DOI https://doi.org/10.5281/zenodo.10049652#137

TWIST, 2024, Vol. 19, Issue 2, pp. 110-119

TWIST

Journal homepage: www.twistjournal.net

Designing Model to Improve Hepatitis Prediction by Using Data Mining and Machine Learning Algorithms

Douaa Ibrahim Alsaadi*
Software Computer Engineering Department,
Higher Health Institute, Najaf, Iraq
[**Corresponding author*]

Hind Abdulrazzaq Mohammed Ali
Civil Engineering Department,
University of Technology-Iraq, Baghdad, Iraq

Asaad Ali Muhsen
Electrical Engineering Department,
University of Wasit, Iraq

Abstract
Hepatitis means inflammation of the liver. The liver is an important organ in the human body that processes nutrients, purifies the blood, and fights infections and viruses that attack the body, so it is a vital organ. When the liver is affected, it affects its performance and functions. Some types affect children between the ages of 12-23 months, as well as children from 2-18 years who did not receive the hepatitis vaccine, and some affect ages over 19 years, therefore, studies have previously attempted to pre-diagnose and predict this disease in order to reduce the risk of contracting the disease and minimizing mortality, as they used many data mining and machine learning techniques for classification. In this paper, a model consisting of a set of techniques was used on a hepatitis data set, where appropriate algorithms were selected for the type of data at the classification stage to obtain high accuracy.

Keywords
Data mining, Machine learning, Hepatitis, Classification, Supper Vector Machine

INTRODUCTION

Hepatitis has become common and noticed recently due to the causes of this disease such as excessive alcohol consumption, taking certain medications or due to the presence of a virus, hepatitis is categorized into: B (HBV), C (HCV) and D (HDV). Viral hepatitis affects approximately 400 million people worldwide, with approximately 1.2 million people chronically infected with hepatitis HBV virus in the United States alone and more than 2.3 million people infected with hepatitis HCV, often the disease is not diagnosed because symptoms do not appear early in the disease and take a long time to appear [1]. When the infection becomes chronic, the disease can be diagnosed through some symptoms such as physical fatigue, loss of appetite, etc., or through blood tests, however, it is difficult to diagnose the infection if it is hepatitis or other viral infections of hepatitis [1]. According to the United Nations hepatitis in light of the current Covid-19 crisis, causes one person to die every 30 seconds.

So, a model can be used to predict the disease in its early stages by improving the data set by using data mining techniques such as data cleaning, classification, predication, etc., data mining techniques are efficiently used in biomedical data analysis, to get accuracy in diagnosing hepatitis.

Rapid Miner was used as a data mining tool because it is commonly used and contains a lot of data mining and machine learning algorithms for classification and prediction. Algorithms were used for preprocessing as well as for classification such as KNN, SVM, Decision Tree, Linear Regression, Naïve Bayes, etc., it is possible to obtain high accuracy in the results of prediction and classification.

Several previous studies have attempted to predict the life of HCV patients, in 2012, a Study [2] they proposed a model based on Data Mining Algorithm and Optimal feature selection approach to classify and predict life prognosis for

HCV by calculating the classification accuracy of Support Vector Machine with and without feature identification which helps reduce the number of clinical procedures by 60% This prediction model assists the medical practitioner in effective decision-making process with fewer attributes as the improved accuracy cancels out 83.12% for this model.

In 2020, [1] the researchers proposed a new hybrid algorithm to help predict prognosis in hepatitis patients. This algorithm performs a comprehensive analysis of the classification performance of the proposed hybrid model (LSVM-MLP) against MLP and SVM-MLP implemented in commercially available software and AI-based classifiers from Published studies selected in the research where the proposed model (LSVM-MLP) was more accurate, reliable, and faster computing AI-based classifier to predict survival of patients with chronic hepatitis, the new hybrid algorithm (LSVM-MLP) achieved 100% of the rate of Heartbeat in discriminating between survivors and deceased patients with chronic hepatitis using blood and demographic data.

Some studies have used data mining techniques. Authors [3] in 2017, they provided an overview of data mining techniques for diagnosing hepatitis disease. They provided a comparison between the latest different proposals in terms of pre-treatment steps applied, accuracy achieved, and training time. This study helped implement, develop and evaluate clinical decision support systems, as an accurate diagnosis is the most important factor.

The study showed promising results and can help in making the decision in the initial diagnosis of hepatitis, and the study proved that most of the data were taken from the UCI machine learning repository and not compared to the real clinical data and most of the experimental results were conducted only on WEKA.

In 2021 the authors [4] proposed an improved and guaranteed predictive model for Ada-Boost and Random-Forest techniques in HCV detections. Cross-validation and accuracy factors the empirical analysis of the proposed framework enhances the prediction accuracy by 88.7 percent compared to other existing classifiers for KNN model designs, hybrid model, and Naïve-Bayes designs. They also used ECC-Elliptic-curve cryptography, RSA and AES-algorithmic.

In 2019, some researchers [5] proposed a predictive method for diagnosing hepatitis using groups of neuro-fuzz technology in order to obtain high accuracy, the study investigated the effectiveness of group learning techniques to predict hepatitis disease using several parameters, age, gender, steroid, antiviral drugs, Fatigue, "Malaise", "Anorexia", "Liver Big", "Liver Firm", "Spleen Palpable", "SpidersAscites", "Arices", "Bilirubin", "Alk' Phosphate", "Sgot", "Albumin" And “Proteme and Histology” where ANFIS was relied on as the supervised SOM and unsupervised machine learning techniques, and the method was developed for group learning using several types of membership functions in ANFIS technology They obtained 93.06% through data collected from the UCI website

In 2019 [6], data mining techniques were used in electronic health care where they proposed an electronic health care system to predict diseases using mixed data mining technology. The first step was to optimize the data level by defining the optimal data partition (Popt) for each diseased data set, and the second step explored a general predictive model (integrating C4.5 and PRISM learners) on the detected information for effective diagnosis. For the disease. The experimental results demonstrate that the hybrid model outperforms the basic learners in almost all cases for the initial diagnosis of diseases. The proposed DDSS may act as an electronic pathologist.

