APPLICATIONS OF
MODELLING AND SIMULATION
http://arqiipubl.com/ams eISSN 2600-8084 VOL 4, 2020, 89-100

Hepatitis C Virus (HCV) Prediction by Machine Learning Techniques

Satish CR Nandipati[1], Chew XinYing[1*], Khaw Khai Wah[2]

[1]School of Computer Sciences, Universiti Sains Malaysia, Pulau Pinang, Malaysia
[2]School of Management, Universiti Sains Malaysia, Pulau Pinang, Malaysia

*Corresponding author: xinying@usm.my

Submitted 05 February 2020, Revised 09 March 2020, Accepted 11 March 2020.

Abstract: Hepatitis C being as a prevalent disease in the world especially in countries like Egypt. It is estimated that 3-4 million new cases every year, indicating as a public health problem and should be addressed with identification and treatment policies. In the initial stage, it is asymptomatic however when infection progress it leads to chronic conditions such as liver cirrhosis and hepatocellular carcinoma. Some of the various non-invasive serum biochemical markers are used to identify this disease. This study aims to know the performance comparisons between multi and binary class labels of the same dataset, not limited to tool comparison, and to know which selected features play a key role in the prediction of Hepatitis C Virus (HCV) by using Egyptian patient's dataset. The highest accuracy is shown by KNN (51.06%, R) and random forest (54.56%, Python) in multi and binary class label respectively. The overall evaluation metrics comparison shows R as a better tool for this case. On the other hand, the performance score of the binary class shows better that the multiclass label. The multi-feature selection methods did not show any similar arrangement/topology in the ranking order of selected features. Finally, the 12 selected features by principal component analysis show similar performances to complete dataset and also the 21 selected features, thus showing these features may play a role in the prediction of the HCV dataset.

Keywords: Classification; Feature selection; Hepatitis C virus; Machine learning; Prediction multi and binary class labels; Python and R tools.

1. INTRODUCTION

Hepatitis C virus (HCV) is an RNA virus and one of the major blood-born human pathogen called as Hepatitis C. HCV infection is largely asymptomatic with little visible symptoms during infection stage. Without treatment, most of the acute infections progress to chronic ones followed by liver diseases such as cirrhosis and hepatocellular carcinoma [1]. According to the World Health Organization (WHO), it is estimated that nearly 3% or 120-130 million world population are infected with HCV and 3-4 million new cases of infection. Thus, representing as of the leading public health problems in the world, which should be addressed with strong program interventions for identification and treatment [2]. It is known they are 8 and 86 confirmed HCV genotypes and subtypes respectively which are said to country dependent [3]. The highest prevalence has been seen in Egypt country which accounts for nearly 22% of the population. This could be due to the widespread HCV infection by the mass-scale treatment campaign of intravenous antischistosomal injections executed between the periods of 1950 – 1980 [4]. The high prevalence is seen in countries like Asia and Africa, whereas low prevalence is seen in Australia, North America, and Western European countries [2]. Liver biopsy has proven to be a useful tool for the detection and staging of liver fibrosis in pediatric and adults. However, this method is painful, invasive and could lead patients to hospitalization afterward, shown prone to 20–30% errors in diagnosis or staging of fibrosis. These aforementioned complications lead to the search for various non-invasive diagnostic biomarkers such as serum biochemical markers, and various imaging techniques (i.e., ultrasound, magnetic resonance imaging, etc.) for the assessment of liver stiffness, or a combination of aforementioned non-invasive methods.

The most common serum biochemical markers include aspartate aminotransferase (AST), alanine aminotransferase (ALT) and AST-to-platelet ratio (APRI) and fibrosis score (FIB-4) [5]. The person with unexplained high ALT levels or high risk of getting infected is recommended for HCV testing, and the positive HCV RNA test is used as a diagnostic tool. The chronic stage of Hepatitis C can be identified with the detectable levels of HCV RNA at 12 weeks, to prevent ongoing transmission the group needs to undergo drug therapy. The monitoring of HCV treatment can be followed by examining the HCV RNA levels. On the other hand, the genotyping can be useful to determine the duration of treatment and predict the likelihood of treatment response [6]. A large amount of data is produced by different medical centers and hospitals. Even though the data contains potential information but not used in a proper way to study clinical decision support, disease surveillance, and population health management [7]. On the other hand, the invasive methods such as biopsy have some potential risks, thus non-invasive methods such as blood serum markers and imaging have been used to predict the disease status and also to reduce the medical costs. The previous studies show that the integration of clinical decision support and computer-based patient records with different machine learning classification techniques and features selection methods are useful for the prediction of different diseases [8].

2. LITERATURE REVIEW

This section reviews some work related to classification and feature selection methods applied for the prediction of HCV diseases.

The management of HCV infected patients can be monitored by the assessment of liver fibrosis staging in Chronic Hepatitis C (CHC), to check the prognosis of the disease, to establish optimal timing for therapy, and to predict the response to treatment respectively. Even though liver biopsy has been used as a better diagnose method, this method has potential risks such as invasive nature, liable to sampling error and cost of monitoring. To overcome this, an alternative is non-invasive methods such as blood serum markers [alanine aminotransferase (ALT), albumin, alpha-fetoprotein (AFP), aspartate aminotransferase (AST), creatinine, glucose, hemoglobin (Hb), indirect bilirubin, international normalized ratio (INR), platelet count, postprandial glucose test (PC%), quantity of HCV_RNA, serology finding, total bilirubin, white blood cells (WBC) count], along with clinical information [such as age, gender, and body mass index (BMI)), and contain histological findings (such as grade of fibrosis and the activity)]. In view of this, both serum biomarkers and clinical information have been taken into consideration to evaluate various machine learning techniques and develop classification models for the prediction of advanced fibrosis. Thus, four classification algorithms such as decision tree, genetic algorithm, multi-linear regression, and particle swarm optimization were used with Matlab and WEKA Softwares to evaluate the Cohort of 39,567 chronic HCV patients in Egypt. A total of 21 attributes are used for binary classification. Among the four classification algorithms, the decision tree showed the highest accuracy and AUROC of 84% and 0.76 respectively and shows the four selected features age, AST, albumin, and platelet count which can be used for further studies [9].

In another study, one of the complications in chronic liver disease is esophageal varices. Hence is it necessary to know the presence of the esophageal varices via upper gastrointestinal endoscopy to avoid bleeding. This method increases the workload of endoscopy units since varies are present in less than 50% of patients with cirrhosis, and also this procedure is uncomfortable for many patients. Thus, the prediction of varices by noninvasive methods benefits from upper gastrointestinal endoscopy screening. Previous studies showed that Fibrosis-4 index (FIB-4) can be useful to predict esophageal varices with 66.9 accuracy and 63% Area Under the Curve (AUC) respectively. During the period between 2006 – 2017, a chronic Hepatitis C dataset which consist of 4962 patients with non-invasive methods such as blood serum such as hemoglobin (HBGL), platelet count, white blood cells (WBC) count, and liver function tests [alanine aminotransferase (ALT), albumin, alpha-fetoprotein (AFP), aspartate aminotransferase (AST), indirect bilirubin, international normalized ratio (INR), prothrombin concentration (PC), total bilirubin], clinical information [age, alcohol consumption and tobacco consumption, body mass index (BMI) and gender], laboratory tests [Anti-Nuclear Antibody (ANA), Baseline_PCR, Creatinine, Diabetes, Glucose, Thyroid-stimulating hormone test (TSH)], ultrasonography on liver and spleen along with Transient Elastography to measure liver stiffness (LS), and some variables such as AST-ALT ratio (AAR) and Fibrosis-4 index (FIB-4) were used to evaluate binary classification using six algorithms (Bayesian Network, Decision Tree, Naïve Bayes, Neural Networks, Random Forest and Support Vector Machine). The six feature selection methods are p-value + Feature Subset Selection (CFS), Information Gain, Principal Components Analysis, greedy stepwise, Genetic algorithm, and Particle Swarm Optimization. Among six algorithms the Bayesian Network showed the highest accuracy of 68.9%, followed SVM (67.8%). Among six feature selection methods and features, the p-value + CFS shows the 9 best-selected features such as Gender, Platelet, Albumin, Total Bilirubin, Baseline_PCR, Liver, Spleen, Stiffness, and prothrombin concentration respectively [10].

