



# A systematic method for diagnosis of hepatitis disease using machine learning

Ravi Kumar Sachdeva<sup>1</sup> · Priyanka Bathla<sup>2</sup> · Pooja Rani<sup>3</sup> · Vikas Solanki<sup>1</sup> · Rakesh Ahuja<sup>1</sup>

Received: 12 June 2022 / Accepted: 22 November 2022 / Published online: 6 January 2023  
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2023

## Abstract

Hepatitis is among the deadliest diseases on the planet. Machine learning approaches can contribute toward diagnosing hepatitis disease based on a few characteristics. On the UCI dataset, authors assessed distinct classifiers' performance in order to develop a systematic strategy for hepatitis disease diagnosis. The classifiers used are support vector machine, logistic regression (LR), K-nearest neighbor, and random forest. The classifiers were employed without class balancing and in conjunction with class balancing using SMOTE strategy. Both studies, classification without class balancing and with class balancing, were compared in terms of different performance parameters. After adopting class balancing, the efficiency of classifiers improved significantly. LR with SMOTE provided the highest level of accuracy (93.18%).

**Keywords** Hepatitis · Machine learning · SMOTE · Support vector machine · LR

## 1 Introduction

One of the most thoughtful issues in medicine is disease diagnosis. Disease analysis refers to classifying a disease depending on its symptoms [1]. Hepatitis is among the top life-threatening diseases. The liver cells called hepatocytes become swollen and get damaged because of hepatitis which affects the working of the liver. Hepatitis is of two types: acute and chronic hepatitis. Acute hepatitis is characterized by relatively minor damage to liver cells called Hepatocytes as compared to chronic hepatitis. The swelling of the liver

lasts for more than 6 months in chronic hepatitis as compared to acute hepatitis, where it lasts normally for 1 or 2 months [2]. Table 1 shows the categorization of different types of hepatitis as acute and chronic [3].

It is important to be concerned about all 5 varieties of hepatitis since they are the source of outbreaks and numerous fatalities. In particular, types B and C are the most significant causes of cirrhosis of the liver and cancer, as well as chronic diseases in millions of individuals around the world [4]. The most prevalent viral agents of hepatic illnesses worldwide are the viruses of hepatitis B and hepatitis C, whose infections initially start as acute infections and develop into chronic infections [5]. According to data from the World Health Organization (WHO), 130–150 million individuals worldwide have a chronic hepatitis C disease. Hepatitis is a viral infection that often results in an inflammatory disease of the liver and is responsible for approximately 1.5 million fatalities per year in the world [6].

Hepatitis can be caused by circumstances, including drinking too much alcohol, adverse drug events, and bacterial and viral infection [7]. If hepatitis is detected early, the chance of successful recovery increases significantly. Using machine learning to diagnose hepatitis disease will bring efficiency and assist inexperienced clinicians. Human error may be because of exhaustion, inexperience, and thoughtfulness. Detecting hepatitis virus remains one of the toughest

✉ Vikas Solanki  
solanki.vikas1@gmail.com  
Ravi Kumar Sachdeva  
ravisachdeva1983@gmail.com  
Priyanka Bathla  
priyankabathla85@gmail.com  
Pooja Rani  
pooja.rani@mmumullana.org  
Rakesh Ahuja  
ahuja2305@gmail.com

<sup>1</sup> Chitkara University Institute of Engineering and Technology,  
Chitkara University, Rajpura, Punjab, India

<sup>2</sup> Chandigarh University, Gharuan, Mohali, Punjab, India

<sup>3</sup> MMICTBM, Maharishi Markandeshwar (Deemed to be  
University), Mullana, Ambala, Haryana, India

**Table 1** Categorization of different types of hepatitis

Hepatitis type	Category
A	Acute
B	Chronic
C	Chronic
D	Chronic
E	Acute

challenges for untrained practitioners [8, 9]. The motivation behind this work is to reduce human error in diagnosis.

For disease classification, supervised machine learning approaches are applied. In supervised machine learning techniques, researchers first train the computer using a specified dataset and then apply test data to assess the model accuracy produced during training. In this paper, a method has been proposed for the diagnosis of hepatitis disease using machine learning techniques. A simulation environment was developed in python to test the method. The authors utilized the hepatitis dataset available on the UCI repository to test the method. The method has been proposed to carry the diagnosis accuracy of existing methods to the next level.

The following are the primary contributions of the presented research work:

1. Unbalancing in the data has been handled to improve accuracy.
2. A method for diagnosis of hepatitis disease has been proposed by determining the most effective classifier out of SVM, LR, KNN, and RF.
3. The effectiveness of the proposed method has been assessed and contrasted with that of earlier studies.

The remainder of the research paper is structured as follows. A summary of the relevant work carried out by prior researchers is provided in Sect. 2. Section 3 introduces the materials and research techniques. Section 4 discusses the results. Section 5 provides a conclusion and perspectives on the future.

## 2 Related work

Below is the work of different researchers who have done similar work. Based on their findings, the researchers offered various studies for hepatitis classification.

Anto and Chandramathi have proposed a system that applied hybrid GA-SA and SVM for predicting hepatitis. The hybrid GA-SA served to identify the dataset's maximum meaningful features and optimize SVM. Classification accuracy of SVM increased as the ideal values of C and gamma

were found. The authors achieved 87% accuracy on the UCI dataset of hepatitis [10].

Anggraeny et al. have proposed using a statistical BN to diagnose hepatitis and the ReliefF algorithm to discover the main symptoms. The ReliefF algorithm yielded four key indications. The authors used these key symptoms to build a BN. The dataset was the UCI dataset, and the accuracy achieved was 76.8% [11].

