Predicting Infectious State of Hepatitis C Virus Affected Patient's Applying Machine Learning Methods

Khair Ahammed
IIT
Noakhali Sci. & Tech. University
Noakhali, Bangladesh
khairahmad6@gmail.com

Md. Shahriare Satu
Dept. of MIS
Noakhali Sci. & Tech. University
Noakhali, Bangladesh
shahriar.setu@gmail.com

Md. Imran Khan
Dept. of CSE
Gono University
Dhaka, Bangladesh
imran.khan@gmail.com

Md Whaiduzzaman
School of Information Systems
QUT
Brisbane, Australia
md.whaiduzzaman@qut.edu.au

***Abstract*—Hepatitis C virus is a major cause for happening liver disease all over the world. However, many tools have been build that try to reduce the influence of this virus. In this work, a machine learning based model has been proposed that can classify hepatitis C virus infected patient's stages of liver. We gathered the instances of liver fibrosis disease of Egyptian patients from UCI machine learning repository. To balance instances of multiple categories, synthetic minority oversampling methodology has been used that increases synthetic instances of patients. Later, we applied different feature selection methods to identify significant features of hepatitis C virus in this dataset. Various classifiers has been employed to categorize patients into balanced primary, feature selected and primary HCV instances. After analyzing this results, KNN shows the best 94.40% accuracy than any other classifiers. This result has been useful to scrutinize and take decision in hepatitis C virus infectious disease.**

***Index Terms*—Hepatitis C virus, machine learning, SMOTE, feature selection, classification**

I. Introduction

More than 160 million people are infected by Hepatitis C virus (HCV) where 350000 people are dying from HCV virus in each year [1]. It is an infectious disease that causes chronic hepatitis, hepatocellular carcinoma and liver cirrhosis. Egypt has severely faced liver disease due to chronic HCV that has been shown the highest prevalence rate 13%-15% of it [2].Hepatitis C virus attacks the liver and a buildup scar tissue in the liver is called fibrosis. According to Meta-analysis of Histological Data in Viral Hepatitis (METAVIR) system [3], the stages of liver fibrosis (F0-F4) can be classified into no fibrosis (F0), portal fibrosis (F1), few septa (F2), many septa (F3) and cirrhosis (F4) respectively. Liver biopsy is still considered as the gold standard for diagnosis of liver fibrosis. But, it is more invasive, expensive and vulnerable with sampling residuals and historical assessment that is inconvenient for the patients [4]. In recent years, many non-invasive methods have been proposed to detect fibrosis and cirrhosis of HCV patients. Besides, these methods are safe, easy, reproducible and give accurate results. Two kinds of non-invasive methods are found which are: based on indexes of serum markers such as FIB-4 score and aspartate aminotransferase(AST) to-platelet ratio index (APRI) and based on imaging techniques like Transient Elastography (TE) respectively [5]. Nevertheless, these methods are not shown such reliability in lieu of liver biopsy [6]. Parkes et al. [7] shows that serum markers of liver fibrosis refer to a gorgeous substitute where they are less invasive than biopsy with no risk of difficulties, easy and repeated performs, observer variability and reduce sampling. Recently, machine learning is a emerging field to investigate the condition of affected patients more precisely than previous techniques. We proposed a machine learning model which can explore significant symptoms and analyze the situation of patients. There were found only a small amount of works that can detect the stages of HCV infected liver fibrosis using machine learning. In 2013, Zayed et al. [8] developed a J48 decision tree (DT) model to analyze 3719 Egyptian patients with choronic HCV which showed 73% accuracy. Ayeldeen et al. [9] also used variable selection as well as DT based classifier to investigate HCV data and found 93.7% accuracy. However, Orczyk and Porwik [10] applied various feature selection and classification methods and three best classifier were achieved nearly 70% accuracy. Hashem et al. [11] analyzed 39567 Egyptian patients by associating the serum biomarkers and clinical information and developed an alternating decision tree (ADT) based model which achieves 84.80% accuracy. Barakat et al. [12] investigated over 166 Egyptian children with Chronic HCV where staging of fibrosis, APRI and FIB-4 scores and their areas under the ROC curve (AUROC) have been obtained and predicted 87.5% accurately using the Random Forest (RF).

