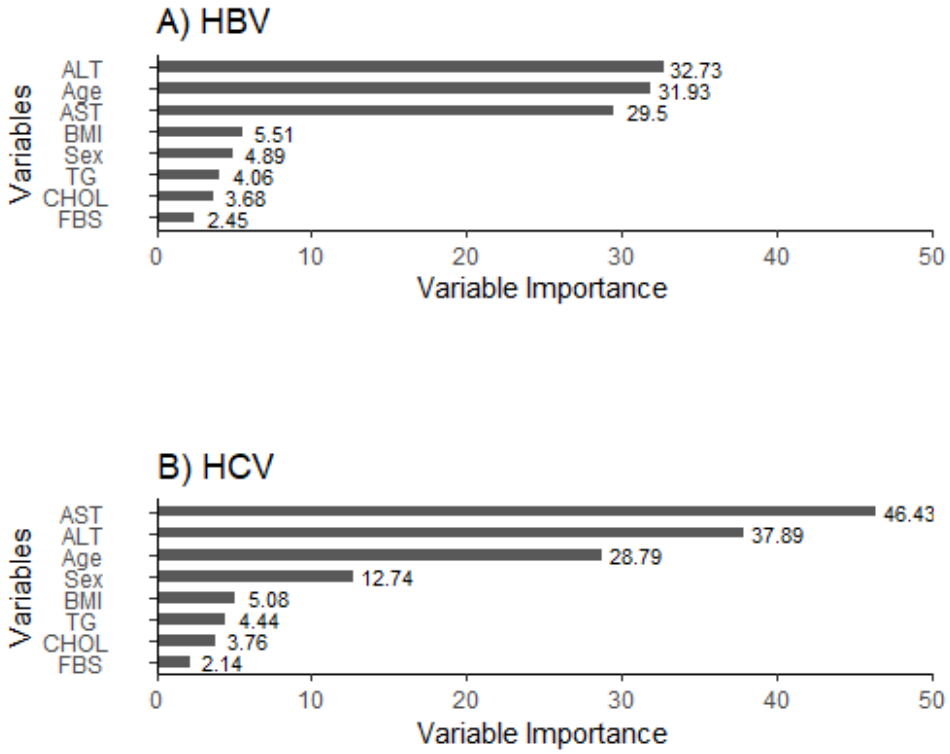
The study involved 534 subjects (267 cases and 267 controls). The case group included 131 patients with HBV, 131 patients with HCV, and 5 patients with HBV and HCV. The control group also included 267 healthy subjects. The characteristics of cases and controls are given in Table I.

Tab. I. Characteristics of the study population.

Variables	Cases N = 267	Controls N = 267	Total N = 534
Age (Year)	38.82 ±10.99	43.05±13.86	40.94 ±12.67
ALT	59.08 ± 58.83	39.33 ± 38.08	49.20 ± 50.49
AST	42.15 ± 33.81	28.06 ± 20.57	35.11 ± 28.83
Sex			
Male	210 (78.7)	157 (58.8)	367 (68.7)
Female	57 (21.3)	110 (41.2)	167 (31.3)
BMI			
Underweight/ Normalweight	89 (33.3)	63 (23.6)	152 (28.5)
Overweight/ Obese	178 (66.7)	204 (76.4)	382 (71.5)
FBS			
≤ 115	252 (94.4)	240 (89.9)	492 (92.1)
> 115	15 (5.6)	27 (10.1)	42 (7.9)
CHOL			
≤ 200	187 (70.0)	177 (66.3)	364 (68.2)
> 200	80 (30.0)	90 (33.7)	170 (31.8)
TG			
≤ 200	189 (70.8)	204 (76.4)	393 (73.6)
> 200	78 (29.2)	63 (23.6)	141 (26.4)

Data are expressed as Mean ± SD and N (%). HBV: Hepatitis B Virus; HCV: Hepatitis C Virus; ALT: Alanine aminotransferase; AST: Aspartate aminotransferase; CHOL: Cholesterol; TG: Triglyceride; FBS: Fasting blood sugar; BMI: Body mass index.

Fig. 1. The variable importance for HBV and HCV.



HBV: Hepatitis B Virus; HCV: Hepatitis C Virus; ALT: Alanine aminotransferase; AST: Aspartate aminotransferase; CHOL: Cholesterol; TG: Triglyceride; FBS: Fasting blood sugar; BMI: Body mass index.

Figure 1 displays the variable importance obtained from random forest for HBV and HVC. The results showed that ALT, age, and AST as the three most important variables for predicting HBV [Fig. 1 (A)]. The random forest also identified that the order of variable importance has belonged to AST, ALT, and age for predicting HCV [Fig. 1 (B)].