The previous studies showed that among different HCV genotypes, the 1 and 3 genotypes show predominantly in entire India, whereas 4 and 6 showed up in some parts of India respectively. The decision tree (DT) is used to classify the genotype a (1 to 6) and genotype 1b. On the other hand, the successful prediction of antiviral therapy in chronic Egyptian patients is analyzed by DT and the important predictor is found to be alpha-fetoprotein (AFP) level. Based on the aforementioned genotype and AFP, the new study is conducted in the Microbiology department of King George's Medical University, Lucknow. This hospital has an Integrated Counselling and Testing Centre (ICTC) facility and Anti-Retroviral Therapy (ART) Center. This hospital provides HIV testing, linkages for medical and psychosocial care and counseling for persons living with HIV infection. During the period January 2007 – July 2008, a consent form is obtained from a total of 350 HIV-infected adults attending the ICTC and ART Centre to enroll them in the study. Pre-designed proforma with 90 attributes were used to interview the subjects about the clinical symptoms. Thus the dataset of 350 observations with 90 attributes was used to evaluate the presence of HCV as a binary class label. The random forest from the R–package is used for the prediction of the model and the accuracy is found to be 98.3% [11].

The limitations of common regression analysis (output, $y = \sum wj.xj + \theta$, where inputs xj is multiplied by a weight wj, and at a constant θ) shows the performance of the candidate is controlled by many factors and thus not going to be a good regression model. These limitations can be overcome by an Artificial Neural Network (ANN) that matches the human brain in solving a problem. Thus, an attempt to use Artificial Neural Network is carried out for a dataset 216 patient's records with 19 attributes (Albumin, Alkaline Phosphatase, Anorexia, Antivirals, Ascites, Bilirubin, Fatigue, Histology, Liver Big, Liver Firm, Malaise, Patient's Age, Patient's Gender, Prothrombin Time, SGOT, Spiders, Spleen Palpable, Steroid, Varices), with 70-30% split of the total data and binary classification have been studied with an artificial neural network. The accuracy is found to be 98.44% [12].

In another study, the cutoff values of alanine aminotransferase/aspartate aminotransferase (AST/ALT) ratio, AST-to-platelet ratio (APRI), and fibrosis score (FIB-4) is used as a scoring system for common fibrosis in adults but not appropriate for children's. Thus there is a need to validate and introduce new cutoff values for pediatrics. So this study consists of the dataset used is 166 Egyptian children with the only HCV attending a pediatrics hospital outpatient. The dataset shows 90 cases belong to all types of fibrosis (F1 + F2 + F3 + F4 + F5) and 76 cases with no fibrosis (F0). The features considered in this study were age, ALT, AST, bilirubin, blood urea nitrogen (BUN), body mass index, cholesterol, gamma-glutamyl transpeptidase (GGT), hemoglobin, International normalized ratio, platelets, sex, steatosis, and white blood cell count. For the prediction of any type of fibrosis, the random forest showed accuracy and ROC of 87.5 % and 0.903 respectively, with bilirubin as a predictor. Similarly, for mild fibrosis, the RF shows 66% accuracy and 0.71 AUC with bilirubin, BUN, GGT and platelets as the most significant features. For the prediction of advanced fibrosis, random forest showed an accuracy (80%) and AUC (0.894) with only 3 features age, bilirubin, and platelets [13].

2.1 Background Study

Based on the aforementioned introduction and literature review sections, it is known that during the last few decades' researchers and clinicians are applying machine learning classification and feature selection algorithms for non-invasive methods such as clinical and biochemical data. To the best of the author's knowledge, the machine learning predictions for multi and binary class labels of the same dataset, and to compare the performance of Python and R tools with respective to 7 classification techniques have not been studied in the literature. In this paper, our objective is to identify the best model, the selected features which play a key role in the prediction of HCV disease, and to compare the performances of Python and R tools by using Hepatitis C Virus (HCV) from Egyptian patient's dataset from UCI in the content of multi and binary class labels.

3. DATA SOURCE

The publicly available UCI machine learning repository is used to retrieve the Hepatitis C Virus (HCV) for Egyptian patients Dataset available at [14]. The multivariate datatype consists of 1385 instances with 29 attributes. The multiclass label (i.e., Baseline histological staging) dataset instances consist of distinct values and frequencies, which are as follows F1 (portal fibrosis without septa, 336), F2 (few septa, 332), F3 (many septa without cirrhosis, 355) and F4 (cirrhosis, 362). The binary class label dataset consists of mild to moderate fibrosis as class label = 0 (F1 and F2) and advanced fibrosis as class label =1 (F3-F4) according to METAVIR score [9], resulting with a balanced dataset (Table 1). Previous studies show that the performance on a balanced dataset is better, this study's purpose is to take note of whether a similar classification performance can be observed between multi and binary class labels of the same dataset.

Table 1. Characteristic of HCV for Egyptian patients dataset used in this study

Multi Class Label	Instances	Binary Class Label	Instances
Portal Fibrosis, F1	336	Class = 0 (Mild to moderate, F1-F2)	668
Few Septa, F2	332		
Many septa, F3	355	Class =1 (Advanced cirrhosis, F3-F4)	717
Cirrhosis, F4	362		

3.1 Description of Attributes

The dataset consists of a total of 29 attributes, out of this 28 are input attributes and 1 is a key attribute referred to as "staging" with multiclass labels (F0-F4) as mentioned in the data source. The description of attributes are shown in Table 2.

Table 2. Attributes description in the HCV dataset

	Description		Description
1	Age	16	ALT 1 (1week) - Alanine transaminase ratio 1 week
2	Gender	17	ALT 4 (4week) - Alanine transaminase ratio 4 weeks
3	BMI (Body Mass Index)	18	ALT 12(12week) - Alanine transaminase ratio 12 weeks
4	Fever	19	ALT 24(24 week) - Alanine transaminase ratio 24 weeks
5	Nausea/Vomiting	20	ALT 36(36week) - Alanine transaminase ratio 36 weeks
6	Headache	21	ALT 48(48week) - Alanine transaminase ratio 48 weeks
7	Diarrhea	22	ALT (after 24week) – Alanine transaminase ratio after 24 weeks
8	Fatigue & generalized bone ache	23	RNA Base
9	Jaundice	24	RNA 4
10	Epigastric pain	25	RNA 12

11	WBC (White blood cells)	26	RNA EOT (RNA End-of-Treatment)
12	RBC (Red blood cells)	27	RNA EF (RNA Elongation Factor)
13	HGB (Hemoglobin)	28	Baseline histological Grading
14	Plat (Platelets)	29	Baseline histological staging (Class labels)
15	AST 1 - Aspartate transaminase ratio		

4. METHODOLOGY

4.1 Data Preprocessing

The missing values present in the dataset are replaced with mode value based on the age attribute. Due to the presence of different measuring units, the data normalization method (bringing values between 0 and 1) has been carried out in two machine learning tools such as Python [preprocessing.normalize(X)] and R [(normalize <- function(x) {return ((x - min (x)) / (max(x) - min(x)))}] to rescale the values of the variables [15,16]. The equation for normalization is

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

where $X_{norm} = i$-th normalized data, $X = (X_1, X_2, \cdots\cdots\cdots, X_n)$, min and max represents the minimum and maximum value of the range respectively.

4.2 Data Analysis

The dataset used to build classification model consists of 1385 instances with 29 attributes (Table 1). A total of seven machine learning (ML) techniques followed by different feature selection methods were used to evaluate the performance of the classifiers on aforementioned dataset. The analytical tools such as Python-based Scikit learn (version 0.21, Spyder) and R-based CARET package (version 3.6.2) were used to perform data analysis.

In Spyder (Python IDE)-based scikit-learn package, before to classification model building, the pre-processing steps such as data normalization, total dataset split into 70–30% as a training and testing data respectively with set.seed (7) was used. The classifiers with default parameters and 10 fold cross-validation (method = "cv", number = 10) were used to evaluate the performance of the classifiers on the aforementioned dataset. Among seven classification models used, the five are Gaussian Naive Bayes (GNB), KNeighborsClassifier (KNN), multilayer perceptron Classifier (NN, with parameters such as solver='sgd', hidden_layer_sizes= (10, 10), activation='relu'), Random Forest (RF) and SVC (for SVM). Adaboost (Boosting) and Bagging are used as an ensemble classifiers.

