Machine learning models for diagnostic classification of hepatitis C tests

Oladosu Oyeibisi Oladimeji^{1*}, Abimbola Oladimeji², Oladimeji Olayanju³

¹Department of Computer Science, University of Ibadan, Ibadan, Nigeria

²Department of Chemistry, University of Ibadan, Ibadan, Nigeria

³Department of Computer Science and Information Technology, University of Bowen, Iwo, Nigeria

Article Info

Article type:
Research

Article History:
Received: 2021-01-22
Accepted: 2021-03-23
Published: 2021-03-26

* Corresponding author:

Oladosu Oyeibisi Oladimeji

Department of Computer Science,
University of Ibadan, Ibadan, Nigeria

Email: oladimejioladosu@gmail.com

Keywords:

Hepatitis C
Machine Learning
Health Data Analytics
Decision Making

ABSTRACT

Introduction: Hepatitis C is a chronic infection caused by hepatitis c virus - a blood borne virus. Therefore, the infection occurs through exposure to small quantities of blood. It has been estimated by World Health Organization (WHO) to have affected 71 million people worldwide. This infection costs individual, groups and government a lot because no vaccine has been gotten yet for the treatment. This disease is likely to continue to affect more people because it's long asymptomatic phase which makes its early detection not feasible.

Material and Methods: In this study, we have presented machine learning models to automatically classify the diagnosis test of hepatitis and also ranked the test features in order to know how they contribute to the classification which help in decision making process by the health care industry. The synthetic minority oversampling technique (SMOTE) was used to solve the problem of imbalance dataset.

Results: The models were evaluated based on metrics such as Matthews correlation coefficient, F-measure, Precision-Recall curve and Receiver Operating Characteristic Area Under Curve. We found that using SMOTE techniques helped raise performance of the predictive models. Also, random forest (RF) had the best performance based on Matthews correlation coefficient (0.99), F-measure (0.99), Precision-Recall curve (1.00) and Receiver Operating Characteristic Area Under Curve (0.99).

Conclusion: This discovery has the potential to impact on clinical practice, when health workers aim at classifying diagnosis result of disease at its early stage.

How to cite this paper

Oyeibisi Oladimeji O, Oladimeji A, Olayanju O. Machine Learning Models for Diagnostic Classification of Hepatitis C Tests. Front Health Inform. 2021; 10: 70. DOI: [10.30699/fhi.v10i1.274](https://doi.org/10.30699/fhi.v10i1.274)

INTRODUCTION

Illnesses or conditions whose effects are persistent or permanent in nature are known to be chronic illnesses [1]. These kinds of diseases are known to have adverse effect on human life and get the huge portion of budgets by individual, government and groups [2, 3]. Hepatitis C is a chronic disease; it is a liver infection caused by the Hepatitis C Virus (HCV) [4, 5]. This HCV is responsible for both severe and chronic hepatitis; hence ranging in severity from slight illness which lasts for few days to a severe, lifetime illness.

Hepatitis C is now a major concern globally; as it is estimated that 71 million people have chronic hepatitis C virus infection globally according to WHO [4, 6]. WHO also affirms that a noteworthy number of individuals infected will develop cirrhosis or liver

cancer. This hepatitis C virus is a blood borne virus; therefore, the most common ways of infection is through exposure to small quantities of blood. This may occur by means of injection drug use, unsafe injection procedures, unsafe health care, transfusion of unscreened blood and blood products, sexual practices and other practices that lead to exposure to blood [4]. It is important to note that HCV does not have effective vaccine yet [4, 6]. However, working is on-going to have. Therefore, as at now the immune system is depended on, to be free of this infection; but treatment is necessary when the infection becomes chronic [4]. Therefore, diagnosis of HCV is of great importance; but the major challenge of new HCV infections is asymptomatic in nature [4]. This could last for decades after which the secondary symptoms begin to show up which leads to liver damage [4]. Another challenge is the diagnostic classification of

the hepatitis tests which is done after a person have been diagnosed with HCV infection, this would aid in determining the degree of liver damage [4], which we intend to propose automated methods to analyze laboratory tests results that will help to make timely decisions.