In 2019, the authors [7] used a model for detection of hepatitis viruses (A, B, C, and E) based on Random Forest, Near K, and Naïve Bayes Classifier. The model was applied to real data for hepatitis patients, and the results were accuracy for three methods (Naïve Bayes, Random Forest, K-nearest) are 93.2%, 98.6%, 95.8% respectively.

DATA COLLECTION AND METHODOLOGY

In our research we designed a model from seven algorithms (K-NN, SVM, Decision Tree, Linear Regression, Neural Net, Logistic Regression, Naïve Bayes), this algorithm applied in to data taken from UCI Machine Learning dataset Repository this data to hepatitis (Type A, B, C) patients that has 155 instance and 20 attributes, the attributes contain numeric values there are no nominal attributes in this data, so some algorithms that work with the nominal values I used converter from numeric to nominal to work correctly, the attributes which are defined shown in Table 1, and the dataset is shown in Fig. 1.

Class	AGE	SEX	STEROID	ANTIVIRAL	MALAISE	ANOREXIA	LIVER BIG	LIVER FIRM	SPLEEN PA	SPIDERS	ASCITES	VARICES	BILIRUBIN	ALK PHOSI	SGOT	ALBUMIN	PROTIME	HISTOLOGY
2	30	2	1	2	2	2	1	2	2	2	2	2	1	85	18	4	?	1
2	50	1	1	2	2	2	1	2	2	2	2	2	0.9	135	42	3.5	?	1
2	78	1	2	2	2	2	2	2	2	2	2	2	0.7	96	32	4	?	1
2	31	1	?	1	2	2	2	2	2	2	2	2	0.7	46	52	4	80	1
2	34	1	2	2	2	2	2	2	2	2	2	2	1	?	200	4	?	1
2	34	1	2	2	2	2	2	2	2	2	2	2	0.9	95	28	4	75	1
1	51	1	1	2	2	1	2	2	1	1	2	2	?	?	?	?	?	1
2	23	1	2	2	2	2	2	2	2	2	2	2	1	?	?	?	?	1
2	39	1	2	2	2	2	2	1	2	2	2	2	0.7	?	48	4.4	?	1
2	30	1	2	2	2	2	2	2	2	2	2	2	1	?	120	3.9	?	1
2	39	1	1	1	2	2	1	1	2	2	2	2	1.3	78	30	4.4	85	1
2	32	1	2	1	2	2	2	1	2	1	2	2	1	59	249	3.7	54	1
2	41	1	2	1	2	2	2	1	2	2	2	2	0.9	81	60	3.9	52	1
2	30	1	2	2	2	2	2	1	2	2	2	2	2.2	57	144	4.9	78	1
2	47	1	1	1	2	2	2	2	2	2	2	2	?	?	60	?	?	1
2	38	1	1	2	1	1	2	2	2	2	1	2	2	72	89	2.9	46	1
2	66	1	2	2	2	2	2	2	2	2	2	2	1.2	102	53	4.3	?	1
2	40	1	1	2	2	2	2	1	2	2	2	2	0.6	62	166	4	63	1
2	38	1	2	2	2	2	2	2	2	2	2	2	0.7	53	42	4.1	85	2
2	38	1	1	1	2	2	1	1	2	2	2	2	0.7	70	28	4.2	62	1
2	22	2	2	1	2	2	2	2	2	2	2	2	0.9	48	20	4.2	64	1

Fig. 1 Datasets of hepatitis patients

Table 1 Attributes In Dataset

Attributes	Value
Class	die (1), live (2)
Age	numerical value
Sex	male (1), female (2)
Steroid	no (1), yes (2)
Antivirals	no (1), yes (2)
Fatigue	no (1), yes (2)
Malaise	no (1), yes (2)
Anorexia	no (1), yes (2)
Liver Big	no (1), yes (2)
Liver Firm	no (1), yes (2)
Spleen Palpable	no (1), yes (2)
Spiders	no (1), yes (2)
Ascites	no (1), yes (2)
Varices	no (1), yes (2)
Bilirubin	0.39, 0.80, 1.20, 2.00, 3.00, 4.00
Alk Phosphate	33, 80, 120, 160, 200, 250
SGOT	13, 100, 200, 300, 400, 500
Albumin	2.1, 3.0, 3.8, 4.5, 5.0, 6.0
Protime	10, 20, 30, 40, 50, 60, 70, 80, 90
Histology	no (1), yes (2)

This data is raw and noisy, must do pre-processing steps on it to allow data mining techniques to work properly.
So, we used missing value and detect outlier data on raw data to clean data. After that applied algorithms of data mining and machine learning to training and testing data to do classification or predication, all previous works used a model consisting of one or two proposed methods of algorithms and applying them to the data, so the accuracy did not exceed 85%.

But in this model, seven algorithms were applied to improve data accuracy and obtain higher results, Fig 2 shows the steps of data mining process.

Fig. 2 Steps of data mining process

We will explain these four steps in more details below:

a. **Raw Data:** in this step we download the dataset and study this data we saw that data have a lot of missing value and outlier and must improve this data by using data mining algorithms.
b. **Pre-Processing**: in this step need to solve missing and outlier data problem by replace missing value by the mean and correct outlier by using k-NN algorithm. In the replace of missing data that mean all.
c. **Missing attribute values** will be replaced with the mean of all values for the particular characteristic.[8]. In correction outlier step by applying K-NN algorithm, the input consists of the k closest training examples in a data set, and the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). In our model k = 5, then the object is simply assigned to the class of those five nearest neighbors.
d. **Analysis (classification) and validation:** this step after solves and correctness the missing and outlier values in dataset then applied the algorithms on the correctness dataset. Systems that construct classifiers are one of the commonly used tools in data mining. Such systems take as input a collection of cases, each belonging to one of a small number of classes and described by its values for a fixed set of attributes, and output a classifier that can accurately predict the class to which a new case belongs [9].

