Intelligent Medicine 2 (2022) 193–198

Contents lists available at ScienceDirect

Intelligent Medicine

journal homepage: www.elsevier.com/locate/imed

Research Article

Applying data mining techniques to classify patients with suspected hepatitis C virus infection

Reza Safdari [1], Amir Deghatipour [2], Marsa Gholamzadeh [1,*], Keivan Maghooli [3]

[1] Health Information Management Department, School of Allied Medical Sciences, Tehran University of Medical Sciences, Tehran, Iran
[2] Health Information Management and Medical Informatics Department, School of Allied Medical Sciences, Tehran University of Medical Sciences, Tehran, Iran
[3] Department of Biomedical Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran

ARTICLE INFO

Keywords:
Data mining methods
Classification
Hepatitis
Hepatitis C virus

ABSTRACT

Background Hepatitis C virus (HCV) has a high prevalence worldwide, and the progression of the disease can cause irreversible damage to severe liver damage or even death. Therefore, developing prediction models using machine learning techniques is beneficial. This study was conducted to classify suspected patients with HCV infection using different classification models.

Methods The study was conducted using a dataset derived from the University of California, Irvine (UCI) Machine Learning Repository. Since the HCV dataset was imbalanced, the synthetic minority oversampling technique (SMOTE) was applied to balance the dataset. After cleaning the dataset, it was divided into training and test data for developing six classification models. These six algorithms included the support vector machine (SVM), Gaussian Naïve Bayes (NB), decision tree (DT), random forest (RF), logistic regression (LR), and K-nearest neighbors (KNN) algorithm. The Python programming language was used to develop the classifiers. Receiver operating characteristic curve analysis and other metrics were used to evaluate the performance of the proposed models.

Results After the evaluation of the models using different metrics, the RF classifier had the best performance among the six methods. The accuracy of the RF classifier was 97.29%. Accordingly, the area under the curve (AUC) for LR, KNN, DT, SVM, Gaussian NB, and RF models were 0.921, 0.963, 0.953, 0.972, 0.896, and 0.998, respectively, RF showing the best predictive performance.

Conclusion Various machine learning techniques for classifying healthy and unhealthy patients were used in this study. Additionally, the developed models might identify the stage of HCV based on trained data.

1. Introduction

Hepatitis is highly prevalent worldwide, an estimated that 58 million people have chronic hepatitis C virus (HCV) infection [1]. Notably, this disease has a higher prevalence in Asian and European regions than in other regions. Viral hepatitis can be caused by various types of viruses, which are hepatitis A virus (HAV), hepatitis B virus (HBV), hepatitis C virus (HCV), hepatitis D virus (HDV), and hepatitis E virus (HEV). Among the types of viral hepatitis, HCV is a global health problem due to its high prevalence [2]. The disease is usually asymptomatic, but the progression of the liver disease can cause irreversible to severe liver damage or even death [3–4]. According to Li et al. [5], the disability-adjusted life year (DALY) trends of HCV have decreased over the years, from 1990 to 2016. However, it still has a high prevalence worldwide. In comparison with HVB, HCV has recorded higher incidence and mortality rates in recent years [6]. Although it can affect individuals of all ages, it is more common in adults. For this reason, eliminating HCV infection, as one of the public health concerns, is one of the main goals of the World Health Organization (WHO) by 2030 [7].

Chronic HCV infection can lead to liver cirrhosis or liver fibrosis. Early diagnosis of fibrosis can lead to better treatment and prevent irreversible complications. Chronic HCV could damage the liver slowly [8]. Unfortunately, patients may be asymptomatic until they develop liver cirrhosis. One of the treatments for patients with liver failure due to cirrhosis is liver transplantation (LTx) [9], and most of such patients may require LTx due to HCV infection. However, early diagnosis of HCV has been effective that most factors contributing to progressive disease are unclear yet [10]. Therefore, developing prediction models using machine learning techniques may be beneficial [11].

Currently, the application of artificial intelligence (AI) has become increasingly common [9]. AI-based techniques are used in various fields, such as the medical sciences [12–14]. Healthcare domains are a data-rich field in all aspects. Analysis of the vast health-related data may

* Corresponding author: Marsa Gholamzadeh, Health Information Management Department, School of Allied Medical Sciences, Tehran University of Medical Sciences, No #17, Farredanesh Alley, Ghods St, Enghelab Ave, Tehran, Iran (Email: m-gholamzadeh@razi.tums.ac.ir).

https://doi.org/10.1016/j.imed.2021.12.003
2667-1026/© 2022 The Author(s). Published by Elsevier B.V. on behalf of Chinese Medical Association. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

prove challenging to humans [15–16]; as a result, clinicians may likely miss intricate information necessary for diagnosis and treatment in clinical practice [17]. This challenge can be overcome by applying computational methods based on machine learning methods that have been used in various domains of the health sciences [18–19].