Bhargav et al. have applied LR, DT, SVM, and NB techniques, for classification of hepatitis disease using UCI dataset. The authors concluded that LR provided the highest accuracy: 87.17% [12].

Hassan and Shah have compared six different machine learning algorithms: LR, DT, NB, KNN, SVM, and RF for classification of hepatitis disease using UCI dataset. The authors have shown using experimental results that RF has given the highest accuracy (85%) for the prediction of hepatitis disease [13].

Nilashi et al. have proposed predicting hepatitis by applying a group of neuro-fuzzy techniques. The authors suggested using ANFIS and SOM techniques. Using NIPALS increased SOM's accuracy. The authors used different types of membership functions in the ANFIS technique to construct the method for ensemble learning. The authors obtained an accuracy of 93.06% when tested on the UCI dataset [6].

Bayrak et al. have used NB, LR, and J48 decision tree for the classification of hepatitis. The authors used the cross-validation technique with  $k = 10$ . The authors applied feature selection methods with before mentioned classifiers and compared the performance of different classifiers on different performance matrices. NB has given the highest accuracy, i.e., 84.51% without feature selection and 88.38% with feature selection [4].

Nivaan and Emanuel proposed a model based on LR for hepatitis prediction. The authors worked on the UCI dataset related to hepatitis disease. As per the authors, they obtained 83.33% accuracy after applying the LR model [14].

Alfyani and Muljono compared the performance of NB and KNN classifiers for hepatitis disease classification using the UCI hepatitis dataset. The authors concluded that NB performs better than KNN with an accuracy level of 74.19% [15].

Basarslan et al. applied attribute selection techniques based on correlation and fuzzy rough attribute selection approaches while presenting an approach for the classification of hepatitis disease. The authors worked on dataset obtained from UCI repository. The authors used KNN, RF, NB, and LR techniques for classification and evaluated the performance of classification for every applied classifier. RF gave the highest accuracy (84.9%) among all applied classifiers [16].

Peng et al. have presented a framework founded on artificial intelligence for hepatitis prediction. The authors have

presented a computer-based diagnosis system for the same. The authors have worked on the hepatitis dataset available on UCI. They have worked with different models like LR, DT, KNN, XGboost, SVM, and RF. The authors concluded that RF has achieved the highest accuracy: 91.9% [17].

Panda et al. have presented a study on various feature reduction and categorization algorithms for predicting hepatitis. The authors have applied KNN, LR, NB, DT, SVM, and RF classifiers for the prediction of hepatitis disease. The hepatitis dataset available on UCI was utilized for experiment purposes. Chi-square and Boruta procedures were used for feature reduction. The authors concluded that LR, Kernel SVM, and KNN performed best with 90.32% accuracy [18].

Mijwil et al. have presented a study on the use of machine learning techniques on various medical datasets. The authors worked on medical datasets related to different diseases available on the UCI repository like breast cancer, hepatitis, chronic kidney disease, COVID-19, immunotherapy, and cryotherapy. The authors worked by applying SVM, C5.0 DT, KNN, and RF classification algorithms. The authors have shown that RF worked best for the hepatitis dataset with 92.88% accuracy [19].

Nayeem et al. have used KNN, NB, SVM, multilayer perceptron (MLP), and RF for prediction of hepatitis disease. The authors carried out their experiments on the hepatitis dataset available on UCI. The authors determined that RF gave the best accuracy, i.e., 92.41% [20].

Yusuf and Akande have worked on hepatitis prediction by applying techniques of machine learning. The authors applied SVM, Gaussian NB, LR, DT, KNN, and MLP techniques for predicting the disease. The authors performed the experiments on the UCI hepatitis dataset. The authors obtained 87% accuracy as the best accuracy by applying MLP and LR classifiers [21].

Das et al. have worked on classifying hepatitis disease by applying velocity enhanced whale optimization algorithm. The authors obtained 92.5% accuracy while working on the hepatitis dataset available on UCI [22].

Different researchers have used various techniques to enhance the effectiveness of classifying models for diagnosing hepatitis. The accuracy of classification methods varies. There is still scope for improving accuracy by using the optimum combination of data preprocessing techniques and classification algorithms. After experimenting with different classifiers, the investigators should choose the classification method that behaves the best on performance indicators. The common issue identified while studying the work of different researchers was the inefficient or absence of handling of missing information. Accuracy might be increased even further by comprehending the effects of missing values. Applying class balancing before using classifiers can act as a critical step in enhancing the performance of diagnosis and classification models.

**Table 2** Features of the hepatitis dataset

S. No.	Feature
1	Age
2	Sex
3	Steroid
4	Antivirals
5	Fatigue
6	Malaise
7	Anorexia
8	Liver big
9	Liver firm
10	Spleen palpable
11	Spiders
12	Ascites
13	Varices
14	Bilirubin
15	Alk phosphate
16	Sgot
17	Albumin
18	Protime
19	Histology

