Indonesian Journal of Electrical Engineering and Computer Science
Vol. 26, No. 3, June 2022, pp. 1564~1572
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v26.i3.pp1564-1572

Diagnosis of hepatitis disease using machine learning techniques

Ibraheem I. Ahmed, Duraid Y. Mohammed, Khamis A. Zidan
College on Engineering, Al-Iraqia University, Baghdad, Iraq

Article Info

Article history:

Received Jan 19, 2022
Revised Mar 18, 2022
Accepted Apr 2, 2022

Keywords:

Adaptive feature selection
Automated hepatitis diagnosis
Decision tree
Hepatitis symptoms
Random forest
Support vector machine

ABSTRACT

Hepatitis is an infection that causes inflammation of liver tissue. Many studies have developed machine learning models for hepatitis disease diagnosis. However, there has been little discussion about the relationship between hepatitis symptoms. The first objective of this study is to provide a brief description of a real-world hepatitis disease symptom dataset. Furthermore, the authors proposed a stand-alone classification platform using random forest, decision tree, and support vector machine into healthy people or hepatitis patients using adaptive wrapper feature selection. It was discovered that there is a strong link between certain characteristics and hepatitis diagnosis. The work presented here may help improve hepatitis diagnosis in the early stages, which may lead to a reduction in the acute effects of hepatitis on human life. It is worth noting that random forest (RF) gave the highest accuracy and stayed slightly consistent through all sets of features in comparison to decision tree (DT) and support vector machines (SVM).

Corresponding Author:

Duraid Y. Mohammed
Departement of Computer Engineering, Al-Iraqia University
Baghdad, Iraq
Email: duraiyehya19@gmail.com

1. INTRODUCTION

Liver is one of the key organs in the human body; it has different functions, such as glycogen storage, chemicals detoxification, drugs metabolization, and energy production also makes proteins important for blood clotting. However, many microorganisms or diseases; such as bacteria and viruses can damage the liver and prevent it from functioning normally [1], [2]. Hepatitis is one of these viruses; it infects the liver tissue cells, causing damage and loss of function the infected cells become dysfunctional. Chronic hepatitis is more than just an illness; it is a pathologic and clinical syndrome. Causes and distinguishes itself through various levels Hepatocellular tissue necrosis and inflammation. In the 1970s, the first chronic classification of hepatitis was proposed [3], [4]. Later, in the 1990s significant progress was made in identifying and classifying Chronic Syndrome Hepatitis; it was divided into five types and named using the alphabet, Hepatitis A Virus (HAV), Hepatitis B Virus (HBV), Hepatitis C Virus (HCV), Hepatitis D Virus (HDV), and Hepatitis E Virus (HEV). All of these viruses attack the liver, each with its own set of symptoms, and the majority of them are successfully treated. HAV typically infects children and is classified as a hepatitis infection. HAV can be avoided by either experiencing the illness or receiving a vaccine. HBV and HCV have no symptoms and are classified as carriers. Hepatitis B is a leading cause of liver disease; it affects the liver by assaulting it [5]. In general, unprotected sex, blood, and reused or shared syringes can all infect the liver. Also, hepatitis can be passed on to a child during pregnancy or after birth. A blood test is typically required once a year for early detection. For most of the cases, routine blood testing is used for hepatitis diagnoses or it is diagnosed during a blood donation [6].

It is worth noting that advances progress in chronic hepatitis therapy occurred simultaneously with identification of categories of hepatitis [7]. Medical diagnostics of hepatitis is a quite difficult task as the number of factors that needs to be considered are many. Alongside clinic tests, machine learning has been utilized for early diagnosis of hepatitis diseases in medicine, as the accuracy of an automatic system can be greatly helpful for the detection of hepatitis as a result will help for making the decision by the physician. With that being said, most of the researchers have focused on extracting and selecting the best number of features and applied in to the best tested algorithms; in the theoretical work it would give great results, but when implementing it, it would require a large number of tests to be done by the patient to get the result which might cost a lot and take some time to get the tests done which is very critical to hepatitis patients.

There are number of Hepatitis features extraction and reduction some of which [2], [8]–[10] Chen performed a hybrid model based on local fisher discriminant analysis and support vector machine (LFDA-SVM) methods, the LFDA is used as a feature extraction tool for dimensionality reduction in order to further enhance the diagnostic precision of the conventional support vector machines (SVM) algorithm, and they reported a classification success rate of 96.77% for 80-20% training-testing partition for a reduced feature subset which contained only two features [9]. Later on, Çalisir and Dogantekin were able to achieve a success rate of 96.12% using principle component analysis (PCA) for feature extraction and reduction which reduced the features to 10 out of 19, and the features are fed to a least square support vector machine (LSSVM) for classification of the disease [2].

