Hepatitis C Virus Prediction Based on Machine Learning Framework: a Real-world Case Study in Egypt

Haba Mamdouh (heba.mamdouh@mu.edu.eg)
Minia University Faculty of Science
Mahmoud Yasin Shams
Kafr el-Sheikh University: Kafrelsheikh University
Tarek Abd El-Hafeez
Minia University Faculty of Science https://orcid.org/0000-0003-1785-1058

Research Article

Keywords: Machine learning, Classification, Feature Selection, Hepatitis C Virus

Posted Date: January 31st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1292024/v1

Hepatitis C Virus Prediction based on Machine Learning Framework: A Real-World Case Study in Egypt

Heba Mamdouh Farghaly[a*] [0000-0002-9672-5115], **Mahmoud Y. Shams**[b][0000-0003-3021-5902] , **and Tarek Abd El-Hafeez**[a,c][0000-0003-1785-1058]

[a]*Department of Computer Science, Faculty of Science, Minia University, EL-Minia, Egypt.*
[b]*Faculty of Artificial Intelligence, Kafrelsheikh University, Egypt*
[c]*Computer Science Unit, Deraya University, EL-Minia, Egypt.*

* Corresponding author.
E-mail address: heba.mamdouh@mu.edu.eg

Abstract:

Prediction and classification of the diseases are essential in medical science, as it attempts to immune the spread of the disease and discover the infected regions from the early stages. Machine Learning (ML) approaches are commonly used for predicting and classifying the diseases that precisely utilized as an efficient tool for the doctors and specialists. To forecast Hepatitis C Virus (HCV) among Egyptian healthcare workers (HCWs), a prediction system based on machine learning methodologies is developed. We used data from the National Liver Institute (NLI), which was formed at Menoufiya University (Menoufiya, Egypt). The dataset includes 859 patients with 12 distinct characteristics. To test the proposed framework's robustness and dependability, we ran two scenarios, one without feature selection and the other with feature selection based on Sequential Forward Selection (SFS). In addition, a feature subset evaluation based on SFS-generated features is carried out. Induction algorithms and classifiers used for model evaluation include Nave Bayes (NB), Random Forest (RF), K Nearest Neighbor (KNN), and Logistic Regression (LR). Then, the effect of parameter tuning on learning techniques is measured. The experimental results indicated that the proposed framework achieved higher accuracies after SFS selection than without feature selection. Moreover, the RF classifier achieved 94.06% accuracy with minimum learning elapsed time 0.54 sec. Finally, after adjusting the hyperparameter values of the RF classifier, the classification accuracy is improved to 94.88% using only four features.

Keywords: Machine learning; Classification; Feature Selection; Hepatitis C Virus.

1. Introduction

2. Hepatitis C Virus (HCV) is a worldwide illness that affects the human population. It's a blood-borne infection that spreads by direct contact with contaminated blood or blood-containing bodily fluids. Hepatitis C is a worldwide illness, according to the World Health Organization (WHO). According to the WHO, around 58 million individuals have chronic HCV infection, with approximately 1.5 million new infections happening each year [1]. When compared to wealthy nations in Europe and North America, poor developing countries in Asia and Africa have the highest frequency of this virus. Furthermore, the number of persons with chronic illnesses is larger in nations like Pakistan, China, and Egypt [2], [3].

The incidence of HCV among Egyptian health-care personnel is poorly understood (HCWs). HCWs are routinely exposed to a variety of biological hazards while doing their duties, and they are closely monitored. As a result, healthcare workers in Egypt who come into intimate contact with patients are at a higher risk of contracting HCV and other blood-borne viruses [4].

As a result, we are in desperate need of a non-invasive HCV diagnosis that is both accurate and dependable. By capturing complicated and nonlinear interactions in clinical data, machine learning (ML) algorithms are especially good at interpreting medical occurrences. To construct a model to diagnose HCV and identify patients who have been infected with the virus, ML methods such as classification approaches can be used. Incorrect attributes in the attribute set can degrade the classifier's performance [5]. Feature selection defines a subset of features or variables that characterize data in order to produce a more compact and essential representation for the given data while ignoring any other characteristics that are redundant and useless [6]. Feature selection is an effective approach to improve the performance of a classifier while also cutting down on model construction time.

The main objective of this paper is to build ML framework (MHF_HCV) for diagnosing HCV disease among HCWs in Egypt. To predict HCV, the MHF_HCV framework used the classification and feature selection approach on a real world HCV dataset. The accuracy of well-known machine learning algorithms for HCV diagnosis and classification was first evaluated. A hybrid prediction model was then created and deployed using sequential forward selection (SFS)-based wrapper feature selection and classification. The effect of changing parameters on learning techniques is then determined. Finally, we examined the performance of different classifiers in terms of accuracy, recall, precision, and F1-scores.

The main influence of this study can be summarized as follows:

- An effective model is designed to predict HCV as accurately as possible.
- The performance of well-known classifiers is measured using all features from of real world HCV dataset for HCWs in Egypt.
- The impact of data splitting on ML models' performance is measured.
- The best attributes are selected using SFS based wrapper feature selection algorithm.
- The performance of classifiers is measured on all features as well as selected features from of real world HCV dataset.
- Tweaking the hyper-parameters improves the classifier's performance.
- The best feature set and classifier, as well as its customizable hyper-parameters, are returned for precise HCV classification.
- Expert decision-making might be aided by the proposed approach, which could be employed and integrated into real-world applications.

This paper is organized as follows: Section 2 presents the motivation and problem formulation, while the details discussion of the related work is investigated in Section 3. Section 4 explains our proposed method and Section 5 demonstrates the

experimental results. Section 6 presents the discussion of the results. Finally, Section 7 presents the conclusion and future work.

3. Motivation and Problem Formulation

Sharps injuries, needles, and scalpels are a high-risk population for HCWs when performing their health-care tasks. As a result of caring for patients afflicted with HCV, HCWs are at risk of infection. This has prompted academics and researchers to develop a methodology for predicting HCV illness in HCWs at an early stage.

The National Liver Institute (NLI), based at Menoufiya University in Menoufiya, Egypt, provided the HCV dataset for HCWs, which was used in the proposed study to construct the ML framework. The HCV dataset D, along with output class Y, consists of feature set $X = \{X_1; X_2; \_ \_ \_; X_n\}$ with n features and the instances $J = \{J_1; J_2; \_ \_ \_; J_m\}$ corresponding to m subjects (cases).

Definition 1 "A dataset D is composition of instances $J = \{J_k \mid 1 \leq k < m\}$, where m is the total number of subjects (cases) and feature set $X = \{X_i \mid 1 \leq i < n\}$, where n is the number of features".

Definition 2 "An instance J_k is represented by feature values X_i, such as $J_k = \{X_i\ 1 \leq i < n\}$ and n is the number of features in D. The value X_i is either categorical or numeric".

As shown in Equation 1, the purpose of building a framework for accurate predictions (PR) using a learning algorithm (C) is to make the predictive model learn to fit by evaluating data behaviour and converges by lowering the error (ER) present in all instances (J).

$$P \rightarrow \left[ER\left\{Y - PR\left[C\left(\sum_{k=1}^{m}\sum_{i=1}^{n} D(X_i, J_k)\right)\right]\right\}\right] \quad (1)$$

The prognoses The accuracy, precision, recall, and F1-scores are used to evaluate the performance (P) of PR from C. Section 4.6 focuses much deeper into the measurements.