### 3 Materials and methods

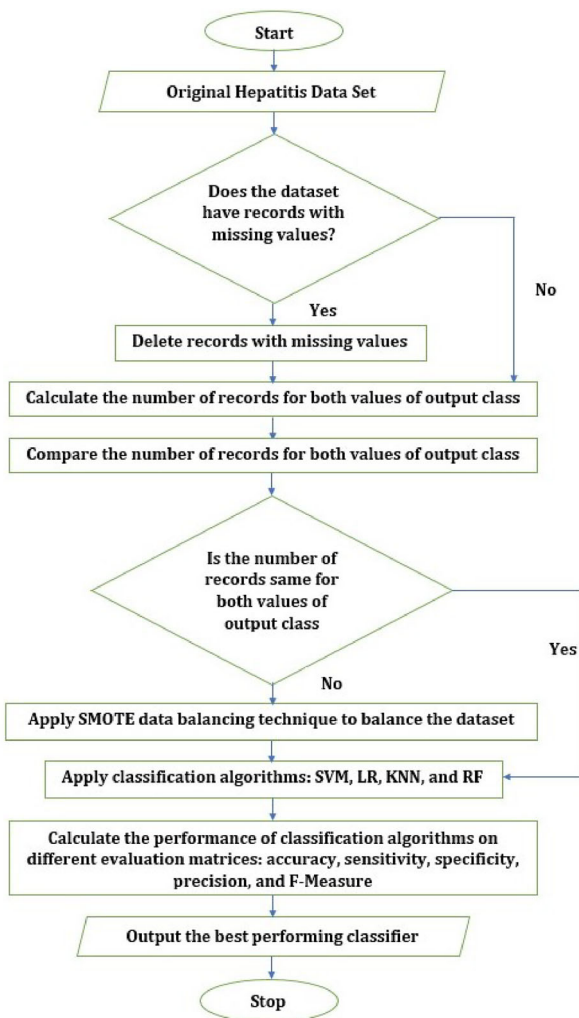
#### 3.1 Dataset

The authors carried out trials with the help of the hepatitis dataset available on UCI. There are nineteen prediction parameters and one class parameter in the dataset. The class attribute in the dataset has two values: 1 for Die and 2 for Live. There are 155 records in all [23]. Table 2 shows the prediction attributes of the dataset.

#### 3.2 Hepatitis disease diagnosis method

Figure 1 depicts flowchart of the proposed hepatitis disease diagnosis method. Figure 2 shows the proposed method. The authors used class balancing to preprocess the dataset before using classification algorithms. After removing records with unknown values of some fields there were 13 occurrences of Die and 67 occurrences of Live, so the dataset was unbalanced. The authors applied SMOTE for class balancing. There were 67 instances of both classes created. After class balancing, SVM, LR, KNN, and RF classifiers were used to diagnose hepatitis disease.

Below is the pseudocode used for the implementation of the method shown in Fig. 2.



**Fig. 1** Flowchart for diagnosis of hepatitis disease using machine learning



**Fig. 2** Proposed hepatitis disease diagnosis method

Pseudocode: Hepatitis Disease Classification

Input: Hepatitis dataset  $HD = (r_1, r_2, \dots, r_n)$

Output: Classification results

Step 1: Start

Step 2: Import the packages required

Step 3: Read the Hepatitis dataset using the Pandas

`#pd.read_csv("hepatitisdataset.csv", header = 0)`

Step 4: Remove records with missing data from HD

`#data = data.dropna()`

Step 5: Apply SMOTE on HD

Step 6: Apply classification using SVM, LR, KNN, and RF classifiers

Step 7: Import the Stratified K-fold module and apply the tenfold validation method

Step 8: Calculate the accuracy, sensitivity, specificity, precision, and F-measure of all classifiers

Step 9: End

### 3.3 Classifiers

The proposed method has used four classifiers to classify hepatitis disease:

**Support vector machines (SVM)** In an SVM classifier, each item in the dataset is drawn as a point on a space with dimensions equal to the number of features, and then a hyperplane is located that divides the data item's classes across its edges. Each feature in the  $n$ -dimensional space represents a coordinate [24].

**Logistic regression (LR)** LR is an algorithm for supervised classification that addresses classification problems. It is used when the value to be predicted is categorical in nature. Logistic regression allows the researcher to estimate the likelihood that one belongs to the given class [25]. It uses a regression model for classification. Capacity for nonlinear problem solving, susceptibility to overfitting, computational efficiency, and simplicity of implementation are some of the important features of LR [26].

**K-nearest neighbor (KNN)** When using the KNN technique to predict an output class for a set of input data, the  $k$ -nearest neighbors from the training dataset drawn on the feature space are taken into account. The most widely employed distance metric is Euclidean distance. For categorization, freshly entered data is given the class label which most neighbors among the  $k$ -nearest neighbors are having. Before using KNN, missing values must also be addressed because it is impossible to calculate the distance from a point that represents a data item with missing values. Since we must calculate the separation between each test case and each training sample, the computation cost is very large [27, 28].

**Table 3** Parameters for different classifiers

Classifier	Parameters
SVM	C = 100, gamma = 0.001, Kernel = 'rbf'
LR	Max_iter = 1000, solver = 'lbfgs', verbose = 0
KNN	n_neighbors = 1
RF	n_estimators = 10, random_state = 1, criterion = 'gini', max_depth = 'None', min_samples_split = 2

*Random forest (RF)* RF has multiple decision trees. Each of them is trained using a random sample taken from the entire dataset. After training is complete, every decision tree produces a separate set of outcomes. As a result, the model's forecast can be made using a majority vote [29].

### 3.4 Evaluation matrices

Performance was measured using the following evaluation matrices. The evaluation matrices have been calculated based on correctly classified positive classes called true positives, correctly classified negative classes called true negatives, incorrect classifications done for positive classes called false positives, and incorrect classifications done for negative classes called false negatives [30, 31].

- **Accuracy** The percentage of samples that are correctly classified across all samples is known as accuracy.

$$\text{Accuracy} = \frac{\text{Correctly classified samples}}{\text{Total samples}} * 100$$

- **Sensitivity** Sensitivity defines the system's ability for accurate positive predictions.