In current clinical practice, clinicians can take advantage of such techniques to analyze large datasets. Different data mining (DM) methods, such as support vector machines (SVMs), decision tree (DT), random forest (RF), neural networks (ANN), linear regression (LR), logistic regression, and k-nearest neighbors (KNN) algorithm, have been applied in various fields of medicine [20–23]. In this regard, some valuable studies have been conducted on chronic hepatitis C. Various studies have compared the performance of different data mining techniques, but their results were inconsistent [11,24–26].

Therefore, this study was conducted to assess risks and factors that influence disease progression. Additionally, various techniques were used to classify the level of susceptibility to HCV infection. Generally, most studies on HCV infection used blood donor information to report the frequency and risk factors of HCV [27]. Therefore, this study used a dataset with data on healthy blood donors and patients with HCV infection at different stages to develop classification models.

2. Methods

Throughout this section, the process involved in the comparison of model performance is described. Figure 1 illustrates the schematic presentation of the steps of enquiry. Google Colab was used as a computational instrument to facilitate this process. The Python programming language was used for developing the proposed framework and developing the machine learning algorithms used in this survey. The Numpy and Pandas modules were used for data pre-processing, and the sci-kit learn library was used for developing supervised classifier algorithms (NumPy is an open-source Python library that facilitates efficient numerical operations on large quantities of data. Pandas is a high-level data manipulation tool that is built on the NumPy package).

2.1. Data source

The HCV dataset from the University of California at Irvine (UCI) Machine Learning Repository was used in this study [28]. This dataset comprised clinical laboratory and demographic data of 615 blood donors and patients with hepatitis C. A total of 13 variables were defined for each patient record. These tests included albumin (ALB), alkaline phosphatase (ALP), bilirubin (BIL), cholesterol (CHOL), creatinine blood test (CREA), choline esterase (CHE), γ-glutamyl-transferase (GGT), aspartate aminotransferase (AST), alanine aminotransferase (ALT), and total protein test (PROT). The demographic characteristics included age and sex. In this dataset, the final diagnosis was characterized by five outcomes of interest: blood donors, suspected blood donors, hepatitis C, fibrosis, and cirrhosis. The patients diagnosed with HCV ranged from chronic hepatitis C infection without fibrosis to end-stage liver cirrhosis with a need for LTx.

2.2. Data preprocessing

Usually, the data collected from patients' records are not completely clear. Therefore, data cleaning is an essential step for developing machine learning models. Data preprocessing involves converting raw data into a logical or understandable format to ensure that the data have the same range of values and the features are comparable. Hence, the raw data were first normalized and converted into appropriate formats, which were more suitable for the different machine learning estimators. First, the ID column was removed. The missing values in our dataset were replaced with the mean value of each variable. Due to the presence of different measuring units, the data normalization method was performed using the StandardScaler function, which standardizes variables by scaling to unit variance.

The dataset was not balanced, which means that most of the records belonged to the same category. Classification of imbalanced data is biased towards the large categories. The symmetric minority oversampling technique (SMOTE) was applied to the HCV dataset to facilitate the performance of various classifiers. This method uses the KNN algorithm in creating new synthetic samples to balance the class distribution of the dataset [29].

2.3. Applied classification algorithms

Classification models, known as supervised methods, were applied in this study to classify existing data. The dataset was divided into a training set (80%) and test set (20%). Thereafter, each classifier model was trained using the balanced training data. After the training process, the classifier classified patients based on the records in the test set. The performance of the classifiers was evaluated with the test data, and the performance of each model was calculated with different metrics. Model development is done by different classifiers, which were described in the following using Python v.3.5.

2.3.1. Logistic regression

The logistic regression (LR) model is one of the most favorable classification methods in supervised machine learning algorithms [30]. It is used to predict the categorical dependent variable using a given set of independent variables. The model is developed based on the probability concept. By mapping probabilities, the output is produced through the logistic sigmoid function [31]. It uses the following formula for creating a logistic function:

$$P = \frac{1}{1 + e^{\bar{y}}}$$

This Sigmoid function maps the predicted values of probabilities between 0 and 1. With this ability, the sigmoid function is widely used in healthcare to analyze multivariate regression problems [32]. Based on its ability to classify categorical data, the LR model was applied in our survey as one of the classifiers.

2.3.2. Decision tree

It is a type of simple classification technique used in developing classifiers based on datasets. It uses a set of techniques in a tree structure to separate data based on different characteristics [33]. As a simple supervised learning method, it can be used for solving classification problems by learning simple decision rules. It applies different methods, such as information theory and entropy. The decision tree classifier applies entropy as an indicator of messy data. The entropy of each variable is calculated. Based on the results, the root is split into branches based on maximum information gain and some rules [23]. The entropy is calculated using the following formula:

$$Entropy = -\sum_{i=1}^{c} p_i \log_2 p_i$$

2.3.3. Naïve bayes

This method is a type of supervised machine learning technique for classifying data based on the probabilistic Bayes' theorem. Bayes' theorem considers all features as independent variables [34]. The naïve Bayesian classification was conducted based on the following formula:

$$\mathrm{P}(C|X) = \frac{\mathrm{P}(X|C)\ P(C)}{P(X)}$$

It is a special case of the Bayesian network, which is based on the hypothesis that an event can be predicted based on other variables.