4. Related Work

The current studies proves that 20 percent of people with viral hepatitis "C" develop symptoms of influenza-like symptoms of the disease, while 80 percent of other people with the disease do not feel any symptoms, but the disease infection remain [7], [8]. Therefore, it is important to undergo frequent tests to ensure that the body is free of infection, especially if the patient is receiving medications through needles, which help transmit the disease. Egypt recorded the highest prevalence of hepatitis C in the world, and this epidemic is expected to reach its peak soon [9]. According to a 2010 study, an estimated more than half a million people contract the virus for the first time each year. While the estimates of the Egyptian Ministry of Health and Population indicated that the number of cases of HIV infection annually is 100,000 people. Studies have shown that the rate of infection with HCV in Egypt is the highest in the world, as it is ten times higher than in Europe and America [10]–[13].

Nandipati et al. [14] presents HCV prediction model using ML approaches based on Random Forest (RF) and k-Nearest Neighbors Algorithm (KNN) classifiers for the applied dataset (HCV) founded in UCI-ML repository [15]. They utilized 668 instances with mild to moderate class 0, and cirrhosis with class 1. They used Python and R for programming with different number of features and attributes. The combination of features can be utilized more efficiently based on ML to provide improved insights into antibody sequences which affected the HCV response [16]. They used ML approaches to predict the clinical groups that combine the features to identify the most significant features using RF classification.

A framework that integrate the data mining with Decision Tree (DT) and Fuzzy logic is presented by Ali et al. [17] to manage and predict the HCV cases. They utilized the Trapezoidal Fuzzy Number (TFN) that achieves 98.1% compared with 92.5% predication results by DT.

The laboratory analysis study of 4962 HCV patients in Egypt between 2006 and 2017 based on ML were performed by Abd El-Salam et al. [18]. They used 24 clinical laboratory variables and the results investigated that 2218 patient infected with Esophageal Varices and not present in 2,744 patients. Their model used six well-known classifiers including Neural Networks (NN), Naïve Bayes (NB), DT, Support Vector Machine (SVM), RF and Bayesian Network (BN). They utilized data collected from the Egyptian National Committee to Combat Viral Hepatitis in the national treatment program for patients with viral hepatitis in Egypt under the supervision of the Ministry of Health. The accuracies obtained are 67.8%, 66.3%, 67.2%, 65.6%, 66.7%, and 68.9% using Support Vector Machine (SVM), RF, C4.5, Multi-Layer Perceptron (MLP), NB, and BN, respectively.

Hashem et al. [19] presents ML approaches to predict Hepatocellular Carcinoma with HCV-related Chronic Liver Disease. They present a set of input variables that are filtered to get the optimal variable subset based on LR, DT and Classification and Regression Tree (CART). The accuracies abstained were 96%, 99%, and 95.5% using LR, DT, and CART, respectively.

The prediction of HCV virus results from viral nucleotides using several combination of ML approaches are presented by KayvanJoo et al. [20]. They used DT, SVM, NB, and NN to predict the Interferon-alpha (IFN-alpha) and ribavirin (RBV) therapy respond based on processed features. They produce 10 attribute weighting models from overall 76 attributes for the initial dataset. These eleven attributes include Chi-square, Gini index, Deviation, Info-Gain, Info-Gain Ratio, SVM, PCA, Uncertainty, Relief, and Rule. The eleven attributes are then classified based on SVM, NB, NN, and DT and the average accuracy is 85%. The summery of the most recent methodologies and efforts for predicting the HCV cases are investigated in Table 1. The table refers to the authors, the utilized dataset, number of instances or patients or collected cases, the feature selections that refer to which the authors used feature selection or not. Further, the authors ML methodologies, and the resulting performance metrics.

Table 1. Summery of the related HCV work

Author	Dataset Utilized	No. of Instances	Feature Selection	Methodology	Performance Metrics %	
Nandipati et al. [14]	HCV UCL-ML Repository	668	Yes 4-Multi feature selection methods	**KNN** **SVM** **RF** **NB** **NN** **Bagging** **Boosting**	Average Accuracy	50.59
					Precision	48.19
					Recall	41.07
Ali et al. [17]	Laboratory examinations HCV data in Egypt 2008–2012	200	No	**DT**	Accuracy	92.50
				TFN	Accuracy	98.10
Abd El-Salam et al. [18]	Egyptian National Committee under the supervision of the Ministry of Health	4962	Yes Filter Warper	**RF**	Accuracy	66.30
				C4.5	Accuracy	67.20
				MLP	Accuracy	65.60
				NB	Accuracy	66.70
				BN	Accuracy	68.90
Hashem et al. [19]	Egyptian National Committee for the Control of Viral Hepatitis Kasr Al-Aini Hospital	4423	Yes Variable Selection	**LR**	Accuracy	96.00
				DT	Accuracy	99.00
				CART	Accuracy	95.50

From the mentioned studies, we can conclude that the efforts made to predict viral hepatitis especially in Egypt are the most interesting and required to address the spread of the virus. The related work focused only on patients in general and not dealing with hospital staff in mind. Moreover, the accuracy based on known traditional classifiers must be improved. The diversity of the database used and the number of instances generated is not a fair comparison between the results obtained. Therefore, in this study, we used a data set collected from real working hospital patients to predict HBV infection incidence from registered cases. We also identified the most relevant features using the SFS feature selection algorithm that achieved accuracy compared to the accuracy obtained before feature selection.

5. Methodology

MHF_HCV is our recommended method for distinguishing between HCV-infected and non-infected people. In Egypt, we concentrated on strategies for improving the accuracy of machine learning (ML) classification algorithms for HCV prediction among HCWs. The classifiers were evaluated on all attributes and chosen features to compare the accuracy they achieved. An SFS-based wrapper feature selection strategy was used to locate the ideal feature subset that affects the class discovery process and increases classification accuracy in order to identify the key features and enhance the efficiency of the classification process. Popular ML classifiers were utilized to classify these chosen features into multiple classes using popular ML classifiers. The methodology of the proposed system is structured into six stages which include: (1) Data gathering; (2) Data pre-processing; (3) Data splitting; (4) Feature selection; (5) ML classifiers; (6) classifier's performance evaluation. Figure 1 shows framework for predicting the HCV disease.