Sensitivity

$$= \frac{\text{True positive classifications}}{\text{True positive classifications} + \text{False negative classifications}} * 100$$

**Table 4** Classifiers' performance before class balancing

Classifier	Accuracy	Sensitivity	Specificity	Precision	F-measure
SVM	83.75	46.15	91.04	50.00	48.00
LR	85.00	38.46	94.02	55.55	45.45
KNN	81.25	46.15	88.05	42.85	44.44
RF	85.00	38.46	94.02	55.55	45.45

- **Specificity** The system's ability for accurate negative predictions is known as specificity.

Specificity

$$= \frac{\text{True negative predictions}}{\text{True negative predictions} + \text{False positive predictions}} * 100$$

- **Precision** The percentage of correctly categorised positive samples to all projected positive samples is called precision.

Precision

$$= \frac{\text{True positive predictions}}{\text{True positive predictions} + \text{False positive predictions}} * 100$$

- **F-measure** The harmonic mean of sensitivity and precision is known as the F-measure.

$$\text{F - measure} = 2 * \frac{\text{Sensitivity} * \text{Precision}}{\text{Sensitivity} + \text{Precision}}$$

## 4 Results and discussion

During the research, the authors used four classifiers: SVM, LR, KNN, and RF, to diagnose hepatitis disease. Classification experiments without class balancing and classification experiments with class balancing were both run on the UCI dataset. Firstly, the classification techniques were applied without class balancing. Secondly, the class balancing technique was used before applying classification techniques to diagnose the disease. Table 3 shows values of different parameters used while applying different classifiers. Optimal parameters have been selected by experimenting with several parameters.

The performance of the classifiers with and without class balancing was compared. The authors used the tenfold validation procedure to verify the results. Table 4 includes the values of performance characteristics of classifiers without applying class balancing. SVM achieved 83.75% accuracy, 46.15% sensitivity, 91.04% specificity, 50.00% precision, and 48.00% F-measure. LR obtained 85.00% accuracy, 38.46% sensitivity, 94.02% specificity, 55.55% precision,



**Table 5** Classifiers' performance after class balancing

Classifier	Accuracy	Sensitivity	Specificity	Precision	F-measure
SVM	90.43	91.04	89.55	89.70	90.37
LR	93.18	98.50	88.05	89.18	93.61
KNN	85.93	94.02	77.61	80.76	86.89
RF	92.52	95.52	89.55	90.14	92.75

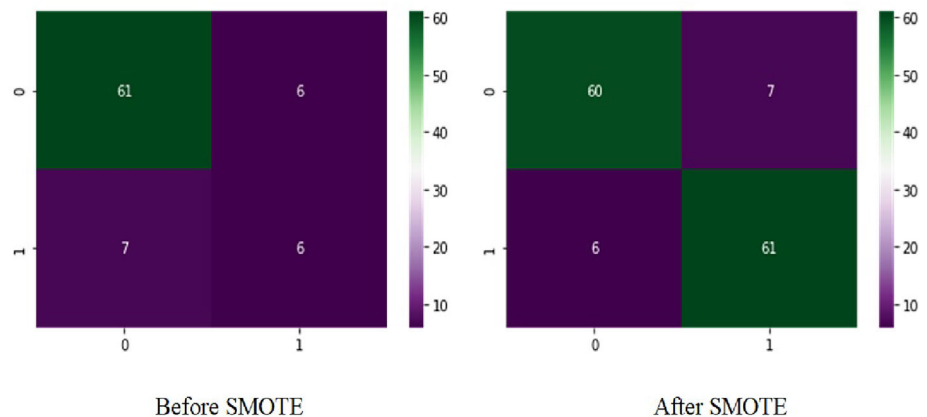
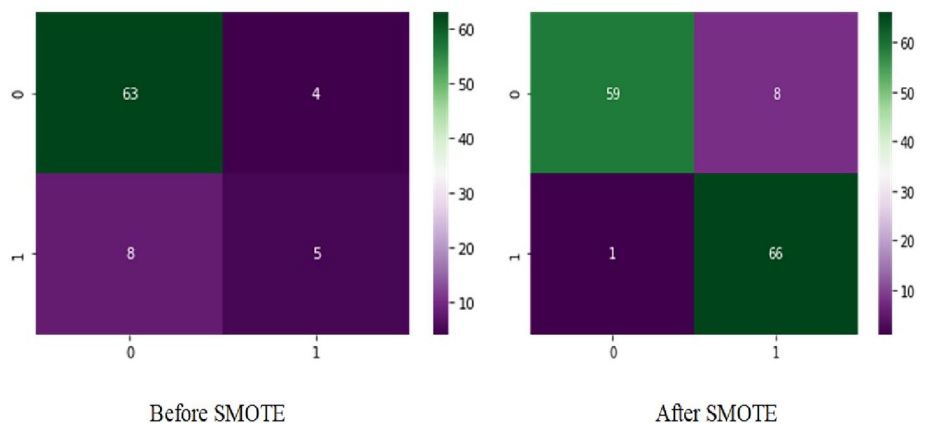
and 45.45% F-measure. KNN got accuracy of 81.25%, sensitivity of 46.15%, specificity of 88.05%, precision of 42.85%, and F-measure of 44.44%. RF achieved 85.00% accuracy, 38.46% sensitivity, 94.02% specificity, 55.55% precision, and 45.45% F-measure.