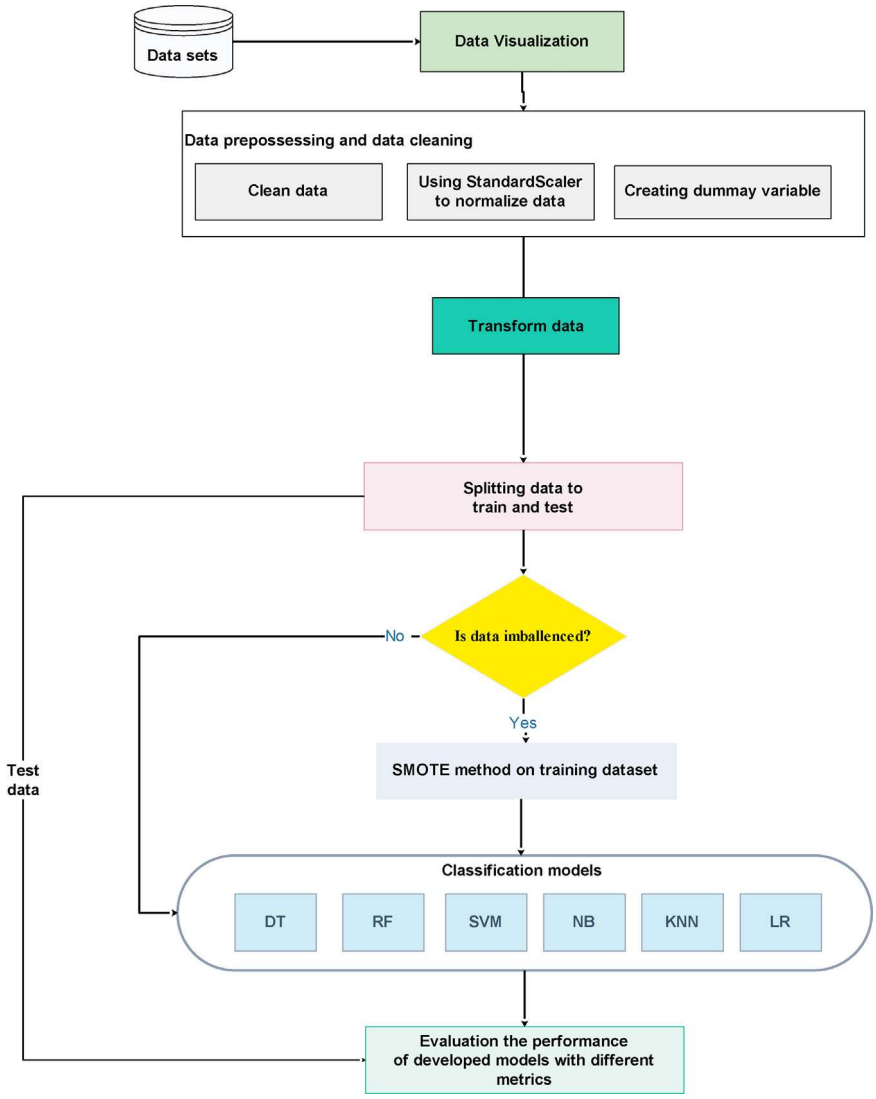

Figure 1. Schematic diagram of the proposed method. DT: decision tree; RF: random forest; SVM: support vector machine; NB: Gaussian Naïve Bayes; KNN: K-nearest neighbors; LR: logistic regression.

2.3.4. Random forest (RF)

This method is another type of supervised machine learning technique. RF used multiple decision trees based on trained data for classifying data. Repeated sampling was performed in this classifier to form a decision tree for each sample [35]. Some data scientists consider RF as a type of ensemble machine learning method based on the decision tree model [36].

2.3.5. Support vector machine (SVM)

This method is a supervised machine learning method that can be used for prediction and classification. It can be used for both linear and nonlinear data. The SVM model tries to find the best hyperplane by maximizing the marginal distance. It can be used as a kernel method for finding the best line for sample division [37].

2.3.6. K-nearest neighbor (KNN) algorithm

It is the simplest and most used classification method applied in machine learning domains. With the KNN method, the test data were labeled based on their similarities [38]. The K-value is the main factor in the KNN method. In this lazy algorithm, the chosen number of K-values indicates the number of neighbors that should be selected. In our survey, the best K-value was identified before fitting the model. The Euclidean distance between two points is calculated to obtain the nearest neighbors [39–40]. In this analysis, the value of K was set at 5, based on a trial-and-error method [41].

$$E(x,\ y) = \sqrt{\sum_{a=1}^{m} \left(x_a - y_a\right)^2}$$

where E is the Euclidean distance between points x and y.