Feature selections methods (FS) with libraries and parameters used are as follows, correlation matrix (library “pandas” and “seaborn”), Rank Feature Importance method with Extra Classifier Tree model (RFI-ECT), recursive feature elimination method with Logistic Regression model (RFE-LR), and SelectKBest (score_func=chi2, k=5) are addressed using Python-based scikit-learn package [17].

In R-caret package (https://cran.r-project.org/web/packages/), before classification model building, the pre-processing steps such as data normalization, total dataset split into 70–30% as a training and testing data respectively with set.seed (123) was used. The five classification algorithms with library and parameters are as follows K-Nearest Neighbor (KNN, library “caret”, method = 'knn', tuneLength = 10), Naïve Bayes (NB, library “e1071”, method = ‘naïve_bayes’), Neural Network (NN, library “nnet”, method = 'nnet', trace= FALSE), Random Forest (RF, library “randomForest”, method= ‘rf’, ntree=500, importance=TRUE) and Support Vector Machine (SVM, library “caret”, method = 'svmLinear', tuneLength =10). The two ensemble classifiers with library and parameters are as follows, Adabag (Bagging, library “adabag”, method = ‘adabag’) and boosting (library “gbm”, method= ‘gbm’, verbose=FALSE). The 10 fold cross-validation (method = "cv", number = 10) was used to evaluate the performance of the classifiers on the aforementioned training and testing data.

Feature selections methods with libraries and parameters are as follows, correlation matrix (library “mlbench”), variable importance estimations such as regression method (VImp-Reg, library “mlbench”) and random forest (VImp-RF, library “randomForest”, nodesize=25, ntree=200, importance = TRUE) respectively, rank feature by Importance (library “class” and “mlbench”) with SVMpoly model (RFI-SVMpoly) and vector quantization model (RFI-LVQ, library “class” and “mlbench”) respectively, and recursive feature elimination method with random forest algorithm ((RFE-RF, library “mlbench”, functions=rfFuncs, method="cv", number=10) and finally a Principal Component Analysis (PCA, library “factoextra”) with prcomp function are addressed using R tool [18].

The model performance on test data is calculated by an evaluation metrics such as an accuracy, precision and recall (macro average) in both multi and binary class label datasets. The overall average scores of evaluation metrics is taken to compare Python and R tools performance. The description of the evaluation metrics are given as:

Accuracy: It is the performance measure ratio of correctly predicted observation of the total observations [19].

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

Precision: It is the ratio of correctly predicted positive observations of the total predicted positive observations [20].

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall: It can be also called sensitivity or true positive rate (TPR). It is the ratio of correctly predicted positive observations to all observations in an actual class [20].

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

where TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative.

5. RESULTS

5.1 Multi and binary class label performance with 29 features/ complete dataset

To the best of author’s knowledge, most of the classification model studies to the HCV dataset have been carried out with binary classification. Thus, comparison studies between multi and binary class labels of the same dataset with the seven machine learning techniques using 2 different tools were not addressed. This study aimed to understand the performances of classifiers and tools on multi and binary class labels of the same HCV datasets. To measure the performance of each classification algorithm the accuracy has been taken into accordance.

In Python, the multiclass dataset shows the highest accuracy (28.36%) in a random forest, followed by KNN with 26.44%. The similar accuracy performance of KNN accuracy has been observed in NN, but when coming to precision and recall the KNN shows a better performance. Similarly, the binary dataset shows the highest accuracy (53.12%) in NN with the exemption to precision in comparison with SVM where the accuracy shows 52.64%. The individual performances of each classifier to accuracy, precision, and recall in binary class label almost shows the double score of multiclass label respectively (Table 3A).

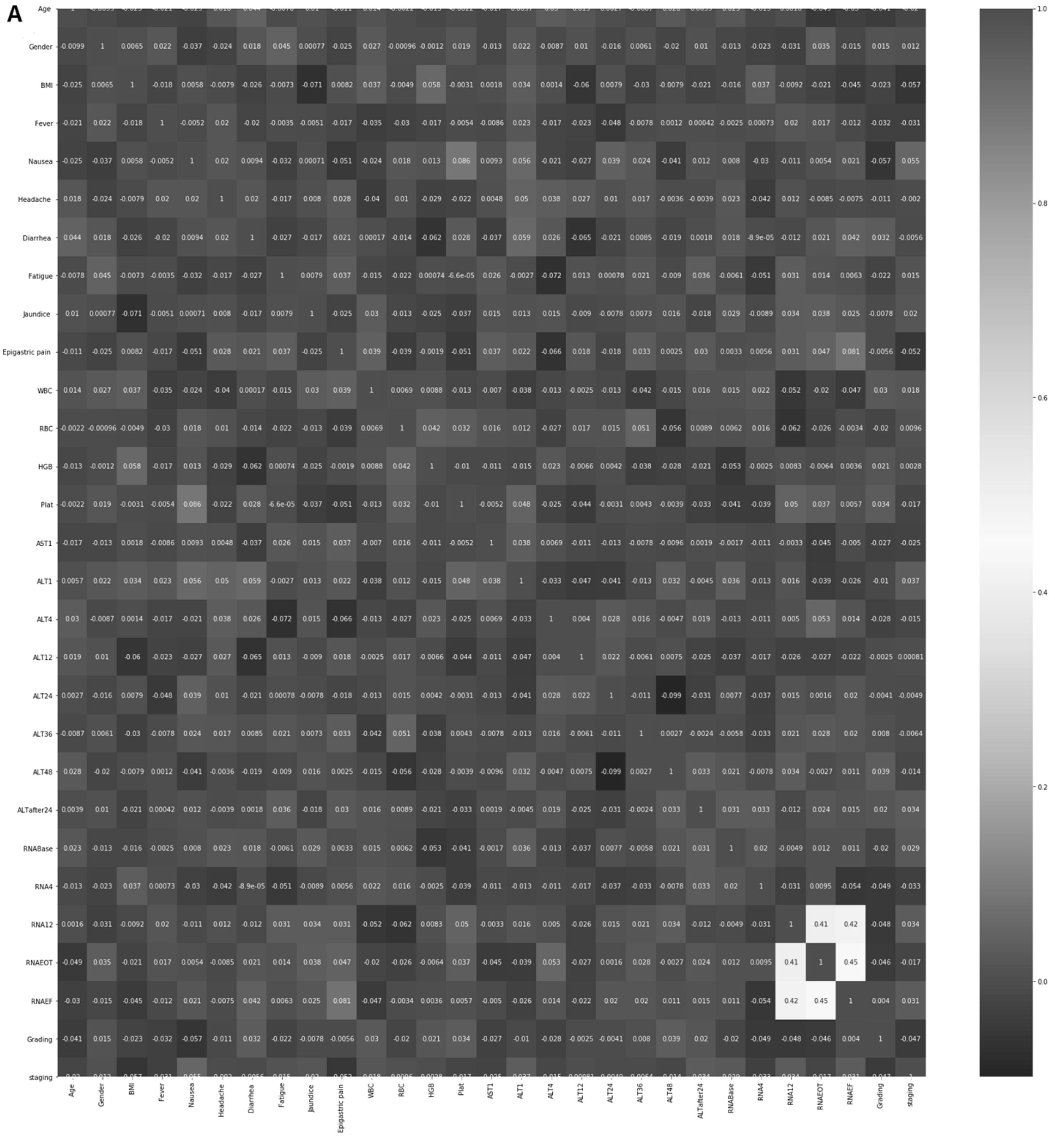
On the other hand, the R multiclass dataset shows the highest accuracies with similar performances in SVM and RF (51.31%) followed by KNN (50.83%), and NB (50.65%). In the binary class label, boosting shows the highest accuracy (54.23%), followed by KNN (53.06%). Similar performance of accuracy also been noticed in NN and Bagging (51.73%) respectively. The accuracies of multiclass and binary class labels show similar performance, but whereas precision and recall show different performances (Table 3B).