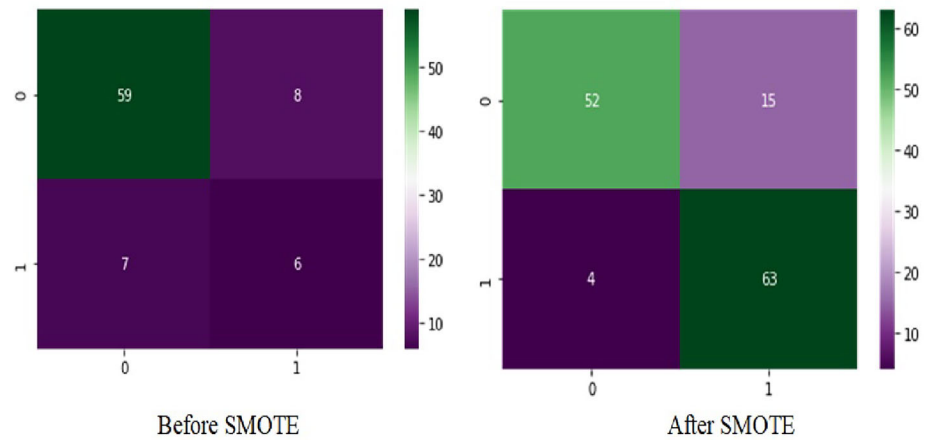
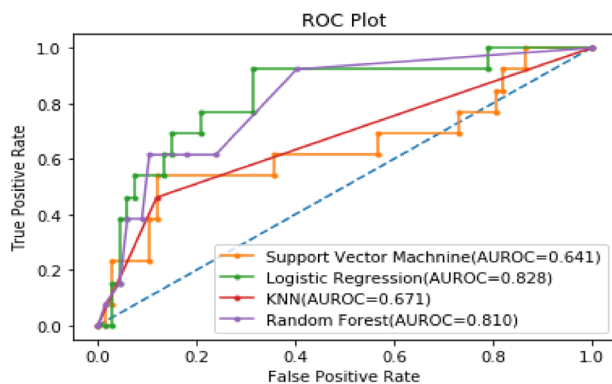
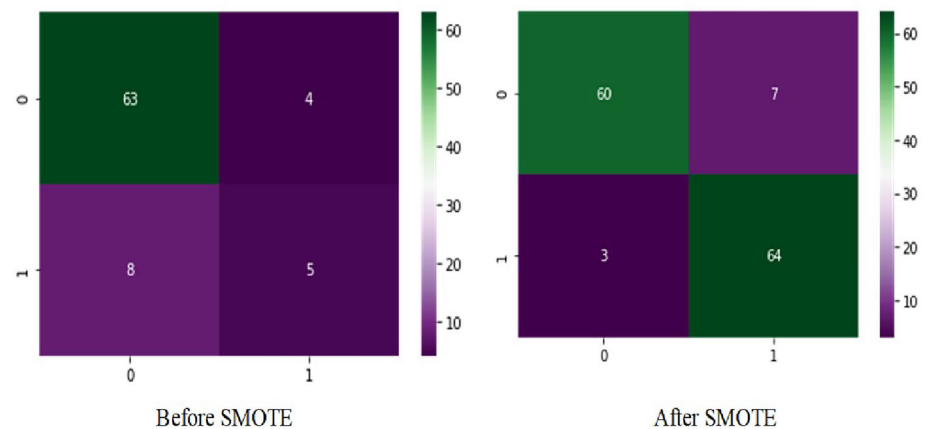
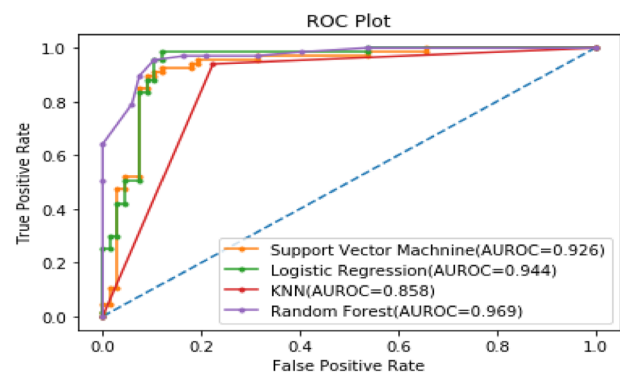
Table 5 displays the results of the performance characteristics of classifiers following class balancing. Applying class balancing improved the classifiers' performance. Table 5 shows the values of performance characteristics of classifiers after applying class balancing. SVM achieved 90.43% accuracy, 91.04% sensitivity, 89.55% specificity, 89.70% precision, and 90.37% F-measure. LR obtained 93.18% accuracy, 98.50% sensitivity, 88.05% specificity, 89.18%

precision, and 93.61% F-measure. KNN got accuracy of 85.93%, sensitivity of 94.02%, specificity of 77.61%, precision of 80.76%, and F-measure of 86.89%. RF achieved 92.52% accuracy, 95.52% sensitivity, 89.55% specificity, 90.14% precision, and 92.75% F-measure. LR saw the most significant improvement in values of performance parameters.

Figures 3, 4, 5, and 6 illustrate the confusion matrix of SVM, LR, KNN, and RF, respectively, before and after applying SMOTE for class balancing.