For the past few decades till this present time, machine learning as a research area has been gaining attention in noticeable researches virtually in all areas of human activities of which the health sector is not left out. Machine learning is a branch of artificial intelligence that deals with studies of scientific algorithms and scientific models that computer system uses to execute tasks efficiently without using explicit instructions, but depending on patterns and inference instead [7-9]. A lot of data have been accumulated by the healthcare industry over the years [10]; which will be of great help in bringing insight into big data for diagnostic, disease prevention, prediction and policy-making purposes in healthcare industry through machine learning and data analytics [11, 12] by making use of the clinical data.

Data analytics is evolving in healthcare industry which has been of great help in diagnostics tasks. Therefore, this work aims at creating a machine learning models for the diagnostic classification of Hepatitis C tests and ranking the test features which will aid in decision making such as the treatment procedure to be used and procedure of managing the infection. The second section of this work discusses the methodology used in this research; while the third section showcases the result, followed by the discussion of the result and finally the conclusion is drawn at the fourth section.

MATERIAL AND METHODS

This section describes the various stages that composed the proposed methodology: Data description, feature ranking and classification task.

Data Description

The dataset that was used to pinpoint this research was acquired from University of California, Irvine (UCI) Machine Learning Repository [13] which is HCV dataset by [14]. The dataset consists of Hepatitis C test records of 615 patients. The patients' record consisted of 238 women and 377 men with the age bracket of 19 to 77 years. The dataset contains 13 features, which present test information of each patient as shown in Table 1.

Table 1: Description of the Features

S/N	Features	Description	Range	Mean	Standard Deviation
1	id	Patient's ID	1 – 615	308	177.679
2	Age	Patient's Age in years	19,...,77	47.408	10.055
3	Sex	Patient's sex	M, F		
4	ALB	albumin	14.9,...,82.2	41.62	5.781
5	ALP	alkaline phosphatase	11.3,...,416.6	68.284	26.028
6	ALT	alanine amino-transferase	0.9,...,325.3	28.451	25.47
7	AST	aspartate amino-transferase	10.6,...,324	34.786	33.091
8	BIL	bilirubin	0.8,...,254	11.397	19.673
9	CHE	choline esterase	1.42,...,16.41	8.197	2.206
10	CHOL	choline	1.43,...,9.67	5.368	1.133
11	CREA	creatinine	8,...,1079.1	81.288	49.756
12	GGT	γ-glutamyl-transferase	4.5,...,650.9	39.533	54.661
13	PROT	total protein	44.8,...,90	72.044	5.403
14	category	Hepatitis C category	0,0s,1,2,3		

The range of the category which is the target class is described in Table 2 below

Table 2: Description of the category

Code	Hepatitis C category	Frequencies
0	Blood donor	533
0s	Suspect blood donor	7
1	Hepatitis	24
2	Fibrosis	21
3	Cirrhosis	30

Data Preprocessing

First step in data mining is data cleaning which involves data pre-processing processes [15, 16]. In this process, id feature was removed from the dataset as it does not contribute to anything to the classification; it is only as means of identifying the patients. Also, it was discovered that the dataset was highly skewed (imbalanced), the Synthetic Minority Oversampling Technique (SMOTE) which is an over-sampling method [17] was used to alleviate the class imbalance problem.

Classification Algorithms

For the classification task, we selected 5 of the most used classifiers in the machine learning and these classifiers are Decision Tree (DT) [18, 19], K-Nearest Neighbors (KNN) [7, 20] Random Forest (RF) [21], Naïve Bayes [22] and Logistic Regression.