In previous works, DT based model was focused on mostly. The systematic studies with various classifiers were missing and not explored the outcomes using multiple evaluation metrics. In present work, the impact of liver fibrosis is analyzed using various feature selection and classification methods. So, the findings of the experiment can be scrutinized elaborately which helps us to identify the best machine learning model for detecting HCV. This work is also needed to the scientists and health care community to explore important symptoms of

978-1-7281-7366-5/20/$31.00 ©2020 IEEE

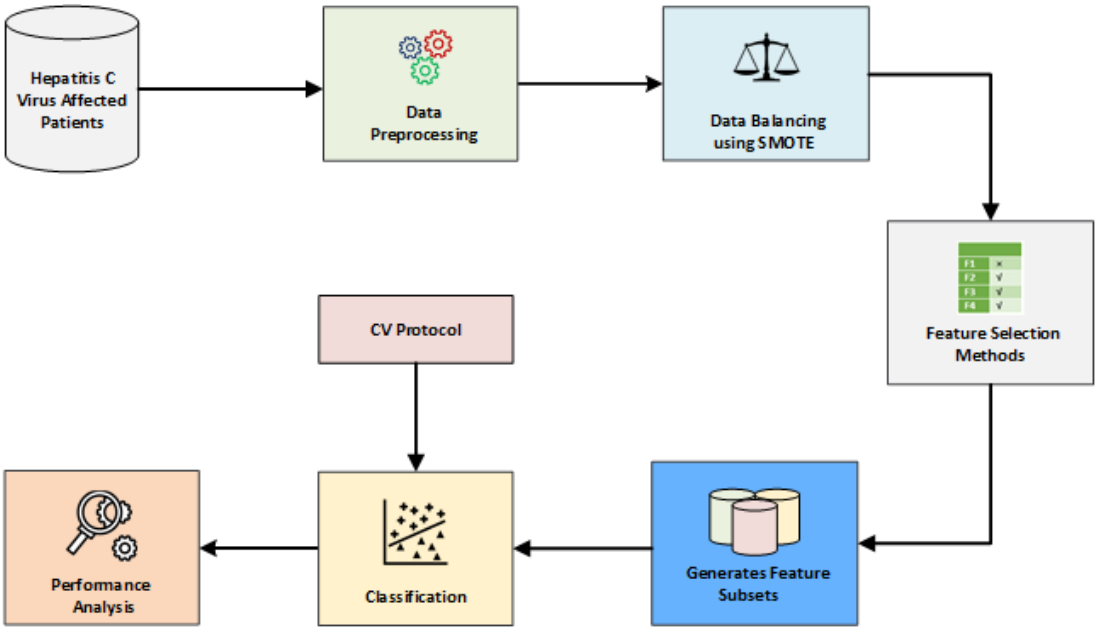

Fig. 1. Proposed Methodology

HCV and enrich awareness about it.

II. Proposed Methodology

In this work, a supervised machine learning approach is proposed where HCV instances were pre-processed and classified to reach the best results. Figure 1 illustrates the steps about how machine learning model can detect the stages of liver fibrosis by analyzing raw HCV patient's data. Hence, this approach is described more elaborately as follows:

A. Dataset Description

This study uses almost 1385 patient's records who are infected by HCV. Patient's data was gathered by the researchers of the School of Information and Computer Science, University of California and later uploaded it from UCI machine learning repository [2]. It demonstrates HCV Liver fibrosis dataset where are also includes under treatment patients. There are remained 1385 patients where 707 (51%) were male patients and 678 (49%) were female patients. This dataset also contains 29 features which contains various information such as age, gender, body mass index, fever, nausea/vomiting, headache, diarrhea, fatigue & generalized, bone ache, jaundice, epigastric pain, white blood cell (WBC), red blood cells (RBC), hemoglobin (HBC), platelets, aspartate transaminase ratio for 1 week (AST), alanine aminotransferase ratio (ALT) of 1,4,12,24,36, 48 and after 36 weeks, RNA base/4/12/end-of-treatment (EOT) / Elongation factor (EF), baseline historical grading and baseline histological staging as class level. The class level illustrates different forecast levels of fibrosis which contains no fibrosis (F0), portal fibrosis (F1), few septa (F2), many septa (F3) and cirrhosis (F4). This dataset does not contain any missing values. In addition, There are found 336 (24.25%) F1 staged patients, 332 (23.97%) F2 staged patients, 355 (25.63%) F3 and 362 (26.13%) F4 staged patients have in this dataset.

B. Data Preprocessing & Balancing

Firstly, we seek any missing and wrong instances in this HCV patient's dataset. But, there was not found any missing or wrong instances in here. In this circumstance, this dataset contains almost 1385 patients and this dataset is slight imbalance. But, it does not contain more instances so that we are perceived low accuracy as expectation. Synthetic Minority Oversampling Techniques (SMOTE) is an oversampling data balancing technique that is used to procreate equivalent instances using the raw records. After using SMOTE 8 times, it generated the new samples based on nearest neighbor strategy. Hence, we balanced and oversampled instances by using SMOTE and generating 5540 where 1344, 1328, 1420 and 1448 patients are in portal fibrosis, few septa, many septa and cirrhosis stage respectively.