The receiver operating characteristic (ROC) curve is another useful performance statistic for demonstrating the

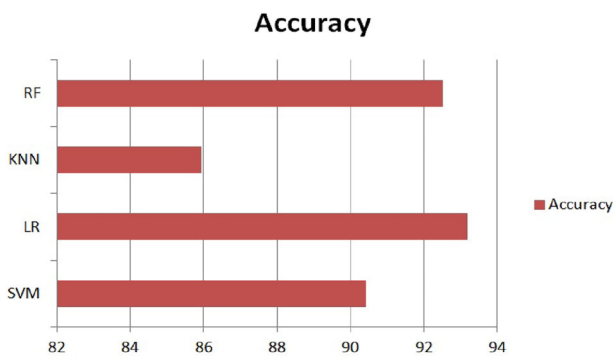
**Fig. 3** SVM confusion matrix before and after SMOTE**Fig. 4** LR confusion matrix before and after SMOTE

**Fig. 5** KNN confusion matrix before and after SMOTE**Fig. 6** RF confusion matrix before and after SMOTE**Fig. 7** ROC curve before SMOTE**Fig. 8** ROC curve after SMOTE

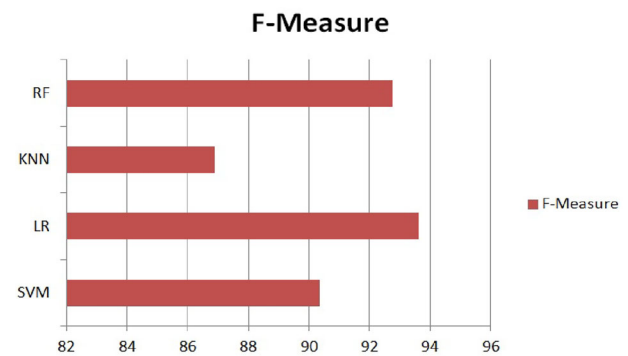
performance of classification algorithms. Ability of classifiers to distinguish between two classes is indicated by the performance parameter known as area under the ROC curve (AUROC). The AUROC value is between 0 and 1. The optimum value of AUROC is 1, and if the classifier's AUROC score is less than 0.5, it cannot classify data. Figure 7 illus-

trates the ROC curve without SMOTE, and Fig. 8 shows the ROC curve with SMOTE. Results show that the proposed method enhanced the AUROC value.

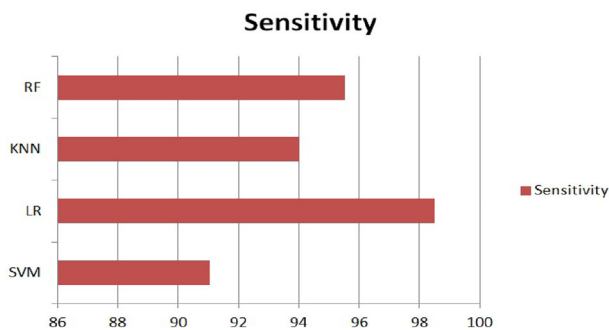
Figure 9 shows the comparison of the classifiers' accuracy after applying SMOTE for class balancing. The accuracy of LR with SMOTE was 93.18%, which was the highest of all



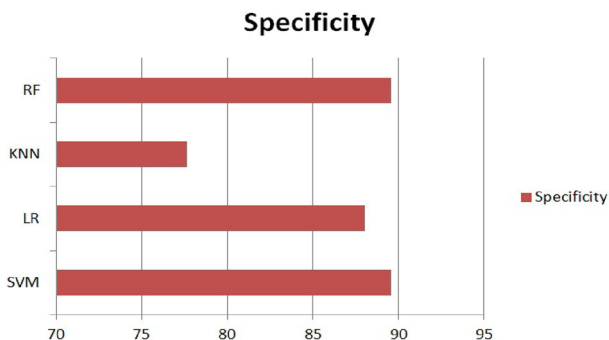
**Fig. 9** Comparison of accuracy of classifiers with SMOTE



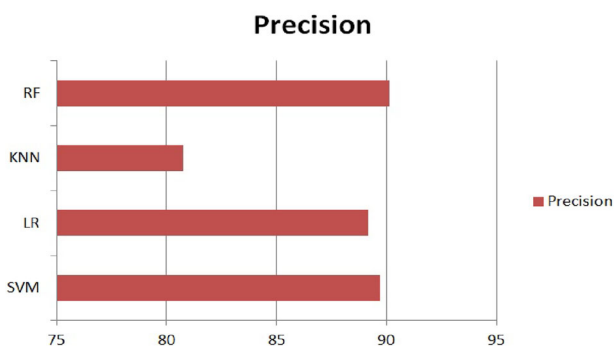
**Fig. 13** Comparison of classifiers' F-measure after applying SMOTE



**Fig. 10** Comparison of sensitivity of classifiers with SMOTE



**Fig. 11** Comparison of specificity of classifiers with SMOTE



**Fig. 12** Comparison of classifiers' precision after applying SMOTE

the classifiers applied. So, doctors may use the method (LR with SMOTE) to classify hepatitis effectively.

Figures 10, 11, 12, and 13 show a comparison of sensitivity, specificity, precision, and F-measure of classifiers after applying SMOTE.

Table 6 compares the recommended method's accuracy to earlier studies reported in Sect. 2.