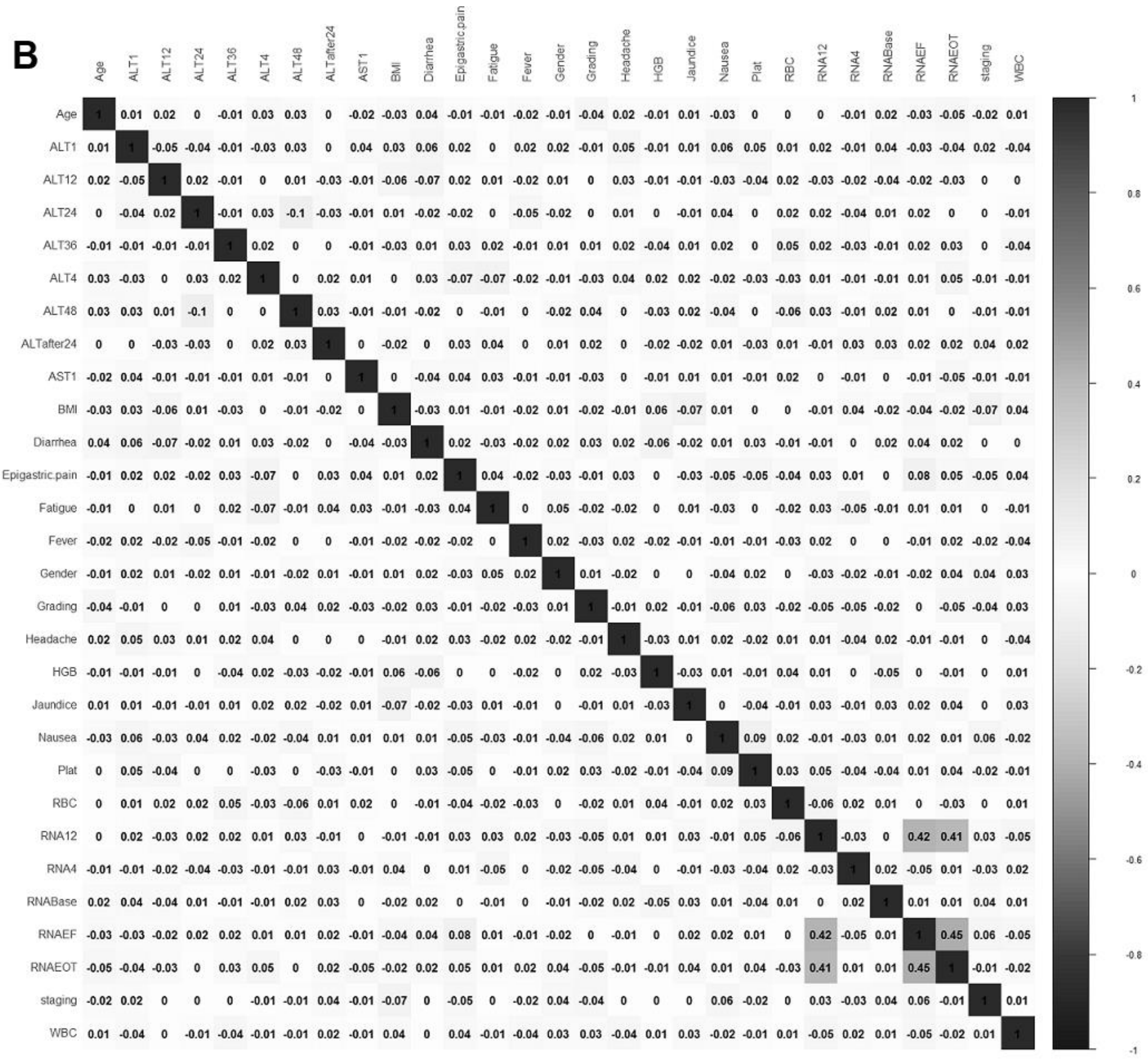
Table 3. The performance comparisons of multi and binary class labels with 29 attributes/features, (A) Python and (B) R - ML techniques

3A : Python	**29 Features**					
	4 Class Labels			**2 Class Labels**		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
KNN	26.44	27.00	27.00	47.35	47.00	47.00
SVM	21.15	22.00	22.00	52.64	53.00	53.00
RF	28.36	30.00	29.00	49.15	50.00	50.00
GNB	24.03	24.00	23.00	52.16	51.00	51.00
NN	26.44	7.00	25.00	53.12	27.00	50.00
Bagging	23.07	24.00	24.00	46.63	47.00	47.00
Boosting (Adaboost)	24.15	24.00	24.00	50.00	50.00	50.00
Overall average	24.80	22.57	24.85	50.15	46.42	49.71
3B : R- Prog						
KNN	50.83	25.78	26.22	53.06	51.61	48.00
SVM	51.31	26.70	26.95	48.54	46.63	45.00
RF	51.31	26.70	26.95	48.54	46.63	45.00
GNB	50.65	25.45	25.94	48.49	46.52	43.50
NN	48.43	11.76	22.67	51.73	50.29	43.00
Bagging (Adabag)	50.13	23.24	25.20	51.73	50.29	43.00
Boosting	48.20	21.95	22.28	54.23	53.03	48.00
Overall average	50.12	23.08	25.17	50.90	49.28	45.07

5.2 Feature Selection

It is known a subset of relevant features could be useful to improve the model performance. In view of this, the correlation analysis with Python and R was performed. The dataset did not show any correlated attributes with a given correlation cutoff at 0.60 as a positive strong correlation [21]. However, in both analytical tools (Figures 1A and 1B) the positive moderate correlation is found between RNA12, RNAEOT, RNAEF (0.41– 0.45) attributes.


A
Age
Gender
BMI
Fever
Nausea
Headache
Diarrhea
Fatigue
Jaundice
Epigastric pain
WBC
RBC
HGB
Plat
AST1
ALT1
ALT4
ALT12
ALT24
ALT36
ALT48
ALTafter24
RNABase
RNA4
RNA12
RNAEOT
RNAEF
Grading
staging
1.0
0.8
0.6
0.4
0.2
0.0