Here, we will explain the classifier in more details:

- **Supper vector machine (SVM):** it is considered a good classifier because of its high generalization performance without the need to add a priori knowledge, even when the dimension of the input space is very high. The aim of SVM is to find the best classification function to distinguish between members of the two classes in the training data. The metric for the concept of the "best" classification function can be realized geometrically. For a linearly separable dataset, a linear classification function Corresponds to a separating hyperplane f(x) that passes through the middle of the two classes, separating the two. Once this function is determined, new data instance f(xn) can be classified by simply testing the sign of the function f (xn); xn belongs to the positive class if f(xn)>0 [10].
- **Decision Tree:** is widely used classification technique, the methodology used here is Divide and conquer. As there were huge amount of data, first we need to divide those data into sub data. The structure of the decision tree is organized in a manner that it contains the root the topmost node in the tree, Branches which are the internal nodes

and leaf node is one which is not further classified. The internal nodes represent a question and the branch which connects the node denotes the solution and the leaf node tries to predict the solution. It is widely used in decision making process [11].

- **Linear Regression:** is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables. The case of one explanatory variable is called simple linear regression, for more than one, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.
- **Neural Net:** It consists of an interconnected group of artificial neurons and processes an information that enters into it, at the same time finding a relationship between inputted data and the "memory" of the network itself, which accumulates during the network training and, in some networks, when the program is executed. Neural networks have been shown to be very promising systems in many forecasting applications and business classification applications due to their ability to “learn” from the data, their nonparametric nature, and their ability to generalize [12].
- **Logistic Regression**: is one of the most important statistical and data mining techniques employed by statisticians and researchers for the analysis and classification of binary and proportional response datasets. Some of the main advantages of LR are that it can naturally provide probabilities and extend to multi-class classification problems. Another advantage is that most of the methods used in LR model analysis follow the same principles used in linear regression. What’s more, most of the unconstrained optimization techniques can be applied to LR [13] .
- **Naive Bayes:** classifier is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions. The algorithm inclines to perform well and learn rapidly in various supervised classification problems. An advantage of the naive Bayes classifier is that it only requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix [14].

In the figures below shows implementation of some algorithms that used in Rapid Miner program to improve the dataset that used, Fig. 3, Fig. 4, Fig. 5, Fig. 6, Fig. 7, Fig. 8, and Fig. 9.