Kaya and Uyar showed that they have obtained a success percentage of 100% for 80-20% training-testing partition using rough set (RS) and extreme learning machine (ELM) hybrid model. The RS is used for feature reduction giving only four used features, and ELM algorithm is used for the classification process [8]. In 2019 Ahmadi *et al.* have successfully obtained 93.06% using a hybrid model, which consists of non-linear iterative partial least squares (NLPALS) to reduce the dimensions of the data as the first stage, self-organizing map (SOM) technique for clustering the data in the experimental dataset as the second stage, and for the last stage, ensembles of neuro-fuzzy inference system is used for classification of the hepatitis diagnosis [10]. Regarding machine learning techniques, there is a wide range of the employed algorithms that have been used to improve performance accuracy of hepatitis prediction, among them are Sartakhti and Mozafari, who they have suggested a machine learning method that uses a hybrid system of SVM and simulated annealing (SA) which have successfully obtained a classification accuracy of 96.25% in predicting hepatitis disease [11]. Although the researchers obtained high accuracy, they removed all of the samples that contain missing data, which is 51% of the dataset, leaving only 80 samples which can highly get poorly trained model.

A comparison has been carried out using a number of ML techniques for specific types of hepatitis by a group of researchers and did a benchmark on the machine learning algorithms. One of these studies was considered with HBV diagnosis, it implemented extreme gradient boosting (XGBoost), random forest (RF), which is regarded as the modern generation of decision tree machine learning and has been utilized by a large number of researchers in recent years [12], decision tree (DT), and logistic regression (LR), they were able to successfully obtain the results 77.9%, 75.2%, 61.9%, and 74.2% respectively [13] 2019. Another study was carried out by Doyle *et al.* on a huge dataset of HCV patients with nearly 10 million patients in the United States (US) which are captured between 2010 and 2016 collected Quintiles and IMS Health Institution (IQVIA), a benchmark on logistic regression (LR), random forest (RF), gradient boosted trees (XGBoost), and stacked ensemble (SE) was made, the best results were obtained using SE with a precision of 97%, followed by XGBoost with a precision of 87%, then only 31% precision for the LR using recall levels greater than 50% [14] 2020.

Elsayad *et al.* demonstrated that using multilayer perceptron neural network (MLPNN) as a classification algorithm fed by the attributes that are selected using the backward linear regression model (LR_Backward) can result in the best performance of 97.3% for the training subset and 88.6% for the testing subset [15]. Recent researchers pointed out some of the algorithms that were less efficient compared to other algorithms used such as K-dimensional trees and neural networks (NN) having a success rate of 59% and 72% respectively [16]. Singh *et al.* on the other hand demonstrated the implementation of combined K-Means clustering and improved ensemble-driven learning, they were able to achieve the rates for accuracy, true positive rate, precision, mean absolute error, F measure, kappa statistic, and root mean squared error rates of the proposed model are 99.35%, 99.4%, 99.4%, 99.4%, 98.63%, 0.79%, and 8.04%, respectively [17].

All of the aforementioned work except [13], [14] had utilized the dataset from the UCI machine learning repository [6]. Despite the fact that extensive research has been done on general feature selection, one set of features to be run on all training models. Hence the goal of this paper is for understanding each of the given features in the dataset and how they are correlated to build a training model using the optimal set of features. This information would be that start needed to build a stand-alone platform with a mobile

application that can be used by the patient with minimal tests required to reduce the cost by removing unnecessary tests, and reduce the time needed to fill all of the tests.

2. METHOD

2.1 Data description

In this study, a real-world dataset is used to perform the experiments on, which is captured from the UCI repository of machine learning databases [6]. It contains 155 samples, each containing 19 attributes. There are two classes in the dataset, "die" with 32 (20.6%) samples and "live" with 123 (79.4%) samples. The dataset used to indicate the presence or absence of the hepatitis disease using a number of medical tests results carried out on a patient. This dataset has many missing values, such as the protime attribute with 43% missing, a full description of the attributes along with their domain value and the missing percentage is stated in the given Table 1.