2.4. Model evaluation

The models were developed in two main stages, including model fitting and model performance calculation. To evaluate the performance of developed models, a confusion matrix was used for each model. It shows the performance of each model by plotting four possible outcomes: true positive, true negative, false positive, and false negative [42–43].

Additionally, the performances of developed classifiers were compared by calculating the accuracy, recall (sensitivity), specificity, precision, and F-measure (F1 score) for each model. These metrics were calculated from the confusion matrix. The formulas for these metrics are as follows [44]:

$$Accuracy = \frac{True\ Positive + True\ negative}{All\ predictions\ outcomes}$$

Table 1 Description of variables for each patient in the dataset

Variables	Description	Type	Mean	Range
Age	Age of patient (years)	Integer	47.40813	19–77
Sex	Gender (Male or Female)	Categorical	1.386992	F and M
ALB (g/L)	It measures the amount of albumin in your blood.	Real	41.6202	14.9–82.2
ALP (U/L)	It measures the amount of alkaline phosphatase enzyme in your bloodstream.	Real	68.28392	11.3–416.6
ALT (U/L)	Aanine aminotransferase, it indicates liver damage from hepatitis, infection, cirrhosis, liver cancer, or other liver diseases	Real	28.45081	0.9–325.3
AST (U/L)	Aspartate aminotransferase is an enzyme that is found mostly in the liver, but also in muscles.	Real	34.78634	10.6–324
BIL (μmol/L)	A bilirubin test measures the amount of bilirubin in the blood.	Real	11.39675	0.8–254
CHE (U/L)	Serum cholinesterase is an enzyme synthesized by hepatocytes and its levels reflect the synthetic function of the liver.	Real	8.196634	1.42–16.41
CHOL (U/L)	It can measure the amount of cholesterol and triglycerides in your blood	Real	5.368099	1.43–9.67
CREA (μmol/L)	A creatinine test is a measure of how well kidneys are performing their job of filtering waste from your blood.	Real	81.28781	8–107.9
GGT (U/L)	An elevation of gamma-glutamyl transferase activity is seen in many forms of liver disease	Real	39.53317	4.5–650
PROT (g/L)	A total protein test measures the amount of protein in your blood	Real	72.04414	44.8–90

ALB: albumin; ALP: Alkaline phosphatase; ALT: alanine aminotransferase; AST: aspartate aminotransferase; BIL: bilirubin; CHE: choline esterase; CHOL: cholesterol; CREA: creatinine blood test; GGT: γ-glutamyl-transferase; PROT: total protein test.

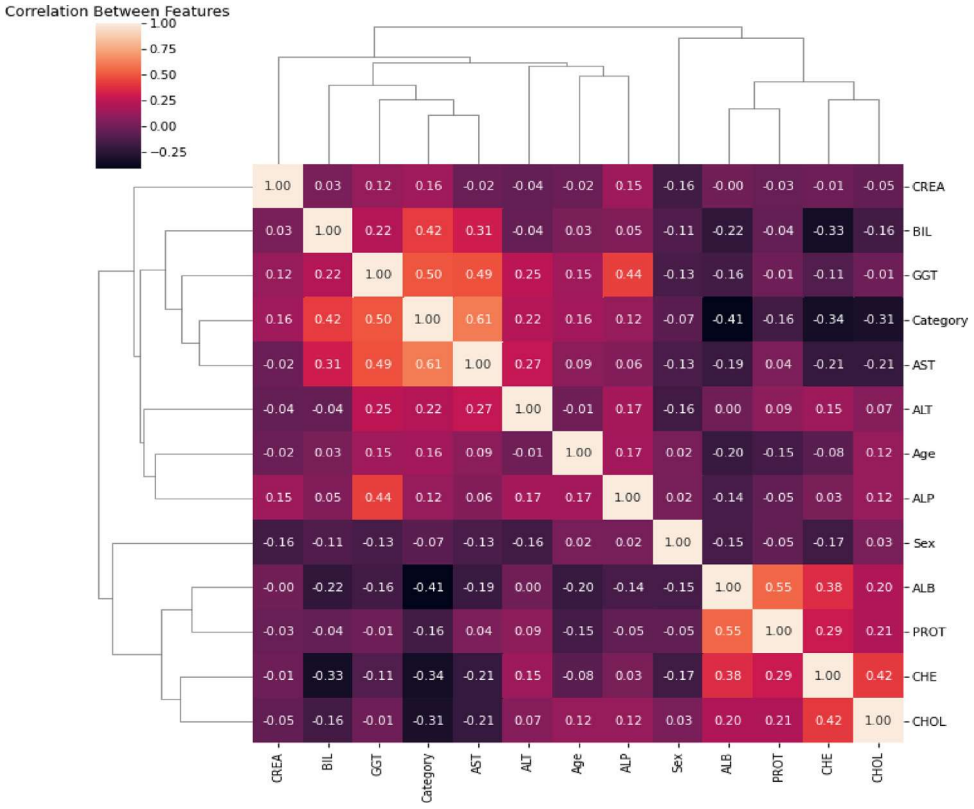

Figure 2. Heat map representative of independent variables.