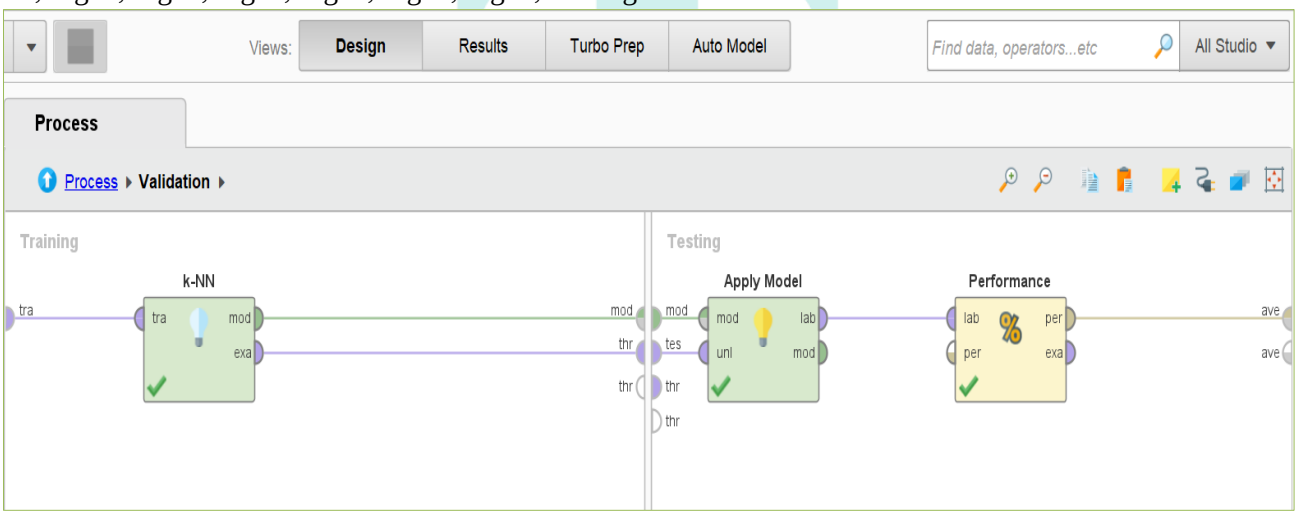

Fig. 3 The implementation of K-NN algorithm

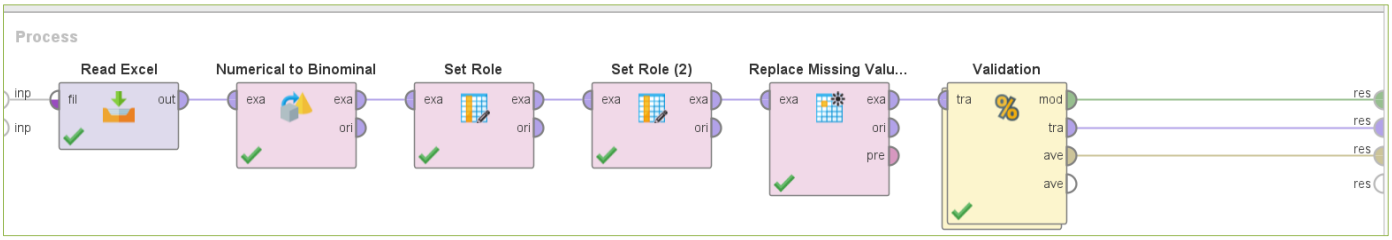

Fig. 4 Reading data from excel sheet and replacing the missing values

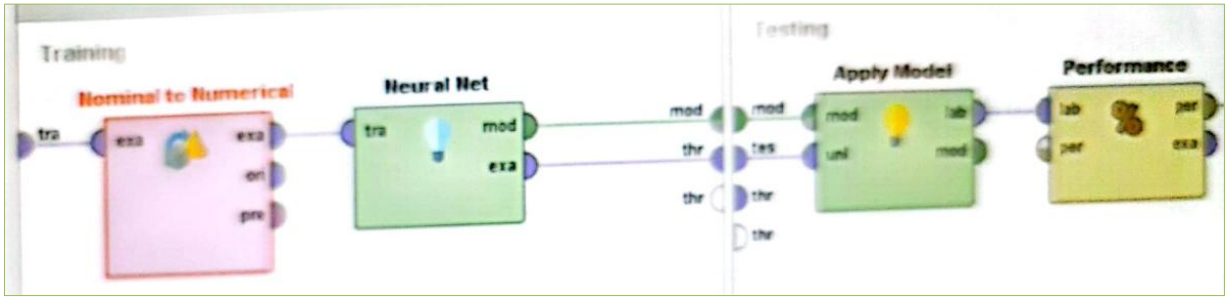

Fig. 5 The implementation of Neural Net algorithm

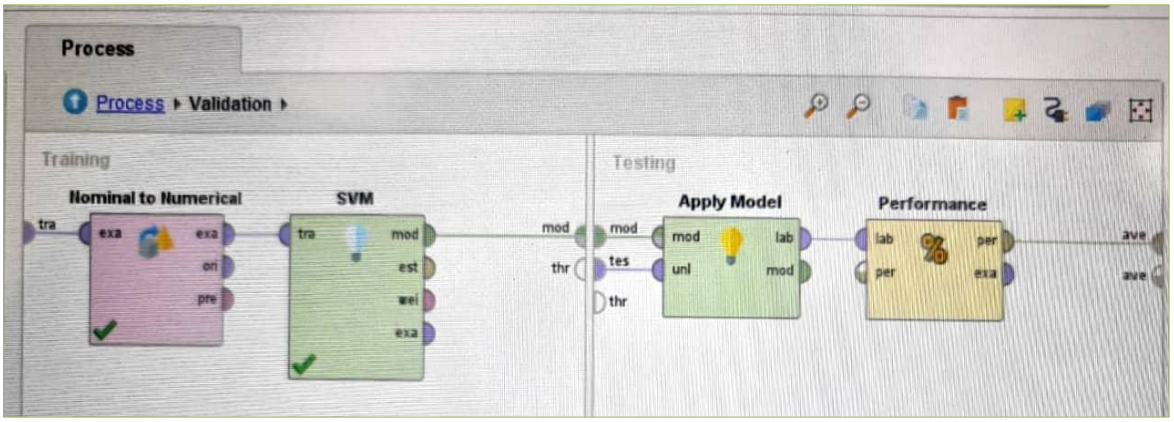

Fig. 6 The implementation of SVM algorithm

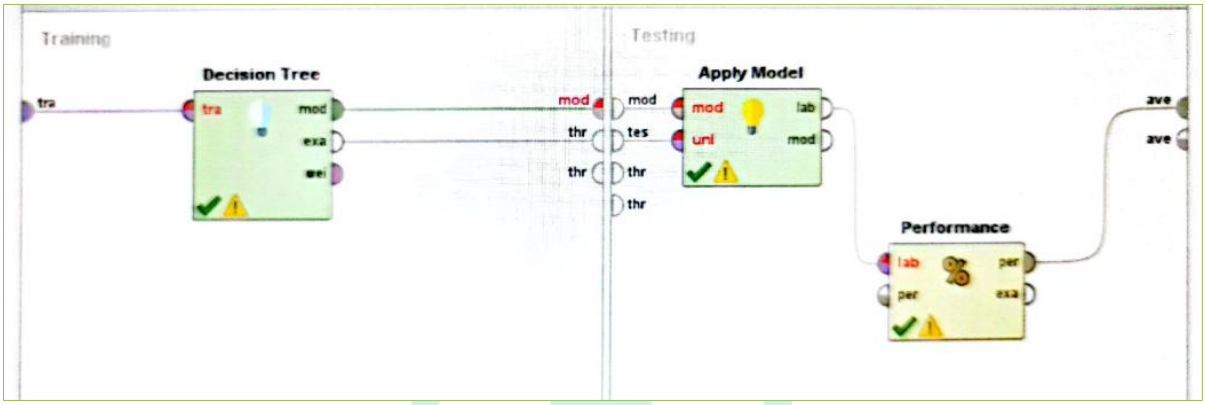

Fig. 7 The implementation of Decision Tree algorithm

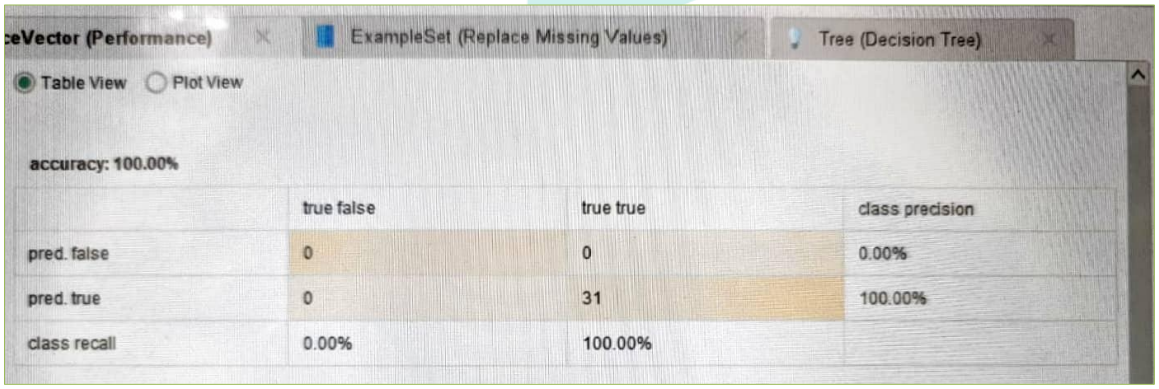

Fig. 8 The accuracy of Decision Tree algorithm

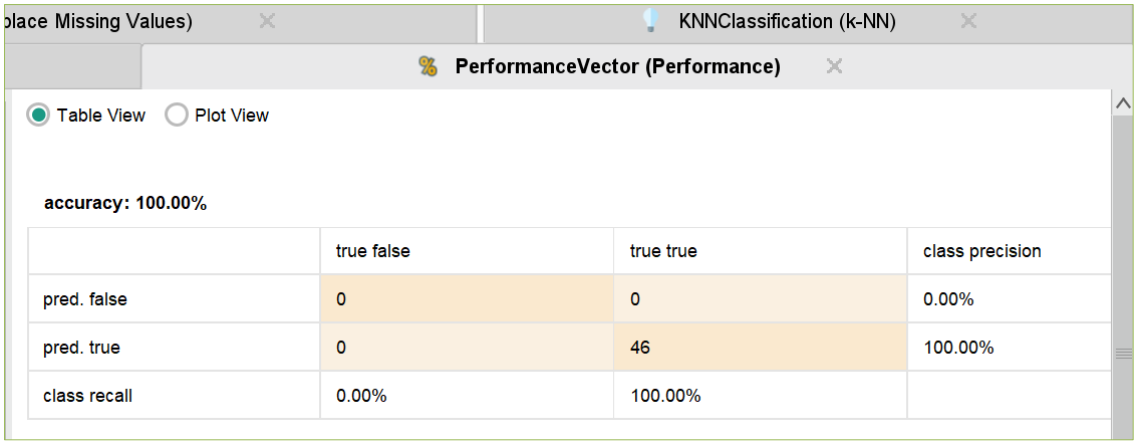

Fig. 9 The accuracy of K-NN algorithm

RESULT

To implement the previously mentioned algorithms (KNN, SVM, DT, LR, NN, logistic regression, Naive Bayes), the Rapid Miner program was used as a data mining tool to deal with the data set, where the basic parameters of accuracy, precision, recall and F-measure were calculated.

Precision is a metric that quantifies the number of correct positive predictions made, and calculated by this formula (1):

$$\text{Precision} = (TP) / (TP + FP) \quad (1)$$

while TP is True Positive, FP is False Positive

Recall is a metric that quantifies the number of correct positive predictions made out of all positive predictions that could have been made, and calculated from formula (2):

$$\text{Recall} = TP / P \quad (2)$$

while TP is True Positive