Decision Tree

This is one of the most used classification algorithms. This algorithm is centered on dividing the data into different subcategories based on a sequence of questions. The process begins with the primary or root node, which is known as the root of the tree based on the highest entropy and contains all samples. Each node is then split into secondary or leaf nodes in either a multi-split or binary form. The decision model is a tree structure that includes the collection of nodes. It includes the decision nodes (split node with the condition) and leaf nodes.

Logistic Regression (LR)

This is a predictive model that is used to get the association between one dependent binary variable and one or more independent variables. LR is differentiated by not assuming a linear relationship between the dependent and independent variables but by displaying association between the output and predictive values [7]. The logistic curve that results from the logistic regression is between 0 and 1 by using natural logarithm for the curve creation.

K-Nearest Neighbor (KNN)

This also is a classification algorithm that classifies the new sample based on a measure known as similarity or distance measure. This measure comprises of three distance measures: Euclidean distance, Manhattan, and Minkowski. KNN first stores the feature sample and class label of the samples which is the training step. The second step is the classification step where user defines a 'k' values in order to classify the unknown sample for the k number of the class labels, so as to classify the unknown sample into the defined class depending on the similar features.

Random Forest (RF)

It is supervised learning algorithm that can be used for both classification and Regression tasks. Random forest [23, 24] is based on bagging technique to create random sample features. The difference between the decision tree and the random forest is the method of finding the root node and splitting the feature node which will run randomly. The first step in random forest is to load the data which consist of 'n' features. Then the algorithm is trained by selecting m from the n features based on bagging technique in order to determine the unbiased OOB error. After that, node 'd' is calculated using the best split, then split the node into sub nodes. These steps are repeated to find m number of trees, after which the total number of vote for each tree for the target class is calculated and the highest class is the final prediction.

Naïve Bayes (NB)

This is a classification algorithm [22] based on Bayes theorem with the objectivity assumption between the

predictors. It takes the dataset as input and then analysis is performed; after which class label is predicted with the use of Bayes' Theorem. It calculates a probability of class in input data and helps to predict the class of the unknown data sample. It is a powerful classification algorithm suitable for large datasets.

Classification

Decision tree [J48], Random forest, Naïve bayes, logistic regression and K-nearest neighbors algorithms were implemented using Waikato Environment for Knowledge Analysis (WEKA). It is a reliable open source software for knowledge analysis developed at the University of Waikato, New Zealand [25]. Cross validation is used as the test mode option with 10 as the number of fold. Class attribute was set as the target to be predicted for the classification. This process was done 5 times coupled with changing the random seed starting from 1 -5 for the process for validation purposes.

RESULTS

After the SMOTE was applied the dataset was increased to 2181, the details are given in the Table 3 below.

Table 3: Description of the category after applying SMOTE

Code	Hepatitis C category	Frequencies
0	Blood donor	533
0s	Suspect blood donor	448
1	Hepatitis	384
2	Fibrosis	336
3	Cirrhosis	480

This section first describes the results that was obtained for the Hepatitis C test classification; before applying SMOTE followed by the results that was obtained after SMOTE was applied. The algorithms were executed as stated in the previous section. The performance measures such as Matthews correlation coefficient (MCC) [26], Receiver Operating Characteristic Area Under Curve (ROC AUC), Precision-Recall Area Under curve (PR AUC), F-Measure and Accuracy that are derived from the confusion matrix which is used to decide how well a classification has performed [27] by reporting the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

Table 4 and Table 5 shown below provide the details of the average performance of the classification based on F-Measure, MCC, AOC RUC, PR AUC and accuracy after the process was repeated five times coupled with changing of random seed values from 1 -5.

Table 4: Summary of performance measure of the classification before applying SMOTE.

Method	F-Measure	MCC	ROC AUC	PRC AUC	Accuracy
J48	0.9002	0.7254	0.8644	0.8802	90.60162
Random Forest	0.916333	0.767	0.9872	0.9542	92.61786
KNN	0.8828	0.5754	0.7208	0.8378	90.21138
LR	0.9226	0.7766	0.9622	0.9404	92.61786
NB	0.909	0.7036	0.966	0.9346	90.79674

Table 5: Summary of performance measure of the classification after applying SMOTE.