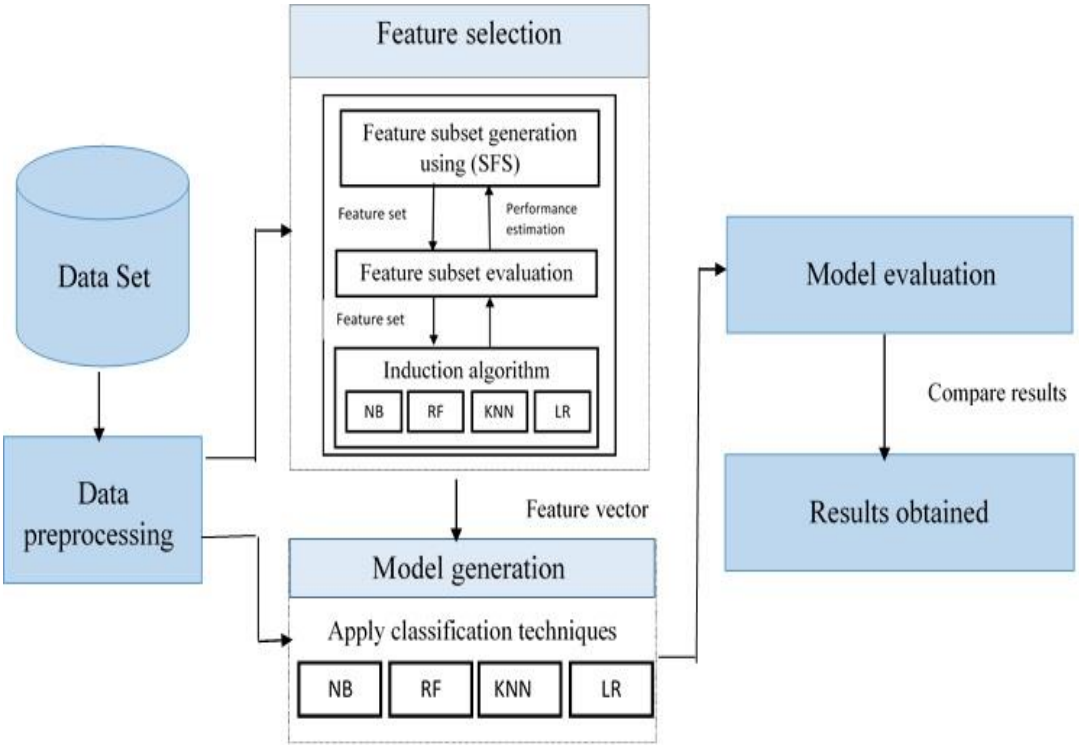


Figure 1. Framework for predicting HCV disease.

5.1. Description of the Dataset

In this study, we used a real world dataset of hepatitis C prevalence among HCWs in Egypt, where there is the highest prevalence of HCV [2]. This dataset was obtained from The NLI institute, founded at Menoufiya University (Menoufiya, Egypt). The dataset consists of 859 records (patients) with 12 features, which present test information of each patient.11 attributes of these features are taken as diagnosis inputs, whereas the 'HCV_PCR' attribute is selected as output. Table 2 shows a brief description of the HCV dataset used in our study.

Table 2. Description of HCV dataset

Feat. #	Feature	Feature Description	Feature Type	Feature Data Range
1	**Gender**	Sex	Binary	0 :female; 1: male.
2	**Residence**	Place of residence.	Binary	1: urban residence; 2: rural residence.
3	**Job**	Occupational category for HCW.	Nominal	1: Doctor (Faculty member); 2: Doctor (Student concession) 3: Surgeon(Faculty member); 4: Surgeon (Student concession); 5:Dentist; 6: Medical student; 7: Nursing supervisor; 8: Nurse; 9: Nursing student; 10: Laboratory Technician; 11: worker; 12: paramedic; 13: Laundry worker; 14: other.
4	**Schisto**	Schistosomiasis infection.	Binary	0: no; 1: yes.
5	**Dealing with syring**	In the three months leading up to their enrolling, they had handled syringes.	Nominal	0: No deal; 1: dealing once; 3: dealing twice; 4: dealing from 3 to 5 times; 5: dealing more than 5 times.
6	**ALT**	Serum alanine aminotransferase, this enzyme test is measured to see if the liver is damaged or diseased.	Numeric	Normal ranges Female: 32, Male: 42.
7	**Neadlestick**	History of neadlestick injury in the year prior to enrolment.	Nominal	0: No needling; 1:needling once; 3: needling twice; 4: needling from 3 to 5 times; 5: needling more than 5 times.
8	**HCV_ELISA**	Antibody to hepatitis C is detected using a blood test. When you are infected with HCV, your body creates this antibody.	Binary	0: negative; 1: positive.
9	**AST**	An aspartate aminotransferase, this enzyme test is measured to Check for liver damage and Check on the success of treatment for liver disease.	Numeric	Normal ranges Female: 32; Male: 42.
10	**HBsAg_ELISA**	Hepatitis B Virus Surface Antigen, this test identifies active infection by the hepatitis B virus.	Binary	0: negative; 1: positive.
11	**Age**	Age in Year	Numeric	[21:64]
12	**HCV_PCR**	Test used to determine whether the HCV exists in your bloodstream.	Binary	1: for patients infected with HCV; 0: for non-infected patients.

5.2. Data Preprocessing

The initial phase in the proposed system is data preparation, which involves removing noisy values and replacing missing values for specific characteristics. It is expected that missing, inconsistent, and duplicate data have been handled in the experimental dataset. As demonstrated in Table 2, the majority of the medical characteristics were converted to categorical data from numerical data.

5.3. Data Splitting

This section presents details explanation of the different splitting algorithms used in our study.

5.3.1. K-fold cross-validation

In machine learning, a cross-validation algorithm is a popular technique. The basic goal of cross-validation is to obtain a reliable and stable model performance estimate. Data is divided into k separate pieces for K-fold cross-validation. Each iteration uses k-1 portions to train the model and the rest as a validation set. According to the amount of folds, the procedure is iterated. The average of the calculated scores represents the model's generalization ability [21].

5.3.2. Train-Test Split

Train-test splits dataset into a random train and test subsets. This method depends on the size of the dataset [22].

5.4. Feature Selection

The main purpose of the feature selection process is to select a subset of features from the original features that maximize classification accuracy [23]. Because of its simplicity and empirical performance, an SFS algorithm based on wrapper selection strategy for feature selection was developed in this work. To measure the excellent features subset, the wrapper model technique employs a classifier as an induction mechanism [24]. Wrapper methods generally achieve better accuracy rate and use cross-validation to avoid over-fitting. Here, 10-fold cross-validation is used on the training set for classification to evaluate the selected feature. SFS is the simplest greedy search algorithm [25]. It is a bottom-up search method that starts with an empty set of features and sequentially adds features that are selected based on some evaluation function which reduces the mean square error rate.

5.5. Classification Process

This work use classification algorithms to appropriately assign a class to an unseen record. Furthermore, the well-known classifiers NB, RF, KNN, and LR are used to explore the contributions of the SFS method's chosen features on classification accuracy.

5.5.1. **Naïve Bayes (NB)**

NB [26], [27] is a method that is widely used for classification and is particularly suitable when the input dimensionality is high. Despite its simplicity, NB can oftentimes outperform more complex classification techniques. It measures the probability of each input feature (attribute) for a predictable state.

To compute the posterior probability for each class c_i, the Bayesian classifier employs Bayes' rule. The NB is based on the simple premise that the characteristics, y, are independent of the class, hence the probability may be determined by multiplying the conditional probabilities of each feature by the class. So, the posterior probability, $P(C_i|y)$, is expressed as:

$$P(C_i|y) = P(y|C_i)\ P(C_i)/P(y). \quad (2)$$

where

$$P(y) = \sum_j P(y|C_j)\ P(C_j). \quad (3)$$

where

$P(C_i)$: the Apriori likelihood of class C_i.