Table 1. The attributes of UCI repository of machine learning hepatitis disease dataset

Feature	Value	Missing	Description
Age	Numeric	0%	The age of the patient
Sex	Male, Female	0%	The gender of the patient.
Steroid	Yes, No	1%	A biologically active organic compound with four rings arranged in a specific molecular configuration. It may damage the liver and cause hepatitis directly.
Antivirals	Yes, No	0%	They are a class of medications, used to treat viral infections.
Fatigue	Yes, No	1%	A term used to describe an overall feeling of lack of energy. It is common for people who have hepatitis.
Malaise	Yes, No	1%	It is a general feeling of discomfort or illness.
Anorexia	Yes, No	1%	Eating disorder shown by an abnormally low body weight.
Liver Big	Yes, No	6%	Enlarged liver is a sign of an underlying issue, such as liver disease, congestive heart failure or cancer.
Liver Firm	Yes, No	7%	Liver can be tender with acute hepatitis or feel hard and irregular (bumpy) with cancer of the liver.
Spleen Palpable	Yes, No	3%	It's part of the immune system, that we can survive without it. This is because the liver can take over many of the spleen's functions. If the spleen does not work properly, will cause the removing of the healthy blood cells.
Spiders	Yes, No	3%	Spider nevus is a common benign vascular anomaly that may appear as solitary or multiple lesions, they are usually less than 2 cm in diameter, a large spider might indicate in a patient with hepatitis C virus.
Ascites	Yes, No	3%	Fluid that accumulates in the abdomen might become infected and require treatment with antibiotics. Ascites are usually a sign of advanced alcoholic hepatitis or cirrhosis.
Varices	Yes, No	3%	It is an abnormally dilated vessel with a tortuous course, they are a complication of late-stage hepatitis C.
Bilirubin	Numeric	4%	Bilirubin is a yellow mixture that can cause jaundice and dark urine to accumulate when the liver is injured.
Alk Phosphate	Numeric	19%	An enzyme associated with the bile ducts but also located in other muscles in the body. It is raised when bile ducts are prevented but may also be risen with bone disorders.
Sgot.	Numeric	3%	An enzyme located in the heart, liver, and other muscles. The test is valuable in identifying liver injury due to hepatitis and may be advanced more than ALT with exposure to drugs toxic to the liver, cirrhosis, or alcoholism.
Albumin	Numeric	10%	The main protein produced by the liver, its level can reduce with failure of liver function though, this typically happens only when the liver has been seriously harmed.
Protime	Numeric	43%	The prothrombin time (PT) is a test that helps evaluate the ability to appropriately form blood clots, A high PT usually means that there is serious liver damage or cirrhosis.
Histology	Yes, No	0%	Routine lab technique used to evaluate the morphology and structure of cells, tissues, and organs under the microscope.
Class	Die, Live	0%	The indication of the presence of hepatitis disease.

2.2. Training framework architecture

As previously stated, many studies have focused on selecting features independently of the training model. Using traditional classification schemes, the previous systems frequently perform a feature selection technique which can then be used in all algorithms to classify the hepatitis disease. This may result in a variable performance for each model depending on what type of algorithm is used and what is the representation of the selected features is, some algorithms might perform worse just because the selected features are not the best to apply on the specified algorithm. In order to address the feature selection issue, the following sub-sections propose and justify a hepatitis disease diagnosis stand-alone platform, starting from the training framework

architecture then testing framework architecture to the real-time diagnosis platform. The full training framework architecture is shown in the Figure 1, each section will be explained in details. The entire process is repeated for all three selected algorithms SVM [18], DT [19]. As described in the previous section, the used dataset in this research is from UCI repository of machine learning databases [6].