$$Specificity = \frac{True\ negative}{True\ negative + False\ Positive}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Accuracy is the best evaluation metric for machine learning models. However, it only works well when all classes are equal in number. For imbalanced datasets, other metrics should be used with accuracy. Alongside other metrics, the area under the receiver operating characteristic (ROC) curve is an effective performance measure for comparing machine learning algorithms [45]. The area under ROC (AUC) with other metrics is commonly used as an evaluation metric to compare the performance of classification algorithms in imbalanced data [46].

3. Results

3.1. Data set description

After preprocessing the data, the dataset included 614 instances. Each patient record was characterized by 12 variables. All variables with their description are presented in Table 1. All variables, except disease category and sex, are numerical. Ten routine diagnostic test findings for liver diseases were available for each record. The average age of all the patients was (47.40 ± 10.05) years (range: 19–77 years). The outcome for classification was disease category. Blood donors, suspected blood donors, patients with hepatitis, patients with fibrosis, and patients with cirrhosis formed 86.67%, 1.14%, 3.90%, 3.41%, and 3.41% of participants, respectively. Since the objective of this dataset was to classify

Table 2 The comparison of applied classifiers and their evaluation metrics (%)

Models	Accuracy	Recall	Precision	Specificity	F-Score
Logistic regression	95.67	90.06	90.06	82.60	90.06
Naïve Bayes	92.43	80.75	83.27	65.21	81.94
Support vector machines	94.59	87.27	88.64	73.91	87.10
K-nearest neighbors ($K = 5$)	95.67	91.93	89.06	86.95	90.42
Decision tree	96.75	90.68	94.01	82.60	92.26
Random forest	97.29	90.99	96.28	82.60	93.42

healthy individuals and patients with hepatitis, the outcome variable was considered as a binary variable, which could either be non-hepatitis (blood donors and suspected blood donors) or hepatitis (at three different stages) for best performance.

In developing supervised learning models, the correlation between the outcome variable and all clinical variables were calculated. Correlation coefficients describe the relationship between variables and dependent groups. The heat map, shown in Figure 2, indicates the correlation between different variables in this dataset.

The ratio of trained (80%) and test (20%) data was 1:3. The distribution of patient records based on outcome variables showed that the dataset was imbalanced. Through preprocessing, the SMOTE algorithm was used to address this problem. After applying SMOTE, the new sample had equal amounts of data for the outcomes, and was ready to be modeled. To avoid data leakage and minimize model overfitting, SMOTE was applied only on the training set.

3.2. Model development

After the developed models (SVM, KNN, LR, RF, NB, and DT) were developed with the trained dataset, the developed classification models were evaluated with predefined metrics. The final performance of the model was measured as the aggregate of results obtained from the testing set. First, the confusion matrix was created for each classifier. Then, the accuracy, recall (sensitivity), precision, and F1 score for each model were calculated based on the confusion matrix. The comparison of the model metrics is shown in Table 2.

To choose the best performing model for use as a proposed mode, the performance of each model but not the average performance received much focus. In addition to other metrics, in comparing the ROC curves for the LR, SVM, KNN, RF, NB, and DT models using their true positive rate (TPR) and false-positive rate (FPR) values at the optimal cutoff points, as shown in Figure 3, RF had the highest AUC. This means that RF was the best performing model.

From the results of the evaluation metrics and other findings, the RF classifier had the best performance among the classifiers based on the confusion matrix of the other methods. The accuracy of the RF classifier was 97.29%, which was the highest among the accuracies of all the other classifiers. According to the evaluation metrics, a high F-score (93.42) represented a high model performance. Regarding F-score, RF had the highest score. In terms of AUC, the RF had the best performance among all the other algorithms. It shows the power of this model in classifying patients.

4. Discussion

Through this survey, the dataset of patients suspected of HCV infection was investigated. Various machine learning techniques were used to classify the patients as healthy and unhealthy. Additionally, the developed models could identify the stage of HCV infection based on trained data. The performance and accuracy of classifiers were reported in this study. Classification models are very popular among physicians [47-48]. Determining the stage of disease and state of health for each patient is crucial for physicians [49-50]. Machine learning techniques enable clinicians to quickly classify patients in the best way without extensive knowledge of the algorithms that make these models [51].