F-Measure provides a way to combine both precision and recall into a single measure that captures both properties, and calculated from formula (3):

$$\text{F- Measure} = 2 \text{ x Precisio x Recall/ Precision + Recall} \quad (3)$$

Accuracy It's the ratio of the correctly labeled subjects to the whole pool of subjects, and accuracy is the most intuitive one, and its formula (4):

$$\text{Accuracy} = (TP+TN)/ (TP+FP+FN+TN) \quad (4)$$

while TP is True Positive, FP is False Positive, TN is True Negative, and FN is False Negative

Table 2 shows the result of four parameters explained above for all proposed algorithms to solve the missing value and outlier:

Table 2 Results of accuracy, precision, Recall, and F-Measure to the proposed model

Classifier	Precision	Recall	Accuracy	F-Measure
K-NN	100	100	100	100
SVM	100	100	100	100
Decision Tree	100	100	100	100
Linear Regression	100	100	100	100
Neural Net	100	100	100	100
Logistic Regression	100	100	100	100
Naïve Bayes	100	100	100	100

In Fig. 10, the chart of obtained results to solve the missing values and outlier are explained:

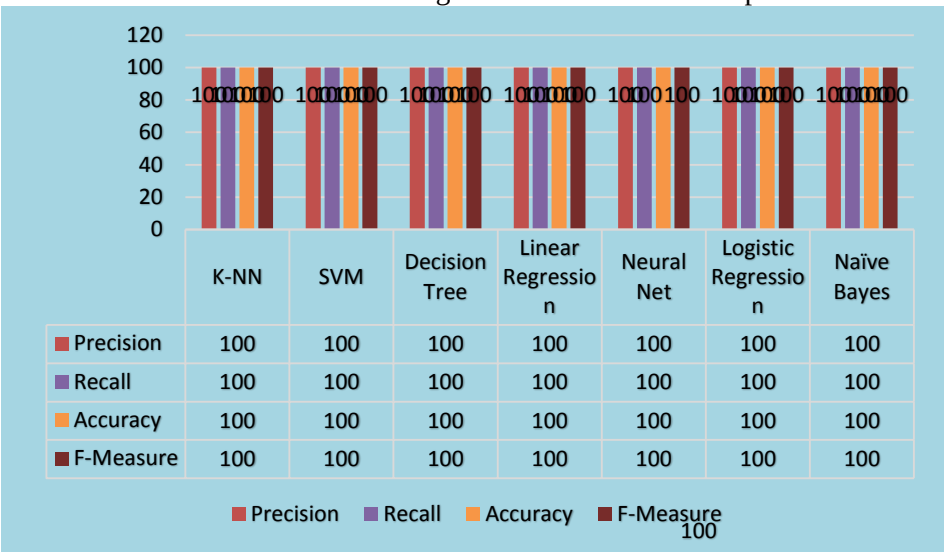

Fig. 10 Chart of results of (accuracy, precision, Recall, F-Measure) to the proposed model

Through the results obtained, the proposed method for improving the data of hepatitis patients proved very high results, as the accuracy was 100%, the precision was 100%, and the recall was 100%. These results are higher than the results of the previous work, where the accuracy did not exceed 85%.