	Age	ALT1	ALT12	ALT24	ALT36	ALT4	ALT48	ALTafter24	AST1	BMI	Diarrhea	Epigastric.pain	Fatigue	Fever	Gender	Grading	Headache	HGB	Jaundice	Nausea	Plat	RBC	RNA12	RNA4	RNABase	RNAEF	RNAEOT	staging	WBC
Age	1	0.01	0.02	0	-0.01	0.03	0.03	0	-0.02	-0.03	0.04	-0.01	-0.01	-0.02	-0.01	-0.04	0.02	-0.01	0.01	-0.03	0	0	0	-0.01	0.02	-0.03	-0.05	-0.02	0.01
ALT1	0.01	1	-0.05	-0.04	-0.01	-0.03	0.03	0	0.04	0.03	0.06	0.02	0	0.02	0.02	-0.01	0.05	-0.01	0.01	0.06	0.05	0.01	0.02	-0.01	0.04	-0.03	-0.04	0.02	-0.04
ALT12	0.02	-0.05	1	0.02	-0.01	0	0.01	-0.03	-0.01	-0.06	-0.07	0.02	0.01	-0.02	0.01	0	0.03	-0.01	-0.01	-0.03	-0.04	0.02	-0.03	-0.02	-0.04	-0.02	-0.03	0	0
ALT24	0	-0.04	0.02	1	-0.01	0.03	-0.1	-0.03	-0.01	0.01	-0.02	-0.02	0	-0.05	-0.02	0	0.01	0	-0.01	0.04	0	0.02	0.02	-0.04	0.01	0.02	0	0	-0.01
ALT36	-0.01	-0.01	-0.01	-0.01	1	0.02	0	0	-0.01	-0.03	0.01	0.03	0.02	-0.01	0.01	0.01	0.02	-0.04	0.01	0.02	0	0.05	0.02	-0.03	-0.01	0.02	0.03	0	-0.04
ALT4	0.03	-0.03	0	0.03	0.02	1	0	0.02	0.01	0	0.03	-0.07	-0.07	-0.02	-0.01	-0.03	0.04	0.02	0.02	-0.02	-0.03	-0.03	0.01	-0.01	-0.01	0.01	0.05	-0.01	-0.01
ALT48	0.03	0.03	0.01	-0.1	0	0	1	0.03	-0.01	-0.01	-0.02	0	-0.01	0	-0.02	0.04	0	-0.03	0.02	-0.04	0	-0.06	0.03	-0.01	0.02	0.01	0	-0.01	-0.01
ALTafter24	0	0	-0.03	-0.03	0	0.02	0.03	1	0	-0.02	0	0.03	0.04	0	0.01	0.02	0	-0.02	-0.02	0.01	-0.03	0.01	-0.01	0.03	0.03	0.02	0.02	0.04	0.02
AST1	-0.02	0.04	-0.01	-0.01	-0.01	0.01	-0.01	0	1	0	-0.04	0.04	0.03	-0.01	-0.01	-0.03	0	-0.01	0.01	0.01	-0.01	0.02	0	-0.01	0	-0.01	-0.05	-0.01	-0.01
BMI	-0.03	0.03	-0.06	0.01	-0.03	0	-0.01	-0.02	0	1	-0.03	0.01	-0.01	-0.02	0.01	-0.02	-0.01	0.06	-0.07	0.01	0	0	-0.01	0.04	-0.02	-0.04	-0.02	-0.07	0.04
Diarrhea	0.04	0.06	-0.07	-0.02	0.01	0.03	-0.02	0	-0.04	-0.03	1	0.02	-0.03	-0.02	0.02	0.03	0.02	-0.06	-0.02	0.01	0.03	-0.01	-0.01	0	0.02	0.04	0.02	0	0
Epigastric.pain	-0.01	0.02	0.02	-0.02	0.03	-0.07	0	0.03	0.04	0.01	0.02	1	0.04	-0.02	-0.03	-0.01	0.03	0	-0.03	-0.05	-0.05	-0.04	0.03	0.01	0	0.08	0.05	-0.05	0.04
Fatigue	-0.01	0	0.01	0	0.02	-0.07	-0.01	0.04	0.03	-0.01	-0.03	0.04	1	0	0.05	-0.02	-0.02	0	0.01	-0.03	0	-0.02	0.03	-0.05	-0.01	0.01	0.01	0	-0.01
Fever	-0.02	0.02	-0.02	-0.05	-0.01	-0.02	0	0	-0.01	-0.02	-0.02	-0.02	0	1	0.02	-0.03	0.02	-0.02	-0.01	-0.01	-0.01	-0.03	0.02	0	0	-0.01	0.02	-0.02	-0.04
Gender	-0.01	0.02	0.01	-0.02	0.01	-0.01	-0.02	0.01	-0.01	0.01	0.02	-0.03	0.05	0.02	1	0.01	-0.02	0	0	-0.04	0.02	0	-0.03	-0.02	-0.01	-0.02	0.04	0.04	0.03
Grading	-0.04	-0.01	0	0	0.01	-0.03	0.04	0.02	-0.03	-0.02	0.03	-0.01	-0.02	-0.03	0.01	1	-0.01	0.02	-0.01	-0.06	0.03	-0.02	-0.05	-0.05	-0.02	0	-0.05	-0.04	0.03
Headache	0.02	0.05	0.03	0.01	0.02	0.04	0	0	0	-0.01	0.02	0.03	-0.02	0.02	-0.02	-0.01	1	-0.03	0.01	0.02	-0.02	0.01	0.01	-0.04	0.02	-0.01	-0.01	0	-0.04
HGB	-0.01	-0.01	-0.01	0	-0.04	0.02	-0.03	-0.02	-0.01	0.06	-0.06	0	0	-0.02	0	0.02	-0.03	1	-0.03	0.01	-0.01	0.04	0.01	0	-0.05	0	-0.01	0	0.01
Jaundice	0.01	0.01	-0.01	-0.01	0.01	0.02	0.02	-0.02	0.01	-0.07	-0.02	-0.03	0.01	-0.01	0	-0.01	0.01	-0.03	1	0	-0.04	-0.01	0.03	-0.01	0.03	0.02	0.04	0	0.03
Nausea	-0.03	0.06	-0.03	0.04	0.02	-0.02	-0.04	0.01	0.01	0.01	0.01	-0.05	-0.03	-0.01	-0.04	-0.06	0.02	0.01	0	1	0.09	0.02	-0.01	-0.03	0.01	0.02	0.01	0.06	-0.02
Plat	0	0.05	-0.04	0	0	-0.03	0	-0.03	-0.01	0	0.03	-0.05	0	-0.01	0.02	0.03	-0.02	-0.01	-0.04	0.09	1	0.03	0.05	-0.04	-0.04	0.01	0.04	-0.02	-0.01
RBC	0	0.01	0.02	0.02	0.05	-0.03	-0.06	0.01	0.02	0	-0.01	-0.04	-0.02	-0.03	0	-0.02	0.01	0.04	-0.01	0.02	0.03	1	-0.06	0.02	0.01	0	-0.03	0	0.01
RNA12	0	0.02	-0.03	0.02	0.02	0.01	0.03	-0.01	0	-0.01	-0.01	0.03	0.03	0.02	-0.03	-0.05	0.01	0.01	0.03	-0.01	0.05	-0.06	1	-0.03	0	0.42	0.41	0.03	-0.05
RNA4	-0.01	-0.01	-0.02	-0.04	-0.03	-0.01	-0.01	0.03	-0.01	0.04	0	0.01	-0.05	0	-0.02	-0.05	-0.04	0	-0.01	-0.03	-0.04	0.02	-0.03	1	0.02	-0.05	0.01	-0.03	0.02
RNABase	0.02	0.04	-0.04	0.01	-0.01	-0.01	0.02	0.03	0	-0.02	0.02	0	-0.01	0	-0.01	-0.02	0.02	-0.05	0.03	0.01	-0.04	0.01	0	0.02	1	0.01	0.01	0.04	0.01
RNAEF	-0.03	-0.03	-0.02	0.02	0.02	0.01	0.01	0.02	-0.01	-0.04	0.04	0.08	0.01	-0.01	-0.02	0	-0.01	0	0.02	0.02	0.01	0	0.42	-0.05	0.01	1	0.45	0.06	-0.05
RNAEOT	-0.05	-0.04	-0.03	0	0.03	0.05	0	0.02	-0.05	-0.02	0.02	0.05	0.01	0.02	0.04	-0.05	-0.01	-0.01	0.04	0.01	0.04	-0.03	0.41	0.01	0.01	0.45	1	-0.01	-0.02
staging	-0.02	0.02	0	0	0	-0.01	-0.01	0.04	-0.01	-0.07	0	-0.05	0	-0.02	0.04	-0.04	0	0	0	0.06	-0.02	0	0.03	-0.03	0.04	0.06	-0.01	1	0.01
WBC	0.01	-0.04	0	-0.01	-0.04	-0.01	-0.01	0.02	-0.01	0.04	0	0.04	-0.01	-0.04	0.03	0.03	-0.04	0.01	0.03	-0.02	-0.01	0.01	-0.05	0.02	0.01	-0.05	-0.02	0.01	1

Figure 1. Correlation plot with 29 attributes (complete features). (A) Plot of Python, (B) Plot of R.

Based on aforementioned correlation data, it is not possible to remove any attributes as a highly correlated feature. To further evaluate, additional feature selected methods from Python and R have been analyzed (see Section 4.2). The similar ranking/topology order of features has not been noticed between different FS methods (Table 4). Thus, indicating furthermore feature selection methods need to be considered, so finally, we opted one such method "principal component analysis".