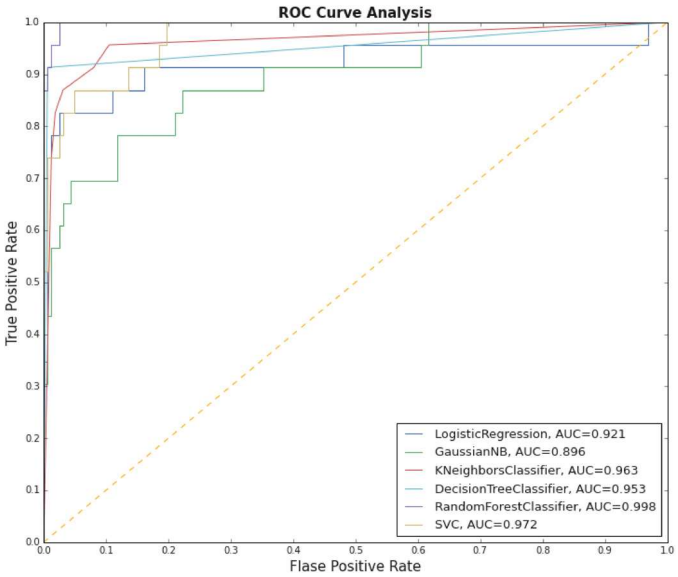

Figure 3. Comparison of the ROC curves for developed models.

Based on our findings, the RF algorithm had the best performance among the implemented algorithms, in terms of the different metrics, which included AUC, F1-score, precision, recall, and accuracy. Since ROC and AUC provide better performance metrics when comparing classifiers of imbalanced datasets, they were applied to compare our models [46].

All classifiers used in this study are types of supervised learning methods. Therefore, all models were trained based on a specific range of data as inputs. To reduce the error of prediction and improve the performance of applied algorithms, preprocessing of data was considered. Accordingly, applying the SMOTE technique to our dataset can address the problem of the imbalanced dataset and enhance the performance of applied models [29].

Since the HCV dataset was published recently in the UCI repository, few studies on the dataset have been reported. Syafa'ah et al. [52] assessed the level of accuracy using the classification data mining techniques to detect HCV infection. Although they classified the patients with high accuracy, their model failed to address the problem of imbalanced data using preprocessing techniques. In another study [53], three classification methods were applied with other techniques to address the imbalance. The study reported a 99.9% accuracy when AUC or ROC as a robust metric for imbalanced data was not reported.

Our study had several limitations. First, the dataset had a small sample size. Second, unknown potentially relevant variables may have been unfortunately missed, as the features included in the model were based on the dataset obtained. Third, the models in our study were developed based on specific datasets, which may not be generalizable for prediction or diagnosis of another status of disease. Fourth, the most common machine learning algorithms were used, and further studies using other algorithms are needed to improve prediction accuracy. Finally, the use of a dataset from a single center within a specific geographic region may limit the external applicability of our findings, since the dataset may not be representative of the whole study population. As a result, the findings may vary in other centers. Developing prediction models based on deep learning methods on large datasets could be the focus of further studies.

Conflicts of interest statement

The authors declare no conflicts of interest.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author contributions

Marsa Gholamzadeh, Reza Safdari and Keivan Maghooli contributed to the conceptualization and design the study. Marsa Gholamzadeh and Amir Deghatipour were responsible for the data preparation and preprocessing. Formal analysis and Methodology were jointly performed by Amir Deghatipour and Marsa Gholamzadeh. Reza Safdari and Keivan Maghooli validated the developed models and results. All authors wrote the paper. Marsa Gholamzadeh, Reza Safdari and Amir Deghatipour edited and reviewed the manuscript.