Table II shows the performance of three ensemble learning methods (bagging, random forest, Adaboost) and logistics regression for predicting HBV and HCV in testing datasets. As seen, the accuracy of bagging, random forest, Adaboost, and logistic regression were 0.65 ± 0.03, 0.66 ± 0.03, 0.62 ± 0.04, and 0.64 ± 0.03, respectively, with random forest showing the best performance for predicting HBV. Also, the performance of random forest compared to other methods was better for predicting HCV.

Tab. II. The performance criteria of methods.

Test	Methods	Accuracy
HBV	Bagging	0.65 ± 0.03
	AdaBoost	0.62 ± 0.04
	Random Forest	0.66 ± 0.03
	Logistic regression	0.64 ± 0.03
HCV	Bagging	0.76 ± 0.03
	AdaBoost	0.75 ± 0.02
	Random Forest	0.77 ± 0.03
	Logistic regression	0.74 ± 0.03

HBV: Hepatitis B Virus; HCV: Hepatitis C Virus.

Discussion

In the current study, the performance of traditional and ensemble learning methods was assessed for predicting HBV and HCV. The results showed that the random forest performs better than other methods for predicting HBV and HCV. This method identified age, ALT, and AST as the top three most important variables for predicting both hepatitis.

According to previous studies, ALT and AST were identified as important variables in discriminating between healthy controls and patients with hepatitis [18, 33-35]. These findings are consistent with our results. The AST was an important variable that was identified by random forest in the present study. Also, the random forest identified ALT as one of the most important variables for predicting HBV and HCV. It seems that due to lifestyle changes and the addition of factors effective in increasing ALT such as BMI and blood lipids, high ALT is not caused by infectious hepatitis in most cases. However, because of the importance of HBV and HCV in endemic areas, it is best to screen individuals for abnormally high levels of transaminases for hepatitis virus.

Based on our findings, age was identified as another important variable in predicting both types of hepatitis. This results in agreement with Yasin et al., who used data mining techniques for the classification of HCV and concluded that age was associated with it [36].

Several studies have been performed in predicting hepatitis disease using machine learning methods. For instance, Karthikeyan and Thangaraju [4] applied six different machine learning methods to classify hepatitis patients. They showed that the Naive Bayes has the highest performance, and the random forest was also relatively good. Syafa'ah et al. [21] also evaluated the performance of classification machine learning methods for predicting HCV. In their study, neural networks and random forests had a good performance. Nandipati et al. [37] compared the performance of different machine learning methods for predicting HCV. They found that random forest had better performance in comparison to other methods in the binary class. Similar results were also reported in a study conducted by Orooji and Kermani [18]. In another study, Kumar and Sikamani [22] showed that the accuracy of random forest was higher than logistic regression to predict hepatitis. Chicco and Jurman [33] used an ensemble learning method for enhanced classification of patients with hepatitis and cirrhosis. The results of their study confirmed the usefulness of random forest for HCV and cirrhosis diagnosis prediction. The results of these studies were in agreement with our study, which indicates that random forest has the best performance.

The main limitations of this study were the small sample size and failure to consider some risk factors associated with hepatitis. Despite these limitations, our study showed that ensemble learning methods perform reasonably well for HCV and HBV prediction. The results could help doctors better identify people at high risk for hepatitis.

In fact, early detection of this dangerous virus can increase the chance of treatment and prevent the complications of hepatitis, including more deaths caused by it.

Conclusions

This study showed that the performance of random forest provided better results compared to other methods for predicting HBV and HCV based on accuracy.

Acknowledgments

The authors would like to express gratitude to the Vice-Chancellor of Research and Technology, Hamadan University of Medical Sciences for the approval and support of this study (Ethical code: IR.UMSHA.REC.1396.330).

Conflict of interest statement

The authors declare that they have no conflicts of interest.

Authors' contributions

M.M., S.N. and H.M. contributed to the study design, analysis, and interpretation of data. A.H. participated in data collection. R.N.V. participated in data analysis and drafting of the manuscript. All authors read and approved the final manuscript.

References

[1] Pawlotsky J-M, Negro F, Aghemo A, Berenguer M, Dalgard O, Dusheiko G, Marra F, Puoti M, Wedemeyer H. EASL recommendations on treatment of hepatitis C: final update of the series. J Hepatol 2020;73:1170-218. https://doi.org/10.1016/j.jhep.2020.08.018.

[2] Bayrak Ea, Kirci P, Ensari T. Performance analysis of machine learning algorithms and feature selection methods on hepatitis disease. IJMSIT 2019;3:135-8.