$P(y)$: the likelihood density for feature y.

$P(y|C_i)$: the class-conditional likelihood density of the feature y that belongs to the C_i class.

$P(C_i|y)$: the posterior probability of the C_i class when observing y.

5.5.2. **Random Forests (RF)**

The RF [28] is defined as an ensemble learning method for classification and regression. Ensemble learning techniques (such as boosting, bagging, and RF) have got a great interest since they are robust to noise and more accurate than single classifiers. RF is a collection of tree structure classifiers. Each tree is trained with a subset of the training data that are randomly selected (i.e. bootstrapped), with the same distribution of samples for all the trees in the forest, and the final classification is then built based on the majority voting of trees in the forest. In other words, RF tries to build several DT with initial variables and various data samples and then combine predictions in order to make the final decision.

The following equation is used to forecast the class label c of a case x by majority voting for an RF with N trees:

$$l(x) = argmax_c \left(\sum_{n=1}^{N} I_{h_n(X)=c} \right). \quad (4)$$

where h_n is the nth tree of the RF and I is the indicator function.

5.5.3. **Logistic Regression (LR)**

The LR [29] is a linear model used for classification problems. LR measures the relationship among the response (dependent) variable and one or more explanatory (independent) variables for a given dataset that indicates the significance and strength of the impact of the explanatory variables on the response variable. The response

variable is a class label that we are trying to predict. However, the explanatory variables are the features or attributes that are used to predict the class label. The output of LR is the probability that given input points belong to a certain class. Typically, LR estimates probabilities using the logistic function, also known as the sigmoid function, which is given by [30]:

$$f(y) = \frac{L}{1 + e^{-k(y-y_0)}}. \quad (5)$$

where e is the natural logarithm base, L is the curve's maximum value, y_0 is the y-value of the sigmoid's midpoint, and k is the logistic growth rate or steepness of the curve.

5.5.4. **K-Nearest Neighbor (KNN)**

The KNN classifier is an instance-based non-parametric classifier [31]. This approach is based on estimating the nearest neighbor. The distance metric is used to classify the new instances based on their similarity measure. The K in KNN stands for the number of closest neighbor data values. The KNN model's primary premise is that a new instance's prediction is created by scanning the complete training set for comparable K neighbor examples and categorizing based on the class of highest occurrences. The Euclidean distance formula is used to find a comparable case.

Euclidean distance is the square root of the sum of squared differences between the new instance (A_i) and the existing instance (B_j) [32].

$$Euclidean_{i,j} = \sqrt{\sum_{k=1}^{n} \left(A_{ik} - B_{jk}\right)^2} \quad (6)$$

5.6. Performance Evaluation

The accuracy of the classification, precision, recall, and F1-scores can all be used to assess the effectiveness of the suggested technique. True positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) are all included in this "confusion matrix" [33], [34].

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (7)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (8)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (9)$$

$$\text{F1} - \text{score} = 2 * \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (10)$$

6. Experimental Results and Analysis

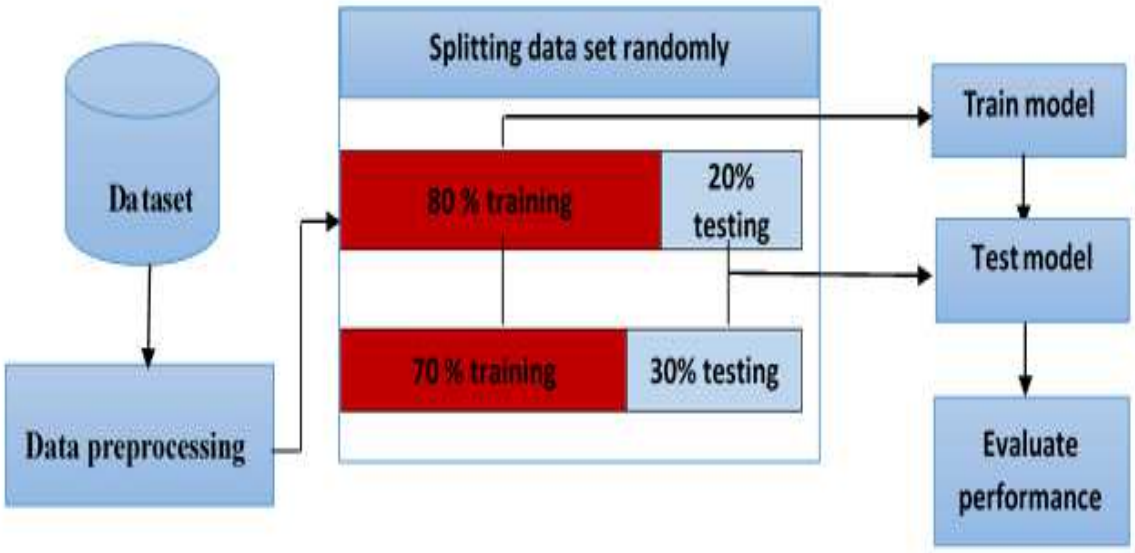
In this section, we have conducted experiments to assess the performance of the ML framework (MHF_HCV) for HCV disease diagnosis among HCWs in Egypt. As mentioned before, the real world data for HCV used for MHF_HCV framework

construction. We conducting our experiments on a 3 GHz i5 computer with a 4GB main memory and 64-bit Windows 7 operating system. The experiment is carried out using the python programming language.

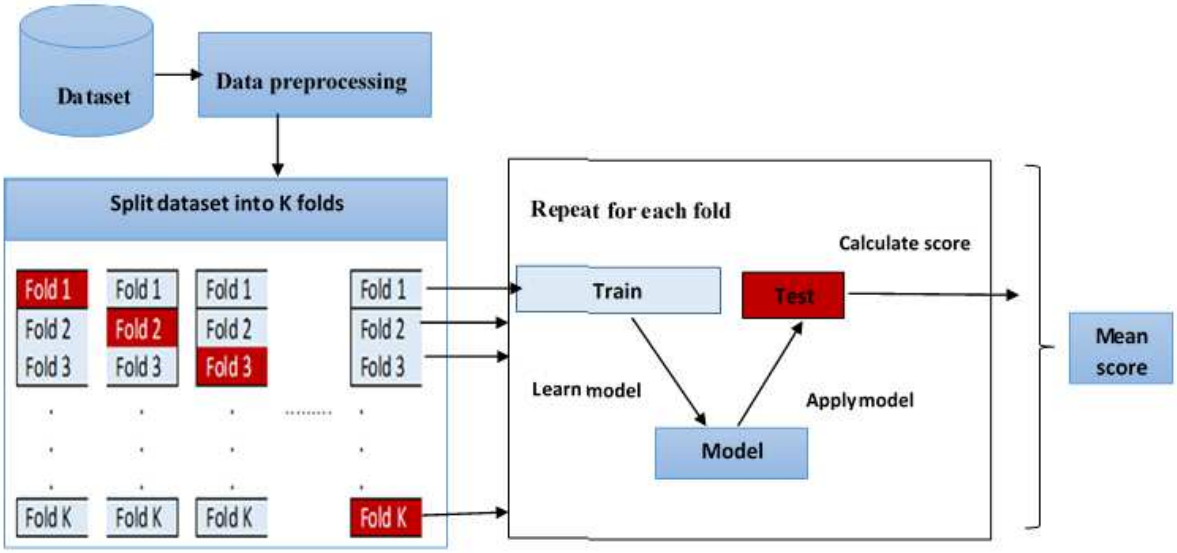
Initially, the focus of the first part of this section is on measuring the effect of using different data splitting techniques on the performance of classification techniques. While in the second part we focus on applying wrapper feature selection method and evaluating the quality of the selected features by conducting a set of experiments. Finally, in the third part we quantify the impact of parameter tuning on learning techniques.