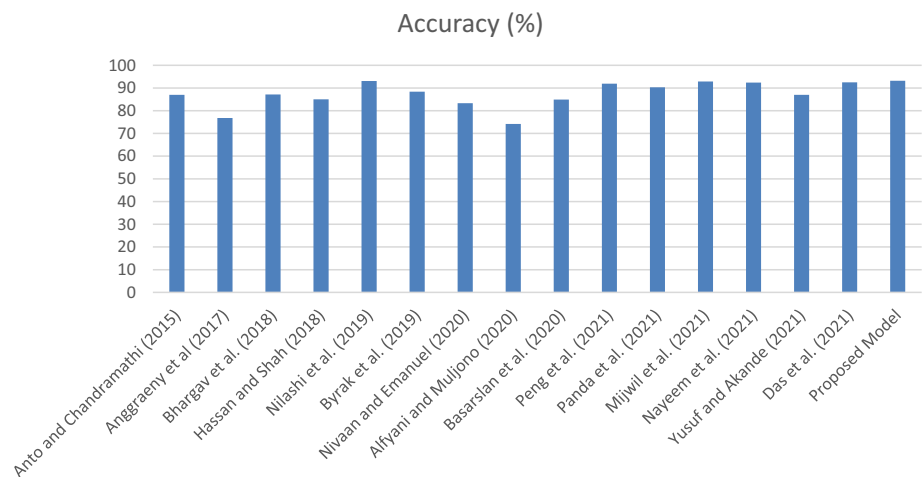
Figure 14 depicts the comparison of the proposed method for the diagnosis of hepatitis disease with earlier studies mentioned in Sect. 2.

## 5 Conclusion

The authors have devised a method for diagnosing hepatitis disease. SMOTE was used to balance classes in the proposed method. The authors performed the classification using four classifiers: SVM, LR, KNN, and RF. The use of SMOTE improved the performance of all of the classifiers. SVM was 90.43% accurate, LR was 93.18% accurate, KNN was 85.93% accurate, and RF was 92.52% accurate. LR with SMOTE for class balancing provided maximum accuracy. Doctors will be able to diagnose hepatitis disease efficiently using the proposed method.

In the future, the authors will apply feature selection with ensemble classification approaches to improve accuracy up to the next level.



**Fig. 14** Comparative analysis of proposed method's accuracy with earlier studies**Table 6** Comparison of the proposed method's accuracy to earlier studies

Authors	Year	Dataset	Method	Accuracy (%)
Anto and Chandramathi	2015	UCI	Hybrid GA-SA and SVM	87.00
Anggraeny et al.	2017	UCI	Statistical BN	76.80
Bhargav et al.	2018	UCI	LR, DT, SVM, and NB	87.17
Hassan and Shah	2018	UCI	LR, DT, NB, KNN, SVM, and RF	85.00
Nilashi et al.	2019	UCI	ANFIS and SOM	93.06
Bayrak et al.	2019	UCI	NB, LR, and J48 decision tree	88.38
Nivaan and Emanuel	2020	UCI	LR	83.33
Alfyani and Muljono	2020	UCI	NB and KNN	74.19
Basarslan et al.	2020	UCI	KNN, RF, NB and LR	84.90
Peng et al.	2021	UCI	LR, DT, KNN, XGboost, SVM and RF	91.9
Panda et al.	2021	UCI	KNN, LR, NB, DT, SVM, RF, Chi-square and Boruta	90.32
Mijwil et al.	2021	UCI	SVM, C5.0 DT, KNN and RF	92.88
Nayeem et al.	2021	UCI	KNN, NB, SVM, multilayer perceptron (MLP), and RF	92.41
Yusuf and Akande	2021	UCI	SVM, Gaussian NB, LR, DT, KNN, and MLP	87
Das et al.	2021	UCI	Velocity enhanced whale optimization	92.5
Proposed method	2022	UCI	SVM, LR, KNN, RF, SMOTE	93.18

## References

1. Cetin O, Temurtas F, Gulgonul S (2015) An application of multilayer neural network on hepatitis disease diagnosis using approximations of sigmoid activation function. *Dicle Med J* 42(2):150–157. <https://doi.org/10.5798/diclemedj.0921.2015.02.0550>
2. Ansari S, Shafi I, Ansari A, Ahmad J, Shah SI (2011) Diagnosis of liver disease induced by hepatitis virus using artificial neural networks. In: *IEEE 14th international multitopic conference*, pp 8–12. <https://doi.org/10.1109/INMIC.2011.6151515>
3. Metwally NF, AbuSharekh EN, Abu-Naser SS (2018) Diagnosis of hepatitis virus using artificial neural network. *Int J Acad Dev* 2(11):1–7