Table 4. The different feature selection methods and their best selected features from Python and R packages

Python	**Best selected features**
RFI - ECT	staging, Plat, RBC, ALT4, ALT24, HGB, ALT48, ALTafter24, ALT12, AST1, RNA4, WBC, ALT36, BMI, RNAEOT, RNABase, RNAEF, Grading, ALT1, Age, RNA12, Gender
SelectkBest	RNAEF, RNABase, RNA12, RNA4, RNAEOT, RBC, Plat, WBC, staging, ALT1, AST1, ALT48, ALT4, BMI, ALT36, Age, ALT24, Grading, ALTafter24, ALT12, Gender, Epigastric pain
RFE - LR	(WBC, RNAEF, Plat), RNAEOT, RNA12, RBC, RNABase, AST1, ALT48, ALT4, staging, ALT1, RNA4, Age, ALT36, BMI, ALT12, Grading, ALT24, ALTafter24, HGB, Gender, Epigastric pain, Fever, Nausea, Jaundice, Headache, Fatigue, Diarrhea
R	**Best selected features**
VImp- Reg	BMI, RNAEF, Nausea, Epigastric pain, RNAEOT, Grading, Gender, ALTafter24,RNABase, Age, RNA4, Plat, RNA12, Fever, ALT1, WBC, AST1,

	ALT48, ALT24, Jaundice, ALT4, RBC, Fatigue, ALT12, HGB, Headache, ALT36, Diarrhea
VImp-RF	RNA4, RNABase, ALT1, RBC, ALT36, ALT48, AST1, RNAEF, WBC, Plat, ALT24, ALT12, Age, ALT4, ALTafter24, RNA12, BMI, RNAEOT, Grading, HGB, Nausea, Fever, Jaundice, Epigastric pain, Gender, Diarrhea, Fatigue, Headache
RFI- SVM Poly	RNA12, RNA4, RNAEF, BMI, Plat, Nausea, Epigastric pain, Grading, RNABase, WBC, RNAEOT, ALT1, ALTafter24, RBC, Fever, AST1, Jaundice, Age, ALT4, Fatigue
RFI- LQV	BMI, RNAEF, Age, ALTafter24, RNABase, RNA4, Grading, Gender, Nausea, Plat, RNA12, AST1, ALT1, ALT48, Fever, ALT4, Jaundice, Epigastric pain, RBC, WBC
RFE - RF	Epigastric pain, ALT48, HGB, Gender, RNA4, RNAEF, Headache, RBC, ALTafter24, Jaundice, RNA12, ALT4, Fatigue, Fever, Nausea, Grading, Age, ALT24, ALT1, BMI, Diarrhea, AST1, ALT36, RNAEOT, RNABase, Plat, ALT12, WBC

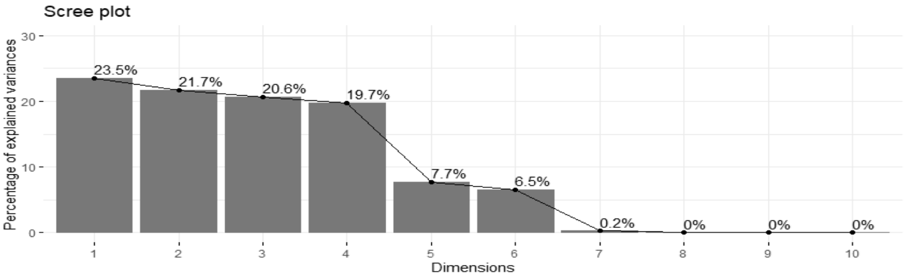


Figure 2. The graphical representation of the percentage of variation retained by each principal component

Table 5. Attributes contribution per each principal component

Dimensions	Contributed Attributes
PC 1	RNAEF, RNAEOT, RNA12
PC 2	Nausea, Plat, Epigastric pain, ALT1, ALT48, RBC, staging, ALT24, WBC, RNA4
PC 3	HGB, ALT1, ALT24, ALT48, Diarrhea, RNABase, BMI, Headache
PC 4	BMI, ALT12, Plat, Jaundice, ALT4, ALT1, ALT24, Headache, Age , staging

The principal component analysis is useful to extract important information, thus reducing the dimensionality of multivariate data to two or three principal components. The minimal loss of information can be visualized graphically. The variation of the features is based on normalized eigenvalue associated with each eigenvector. In this study, the 1st and 2nd principal components (i.e., PC1 and PC2) attributes showed 45.2% variation, 3rd and 4th principal components (i.e., PC3 and PC4) showed 40.3% variation, the first three principal components (i.e., PC1-3) showed 65.8% of the variation. The total of four principal components (PC1-PC4) showed 85.5% of the variation (Figure 2). The features information is enclosed by each principal component (PC), thus the contribution of attributes for each PC has been shown in Table 5.

Since the 1st and 2nd principal components (i.e., PC1 and PC2) attributes showed 45.2% variation which is half of the four PC's (i.e., PC1-PC4, 85.5% variation, Figure 2) and the RNAEOT and RNA12 are positive moderate correlated attributes (Figure 1A and 1B). Furthermore, the consideration of PC1-2 attributes with PC3-4 (i.e., PC1-4, 21 features) is to know the performance of PC1-2 with other attributes of PC3-4 respectively. So we considered 12 selected features (i.e., PC1-2) and 21 selected features for comparison between PC1-2 and PC3-4 (Table 6). A total of 7 attributes are removed (i.e., gender, fever, fatigue and generalized bone ache, AST1, ALT36, ALT after 24 weeks, Baseline histological grading) by this method.

Table 6. Two feature selected datasets

Components	Number	Feature Selected attributes
PC 1 - 2	12	RNAEF , RNAEOT, RNA12, Nausea, Plat, Epigastric pain, ALT1, ALT48, RBC, ALT24, WBC, RNA4
PC 1 - 4	21	RNAEF, RNAEOT, RNA12, Nausea, Plat, Epigastric pain, ALT1, ALT48, RBC, ALT24, WBC, RNA4, HGB, Diarrhea, RNABase ,BMI ,Headache, ALT12, Jaundice, ALT4, Age

5.3 Multi and Binary Class Label Performances on 12 Selected Features

The 12 selected features were taken into consideration to build a model. The highest accuracy in Python - multiclass dataset with a similar performance of the classifiers has been observed in SVM and Boosting (25.96%) respectively. The similar performance of the classifiers has been observed in KNN and RF (25%). Similarly, in a binary dataset, the highest accuracy has been observed in RF (54.56%), and the same goes for precision and recall, followed by NN (53.1%) where precision shows 27% is shown in Table 7A.

In R-multi and binary class label datasets (12 attributes, Table 7B), the highest accuracy is shown by KNN (51.06% and 52.97%) followed by NB (50.57%, 51.56%) respectively. In both class labels, similar performances of accuracy are shown by SVM and RF (49.86 and 50.85%) is shown in Table 7B.

Table 7. The performance comparisons of multi and binary class labels with 12 feature selected attributes. (A) Python, (B) R-ML techniques

7A : Python	**12 Attributes**					
	4 Class Labels			**2 Class Labels**		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
KNN	25.00	26.00	26.00	47.83	48.00	48.00
SVM	25.96	27.00	26.00	48.07	47.00	47.00
RF	25.00	26.00	26.00	54.56	55.00	55.00
NB	21.00	21.00	21.00	52.10	52.00	52.00
NN	21.39	5.00	25.00	53.10	27.00	50.00
Bagging	24.75	25.00	25.00	47.11	48.00	48.00
Boosting (Adaboost)	25.96	25.00	25.00	50.00	50.00	50.00
Overall average	24.15	22.14	24.85	50.39	46.71	50.00
7B : R-prog						
KNN	51.06	26.81	26.56	52.97	51.23	52.00
SVM	49.86	24.63	24.75	50.85	49.17	44.50
RF	49.86	24.63	24.75	50.85	49.17	44.50
NB	50.57	25.93	25.82	51.56	50.00	45.00
NN	48.75	17.50	23.11	50.94	50.00	27.00
Bagging (Adabag)	49.76	18.42	24.62	49.82	47.94	35.00
Boosting	50.38	25.33	25.57	47.19	44.88	39.50
Overall average	50.03	23.32	25.02	50.59	48.91	41.07

5.4 Multi and Binary Class Label Performances on 21 Selected Features

The 21 selected features were taken into consideration to build a model. The highest accuracy in the ‘Python-multiclass dataset has been observed in KNN (26.44%), followed by RF (26.20%) with similar performances in precision and recall (27%). Similarly, in the binary dataset, the highest accuracy has been observed in SVM (52.64%), and the same goes for precision and recall, followed by NB (51.68%) is shown in Table 8A.