References

[1] Lanini S, Easterbrook PJ, Zumla A, et al. Hepatitis C: global epidemiology and strategies for control. Clin Microbiol Infect 2016;22(10):833–8. doi:10.1016/j.cmi.2016.07.035.
[2] Petruzziello A, Marigliano S, Loquercio G, et al. Global epidemiology of hepatitis C virus infection: an up-date of the distribution and circulation of hepatitis C virus genotypes. World J Gastroenterol 2016;22(34):7824–40. doi:10.3748/wjg.v22.i34.7824.
[3] Han R, Zhou J, François C, et al. Prevalence of hepatitis C infection among the general population and high-risk groups in the EU/EEA: a systematic review update. BMC Infect Dis 2019;19(1):655. doi:10.1186/s12879-019-4284-9.
[4] Stanaway JD, Flaxman AD, Naghavi M, et al. The global burden of viral hepatitis from 1990 to 2013: findings from the Global Burden of Disease Study 2013. Lancet 2016;388(10049):1081–8. doi:10.1016/S0140-6736(16)30579-7.
[5] Li M, Wang ZQ, Zhang L, et al. Burden of viral hepatitis caused by specific aetiologies in China, 1990–2016: findings from the GBD 2016. BMC Public Health 2020;20(1):1461. doi:10.1186/s12889-020-09533-4.
[6] Ghobad M, Bakhtiar P, Cyrus A, et al. Incidence, mortality, and burden of hepatitis B and C and geographical distribution in iran during 2008-2015. Iran J Public Health 2019;48(Supple 1):10–19.
[7] Botheju WSP, Zghyer F, Mahmud S, et al. The epidemiology of hepatitis C virus in central Asia: systematic review, meta-analyses, and meta-regression analyses. Sci Rep 2019;9(1):2090. doi:10.1038/s41598-019-38853-8.
[8] Karsdal MA, Krarup H, Sand J, et al. Review article: the efficacy of biomarkers in chronic fibroproliferative diseases–early diagnosis and prognosis, with liver fibrosis as an exemplar. Aliment Pharmacol Ther 2014;40(3):233–49. doi:10.1111/apt.12820.
[9] Wiegand J, Berg T. The etiology, diagnosis and prevention of liver cirrhosis: part 1 of a series on liver cirrhosis. Dtsch Arztebl Int 2013;110(6):85–91. doi:10.3238/arztebl.2013.0085.
[10] Ghany MG, Strader DB, Thomas DL, et al. Diagnosis, management, and treatment of hepatitis C: an update. Hepatology 2009;49(4):1335–74. doi:10.1002/hep.22759.
[11] Wei R, Wang J, Wang X, et al. Clinical prediction of HBV and HCV related hepatic fibrosis using machine learning. EBioMedicine 2018;35:124–32. doi:10.1016/j.ebiom.2018.07.041.
[12] Mir AA, Sarwar A. Artificial intelligence-based techniques for analysis of body cavity fluids: a review. Artif Intell Rev 2021. doi:10.1007/s10462-020-09946-y.
[13] Kaur S, Singla J, Nkenyereye L, et al. Medical diagnostic systems using artificial Intelligence (AI) Algorithms: principles and Perspectives. IEEE Access 2020;8:228049–69. doi:10.1109/ACCESS.2020.3042273.
[14] Guan J. Artificial intelligence in healthcare and medicine: promises, ethical challenges and governance. Chin Med Sci J 2019;34(2):76–83. doi:10.24920/003611.
[15] Maher NA, Senders JT, Hulsbergen AF, et al. Passive data collection and use in healthcare: a systematic review of ethical issues. Int J Med Inform 2019;129:242–7. doi:10.1016/j.ijmedinf.2019.06.015.
[16] Hong L, Luo M, Wang R, et al. Big data in health care: applications and challenges. Data Inf Manag 2018;2(3):175–97. doi:10.2478/dim-2018-0014.
[17] Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J 2019;6(2):94. doi:10.7861/futurehosp.6-2-94.
[18] Manogaran G, Lopez D. A survey of big data architectures and machine learning algorithms in healthcare. Int J Biomed Eng Technol 2017;25(2–4):182–211. doi:10.1504/IJBET.2017.087722.
[19] Fröhlich H, Balling R, Beerenwinkel N, et al. From hype to reality: data science enabling personalized medicine. BMC Med 2018;16(1):1–15. doi:10.1186/s12916-018-1122-7.
[20] Rahman AS, Shamrat FJM, Tasnim Z, et al. A comparative study on liver disease prediction using supervised machine learning algorithms. IJSTR 2019;8(11):419–22.
[21] Yahyaoui A, Jamil A, Rasheed J, et al.. Proceedings of 2019 1st international informatics and software engineering conference (UBMYK). IEEE;2019:1–4.
[22] Sahoo AK, Pradhan C, Das H. Performance evaluation of different machine learning methods and deep-learning based convolutional neural network for health decision making. In: Harff J, Bailey G, Luth F, editors. Nature Inspired Computing for Data Science. Studies in Computational Intelligence, vol 871. Springer, Cham; 2020.
[23] Ramalingam V, Dandapath A, Raja MK. Heart disease prediction using machine learning techniques: a survey. IJET 2018;7(2.8):684–7. doi:10.14419/ijet.v7i2.8.10557.
[24] Hashem S, ElHefnawi M, Habashy S, et al. Machine learning prediction models for diagnosing hepatocellular carcinoma with HCV-related chronic liver disease. Comput Methods Programs Biomed 2020;196:105551. doi:10.1016/j.cmpb.2020.105551.