6.1. Testing Different Data Splitting Methods

The performance ML model depends significantly on the quality of data and the strategy of using the data [35]. As a result, determining the impact of data splitting on the performance of ML models is extremely important, as it will pave the way for improved ML-based modeling by allowing for the selection of an appropriate data splitting strategy. On real-world HCV data, we used all attributes to compare alternative data partitioning approaches. K-fold cross-validation and the Train-test splits method were used to split the data in this investigation. Using the random splitting method, the dataset was divided into two parts, with different ratios: 70: 30 and 60: 40 train/test split as shown in Figure 2a. While performing k-fold cross-validation using k = 5 and k = 10 as shown in Figure 2b.



(a)



(b)

Figure 2. The data splitting impact on the applied classifiers
(a) The dataset divided into two parts randomly with different ratios: 70: 30 and 60: 40 train/test split. (b) Performing k-fold cross-validation using k = 5 and k = 10.

Table 3 and Figure 3 show the performance results for all classifiers with using two data partitioning methods.

Table 3. Comparison of different classifiers using two data partitioning methods in the term of accuracy.

Classifier	Data splitting approach			
	Train-test splits		K-fold cross validation	
	70:30	**60:40**	**k=5**	**k=10**
NB	91.47	92.15	92.08	**92.66**
RF	91.47	93.31	93.13	**94.06**
KNN	90.31	89.24	89.75	**90.8**
LR	91.86	91.27	**93.01**	**92.2**

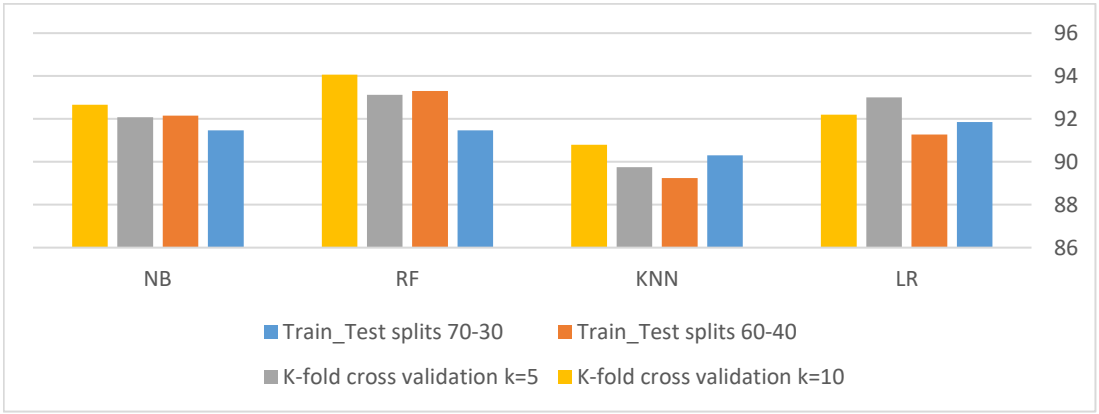


Figure 3. The accuracy of the classifier while using various data splitting strategies.

When applying classification techniques using the k-fold cross-validation (k = 10), their performance is better in most cases as shown in Table 3. AS we can also observe that the best performances occur when using 10-fold cross-validation with the RF classifier. Therefor In this study, 10-fold cross-validation method for splitting the HCV dataset was found as the best option for ML modeling as shown in Figure 3.

6.2. Influence of Feature Selection on the Performance of the ML Models

For the HCV dataset, we used a well-known classification approach, such as NB, RF, KNN, and LR, to see how feature selection affects classifier effectiveness. Figure 4 shows two situations for HCV prediction using classification techniques: 1) without using the suggested feature selection approach and 2) with using the SFS feature selection method. The performance of classification approaches was then measured using the evaluation metrics mentioned in Section 4.6. For each classification approach, the default settings were used. To identify the optimal feature subset, the experiments used a 10 cross-validation approach.

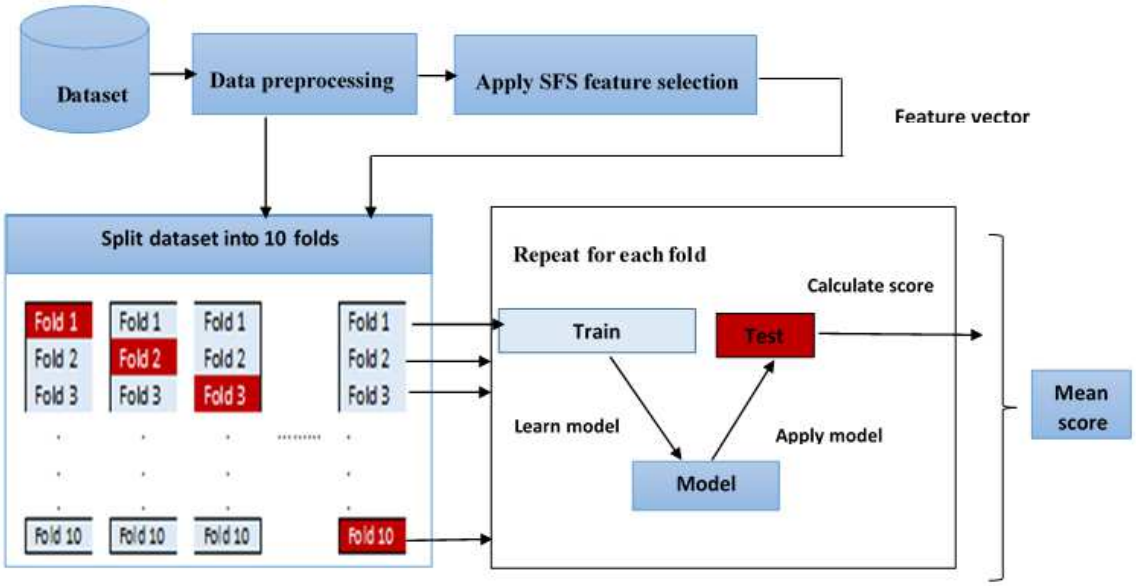


Figure 4. The impact of feature selection on classifier efficiency.

As previously stated, the strategy of selecting wrapper features is dependent on the classification model. SFS is a feature selection strategy that uses a variety of classifiers, including NB, RF, KNN, and LR. Table 4 illustrates how many characteristics were chosen when SFS was wrapped with various classifiers.

Table 4. The subset of features selected using SFS method wrapped with NB, RF, KNN, and LR classifiers.