Method	F-Measure	MCC	ROC AUC	PRC AUC	Accuracy
J48	0.9704	0.9632	0.984	0.9576	97.0289
Random Forest	0.9898	0.9868	1	0.999	98.97294
KNN	0.88	0.8614	0.9232	0.821	88.2531
LR	0.9376	0.9234	0.991	0.9646	93.76434
NB	0.8414	0.8186	0.9804	0.9104	84.96102

DISCUSSION

Based on the results obtained, it was observed the Random Forest algorithm performed best out of the algorithms considered for this study in terms of the performance metrics which includes F-measure, MCC, ROC AUC, PRC AUC, and Accuracy using cross validation method. Furthermore, looking at the ROC AUC and PRC AUC which provide more robust and better performance estimates when comparing classifiers on imbalance dataset, it was observed that the algorithms performed better after applying SMOTE technique to the dataset.

The result of this study is encouraging and promising as these models can be used by the healthcare industry to classify the diagnosis test of Hepatitis C as it is not expensive and time consuming to use. Also from the result obtained from the feature ranking it shows that age and gender contribute little or nothing in the classification of the Hepatitis C test.

Certain that the dataset was recently published in the UCI repository, little related studies have covered it. This includes Chawathe study [6], in which the author proposed various models using different algorithms coupled with feature selection. The results were compared based on accuracy, F-Measure, ROC AUC, classification time and training time in which results

were depicted with graph. The results showed that the ROC AUC is below 0.98. Also Orooji and Kermani study [28], in which the authors proposed various models using three algorithms coupled with techniques of handling imbalanced dataset. Their results also showed that Random forest has the best performance out of the three algorithms with 99.9% accuracy and F-measure of 9.99. The authors did not provide neither the PRC AUC nor ROC AUC.

CONCLUSION

This paper has presented machine learning models for the automatic classification of Hepatitis C test at early stage. It will aid in decision making which can also be extended to others diagnosis task. This study has also shown the importance of machine learning in the healthcare industry in decision making. Conclusively, it is important to note that public should be aware that healthy living such as protection from blood exposure by making provision for personal objects such as razor blade and needle and following of standard medical procedures.

It would be interesting in the future research to get to know whether blood type, genotype could be included in the dataset and check out the roles they play in the classification of Hepatitis C diagnosis classification.

ACKNOWLEDGEMENT

The authors would like to appreciate Rev. Sunday Oladimeji and Rev. Bayo Afolaranmi for the proof reading and language editing of the manuscript

AUTHOR'S CONTRIBUTION

OOO – Conception, design and drafting of the work.

AO – Interpretation of the data and revised the work.

OO – Acquisition and analysis of the data and revised the work.

The authors agree on this final form of the manuscript, and attested that all authors contributed in the final draft of the 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. Alam TM, Iqbal MA, Ali Y, Wahab A, Ijaz S, Baig TI, et al. A model for early prediction of diabetes. *Informatics in Medicine Unlocked*, 2019; 16: 100204.