C. Feature Selection & Classification

Feature selection [13], [14] helps a machine learning model to identify appropriate features that causes low consumption cost, short time, reduce the chance of overfitting and generate more accurate results. Several filter based feature selection methods such as Chi-Square Attribute Evaluation (CSAE), Gain Ratio Attribute Evaluation (GRAE), Info Gain Attribute Evaluation (IGAE) and ReliefF (RFAE) are implemented to rank and identify features from SMOTE generated HCV dataset. In this case, all of these feature selection methods were used ranker algorithm. Then, different classifiers [15], [16] such as Random Forest (RF), K-nearest neighbor (KNN), Decision Tree (DT), Naïve Bayes (NB), Logistic Regression (LR), Xgboost classifier (XGB), Support Vector Machine

TABLE I
EXPERIMENTAL RESULT OF FIVE BEST CLASSIFIERS

Classifier	Accuracy	AUROC	F-Measure	G-Mean	Sensitivity	Specificity	Type I Error	Type II Error
HCV Dataset without applying SMOTE								
KNN	0.2548	0.5037	0.2549	0.4380	0.2548	0.7527	0.5619	0.2472
RF	0.2512	0.4998	0.2477	0.4336	0.2512	0.7484	0.5663	0.2515
NB	0.2476	0.4976	0.2437	0.4302	0.2476	0.7476	0.5697	0.2523
DT	0.2433	0.4950	0.2428	0.4262	0.2433	0.7467	0.5737	0.2532
LR	0.2433	0.4945	0.2389	0.4259	0.2433	0.7457	0.5740	0.2542
AVG	0.2480	0.4981	0.2456	0.4308	0.2480	0.7482	0.5691	0.2517
HCV Dataset after applying SMOTE								
HCV Dataset								
KNN	0.9440	0.9627	0.9440	0.9625	0.9440	0.9814	0.0374	0.0185
RF	0.9216	0.9477	0.9216	0.9473	0.9216	0.9737	0.0526	0.0262
SVM	0.8277	0.8851	0.8277	0.8832	0.8277	0.9424	0.1167	0.0575
DT	0.5767	0.7176	0.5767	0.7037	0.5767	0.8586	0.2962	0.1413
XGB	0.5572	0.7043	0.5563	0.6888	0.5572	0.8515	0.3111	0.1484
AVG	**0.7654**	**0.8435**	**0.7653**	**0.8371**	**0.7654**	**0.9215**	**0.1628**	**0.0784**
CSAE Dataset								
KNN	0.9440	0.9627	0.9440	0.9625	0.9440	0.9814	0.0374	0.0185
RF	0.9256	0.9503	0.9256	0.9500	0.9256	0.9751	0.0499	0.0248
SVM	0.8277	0.8851	0.8277	0.8832	0.8277	0.9424	0.1167	0.0575
DT	0.5848	0.7230	0.5848	0.7097	0.5848	0.8612	0.2902	0.1387
XGB	0.5570	0.7042	0.5562	0.6886	0.5570	0.8514	0.3113	0.1485
AVG	**0.7679**	**0.8451**	**0.7797**	**0.8388**	**0.7679**	**0.9223**	**0.1611**	**0.0776**
GRAE Dataset								
KNN	0.9440	0.9627	0.9440	0.9625	0.9440	0.9814	0.0374	0.0185
RF	0.9256	0.9503	0.9256	0.9500	0.9256	0.9751	0.0499	0.0248
SVM	0.8277	0.8851	0.8277	0.8832	0.8277	0.9424	0.1167	0.0575
DT	0.5810	0.7205	0.5809	0.7069	0.5810	0.8600	0.2930	0.1399
XGB	0.5570	0.7042	0.5562	0.6886	0.5570	0.8514	0.3113	0.1485
AVG	**0.7671**	**0.8446**	**0.7669**	**0.8382**	**0.7671**	**0.9221**	**0.1617**	**0.0779**
IGAE Dataset								
KNN	0.9440	0.9627	0.9440	0.9625	0.9440	0.9814	0.0374	0.0185
RF	0.9256	0.9503	0.9256	0.9500	0.9256	0.9751	0.0499	0.0249
SVM	0.8277	0.8851	0.8277	0.8832	0.8277	0.9424	0.1167	0.0575
DT	0.5736	0.7156	0.5736	0.7014	0.5736	0.8576	0.2985	0.1423
XGB	0.5570	0.7042	0.5562	0.6886	0.5570	0.8514	0.3113	0.1485
AVG	**0.7656**	**0.8436**	**0.7654**	**0.8371**	**0.7656**	**0.9216**	**0.1628**	**0.0783**
RFAE Dataset								
KNN	0.9440	0.9627	0.9440	0.9625	0.9440	0.9814	0.0374	0.0185
RF	0.9256	0.9503	0.9256	0.9500	0.9256	0.9751	0.0499	0.0249
SVM	0.8277	0.8851	0.8277	0.8832	0.8277	0.9424	0.1167	0.0575
DT	0.5805	0.7202	0.5804	0.7065	0.5805	0.8599	0.2934	0.1400
XGB	0.5570	0.7042	0.5562	0.6886	0.5570	0.8514	0.3113	0.1485
AVG	**0.7670**	**0.8445**	**0.7668**	**0.8382**	**0.7670**	**0.9220**	**0.1617**	**0.0779**