Classifier	NO. of features	Selected features
SFS+NB	5	['Residence', 'age', 'HCV_ELISA', 'AST', 'Schisto']
SFS+RF	5	['HCV_ELISA', 'HBsAg_ELISA', 'Job', 'dealing with syring', 'neadlestick']
SFS+KNN	5	['age', 'HCV_ELISA', 'HBsAg_ELISA', 'gender', 'dealing with syring']
SFS+LR	5	['Residence', 'HCV_ELISA', 'HBsAg_ELISA', 'Job', 'neadlestick']

As shown in the Table 4, although the number of features is fixed for each classifier, the selected features are different. This is because the SFS feature selection method uses the classifier as an induction algorithm to select the features that achieve the best performance for each classifier. After that, the selected features are used for classification of testing samples. Table 5 compares the performance of several classification approaches in terms of F1-scores, classification accuracy, recall, and precision when utilizing the first and second scenarios. The highest obtained values are shown in bold font. For each categorization approach, the default settings were utilized. To test the performance of the classifier, the trials used a 10-cross-validation method.

Table 5. Comparison of different classifiers with and without using the SFS feature selection method.

classifier	Evaluation metrics	Without Feature selection	With SFS feature selection
NB	Accuracy	92.66	**93.01**
	F1score	70.02	71.27
	Precision	56.59	57.52
	Recall	96.32	98.035
RF	Accuracy	**94.06**	**94.06**
	F1score	63.71	71.33
	Precision	67.62	62.34
	Recall	61.95	84.52
KNN	Accuracy	90.8	**92.66**
	F1score	36.02	44.52
	Precision	50.66	76.14
	Recall	29.24	33.15
LR	Accuracy	92.2	**93.01**
	F1score	62.83	64.75
	Precision	57.92	62.07
	Recall	72.41	71.39

From Table 5, when applying the classification technique using the first scenario with all features in the HCV dataset, we noticed that RF achieves higher performance with 94.06% accuracy, 63.71 % F1-scores, 67.62% precision and 61.95% Recall.

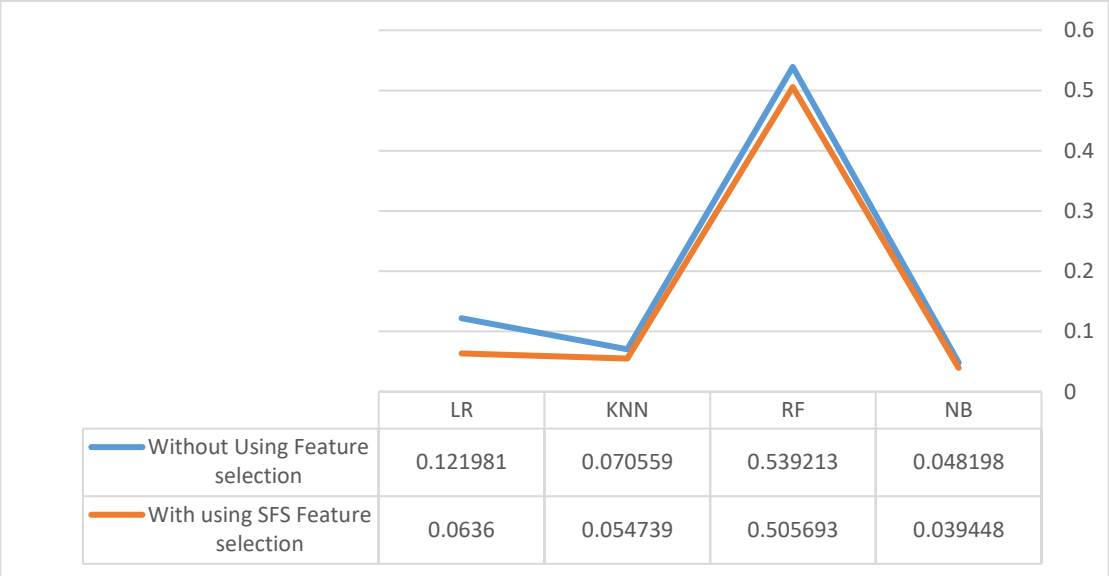
In the second scenario, the aim of this experiment is to evaluate the effectiveness of applying classification techniques with only features selected using the SFS feature selection method shown in Table 4 and compare it with the first scenario.

In Table 5, we noticed that when applying the classification using the second scenario, it can improve the performance of classifiers in terms of F1-scores, precision, recall and accuracy indicating that the features selected by the SFS method are more effective.

Although the accuracy of utilizing the RF algorithm with and without the SFS feature selection approach is the same, the RF algorithm in the second scenario obtained the same accuracy with a lower amount of features than the original set of data. It also has higher F1-scores (71.33%) and Recall (84.52%) values.

These findings also revealed that the RF classifier in the second scenario performs better with only five features out of the whole HCV dataset.

Figure 5 shows the learning time of the model in seconds for the HCV dataset whether utilizing or not using the SFS feature selection approach.



	LR	KNN	RF	NB
Without Using Feature selection	0.121981	0.070559	0.539213	0.048198
With using SFS Feature selection	0.0636	0.054739	0.505693	0.039448

Figure 5. Learning time in seconds.

Figure 5 shows that classifiers who used the second scenario spent less time to build the model than those who used the first. Furthermore, the results demonstrate that developing a model using the NB technique takes less time than with the RF method.

In conclusion, despite the fact that it took longer to develop the model than a model using the original dataset, the RF classifier using a reduced dataset performed the best for the HCV dataset.

6.3. Influence of the Parameter Tuning on the Learning Algorithm

Because default settings may not be suitable for all jobs, parameter tuning is required [36]. As expected, the highest performing classifier was the RF classifier with SFS feature selection applied. We tweaked the classifier's hyperparameters to improve its performance. We employed the grid search approach, as shown in Figure 6, to evaluate a set of hyperparameters and select the optimal parameter values for a specific job based on validation accuracy. This is a more computationally intensive method than just utilizing the model's default parameter settings.