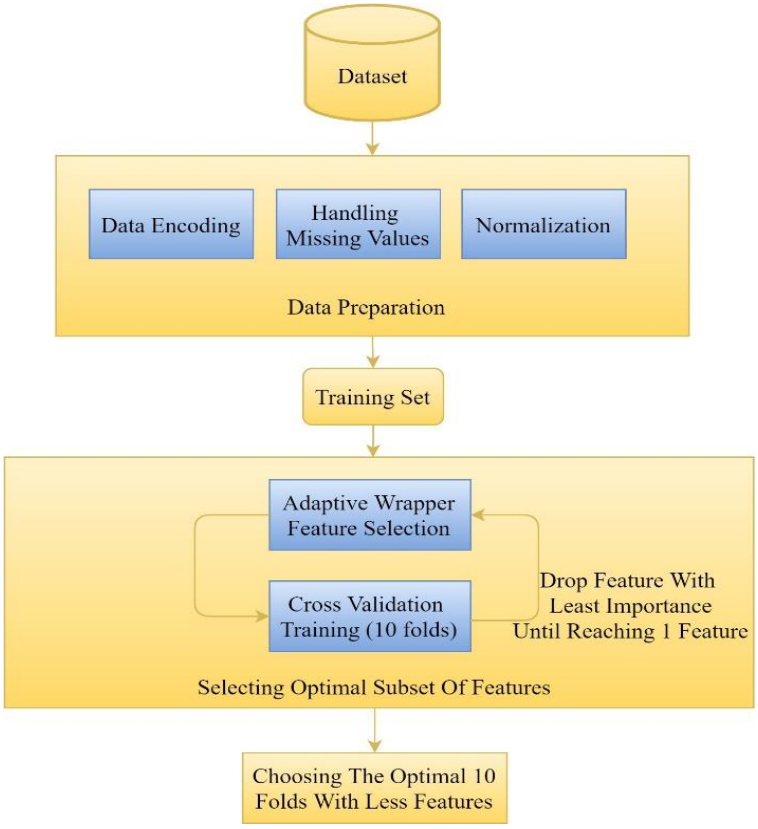

Figure 1. Training framework architecture

2.2.1. Data preparation

Data preparation includes all the required pre-processing steps before start training the model in order to get the best results [20], the preparation consists of three steps:

a) Data encoding: the dataset contains six numerically assigned values and 13 nominally assigned values. In order to perform most of the machine learning algorithms, real numbers must be used. Hence, all of the nominal values must be converted to real numbers, a simple representation of the real numbers 1 and 2 is used. For example, the male class is represented by 1 and the female class is represented by 2.

b) Handling missing values: the dataset contains a high level of missing data with a (48.3%) missing, removal of the missing values can lead to an unusable dataset since it only consists of 155 samples, so it is crucial to remove them. The missing values are replaced with the mean of the values of each class, which is a statistical approach [21].

c) Normalization: the range of values of the dataset vary significantly, especially after encoding the nominal values to 1 and 2 real numbers, in this case the scaling is required to avoid greater numeric ranges dominate the attributes that are in smaller numeric ranges. Also, it makes the execution of any algorithm faster by avoiding the use of wide range of numbers [22]. The data is scaled into the interval of (0–1) according to (1), where x is the original value of the attribute, xnormalized is the scaled value, mina is the minimum value of the attribute a, maxa is the maximum value of the attribute a.

$$x_{normalized} = \frac{x-min_a}{max-min_a}) \quad (1)$$

2.2.2. Selecting optimal subset of feature

The features selection block describes the procedure for selecting the optimal subset of features, which varies according to the algorithm used and the learning performance. This is accomplished by following the procedure outlined below.

a) Adaptive wrapper feature selection: Backward elimination is used in this study to identify the set of optimal features [23], and the results are presented in this paper. First, all of the features for the selected algorithm are used, and then, with each iteration, the importance of each feature is calculated, and the feature with the lowest importance is dropped. This loop is repeated until only one feature is left to be explored. This process is repeated until the diagnosis performance declines significantly as will be discussed in the results section.

b) Cross validation training: Following the aforementioned selection of the features subset with the optimal performance (starting with full feature set in the first iteration). Cross validation [24] with 10 folds is used to calculate the discriminant performance, and all trained models are saved for use in diagnosing unseen samples, as will be demonstrated later. After performing this process for all of the algorithms, the performance of the models is measured by comparing the results that were captured using the cross validation for each subset of features to determine which model is performing the best for each number of features.

c) Choosing the optimal fold with less features: This research is being conducted in order to develop a mobile application that can be used by patients or health care facilities. Thus, one of the required goals is to reduce the number of features, which in turn reduces the cost of the tests while maintaining the highest possible accuracy. This procedure identifies the fewest possible features with the optimal ten folds. The ten output models from each fold are saved to be used later in the testing face.

2.3. Testing framework architecture

As shown in Figure 2, a simulated test was run on an unseen portion of the data set. The testing process is similar to the training process in terms of data preparation. After taking the prepared data, it is fed to a script that runs the prediction on the 10 pre-trained models, and a voting process is used to select the decision with the highest probability [25], except the case if five models predicted effected and five models predicted healthy, the patient is considered as effected with hepatitis disease because we are dealing the a disease and it's best for the patient to check a doctor.

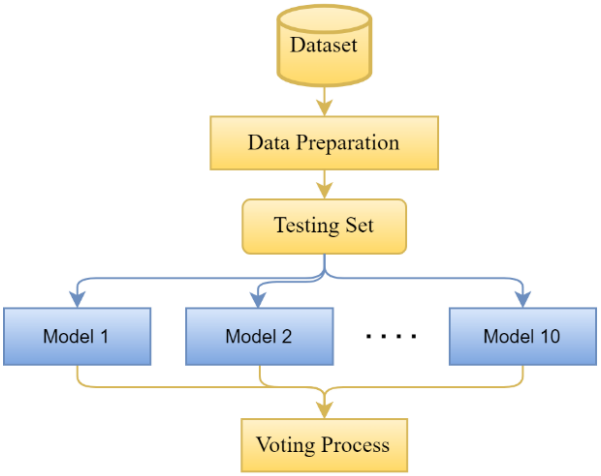

Figure 2. Testing framework architecture

2.4. Real-time diagnosis platform

The system described in this article is based on the framework for real-time hepatitis diagnosis and has been customized through the integration of a mobile application with a Raspberry Pi server, as illustrated in Figure 3. The used feature set is reduced to six features only which are (Spiders, Ascites, Bilirubin, Alk Phosphate, Albumin, and Protime).