In the R- four class label (21 attributes), the highest accuracy is shown by KNN (50.70%) followed by similar performances of SVM and RF (50.34%) respectively. Similarly, in R binary class label, the highest accuracy is shown in bagging (52.20), followed by KNN (51.23%), the similar performances have been seen in SVM and RF (49.28) is shown in Table 8B.

Table 8. The performance comparisons of multi and binary class labels with 21 feature selected attributes. (A) Python, (B) R-ML techniques

8A : Python	**21 Attributes**					
	4 Class Labels			**2 Class Labels**		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
KNN	26.44	27.00	27.00	47.35	47.00	47.00
SVM	24.51	25.00	25.00	52.64	53.00	53.00
RF	26.20	27.00	27.00	50.72	51.00	51.00
NB	24.03	24.00	24.00	51.68	51.00	51.00
NN	26.44	7.00	25.00	46.87	23.00	50.00
Bagging	21.39	21.00	22.00	51.20	52.00	52.00
Boosting (Adaboost)	22.11	22.00	22.00	50.24	50.00	50.00
Overall average	24.44	21.85	24.57	50.10	46.71	50.57
8B : R-prog						
KNN	50.70	25.90	26.04	51.23	49.70	42.00
SVM	50.34	25.26	25.50	49.28	47.30	39.50
RF	50.34	25.26	25.50	49.28	47.30	39.50
NB	49.52	23.60	24.26	47.69	45.50	40.50
NN	48.43	11.76	22.67	50.68	50.98	13.00

Bagging (Adabag)	49.37	22.25	24.05	52.20	51.40	36.50
Boosting	49.71	24.54	24.53	46.63	44.61	43.50
Overall average	49.77	22.65	24.65	49.57	48.11	36.35

Our results show significant differences of accuracies for a multi-class label and insignificance for binary class labels between Python and R tools respectively, among 3 datasets (Table 3, 7 and 8). The highest accuracy (51.06%) is shown by KNN for the multiclass label (R), similarly, for binary class label, the highest accuracy (54.5%) is shown by random forest (Python). Our analysis did not show significance with previous studies this could be due to the nature of sampling, nature of dataset and the selected features (Table 9).

Table 9. Comparing the results of this paper with other predictive methods

Classifier	Instances	Attributes	Selected Features	Accuracy (%)	Reference
Decision tree	39,567	21	age, AST, albumin, and platelet count	84.00	8
Bayesian Network	4962	24	Gender, Platelet, Albumin, Total Bilirubin, Baseline_PCR, Liver, Spleen, Stiffness, and prothrombin concentration	68.90	9
Random Forest	350	90	–	98.30	10
Artificial neural network	216	19	–	98.44	11
Decision tree	859	13	–	84.21	12
Random Forest	166	14	Age, Bilirubin, BUN, GGT, Platelets	66.00 - 87.50	13
KNN (Multi class label)	1385	29	RNAEF , RNAEOT, RNA12, Nausea, Platelets, Epigastric pain, ALT1, ALT48, RBC, ALT24, WBC, RNA4	51.06	Our study (R-prog)
Random forest (Binary class label)	1385	29		54.56	Our study (Python)

5.5 Comparison of Python and R Classifiers in 3 Datasets

In the comparison of Python classifiers between both class labels, the approximate rank order of the classifiers is SVM, RF, KNN, and NN. Similarly, in R, the approximate order is KNN, SVM/RF and NB.

In a comparison of tools, the multi-class dataset overall average performances of precision and recall to 3 datasets (29, 12 and 21 features) in Python show a range of 21.85 – 24.85%, and R with a range of 22.65 – 25.02%. However, the performance in overall average accuracy shows a different scenario where Python shows 24.15 – 24.8% and R shows 49.77 – 50.12% indicating the R tool has shown better accuracy performance. Similarly, binary dataset overall average performances of accuracy, precision, and recall to 3 datasets (29, 12 and 21 features) in Python with a range of 46.42 – 50.57% and R with a range of 36.35 – 50.90%, indicating the performance of both tools agree (Figures 3A and 3B).

On the other hand, the within the comparison of 3 datasets features, the 12 selected features overall average accuracy, precision, and recall shows similar performances to 29 and 21 selected features in both class labels and tools. Thus, indicating 12 selected features can be useful for the prediction of the HCV dataset. Moreover the binary dataset shows similar performances either in Python or R show (Figures 3A and 3B).

6. DISCUSSION

In this study, the multi and binary datasets performances of the HCV dataset have been evaluated with classification and features selection algorithms implemented in Scikit learn package of Python and R-CARET package respectively. The RF with an accuracy of 54.56% in binary classification shows a significant difference with the previous studies classifier (i.e., RF with an accuracy of 66-87.5, 98.3%) [10, 13]. On the other hand, the KNN shows an accuracy of 51.06% in a multi-class label. To the best of author knowledge, this is the first study where multi and binary class labels of the same dataset were evaluated with two analytical tools. The 3 multiclass datasets showed similar results with overall average precision and recall scores in both tools. However, this scenario is different from accuracy where R shows better performance than Python. On the other hand, average performances of evaluation metrics (accuracy, precision, and recall) in binary datasets show insignificance differences to both analytical tools. Based on the aforementioned performances R-ML techniques show better performance. This could be due to the implementation of different packages. In a comparison of three datasets (with 29, 12 and 21 features) and both tools, the 12 selected features show similar performances with the remaining two datasets, indicating these 12 features could play a role in better model performances. On the other hand, the 12 selected features in this study are of new variables when compared with previous studies [8-9, 13]. However, it is noticed that some of the general symptoms of Hepatitis C are fever, fatigue, generalized bone ache are not in the list of selected features, apart from this Jaundice is in the least selected features the one reason could be aforementioned symptoms could be at a stage of early infection unless one go for blood test for further confirmation of the disease.