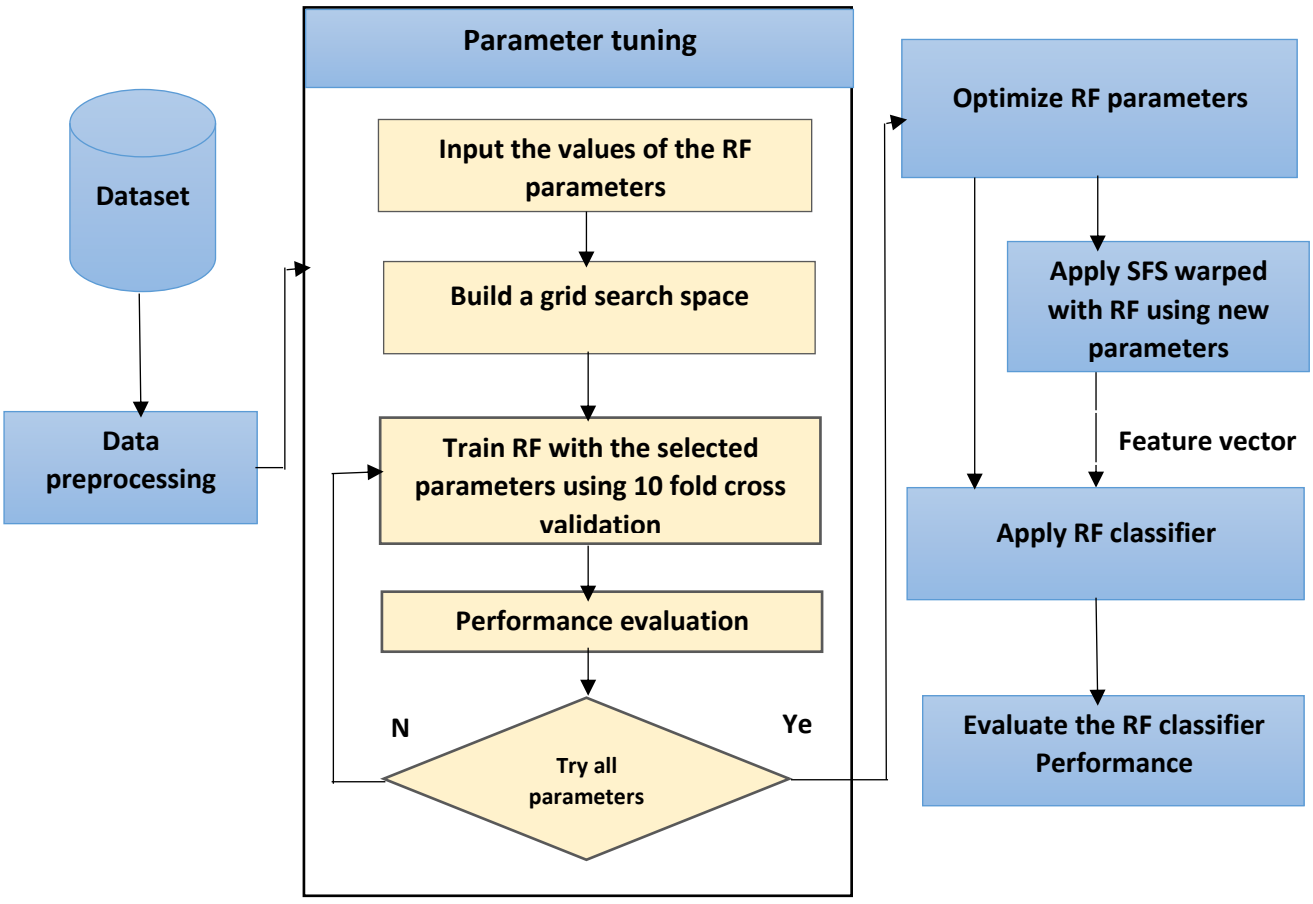


Figure 6. The effect of tweaking hyper parameters on classification performance.

In this study, we used the following hyperparameters for the RF classifier:

- n_estimators: number of trees in the forest.
- max_depth: maximum depth of each tree.
- max_features: The max_features parameter specifies the size of the random subsets of features to consider when splitting a node.
- criterion: The function to measure the quality of a split.

Table 6 shows the best hyperparameters values for the RF classifier.

Table 6. The best hyperparameters values

n_estimators	max_depth	max_features	criterion
200	9	auto	gini

Following the discovery of the optimum parameters, we examine the efficiency of hyperparameter modification on classification performance as follows: First, we use the updated hyperparameter values to retrain the RF classifier with the entire dataset. Then, using the modified parameter values, we run the SFS feature selection technique wrapped with the RF classifier to choose the optimal feature subset. The tests were done using a 10 cross-validation and the new hyperparameter values presented in Table 6 to evaluate the performance of the RF classifier.

Table 7 shows the effect of using hyperparameter tuning on the classification performance.

Table 7: Performances comparison between the RF classifier using default parameters, parameter tuning, and feature selection with the new parameters.

Measure	Default parameters	Tuning parameters	Feature selection using new parameters
Accuracy	94.06	94.29	**94.88**
NO. of features	11	11	4

From Table 7, it is clear that the hyperparameter tuning improves the RF classifier performance. When using the SFS feature selection method wrapped with the RF classifier using the new parameters, the classification accuracy is improved using only five features out of all features of the HCV dataset.

7. Discussion

In this paper, we proposed a ML framework, MHF_HCV, for HCV disease diagnosis among HCWs in Egypt. The aim of this study is to work on methods to enhance the accuracy of ML classification methods in order to diagnose hepatitis C using real world data for HCV. To achieve this, first we measured the influence of the data splitting technique used on the performance of classifiers. Then, we used an SFS based wrapper feature selection approach to find optimal feature subset that influences the class discovery process and improves the classification accuracy. After that we test the performance of the classifiers using all attributes and selected features separately to compare the achieved accuracy and evaluate the quality of the selected features on improving the classification accuracy. Finally, we quantifying the impact of parameter tuning on learning techniques.

The results of the experiments that presented in this work are summarized as follows:

- When applying classification techniques using different splitting methods, we found that the10 fold cross-validation method for splitting the HCV

dataset was found as the best option for ML modeling. AS we can also observe that the best performances occur when using 10 fold cross-validation with the RF classifier.

- When comparing classification techniques with and without the SFS feature selection method, we found that the features selected by SFS improved the performance of the classification techniques. Also, the RF classifier with SFS feature selection method has the best performance using only five features out of all features of the HCV dataset.
- For the HCV dataset, the RF classifier with a reduced dataset had the greatest performance; while it took longer to create the model, it was still faster than building a model with the original dataset.
- When adjusting the hyperparameter of the RF classifier, the performance of the RF classifier was improved. Also, when the SFS feature selection method wrapped with the RF classifier using new parameters, the classification accuracy was improved using only 4 features.

In terms of parameter tuning and SFS feature selection, the RF classifier with 10-fold cross-validation achieved the maximum accuracy score of 94.88. Furthermore, the work given in this study has proven the study's aims, with the findings demonstrating its efficacy.

Threads and limitations

Although the proposed framework achieved superior performance to diagnosis the HCV disease, it still has some limitations. The first one is that the sample selected for this study was specifically for Egyptian HCWs, working in a high-risk environment in NLI, and not for HCV patients in general. The second limitation is the size of the HCV data set, which consists of 859 patients. Using a large HCV dataset to train the model can potentially improve the performance of the proposed HCV framework. The current study used only 11 features that represent the results of laboratory tests required to diagnose HCV. Therefore, more features are required to give more details that may be useful in diagnosing newly infected cases of HCV.

8. Conclusion and Future Work

In this paper, ML framework based on NB, RF, KNN, and LR are proposed to classify and predict the infected patients with HCV for the enrolled features. SFS feature selection is presented to select the most significant features of the applied dataset. To manipulate the enrolled data without feature selection, we also tested the dataset by applying the all the features directly to the NB, RF, KNN, and LR without feature selection. Finally, we tested the impact of parameter tuning on learning techniques. The results indicates that there is a great enhancement of the accuracy obtained after feature selection using SFS with the new parameters in terms of the accuracy. Moreover, the elapsed learning time is very law which mean the speed of learning process of the proposed framework. In the future work, we plan to

recommend a treatment protocol based on genetic algorithm or DNA sequence analysis to help the specialist to treat and handle the spread of HCV in Egypt.