[3] Hussien SO, Elkhatem SS, Osman N, Ibrahim AO. A review of data mining techniques for diagnosing hepatitis. SCCSIT 2017, pp. 1-6.

[4] Karthikeyan T, Thangaraju P. Analysis of classification algorithms applied to hepatitis patients. IJCA 2013;62.

[5] Hashem S, Esmat G, Elakel W, Habashy S, Raouf SA, Elhefnawi M, Eladawy MI, ElHefnawi M. Comparison of machine learning approaches for prediction of advanced liver fibrosis in chronic hepatitis C patients. IEEE/ACM transactions on computational biology and bioinformatics. TCBB 2017;15:861-8. https://doi.org/10.1109/TCBB.2017.2690848

[6] Chen S, Zhang Z, Wang Y, Fang M, Zhou J, Li Y, Dai E, Feng Z, Wang H, Yang Z, Li Y. Using quasispecies patterns of hepatitis B virus to predict hepatocellular carcinoma with deep sequencing and machine learning. J Infect Dis 2021;223:1887-96. https://doi.org/10.1093/infdis/jiaa647

[7] Organization WHO. Global hepatitis report 2017. World Health Organization 2017.

[8] Stanaway JD, Flaxman AD, Naghavi M, Fitzmaurice C, Vos T, Abubakar I, Abu-Raddad LJ, Assadi R, Bhala N, Cowie B, Forouzanfour MH. The global burden of viral hepatitis from 1990 to 2013: findings from the Global Burden of Disease Study 2013. Lancet 2016;388:1081-8. https://doi.org/10.1016/S0140-6736(16)30579-7

[9] Moghadami M, Dadashpour N, Mokhtari AM, Ebrahimi M, Mirahmadizadeh A. The effectiveness of the national hepatitis B vaccination program 25 years after its introduction in Iran: a historical cohort study. BJID 2020;23:419-26. https://doi.org/10.1016/j.bjid.2019.10.001

[10] Petruzziello A, Marigliano S, Loquercio G, Cozzolino A, Cacciapuoti C. Global epidemiology of hepatitis C virus infection: An up-date of the distribution and circulation of hepatitis C virus genotypes. WJG 2016;22:7824. https://doi.org/10.3748/wjg.v22.i34.7824.

[11] Hanafiah KM, Groeger J, Flaxman AD, Wiersma ST. Global epidemiology of hepatitis C virus infection: new estimates of age-specific antibody to HCV seroprevalence. Hepatology 2013;57:1333-42. https://doi.org/10.1002/hep.26141

[12] Gower E, Estes C, Blach S, Razavi-Shearer K, Razavi H. Global epidemiology and genotype distribution of the hepatitis C virus infection. Hepatology 2014;61:S45-S57. https://doi.org/10.1016/j.jhep.2014.07.027

[13] Abtahi S, Sharifi M. Machine learning method to control and observe for treatment and monitoring of Hepatitis B virus. arXiv preprint arXiv:200409751 2020. https://doi.org/10.48550/arXiv.2004.09751

[14] Morozov VA, Lagaye S. Hepatitis C virus: Morphogenesis, infection and therapy. World J Hepatol 2018;10:186. https://doi.org/10.4254/wjh.v10.i2.186

[15] Salehi-Vaziri M, Sadeghi F, Hashiani AA, Fesharaki MG, Alavian SM. Hepatitis B virus infection in the general population of Iran: an updated systematic review and meta-analysis. Hepat Mon 2016;16:e35577. https://doi.org/10.5812/hepatmon.35577

[16] Merat S, Rezvan H, Nouraie M, Jafari E, Abolghasemi H, Radmard AR, Zaer-rezaii H, Amini-Kafiabad S, Maghsudlu M, Pourshams A, Malekzadeh R. Seroprevalence of hepatitis C virus: the first population-based study from Iran. IJID 2010;14:e113-e6. https://doi.org/10.1016/j.ijid.2009.11.032

[17] Jain D, Singh V. Feature selection and classification systems for chronic disease prediction: A review. Egypt Inform J 2018;19:179-89. https://doi.org/10.1016/j.eij.2018.03.002

[18] Orooji A, Kermani F. Machine learning based methods for handling imbalanced data in hepatitis diagnosis. Front Health Inform 2021;10:57. https://doi.org/10.30699/fhi.v10i1.259

[19] Kashif AA, Bakhtawar B, Akhtar A, Akhtar S, Aziz N, Javeid MS. Treatment response prediction in Hepatitis C patients using machine learning techniques. IJTIM 2021;1:79-89. https://doi.org/10.54489/ijtim.v1i2.24

[20] Bhargav KS, Thota D, Kumari TD, Vikas B. Application of machine learning classification algorithms on hepatitis dataset. Int J Appl Eng Res 2018;13:12732-7.