Fig. 11, Fig. 12, Fig. 13, Fig. 14, Fig. 15, Fig. 16, Fig. 17 shows the obtained ROC curves for all classifiers.

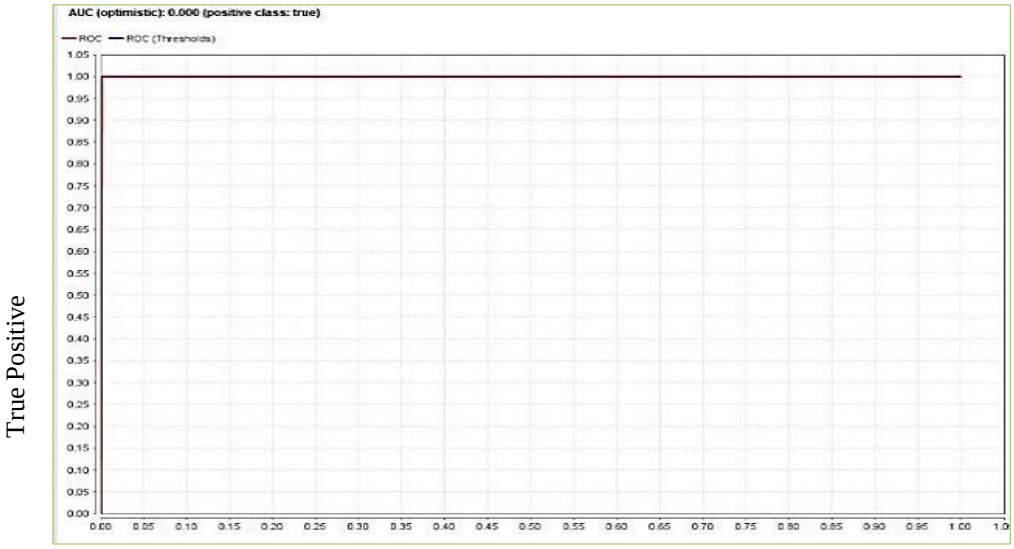

Fig. 11 ROC for the KNN

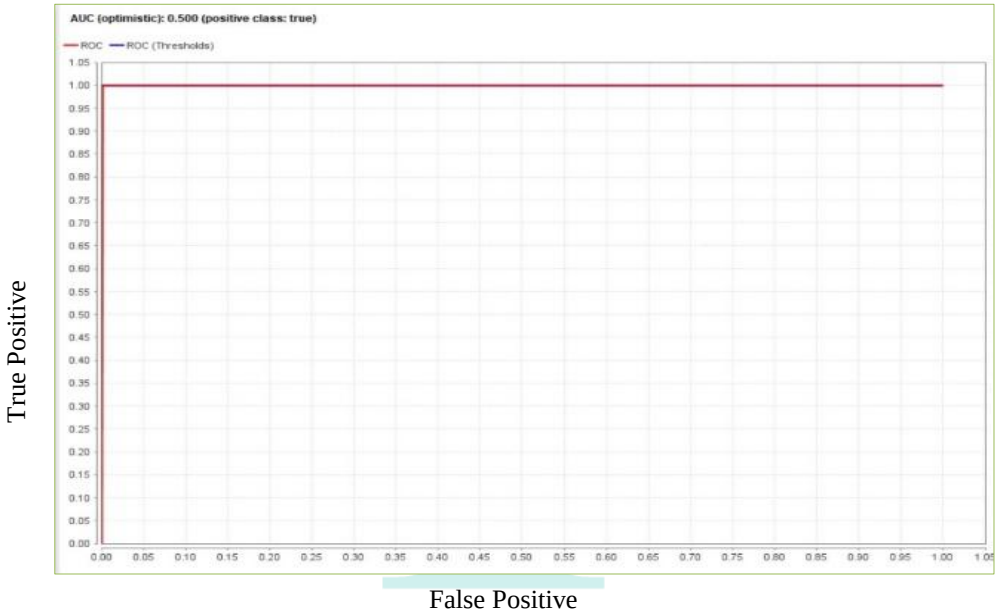

Fig. 12 ROC for the Logistic Regression

Fig. 13 ROC for the Linear Regression

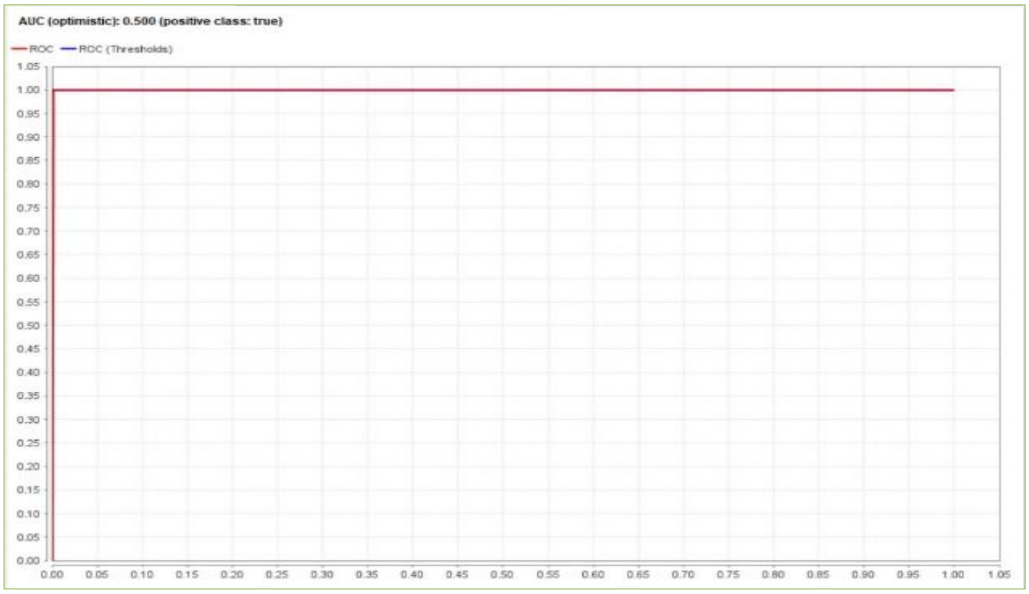

Fig. 14 ROC for the Naïve Bayes

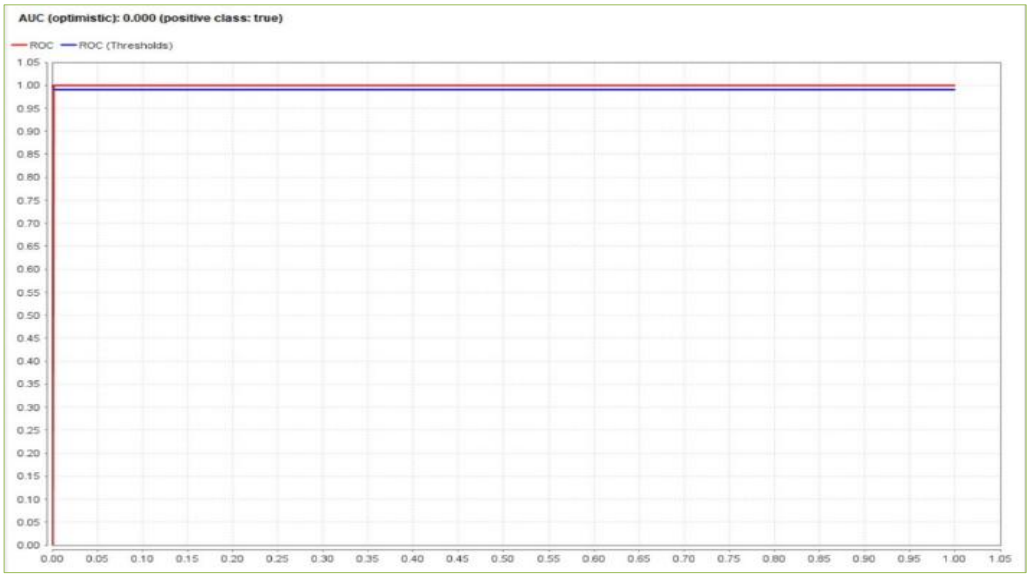

Fig. 15 ROC for the Neural Net

Fig. 16 ROC for the SVM

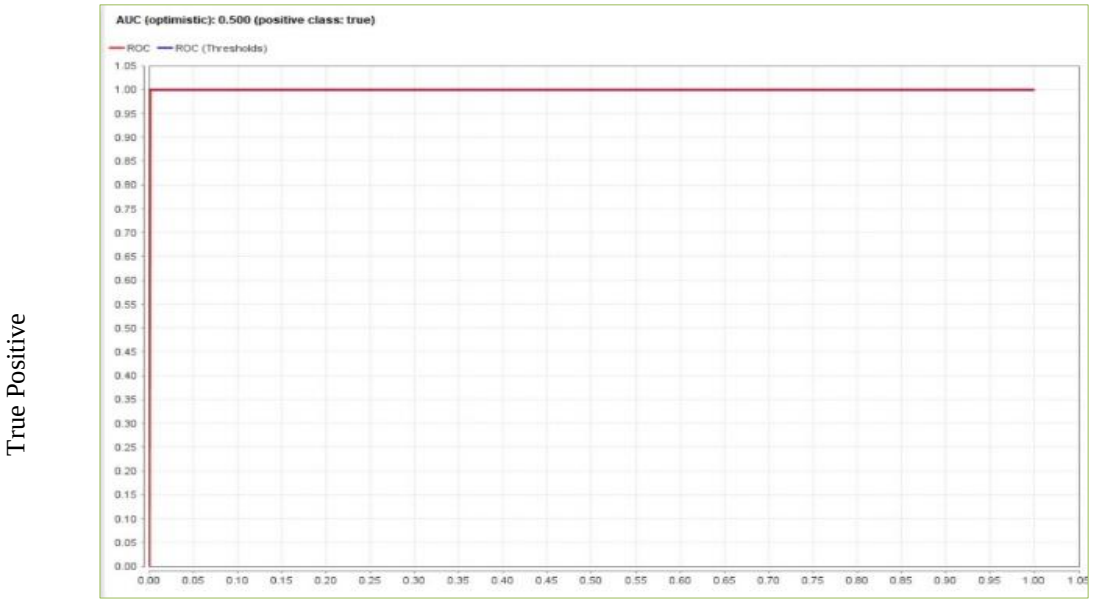

Fig. 17 ROC for the Decision Tree

CONCLUSION

Hepatitis is one of the health problems facing the world that requires expensive treatment. In this study, a proposed method is presented to improve the data of hepatitis patients in terms of solving missing values and processing outliers in the data. In the proposed model, seven data mining algorithms are used, which are among the best algorithms to improve data, train and test the data after cleaning it from missing and outlier values. Previous studies were promising accuracy and could help in decision-making in the initial diagnosis of hepatitis. In proposed model obtained high accuracy 100%, as well as most studies, experimental results were conducted only on WEKA, while in the proposed model, Rapid Miner was used as a data mining tool, which contains more and newer algorithms and techniques, which will make the results more accurate. As for future work, we suggest using the proposed model with diseases other than hepatitis, and also increasing the amount of data.