a) Patient: in terms of patients only need to fill out 6 tests. Its 6 test features like Spiders, Ascites, Bilirubin, Alk Phosphate, Albumin, and Protime mentioned earlier, and transmit the data. The result will be displayed which indicates well done or healthy.
b) Mobile application: the mobile application receives the data and validate it and push it to the server which is running on a Raspberry Pi, and as a result it receives the voting result back from the raspberry and displays it to the patient. Below are some of the screenshots from the mobile application Figure 4 showing that the patient is indeed inffected.
c) Raspberry Pi: the server is located in the Raspberry Pi, it receives the data from the mobile application and loads the trained models from the memory which were previously updated to its storage, then runs the prediction script which is written in Python. To match the cross-validation process, 10 models are retrieved from the storage, each model is trained/tested on a different subset from the dataset. The result is then calculated using a voting process between all of the models' predictions and it's pushed to the mobile application.

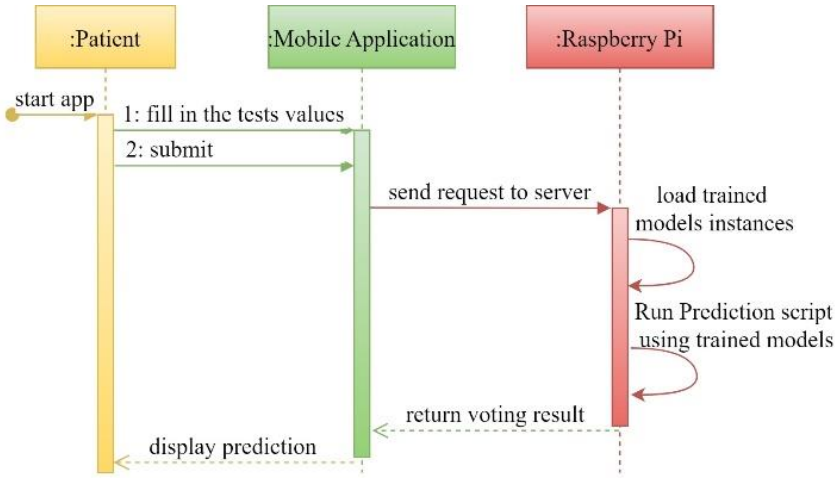

Figure 3. Prediction platform sequence diagram

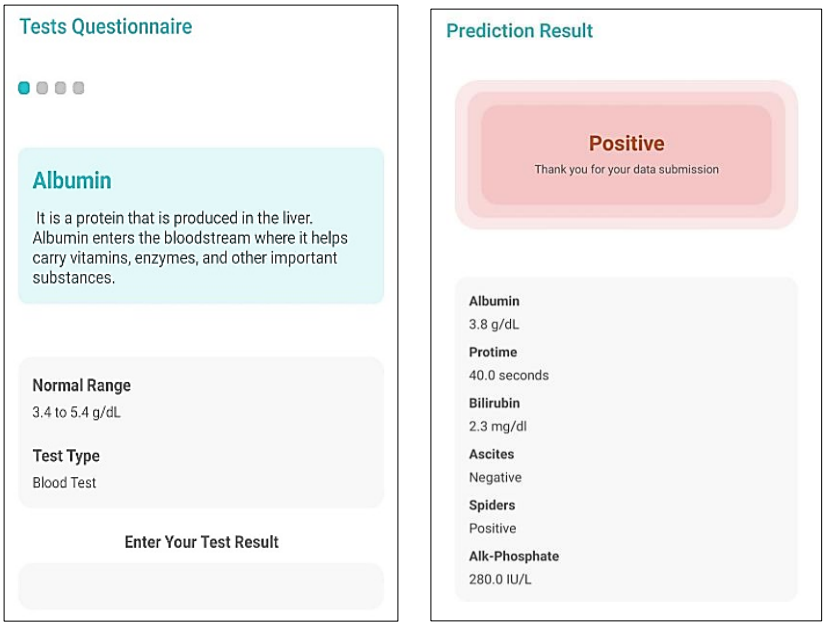

Figure 4. Prediction platform sequence diagram

2.5. Used tools

2.5.1. WEKA

The initial analysis were done using WEKA software. It was used first to analyze the available datasets and helped in the decision of using the UCI dataset. Second, it was used to encode, handle missing data, normalize and shuffle the dataset in order to prepare it for learning. Then it was used to find the best performing models out of the most used models in the previous studies.