[21] Syafa'ah L, Zulfatman Z, Pakaya I, Lestandy M. Comparison of machine learning classification methods in Hepatitis C virus. JOIN 2021;6:73-8. https://doi.org/10.15575/join.v6i1.719

[22] Kumar N, Sikamani K. Prediction of chronic and infectious diseases using machine learning classifiers – A systematic approach. Int J Intell Eng Syst 2020;13:11-20. https://doi.org/10.22266/ijies2020.0831.02

[23] Najafi-Vosough R, Faradmal J, Hosseini SK, Moghimbeigi A, Mahjub H. Predicting hospital readmission in heart failure patients in Iran: a comparison of various machine learning methods. Healthc Inform Res 2021;27:307-14. https://doi.org/10.4258/hir.2021.27.4.307

[24] Breiman L. Bagging predictors. Mach Learn 1996;24:123-40.

[25] Breiman L. Random forests. Mach Learn 2001;45:5-32.

[26] Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. JCSS 1997;55:119-39. https://doi.org/10.1006/jcss.1997.1504

[27] KayvanJoo AH, Ebrahimi M, Haqshenas G. Prediction of hepatitis C virus interferon/ribavirin therapy outcome based on viral nucleotide attributes using machine learning algorithms. BMC Res Notes 2014;7:1-11.

[28] Mokhtari AM, Moghadami M, Seif M, Mirahmadizadeh A. Association of routine Hepatitis B vaccination and other effective factors with Hepatitis B virus infection: 25 years since the introduction of national Hepatitis B vaccination in Iran. IJMS 2021;46:93. https://doi.org/10.30476/ijms.2019.83112.1199

[29] Dolan K, Wirtz AL, Moazen B, Ndeffo-Mbah M, Galvani A, Kinner SA, Courtney R, McKee M, Amon JJ, Maher L, Hellard M. Global burden of HIV, viral hepatitis, and tuberculosis in prisoners and detainees. Lancet 2016;388:1089-102. https://doi.org/10.1016/S0140-6736(16)30466-4

[30] Rezaei N, Asadi-Lari M, Sheidaei A, Gohari K, Parsaeian M, Khademioureh S, Maghsoudlu M, Kafiabad SA, Zadsar M, Motevalian SA, Delavari F. Epidemiology of hepatitis B in Iran from 2000 to 2016: a systematic review and meta-regression analysis. Arch Iran Med 2020;23:189-96.

[31] Agresti A, Kateri M. Categorical data analysis. SpringerBerlin Heidelberg 2011, pp. 206-8.

[32] Team RC. R: a language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. 2020. https://www.R-project.org/

[33] Chicco D, Jurman G. An ensemble learning approach for enhanced classification of patients with hepatitis and cirrhosis. IEEE Access 2021;9:24485-98. https://doi.org/10.1109/ACCESS.2021.3057196

[34] Akkaya O, Kiyici M, Yilmaz Y, Ulukaya E, Yerci O. Clinical significance of activity of ALT enzyme in patients with hepatitis C virus. WJG 2007;13:5481. https://doi.org/10.3748/wjg.v13.i41.5481

[35] Pradat P, Alberti A, Poynard T, Esteban J-I, Weiland O, Marcellin P, Badalamenti S, Trépo C. Predictive value of ALT levels for histologic findings in chronic hepatitis C: a European collaborative study. Hepatology 2002;36:973-7. https://doi.org/10.1053/jhep.2002.35530

[36] Yasin H, Jilani TA, Danish M. Hepatitis-C classification using data mining techniques. Int J Comput Appl 2011;24:1-6.

[37] Nandipati SC, XinYing C, Wah KK. Hepatitis C virus (HCV) prediction by machine learning techniques. Model Simul Eng 2020;4:89-100.

Received on February 11, 2022. Accepted on September 01, 2022.

Correspondence: Hossein Mahjub, Center for Health Sciences, Department of Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran. PO BOX: 65175-4171 - Tel.: +98 81 38380025 - Fax: +98 81 38380509 - E-mail: mahjub@umsha.ac.ir

How to cite this article: Majzoobi MM, Namdar S, Najafi-Vosough R, Hajilooi AA, Mahjub H. Prediction of Hepatitis disease using ensemble learning methods. J Prev Med Hyg 2022;63:E424-E428. https://doi.org/10.15167/2421-4248/jpmh2022.63.3.2515