ACKNOWLEDGEMENT

I would like to express my deep gratitude and appreciation to my research colleagues who contributed a lot to completing this research in the best possible way.

REFERENCES

1. L. Parisi, N. RaviChandran, and M. L. Manaog, "A novel hybrid algorithm for aiding prediction of prognosis in patients with hepatitis," *Neural Comput Appl*, vol. 32, no. 8, pp. 3839–3852, Apr. 2020, doi: 10.1007/s00521-019-04050-x.
2. V. Kumar, "Hepatitis Prediction Model based on Data Mining Algorithm and Optimal Feature Selection to Improve Predictive Accuracy," 2012. [Online]. Available: http://rapid-i.com/content/view/281/225/lang,en
3. *Sudan Conference on Computer Science and Information Technology (SCCSIT)*. IEEE, 2017.
4. D. A. Jadhav, "An enhanced and secured predictive model of Ada-Boost and Random-Forest techniques in HCV detections," in *Materials Today: Proceedings*, 2021, vol. 51, pp. 186–195. doi: 10.1016/j.matpr.2021.05.071.
5. M. Nilashi, H. Ahmadi, L. Shahmoradi, O. Ibrahim, and E. Akbari, "A predictive method for hepatitis disease diagnosis using ensembles of neuro-fuzzy technique," *J Infect Public Health*, vol. 12, no. 1, pp. 13–20, Jan. 2019, doi: 10.1016/j.jiph.2018.09.009.
6. B. K. Sarkar and S. S. Sana, "An e-healthcare system for disease prediction using hybrid data mining technique," *Journal of Modelling in Management*, vol. 14, no. 3, pp. 628–661, Sep. 2019, doi: 10.1108/JM2-05-2018-0069.
7. T. Islam Trishna, G. Imran Hossen Sajal, S. Uddin Emon, S. Kundu, R. Rahman Ema, and T. Islam, "Detection of Hepatitis (A, B, C and E) Viruses Based on Random Forest, K-nearest and Naïve Bayes Classifier."
8. R. Asgarnezhad, K. Ali, and M. Alhameedawi, "MVO-Autism: An effective pre-treatment with High Performance for Improving Diagnosis of Autism Mellitus Keywords: Data mining Pre-processing Machine learning techniques Autism mellitus," *Journal of Electrical and Computer Engineering Innovations*, vol. 10, no. 1, pp. 209–220, 2022, doi: 10.22061/JECEI.2021.8109.480.
9. Rakesh Bhatiya, G. Ravi Kumar, R. Kumar, and R. Verma, "Classification Algorithms for Data Mining: A Survey An Efficient Predict ion of Breast Cancer Dat a using Dat a Mining Techniques Classification Algorithms for Data Mining: A Survey."
10. S. N. #1 and E. Ramaraj, "Classification algorithm in Data mining: An Overview," *International Journal of P2P Network Trends and Technology*, vol. 4, 2013, [Online]. Available: http://www.ijpttjournal.org
11. S. Umadevi and K. S. Jeen Marseline, "A Survey on Data Mining Classification Algorithms."
12. P. Gaur, "Neural Networks in Data Mining." [Online]. Available: www.ijecse.org
13. M. Maalouf, "Logistic regression in data analysis: an overview," 2011
14. V. S and D. S, "Data Mining Classification Algorithms for Kidney Disease Prediction," *International Journal on Cybernetics & Informatics*, vol. 4, no. 4, pp. 13–25, Aug. 2015, doi: 10.5121/ijci.2015.4402

BIOGRAPHIES

Douaa Alsaadi was born in Baghdad, in January 1987. She got B.Sc in Computer Engineering from Almustansyria University of Baghdad in 2009. She got M.Sc from Azad Isfahan University. She is teaching in Higher Health Institute in An-Najaf Al-Ashraf, and got a certificate as a trainer in computer science from health ministry in 2015. Douaa Alsaadi has experience in website designing in HTML5, programing in python and Arduino IDE. She is interesting in programming application, sensors, IoT, database.

Hind Abdulrazzaq Mohammed Ali was born in Baghdad. She got PhD in evolutionary algorithm from University of Bourgogn France Comte, Montbéliard, France in 2018. She obtained M.Sc in Artificial Intelligence from Iraqi Commission for Computers and Informatics, Institute for Post Graduate Studies in Informatics, Baghdad, Iraq, in 2001. Dr. Ali got B.Sc in Computer Science from University of Technology, Baghdad, Iraq. She is experienced in Java Script, Visual Basic.Net. She is serving as a Lecturer in Civil Engineering Department, University of Technology, Baghdad, Iraq. Her Research fields are Evolutionary Algorithm, Rich Problems, Smart Cities, Sensors, IOT, Classifier Learning Systems, Optimization Systems.

Asaad Ali Muhsen was born in 1986 in Iraq. He is currently pursuing PhD in Power Systems from Cukurova University, Adana, Turkey. He has secured Masters in Electrical Engineering from College of Engineering & Technology, University of Wasit. He is currently serving as Assistant Lecturer in Electrical Engineering Department, College of Engineering, Wasit University, Wasit, Iraq. His main research interests are power quality, FACTS, power electronics, Power system operation and control, application of intelligent control techniques.