2.5.2. Python

Python was used as the main processing tool, the models that were found performing the best using WEKA were programmed using Python to perform furthermore analysis, the target of this project was to be able to predict the presence of hepatitis disease for patients using a mobile application with their own test results, so every model performance was tested with a feature sub-selection, starting from 19 features down to 1 feature. The feature importance was calculated in every iteration to drop the least important feature, this resulted in high performance even with very limited number of features.

2.5.3. Spring boot framework

The framework is uded as a server installed on a Raspberry Pi. It handles the requests from the mobile application with the patient's tests which are down to six features only. After the request is received by the server, it loads the trained models and predict the result using 10 models then perform a voting to deliver the result back to the application.

2.5.4. Rasbperry Pi 4

The Raspberry Pi 4 was used in this project, it is considered as the standalone platform that performs the prediction. It was loaded up with the trained models, and the spring boot server was installed on it to be able to receive and process requests from the mobile application. The application connects with the server using the same network in order to access the server.

2.5.5. Android studio

Android studio program is used to build a mobile application. It is considered as a client application which is used by the patients. The patient can enter his/her lab test results and it is connected to the spring boot server in the Raspberry Pi. It can send the request to the server and receive the result instantly through network calls.

3. RESULTS AND DISCUSSION

The described process was implemented using three models depending on the recent studies, these models gave the highest results. The left section of the Table 2 lists the feature being dropped in ever iteration during the learning process, starting with 19 features down to 1 feature. The right side of the table shows the performance of each algorithm for every subset of the selected features.

Table 2. The feature subsets and the cross-validation accuracy of the trained models

No. of features	Feature to drop			Cross validation (10 folds)		
	RF	DT	SVM	RF	DT	SVM
19	Antivirals	Albumin	Age	0.947	0.925	0.922
18	Anorexia	Anorexia	Sex	0.961	0.922	0.922
17	Liver Firm	Alk Phosphate	Steroids	0.954	0.936	0.9
16	Liver Big	Antivirals	Antivirals	0.95	0.943	0.9
15	Sex	Histology	Fatigue	0.943	0.922	0.908
14	Steroids	Liver Big	Malaise	0.95	0.926	0.904
13	Spleen Palpable	Malaise	Anorexia	0.947	0.919	0.894
12	Fatigue	Liver Firm	Liver Big	0.947	0.915	0.904
11	Varices	Spiders	Liver Firm	0.943	0.94	0.904
10	Sgot	Sgot	Spleen Palpable	0.94	0.922	0.911
9	Malaise	Spleen Palpable	Spiders	0.95	0.94	0.897
8	Age	Steroids	Ascites	0.947	0.929	0.901
7	Histology	Varices	Varices	0.954	0.919	0.904
6	Alk Phosphate	Sex	Bilirubin	0.957	0.901	0.904
5	Spiders	Fatigue	Alk Phosphate	0.95	0.926	0.872
4	Ascites	Bilirubin	Sgot	0.936	0.918	0.879
3	Bilirubin	Age	Albumin	0.918	0.908	0.879
2	Albumin	Ascites	Protime	0.911	0.897	0.872
1	Protime	Protime	Histology	0.868	0.879	0.698

The Figure 5 shows the accuracy of the model alongside with the number of selected features, it is worth noting that the consistency of the performance remained similar regardless the number of features used. Below that is the result of ROC curve and AUC of all of the training models is shown in Figure 6. It clearly shows that random forest algorithm achived the highest result of 0.952 followed by support vector machine and decision tree with the results 0.925 and 0.912 respectively.