2. Skyler JS, Bakris GS, Bonifacio, Darsow T, Eckel RH, Groop L, et al. Differentiation of diabetes by pathophysiology, natural history, and prognosis. *Diabetes*. 2017; 66(2): 241-55. PMID: 27980006 DOI: 10.2337/db16-0806 [\[PubMed\]](#)
3. Tao Z, Shi A, Zhao J. Epidemiological perspectives of diabetes. *Cell Biochem Biophys*. 2015; 73(1): 181-5. PMID: 25711186 DOI: 10.1007/s12013-015-0598-4 [\[PubMed\]](#)
4. World Health Organization. Hepatitis C [Internet]. 2020 [cited: 9 Nov 2020]. Available from: www.who.int/news-room/fact-sheets/detail/hepatitis-c
5. Centers for Disease Control and Prevention. Hepatitis [Internet]. 2018 [cited: 9 Nov 2020]. Available from: www.cdc.gov/hepatitis/hcv/cfaq.htm
6. Chawathe SS. Diagnostic classification using hepatitis C tests. International IOT, Electronics and Mechatronics Conference. IEEE; 2020.
7. Bishop C. Pattern recognition and machine learning. Springer; 2006.
8. Awan SE, Bennamoun M, Sohel F, Sanfilippo FM, Dwivedi G. Machine learning based prediction of heart failure readmission or death: Implications of choosing the right model and the right metrics. *ESC Heart Fail*. 2019; 6(2): 428-35. PMID: 30810291 DOI: 10.1002/ehf2.12419 [\[PubMed\]](#)
9. Oladimeji OO, Oladimeji O. Predicting survival of heart failure patients using classification algorithms. *Journal of Information Technology and Computer Engineering*. 2020; 4(2): 90-4.
10. Joloudari JH, Saadatfar H, Dehzangi A, Shamshirband S. Computer-aided decision-making for predicting liver disease using PSO-based optimized SVM with feature selection. *Informatics in Medicine Unlocked*. 2017; 17: 100255.
11. Metzge BE, Lowe LP, Dyer AR, Trimble ER, Chaovarindr U, Coustan DR, et al. Hyperglycemia and adverse pregnancy outcomes. *N Engl J Med*. 2008; 358(19): 1991-2002. PMID: 18463375 DOI: 10.1056/NEJMoa0707943 [\[PubMed\]](#)
12. Sneha N, Gangi T. Analysis of diabetes mellitus for early prediction using optimal features selection. *Journal of Big Data*. 2019; 6: 13.
13. UCI. Machine learning repository [Internet]. 2007 [cited: 4 Nov 2020]. Available from: <https://archive.ics.uci.edu/ml/index.php>
14. Hoffmann G, Bietenbeck A, Lichtinghagen R, Klawon F. Using machine techniques to generate laboratory diagnostic pathways: A case study. *Journal of Laboratory and Precision Medicine*. 2018; 3(6): 58-67.
15. Han J, Kamber M, Pei J. Data mining: Concepts and techniques. 3rd ed. Elsevier; 2001.
16. Larose DT, Larose CD. Introduction to data mining and knowledge discovery. John Wiley & Sons; 1996.
17. Hu G, Xi T, Mohammed F, Miao H. Classification of wine quality with imbalanced data. *International Conference on Industrial Technology*. IEEE; 2016.
18. Burkov A. The hundred page machine learning book. Andriy Burkov; 2019.
19. Alshamlan H, Badr G, Alohal Y. Gene selection and cancer classification method using artificial bee colony and SVM algorithms (ABC-SVM). *International Conference on Data Engineering*. Springer; 2015.
20. Breiman L. Random forest. *Machine Learning*. 2001; 45: 5-32.
21. Marsland S. Machine learning: An algorithmic perspective. 2nd ed. CRC Press; 2015.
22. Martinez-Arroyo M, Sucar L. Learning an optimal Naïve Bayes classifier. *International Conference on Pattern Recognition*. IEEE; 2006.
23. Lee JW, Lee JB, Park M, Song SH. An extensive evaluation of recent classification tools applied to microarray data. *Computational Statistics and Data Analysis*. 2005; 48: 869-85.
24. Yeung KY, Bumgarner RE, Raftery AE. Bayesian model averaging: Development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics*. 2005; 21(10): 2394-402.
25. WEKA. The workbench for machine learning [Internet]. 2015 [cited: 3 Nov 2020]. Available from: www.cs.waikato.ac.nz/ml/weka
26. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta (BBA)-Protein Structure*. 1975; 405(2): 442-51.
27. Diez P. Smart wheelchairs and brain-computer interfaces. Academic Press; 2018.
28. Orooji A, Kermani F. Machine learning based methods for handling imbalanced data in hepatitis diagnosis. *Frontiers in Health Informatics*. 2021; 10: 57.