Acknowledgements

Authors sincerely acknowledge Computer Science Department in Faculty of Science, Minia University for the facilities and support.

Declarations:

Funding: Not applicable.

Disclosure of potential Conflict of Interest: The authors declare that they have no conflict of interest.

Ethical Statement: "All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards."

Consent Statement: "Informed consent was obtained from all individual participants included in the study."

References

[1] WHO, "Hepatitis C," https://www.who.int/news-room/fact-sheets/detail/hepatitis-c, 2021.
[2] 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 J. Hepatol., vol. 7, no. 26, p. 2676, 2015.
[3] R. Huang et al., "Noninvasive measurements predict liver fibrosis well in hepatitis C virus patients after direct-acting antiviral therapy," Dig. Dis. Sci., vol. 65, no. 5, pp. 1491–1500, 2020.
[4] C. Westermann, C. Peters, B. Lisiak, M. Lamberti, and A. Nienhaus, "The prevalence of hepatitis C among healthcare workers: a systematic review and meta-analysis," Occup. Environ. Med., vol. 72, no. 12, pp. 880–888, 2015.
[5] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," in Machine learning proceedings 1994, Elsevier, 1994, pp. 121–129.
[6] E. Triantaphyllou and G. Felici, Data mining and knowledge discovery approaches based on rule induction techniques, vol. 6. Springer Science & Business Media, 2006.
[7] J. H. Hoofnagle and A. M. Di Bisceglie, "The treatment of chronic viral hepatitis," N. Engl. J. Med., vol. 336, no. 5, pp. 347–356, 1997.
[8] E. Jaeckel et al., "Treatment of acute hepatitis C with interferon alfa-2b," N. Engl. J. Med., vol. 345, no. 20, pp. 1452–1457, 2001.
[9] C. Frank et al., "The role of parenteral antischistosomal therapy in the spread of hepatitis C virus in Egypt," The Lancet, vol. 355, no. 9207, pp. 887–891, 2000.
[10] N. Antaki et al., "The neglected hepatitis C virus genotypes 4, 5 and 6: an international consensus report," Liver Int., vol. 30, no. 3, pp. 342–355, 2010.
[11] P. Burra et al., "Liver transplantation for alcoholic liver disease in Europe: a study from the ELTR (European Liver Transplant Registry)," Am. J. Transplant., vol. 10, no. 1, pp. 138–148, 2010.
[12] S. Bruno et al., "Sustained virologic response prevents the development of esophageal varices in compensated, Child-Pugh class A hepatitis C virus–induced cirrhosis. A 12-year prospective follow-up study," Hepatology, vol. 51, no. 6, pp. 2069–2076, 2010.
[13] E. J. Bini and P. V. Perumalswami, "Hepatitis B virus infection among American patients with chronic hepatitis C virus infection: prevalence, racial/ethnic differences, and viral interactions," Hepatology, vol. 51, no. 3, pp. 759–766, 2010.

[14] S. C. Nandipati, C. XinYing, and K. K. Wah, “Hepatitis C virus (HCV) prediction by machine learning techniques,” Appl. Model. Simul., vol. 4, pp. 89–100, 2020.
[15] UCI-ML repository HCV, “UCI-ML repository,” https://archive.ics.uci.edu/ml/datasets/Hepatitis+C+Virus+%28HCV%29+for+Egyptian+patients.
[16] S. Eliyahu et al., “Antibody repertoire analysis of hepatitis C virus infections identifies immune signatures associated with spontaneous clearance,” Front. Immunol., vol. 9, p. 3004, 2018.
[17] M. M. R. Ali, Y. Helmy, A. E. Khedr, and A. Abdo, “Intelligent Decision Framework to Explore and Control Infection of Hepatitis C Virus,” in International Conference on Advanced Machine Learning Technologies and Applications, 2018, pp. 264–274.
[18] S. M. Abd El-Salam et al., “Performance of machine learning approaches on prediction of esophageal varices for Egyptian chronic hepatitis C patients,” Inform. Med. Unlocked, vol. 17, p. 100267, 2019.
[19] S. Hashem et al., “Machine learning prediction models for diagnosing hepatocellular carcinoma with HCV-related chronic liver disease,” Comput. Methods Programs Biomed., vol. 196, p. 105551, 2020.
[20] A. H. KayvanJoo, M. Ebrahimi, and G. Haqshenas, “Prediction of hepatitis C virus interferon/ribavirin therapy outcome based on viral nucleotide attributes using machine learning algorithms,” BMC Res. Notes, vol. 7, no. 1, pp. 1–11, 2014.
[21] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in Ijcai, 1995, vol. 14, no. 2, pp. 1137–1145.
[22] H. Liu and M. Cocea, “Semi-random partitioning of data into training and test sets in granular computing context,” Granul. Comput., vol. 2, no. 4, pp. 357–386, 2017.
[23] G. Chandrashekar and F. Sahin, “A survey on feature selection methods,” Comput. Electr. Eng., vol. 40, no. 1, pp. 16–28, 2014.
[24] R. Kohavi and G. H. John, “Wrappers for feature subset selection,” Artif. Intell., vol. 97, no. 1–2, pp. 273–324, 1997.
[25] R. Gutierrez-Osuna, “Pattern analysis for machine olfaction: A review,” IEEE Sens. J., vol. 2, no. 3, pp. 189–202, 2002.
[26] P. Langley, “Selection of relevant features in machine learning,” in Proceedings of the AAAI Fall symposium on relevance, 1994, vol. 184, pp. 245–271.
[27] M. Gopal, Applied machine learning. McGraw-Hill Education, 2019.
[28] L. Breima, “Random Forests. Machine Learning,” 2010.
[29] S. Le Cessie and J. C. Van Houwelingen, “Ridge estimators in logistic regression,” J. R. Stat. Soc. Ser. C Appl. Stat., vol. 41, no. 1, pp. 191–201, 1992.
[30] J. Shu et al., “Clear cell renal cell carcinoma: CT-based radiomics features for the prediction of Fuhrman grade,” Eur. J. Radiol., vol. 109, pp. 8–12, 2018.
[31] R. Kumari and J. Jose, “Seizure detection in EEG using Biorthogonal wavelet and fuzzy KNN classifier,” Elixir Hum Physiol, vol. 41, pp. 5766–5770, 2011.
[32] O. Altay and M. Ulas, “Prediction of the autism spectrum disorder diagnosis with linear discriminant analysis classifier and K-nearest neighbor in children,” in 2018 6th International Symposium on Digital Forensic and Security (ISDFS), 2018, pp. 1–4.
[33] M. Sokolova, N. Japkowicz, and S. Szpakowicz, “Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation,” in Australasian joint conference on artificial intelligence, 2006, pp. 1015–1021.
[34] S. Raschka, “An overview of general performance metrics of binary classifier systems,” ArXiv Prepr. ArXiv14105330, 2014.
[35] P. G. Asteris et al., “On the metaheuristic models for the prediction of cement-metakaolin mortars compressive strength,” 1, vol. 1, no. 1, p. 063, 2020.
[36] J. Hartmann, J. Huppertz, C. Schamp, and M. Heitmann, “Comparing automated text classification methods,” Int. J. Res. Mark., vol. 36, no. 1, pp. 20–38, 2019.