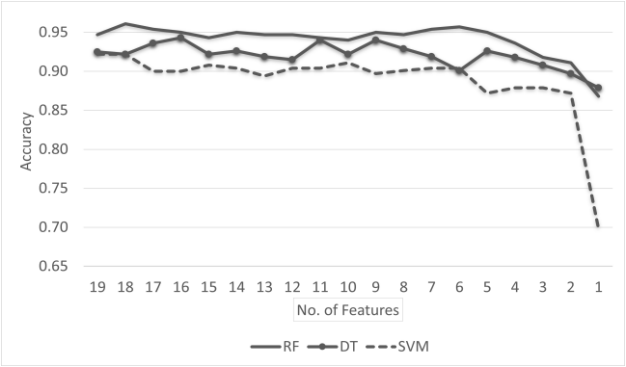

Figure 5. Model accuracy verses number of features used

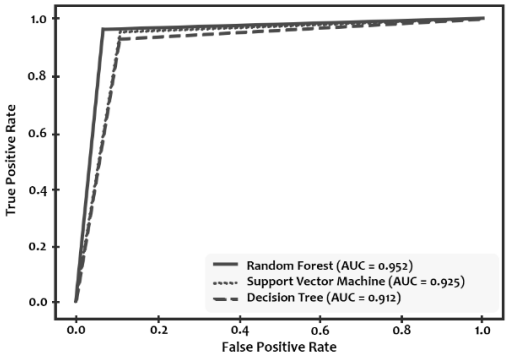

Figure 6. Roc curve and AUC

4. CONCLUSION

The used machine learning algorithms provided an accuracy of 96.1 in random forest with 18 features and 95.7% with only 6 features. Decision tree classifier performed 94.3% with 16 features, and 94% with only 9 features. As for support vector machine classifier, it performed 92.2 with 18 features minimum. Having different set of features for each model has not only proven that it could provide significant improvement in results, but it also proved that it could keep the performance slightly consistent regardless the number of features being used to train the model with minimal loss in accuracy with lower number of features in comparison with using one set of features in all training models.

REFERENCES

[1] K. Polat and S. Güneş, "Hepatitis disease diagnosis using a new hybrid system based on feature selection (FS) and artificial immune recognition system with fuzzy resource allocation," *Digital Signal Processing: A Review Journal*, vol. 16, no. 6, pp. 889–901, 2006, doi: 10.1016/j.dsp.2006.07.005.

[2] D. Çalişir and E. Dogantekin, "A new intelligent hepatitis diagnosis system: PCA-LSSVM," *Expert Systems with Applications*, vol. 38, no. 8, pp. 10705–10708, 2011, doi: 10.1016/j.eswa.2011.01.014.

[3] S. M. Feinstone, A. Z. Kapikian, R. H. Purcell, H. J. Alter, and P. v. Holland, "Transfusion-associated hepatitis not due to viral hepatitis type A or B," *New England Journal of Medicine*, vol. 292, no. 15, pp. 767–770, 1975, doi: 10.1056/nejm197504102921502.

[4] L. B. Seeff *et al.*, "A randomized, double blind controlled trial of the efficacy of immune serum globulin for the prevention of post-transfusion hepatitis: A veterans administration cooperative study," *Gastroenterology*, vol. 72, no. 1, pp. 111–121, 1977, doi: 10.1016/S0016-5085(77)80313-2.

[5] A. Cropley and M. Weltman, "The use of immunosuppression in autoimmune hepatitis: A current literature review," *Clinical and molecular hepatology*, vol. 23, no. 1, pp. 22–26, Mar. 2017, doi: 10.3350/CMH.2016.0089.

[6] G. Gong, "UCI machine learning repository: Hepatitis data set," 1988. https://archive.ics.uci.edu/ml/datasets/hepatitis.

[7] F. Zoulim and D. Durantel, "Antiviral therapies and prospects for a cure of chronic hepatitis B," *Cold Spring Harbor Perspectives in Medicine*, vol. 5, no. 4, pp. 1–22, 2015, doi: 10.1101/cshperspect.a021501.

[8] Y. Kaya and M. Uyar, "A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease," *Applied Soft Computing Journal*, vol. 13, no. 8, pp. 3429–3438, 2013, doi: 10.1016/j.asoc.2013.03.008.

[9] H. L. Chen, D. Y. Liu, B. Yang, J. Liu, and G. Wang, "A new hybrid method based on local fisher discriminant analysis and support vector machines for hepatitis disease diagnosis," *Expert Systems with Applications*, vol. 38, no. 9, pp. 11796–11803, 2011, doi: 10.1016/j.eswa.2011.03.066.

[10] M. Nilashi, H. Ahmadi, L. Shahmoradi, O. Ibrahim, and E. Akbari, "A predictive method for hepatitis disease diagnosis using ensembles of neuro-fuzzy technique," *Journal of Infection and Public Health*, vol. 12, no. 1, pp. 13–20, 2019, doi: 10.1016/j.jiph.2018.09.009.

[11] J. S. Sartakhti, M. H. Zangooei, and K. Mozafari, "Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA)," *Computer Methods and Programs in Biomedicine*, vol. 108, no. 2, pp. 570–579, 2012, doi: 10.1016/j.cmpb.2011.08.003.

[12] D. Mohammed, K. A. Al-Karawi, P. Duncan, and F. F. Li, "Overlapped music segmentation using a new effective feature and random forests," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 8, no. 2, p. 181, Jun. 2019, doi: 10.11591/ijai.v8.i2.pp181-189.