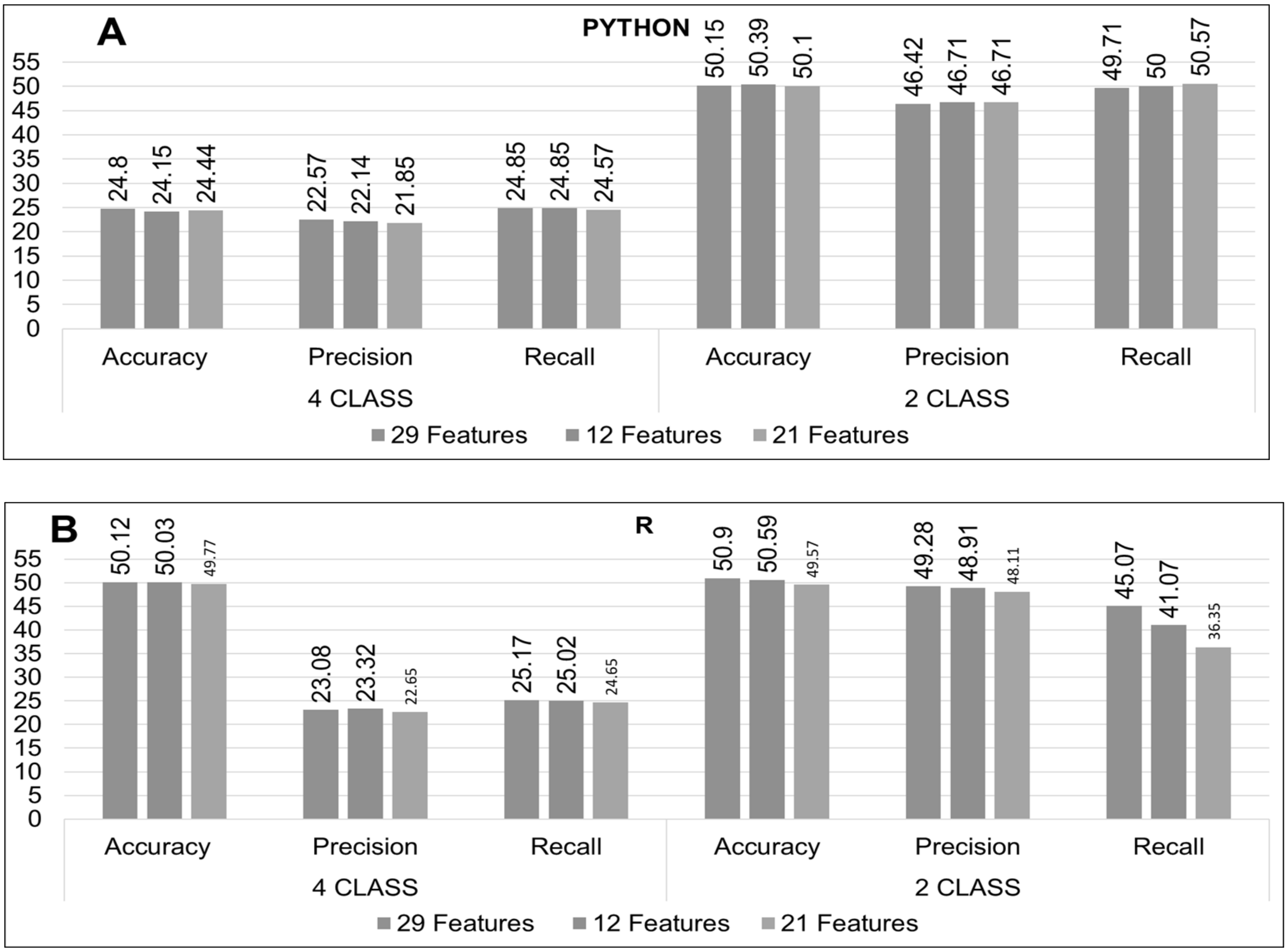


Figure 3. The average accuracy, precision and recall comparison with 3 datasets in multi and binary datasets. (A) Python, (B) R-ML

6. DISCUSSION

In this study, the multi and binary datasets performances of the HCV dataset have been evaluated with classification and features selection algorithms implemented in Scikit learn package of Python and R-CARET package respectively. The RF with an accuracy of 54.56% in binary classification shows a significant difference with the previous studies classifier (i.e., RF with an accuracy of 66-87.5, 98.3%) [10, 13]. On the other hand, the KNN shows an accuracy of 51.06% in a multi-class label. To the best of author knowledge, this is the first study where multi and binary class labels of the same dataset were evaluated with two analytical tools. The 3 multiclass datasets showed similar results with overall average precision and recall scores in both tools. However, this scenario is different from accuracy where R shows better performance than Python. On the other hand, average performances of evaluation metrics (accuracy, precision, and recall) in binary datasets show insignificance differences to both analytical tools. Based on the aforementioned performances R-ML techniques show better performance. This could be due to the implementation of different packages. In a comparison of three datasets (with 29, 12 and 21 features) and both tools, the 12 selected features show similar performances with the remaining two datasets, indicating these 12 features could play a role in better model performances. On the other hand, the 12 selected features in this study are of new variables when compared with previous studies [8-9, 13]. However, it is noticed that some of the general symptoms of Hepatitis C are fever, fatigue, generalized bone ache are not in the list of selected features, apart from this Jaundice is in the least selected features the one reason could be aforementioned symptoms could be at a stage of early infection unless one go for blood test for further confirmation of the disease.

7. CONCLUSION

This research shows the evaluation of the HCV by Egyptian patient's dataset to multi and binary class labels, performance comparison of Python and R tools, and the 12 selected features for better model building. Despite class labels and tools used, average performances of evaluation metrics of the classifiers are in the order of SVM, RF, KNN, NN, and NB. Similar performances have been shown by both analytical tools to the individual class labels for 3 datasets. The less variation in performance differences between 12 selected features and remaining datasets (29 and 21 selected features) indicates the selected features can be useful for the prediction of the HCV dataset. However, it is noticed that some attributes which are referred to as general symptoms are not selected by feature selection methods. Apart from this, it is necessary to keep track of these reduced features while performing classification modeling since no single feature selection method has shown the consistent topology of selected features. Based on this study, it is known that analytical techniques and tools for multi and binary class labels play a role in model classification accuracies along with nature of dataset and selected features.

ACKNOWLEDGEMENTS

This work was supported by the Universiti Sains Malaysia, Short Term Grants [304/PKOMP/6315170].

CONFLICTS OF INTEREST

There is no conflict of interest.

REFERENCES

[1] A. Elgharably, A. I. Gomaa, M. M. Crossey, P. J. Norsworthy, I. Waked and S. D. Taylor-Robinson, Hepatitis C in Egypt - past, present, and future, *International Journal of General Medicine*, 10, 2017, 1-6.

[2] A. M. Vladimir and L. Sylvie, Hepatitis C virus: Morphogenesis, infection and therapy, *World Journal of Hepatology*, 10(2), 2018, 186-212.

[3] C. W. Spearman, G. M. Dusheiko, M. Hellard and M. Sonderup, Hepatitis C, *Lancet*, 394(10207), 2019, 1451-1466.

[4] D. Omran, M. Alboraie, R. A. Zayed, M. N. Wifi, M. Naguib, M. Eltabbakh, M. Abdellah, A. F. Sherief, S. Maklad, H. H. Eldemellawy, O. K. Saad, D. M. Khamiss and M. E. Kassas, Towards hepatitis C virus elimination: Egyptian experience, achievements and limitations, *World Journal of Gastroenterology*, 24(38), 2018, 4330-4340.

[5] X. Li, H. Xu, P. Gao, Fibrosis index based on 4 factors (fib-4) predicts liver cirrhosis and hepatocellular carcinoma in chronic Hepatitis C virus (HCV) patients, *Medical Science Monitor*, 25, 2019, 7243-7250.

[6] J. L. Horsley-Silva and H. E. Vargas, New therapies for hepatitis C virus infection, *Gastroenterology and Hepatology*, 13(1), 2017, 22-31.

[7] V. Palanisamy and R. Thirunavukarasu, Implications of big data analytics in developing healthcare frameworks – A review, *Journal of King Saud University–Computer and Information Sciences*, 31(4), 2019, 415-425.

[8] N. Satish Chandra Reddy, S. N. Song, Z. M. Lim and C. Xin Ying, Classification and feature selection approaches by machine learning techniques: Heart disease prediction, *International Journal of Innovative Computing*, 9(1), 2019, 39-46.

[9] S. Hashem, G. Esmat, W. Elakel and H. Shahira, Comparison of machine learning approaches for prediction of advanced liver fibrosis in chronic hepatitis C patients, *IEEE/ACM Trans Computational Biology and Bioinformatics*, 15(3), 2018, 861-868.

[10] S. M. El-Salam, M. M. Ezz, S. Hashem, W. Elakel, R. M. Salama, H. Elmakhzangy and M. ElHefnawi, Performance of machine learning approaches on prediction of esophageal varices for Egyptian chronic hepatitis C patients, *Informatics in Medicine Unlocked*, 17, 2019, 1-7.

[11] G. G. Agarwal, A. K. Singh, V. Venkatesh and N. Wal, Determination of risk factors for hepatitis C by the method of random forest. *Annal of Infectious Disease and Epidemiology*, 4(1), 2019, 1-4.

[12] N. Metwally, E. AbuSharekh and S. Abu-Naser, Diagnosis of hepatitis virus using artificial neural network, *International Journal for Academic Development*, 2, 2018, 1-7.

[13] 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*, 25(3), 2019, 173-181.

[14] https://archive.ics.uci.edu/ml/datasets/Hepatitis+C+Virus+%28HCV%29+for+Egyptian+patients

[15] https://scikit-learn.org/stable/modules/preprocessing.html#normalization

[16] https://www.rdocumentation.org/packages/SciencesPo/versions/1.3.5/topics/normalize

[17] Scikit-learn, “Scikit-learn: Machine Learning in Python,” 2016.

[18] http://topepo.github.io/caret/index.html.

[19] https://en.wikipedia.org/wiki/Confusion_matrix

[20] https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.

[21] C. Ray and A. Ray, Intrapartum cardiotocography and its correlation with umbilical cord blood pH in term pregnancies: a prospective study, *International Journal of Reproduction, Contraception, Obstetrics and Gynecology*, 6, 2017, 2745-2752.