(SVM) using radial basis function (RBF) kernel and Gradient Boosting (GB) are used to analyze HCV dataset using 10-fold cross validation. The classifier's performance were measured by different evaluation metrics such as accuracy, AUROC, f-measure, g-mean, sensitivity, specificity, type I error and type II error rate respectively.

III. EXPERIMENTAL RESULT & DISCUSSION

In this experiment, HCV patient's records was investigated using various machine learning techniques to detect the stages of the HCV patient. To identify significant features, Waikato Environment for Knowledge Analysis (WEKA) was used and various classification was implemented into HCV dataset using python language based sci-kit learn library. Table I represents top five best sorted results of different classifiers based on various evaluation criteria. In these circumstance, GB, NB and LR shows their accuracy less than 50%, so they are discarded in this section. In primary HCV patient's data, these classifiers are represented much poor results (the highest accuracy is shown by KNN (25.48%). Thus, SMOTE was applied 8 times in the raw HCV dataset and generated more synthetic instances like existing instances. Then, various feature selection techniques were used to the SMOTE generated HCV patient's instances and produce datasets. The primary HCV data was

not used because the classifiers do not find better results in it. Therefore, feature selection based datasets also kept all features in a rank and zero / negative ranked features can not removed because it lessen the performance of this model.

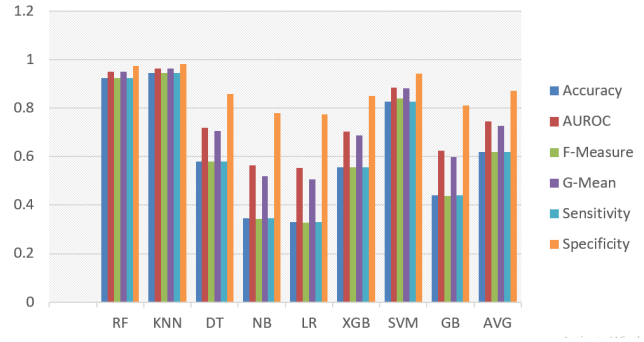

Fig. 2. Experimental Results of Different Classifiers

The experimental performance of classifiers have been increased rapidly for SMOTE generated raw HCV data. SMOTE generates more synthetic instances as well as balanced its categories using nearest neighbour strategy. When raw HCV dataset has been explored, KNN shows better result than other classifiers as well as KNN shows the highest accuracy (94.40%) in SMOTE generated data. Later, RF shows the second and SVM shows the third highest performance in this work. Therefore, GRAE and IGAE dataset show the almost same result for each metrics. On average, KNN (94.40%), RF (92.48%) and SVM (82.77%) shows better outcomes than other classifiers. Figure 2 presents the average results of different classifiers in this experiment.

When the experimental result is observed, KNN represents the highest result for different metrics for all dataset. If we explore the performance of previous works in section I, this work shows the highest result than others. Many works also analyzed HCV dataset using only various DT based model (e.g., J48, DT, RF) [9], [11], [12]. But in current work, it represents a details systematic analysis in HCV detection.

IV. Conclusion & Future Works

In this work, HCV patient's dataset was investigated using data balancing, feature selection and classification techniques. We observed that KNN shows the best result to classify the stages of HCV patients. Besides, RF and SVM show the feasible results and indicate that they work better at such kind of analysis. Some limitations have been found when we analyze HCV patient records. The experimental ECV raw dataset does not contain more instances and features to analyze it. Furthermore, this should be done by using more feature selection, transformation as well as more classification approaches. In future, we will mitigate this limitations by gathering more instances and integrating Internet of Things (IoT) based modules to detect HCV automatically.