[25] Konerman MA, Beste LA, Van T, et al. Machine learning models to predict disease progression among veterans with hepatitis C virus. PLoS ONE 2019;14(1):e0208141. doi:10.1371/journal.pone.0208141.
[26] Spann A, Yasodhara A, Kang J, et al. Applying machine learning in liver disease and transplantation: a comprehensive review. Hepatology 2020;71(3):1093–105. doi:10.1002/hep.31103.
[27] Sy T, Jamal MM. Epidemiology of hepatitis C virus (HCV) infection. Int J Med Sci 2006;3(2):41. doi:10.7150/ijms.3.41.
[28] Dua D, Graff C. HCV data Data Set. UCI machine learning repository. 2020. Available from https://archive.ics.uci.edu/ml/datasets/HCV+data.
[29] Chawla NV, Bowyer KW, Hall LO. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321–57. doi:10.1613/jair.953.
[30] Ahmed H, Younis EM, Hendawi A, et al. Heart disease identification from patients' social posts, machine learning solution on Spark. Future Gener Comput Syst 2020;111:714–22 doi10.1016/j.future.2019.09.056.
[31] Lever J, Krzywinski M, Altman N. Logistic regression. Nat Methods 2016;13(7):541–2. doi:10.1038/nmeth.3904.
[32] Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. J Biomed Inform 2002;35(5):352–9. doi:10.1016/s1532-0464(03)00034-0.
[33] Amin MS, Chiam YK, Varathan KD. Identification of significant features and data mining techniques in predicting heart disease. Telemat Inform 2019;36:82–93. doi:10.1016/j.tele.2018.11.007.
[34] Dulhare UN. Prediction system for heart disease using Naive Bayes and particle swarm optimization. Biomedical Research 2018;29(12).
[35] Subasi A, Alickovic E, Kevric J. Diagnosis of chronic kidney disease by using random forest. CMBEBIH 2017. Springer;2017:589–94. doi:10.1007/978-981-10-4166-2_89.
[36] Shaik AB, Srinivasan S. Proceedings of International conference on innovative computing and communications. Springer;2019:253–60.
[37] Kotsiantis SB, Zaharakis I, Pintelas P. Supervised machine learning: a review of classification techniques. Informatica 2007;160(1):3–24. doi:10.1016/j.compbiolchem.2009.04.004.
[38] Mohamed AE. Comparative study of four supervised machine learning techniques for classification. Int J Appl Sci Technol 2017;7(2):5–18.
[39] Medjahed SA, Saadi TA, Benyettou A. Breast cancer diagnosis by using k-nearest neighbor with different distances and classification rules. Int J Comput Appl 2013;62(1). doi:10.5120/10041-4635.
[40] Zhang S, Cheng D, Deng Z, et al. A novel kNN algorithm with data-driven k parameter computation. Pattern Recognit Lett 2018;109:44–54. doi:10.1016/j.patrec.2017.09.036.
[41] Goel R, Mehmood I, Ugail H. A study of deep learning-based face recognition models for sibling identification. Sensors 2021;21(15):5068 Basel. doi:10.3390/s21155068.
[42] Asadi H, Dowling R, Yan B, et al. Machine learning for outcome prediction of acute ischemic stroke post intra-arterial therapy. PLoS One 2014;9(2):e88225. doi:10.1371/journal.pone.0088225.
[43] Deng X, Liu Q, Deng Y, et al. An improved method to construct basic probability assignment based on the confusion matrix for classification problem. Inf Sci 2016;340:250–61. doi:10.1016/j.ins.2021.11.018.
[44] Radja M, Emanuel AWR. Proceedings of 5th International Conference on Science in Information Technology (ICSITech). IEEE;2019:252–8.
[45] Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett 2006;27(8):861–74. doi:10.1016/j.patrec.2005.10.010.
[46] Brzezinski D, Stefanowski J. Prequential AUC: properties of the area under the ROC curve for data streams with concept drift. Knowl Inf Syst 2017;52(2):531–62. doi:10.1007/s10115-017-1022-8.
[47] Marshall RJ. The use of classification and regression trees in clinical epidemiology. J Clin Epidemiol 2001;54(6):603–9. doi:10.1016/s0895-4356(00)00344-9.
[48] Nguyen Q, Valizadegan H, Hauskrecht M. Learning classification models with soft-label information. J Am Med Inform Assoc 2014;21(3):501–8. doi:10.1136/amiajnl-2013-001964.
[49] Das H, Naik B, Behera HS. Classification of diabetes mellitus disease (DMD): a data mining (DM) approach. Singapore: Springer Singapore;2018:539–49.
[50] Abd AL-Nabi DL, Ahmed SS. Survey on classification algorithms for data mining:(comparison and evaluation). Comput Eng Intell Syst 2013;4(8):18–27.
[51] Yoo I, Alafaireet P, Marinov M, et al. Data mining in healthcare and biomedicine: a survey of the literature. J Med Syst 2012;36(4):2431–48. doi:10.1007/s10916-011-9710-5.
[52] Syafa'ah L, Zulfatman Z, Pakaya I, et al. Comparison of machine learning classification methods in hepatitis C virus. J Online Inform 2021;6(1):73–8. doi:10.15575/join.v6i1.719.
[53] Orooji A, Kermani F. Machine learning based methods for handling imbalanced data in hepatitis diagnosis. Front Health Inform 2021;10(1):57.