[13] Y. Wang, Z. Du, W. R. Lawrence, Y. Huang, Y. Deng, and Y. Hao, "Predicting hepatitis b virus infection based on health examination data of community population," *International Journal of Environmental Research and Public Health*, vol. 16, no. 23, 2019, doi: 10.3390/ijerph16234842.
[14] O. M. Doyle, N. Leavitt, and J. A. Rigg, "Finding undiagnosed patients with hepatitis C infection: an application of artificial intelligence to patient claims data," *Scientific Reports*, vol. 10, no. 1, pp. 1–10, 2020, doi: 10.1038/s41598-020-67013-6.
[15] A. M. Elsayad, A. M. Nassef, and M. Al-Dhaifallah, "Diagnosis of hepatitis disease with logistic regression and artificial neural networks," *Journal of Computer Science*, vol. 16, no. 3, pp. 364–377, 2020, doi: 10.3844/JCSSP.2020.364.377.
[16] D. F. Santos-Bustos and H. E. Espitia-Cuchango, "Hepatitis diagnosis using optimized KD-Trees and Neural Networks," *International Journal of Engineering Research and Technology*, vol. 13, no. 9, pp. 2269–2274, 2020, doi: 10.37624/ijert/13.9.2020.2269-2274.
[17] A. Singh, J. C. Mehta, D. Anand, P. Nath, B. Pandey, and A. Khamparia, "An intelligent hybrid approach for hepatitis disease diagnosis: Combining enhanced k-means clustering and improved ensemble learning," *Expert Systems*, vol. 38, no. 1. 2021, doi: 10.1111/exsy.12526.
[18] Y. S. Li and P. H. Peiyi, "SVM classification: its contents and challenges," 2003.
[19] H. H. Patel and P. Prajapati, "Study and analysis of decision tree based classification algorithms," *International Journal of Computer Sciences and Engineering*, vol. 6, no. 10, pp. 74–78, Oct. 2018, doi: 10.26438/ijcse/v6i10.7478.
[20] S. Gopal, K. Patro, and K. Kumar Sahu, "Normalization: A preprocessing stage," [Online]. Available: www.kiplinger.com,
[21] L. Wohlrab and J. Fürnkranz, "A review and comparison of strategies for handling missing values in separate-and-conquer rule learning," *Journal of Intelligent Information Systems*, vol. 36, no. 1, pp. 73–98, Feb. 2011, doi: 10.1007/s10844-010-0121-8.
[22] J.-M. Jo, "Effectiveness of normalization pre-processing of big data to the machine learning performance," *The Journal of the Korea institute of electronic communication sciences*, vol. 14, no. 3, pp. 547-552, 2019, doi: 10.13067/JKIECS.2019.14.3.547.
[23] K. Z. Mao, "Orthogonal forward selection and backward elimination algorithms for feature subset selection," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 34, no. 1, pp. 629–634, Feb. 2004, doi: 10.1109/TSMCB.2002.804363.
[24] D. Berrar, "Cross-validation," in *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, vol. 1–3, Elsevier, 2018, pp. 542–545. doi: 10.1016/B978-0-12-809633-8.20349-X.
[25] Institute of Electrical and Electronics Engineers, *2019 2nd International Conference on New Trends in Computing Sciences (ICTCS) : proceedings : Amman, Jordan, 9-11 October 2019*, 2019.

BIOGRAPHIES OF AUTHORS

Ibraheem Ismael Ahmed Received a B.Sc. degree in software engineering from the University of Al-Iraqia, Iraq. My research currently focuses on the prediction of the most common diseases in Iraq. I'm interested in end-to-end research and development projects, starting with research ending with an application that can be utilized by the beneficiary. He can be contacted at email: ibraheem.azzawi94@gmail.com.

Asst. Prof. Dr. Duraid Y. Mohammed in DSP and Machine Learning from University of Salford Manchester, United Kingdom. My research currently focuses on the Audio Indexing and automated metadata generation. In particular, I am interested in the non-exclusive classification which will be a potential solution where the overlap takes place between speech, music and/or the other events. I have also contributed as a researcher in the Innovation Fellowship of Salford university by developing a toolbox for this issue and won as the best project, which aims to categorize speech and music in both cases (pure and overlapped) and detects the start and end time of each segment in the arbitrary audio soundtrack recording. He can be contacted at email: duraidyehya19@gmail.com.

Prof. Dr. Khamis A. Zidan is the chairman and membership of many scientific committees and supervisory promotions, scientific committees and investigative, test scores, engineering, development the preparatory committees of scientific conferences held in the Ministry of Higher Education and Iraqi universities. He had participated in many training and development courses in the field of computers and information technology and communications inside and outside Iraq. He had received numerous certificates of appreciation in the field of computers and information and communication technology from universities and from international centers inside and outside Iraq. He can be contacted at email: khamis_zidan@yahoo.com.