Acknowledgment

The authors acknowledge that this research is partially supported through the Australian Research Council Discovery Project: DP190100314, "Re-Engineering Enterprise Systems for Microservices in the Cloud".

References

[1] A. A. Mohamed, T. A. Elbedewy, M. El-Serafy, N. El-Toukhy, W. Ahmed, and Z. A. El Din, "Hepatitis c virus: A global view," *World journal of hepatology*, vol. 7, no. 26, p. 2676, 2015.

[2] M. Nasr, K. El-Bahnasy, M. Hamdy, and S. M. Kamal, "A novel model based on non invasive methods for prediction of liver fibrosis," in *2017 13th International Computer Engineering Conference (ICENCO)*. IEEE, 2017, pp. 276–281.

[3] P. Bedossa and T. Poynard, "An algorithm for the grading of activity in chronic hepatitis c," *Hepatology*, vol. 24, no. 2, pp. 289–293, 1996.

[4] N. Chalasani, Z. Younossi, J. E. Lavine, M. Charlton, K. Cusi, M. Rinella, S. A. Harrison, E. M. Brunt, and A. J. Sanyal, "The diagnosis and management of nonalcoholic fatty liver disease: practice guidance from the american association for the study of liver diseases," *Hepatology*, vol. 67, no. 1, pp. 328–357, 2018.

[5] S. Hashem, G. Esmat, W. Elakel, S. Habashy, S. A. Raouf, M. Elhefnawi, M. I. Eladawy, and M. ElHefnawi, "Comparison of machine learning approaches for prediction of advanced liver fibrosis in chronic hepatitis c patients," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 15, no. 3, pp. 861–868, 2017.

[6] J. Cobbold, M. Crossey, P. Colman, R. Goldin, P. Murphy, N. Patel, J. Fitzpatrick, W. Vennart, H. Thomas, I. Cox *et al.*, "Optimal combinations of ultrasound-based and serum markers of disease severity in patients with chronic hepatitis c," *Journal of viral hepatitis*, vol. 17, no. 8, pp. 537–545, 2010.

[7] J. Parkes, I. N. Guha, P. Roderick, and W. Rosenberg, "Performance of serum marker panels for liver fibrosis in chronic hepatitis c," *Journal of hepatology*, vol. 44, no. 3, pp. 462–474, 2006.

[8] N. Zayed, A. B. Awad, W. El-Akel, W. Doss, T. Awad, A. Radwan, and M. Mabrouk, "The assessment of data mining for the prediction of therapeutic outcome in 3719 egyptian patients with chronic hepatitis c," *Clinics and research in hepatology and gastroenterology*, vol. 37, no. 3, pp. 254–261, 2013.

[9] H. Ayeldeen, O. Shaker, G. Ayeldeen, and K. M. Anwar, "Prediction of liver fibrosis stages by machine learning model: A decision tree approach," in *2015 Third World Conference on Complex Systems (WCCS)*. IEEE, 2015, pp. 1–6.

[10] T. Orczyk and P. Porwik, "Liver fibrosis diagnosis support system using machine learning methods," in *Advanced Computing and Systems for Security*. Springer, 2016, pp. 111–121.

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

[12] N. H. Barakat, S. H. Barakat, and N. Ahmed, "Prediction and staging of hepatic fibrosis in children with hepatitis c virus: A machine learning approach," *Healthcare Informatics Research*, vol. 25, no. 3, pp. 173–181, 2019.

[13] M. Dash and H. Liu, "Feature selection for classification," *Intelligent data analysis*, vol. 1, no. 3, pp. 131–156, 1997.

[14] M. S. Satu, F. Tasnim, T. Akter, and S. Halder, "Exploring significant heart disease factors based on semi supervised learning algorithms," in *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*. IEEE, 2018, pp. 1–4.

[15] M. S. Satu, S. Ahamed, F. Hossain, T. Akter, and D. M. Farid, "Mining traffic accident data of n5 national highway in bangladesh employing decision trees," in *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*. IEEE, 2017, pp. 722–725.

[16] M. S. Satu, T. Akter, and M. J. Uddin, "Performance analysis of classifying localization sites of protein using data mining techniques and artificial neural networks," in *2017 International Conference on Electrical, Computer and Communication Engineering (ECCE)*. IEEE, 2017, pp. 860–865.