4. Bayrak EA, Kirci P, Ensari T (2019) Performance analysis of machine learning algorithms and feature selection methods on hepatitis disease. *Int J Multidiscip Stud Innov Technol* 3(2):135–138
5. Perz JF, Armstrong GL, Farrington LA, Hutin YJF, Bell BP (2006) The contributions of hepatitis B virus and hepatitis C virus infections to cirrhosis and primary liver cancer worldwide. *J Hepatol* 45(4):529–538. <https://doi.org/10.1016/j.jhep.2006.05.013>
6. Nilashi M, Ahmadi H, Shahmoradi L, Ibrahim O, Akbari E (2019) A predictive method for hepatitis disease diagnosis using ensembles of neuro-fuzzy technique. *J Infect Public Health* 12(1):13–20. <https://doi.org/10.1016/j.jiph.2018.09.009>
7. Neshat M, Sargolzaei M, Toosi AN, Masoumi A (2012) Hepatitis disease diagnosis using hybrid case based reasoning and particle swarm optimization. *ISRN Artif Intell* 2012:1–6. <https://doi.org/10.5402/2012/609718>
8. Avci D (2016) An automatic diagnosis system for hepatitis diseases based on genetic wavelet kernel extreme learning machine. *J Electr Eng Technol* 2012:993–1002. <https://doi.org/10.5370/JEET.2016.11.4.993>
9. Dogantekin E, Dogantekin A, Avci D (2009) Automatic hepatitis diagnosis system based on linear discriminant analysis and adaptive network based on fuzzy inference system. *Expert Syst Appl* 36(8):11282–11286. <https://doi.org/10.1016/j.eswa.2009.03.021>
10. Anto S, Chandramathi S (2015) An expert system based on SVM and Hybrid GA-SA optimization for hepatitis diagnosis. *Int J Comput Eng Res Trends* 2(7):437–443
11. Anggraeny FT, Purbasari IY, Suryaningsih E (2017) Relief feature selection and Bayesian network model for hepatitis diagnosis. In: 3rd International conference on information technology and business, lampung, pp 113–118
12. Bhargav KS, Kumari TD, Thota DSSB, Vikas B (2018) Application of machine learning classification algorithms on hepatitis dataset. *Int J Appl Eng Res* 13(16):12732–12737
13. Hassan CAU, Khan MS, Shah MA (2018) Comparison of machine learning algorithms in data classification. In: 24th International conference on automation and computing (ICAC), pp 1–6. <https://doi.org/10.23919/ICAC.2018.8748995>
14. Nivaan GV, Emanuel AWR (2020) Analytic predictive of hepatitis using the regression logic algorithm. In: 3rd International seminar on research of information technology and intelligent systems (ISRITI), pp 106–110. <https://doi.org/10.1109/ISRITI51436.2020.9315365>
15. Alfiani R (2020) Comparison of Naïve Bayes and KNN algorithms to understand hepatitis. In: International seminar on application for technology of information and communication (iSemantic), pp 196–201. <https://doi.org/10.1109/iSemantic50169.2020.9234299>
16. Basarslan MS, Bakir H, Yücedağ İ (2020) Fuzzy logic and correlation-based hybrid classification on hepatitis disease data set. In: Hemanth D, Kose U (eds) Artificial intelligence and applied mathematics in engineering problems. ICAIAME 2019. Lecture notes on data engineering and communications technologies, vol 43. Springer, Cham, pp 787–800. [https://doi.org/10.1007/978-3-030-36178-5\\_68](https://doi.org/10.1007/978-3-030-36178-5_68)
17. Peng J et al (2021) An explainable artificial intelligence framework for the deterioration risk prediction of hepatitis patients. *J Med Syst* 45(5):1–9. <https://doi.org/10.1007/s10916-021-01736-5>
18. Panda N, Satapathy SK, Mishra S, Mallick PK (2021) Empirical study on different feature selection and classification algorithms for prediction of hepatitis disease. In: Tripathy HK, Mishra S, Mallick PK, Panda AR (eds) Technical advancements of machine learning in healthcare. Studies in computational intelligence, vol 936. Springer, Singapore, pp 75–86. [https://doi.org/10.1007/978-981-33-4698-7\\_4](https://doi.org/10.1007/978-981-33-4698-7_4)
19. Mijwil MM, Salem IE, Abttan R (2021) Utilisation of machine learning techniques in testing and training of different medical datasets. *Asian J Comput Inf Syst* 9:29–34
20. Nayeem MJ, Rana S, Alam F, Rahman MA (2021) Prediction of hepatitis disease using K-nearest neighbors, Naive Bayes, support vector machine, multi-layer perceptron and random forest. In: International conference on information and communication technology for sustainable development (ICICT4SD), pp 280–284. <https://doi.org/10.1109/ICICT4SD50815.2021.9397013>
21. Yusuf A, Akande O (2021) Hepatitis diseases prediction using machine-learning techniques. *Fudma J Sci* 5(3):1–8. <https://doi.org/10.33003/fjs-2021-0503-515>
22. Das S, Nayak M, Senapati MR, Satapathy J (2021) Medical data classification using velocity enhanced whale optimization algorithm. In: First international conference on advances in computing and future communication technologies (ICACFCT), pp 18–22. <https://doi.org/10.1109/ICACFCT53978.2021.9837345>
23. Hepatitis data set [Online]. <https://archive.ics.uci.edu/ml/datasets/hepatitis>
24. Tiwari H (2022) Early prediction of heart disease using deep learning approach. In: Gupta D, Kose U, Khanna A, Balas VE (eds) Deep learning for medical applications with unique data. Academic Press, Cambridge, pp 107–122. <https://doi.org/10.1016/B978-0-12-824145-5.00014-9>
25. Książek W, Gandor M, Pławiak P (2021) Comparison of various approaches to combine logistic regression with genetic algorithms in survival prediction of hepatocellular carcinoma. *Comput Biol Med* 134:1–13. <https://doi.org/10.1016/j.combiomed.2021.104431>
26. Mahmood I, Abdulazeez AM (2021) The role of machine learning algorithms for diagnosing diseases. *J Appl Sci Technol Trends* 2(1):10–19
27. Geler Z, Kurbalija V, Ivanović M, Radovanović M (2020) Weighted kNN and constrained elastic distances for time-series classification. *Expert Syst Appl*. <https://doi.org/10.1016/j.eswa.2020.113829>
28. Srim S (2019) A review on object tracking based on KNN classifier. *Int Res J Eng Technol* 6(12):924–927
29. Rani P, Kumar R, Jain A, Chawla SK (2021) A hybrid approach for feature selection based on genetic algorithm and recursive feature elimination. *Int J Inf Syst Model Des* 12(2):17–38. <https://doi.org/10.4018/IJISMD.2021040102>
30. Sachdeva RK, Bathla P (2022) A machine learning-based framework for diagnosis of breast cancer. *Int J Softw Innov* 10(1):1–11. <https://doi.org/10.4018/IJSI.301221>
31. Sachdeva RK, Bathla P, Rani P, Kukreja V, Ahuja R (2022) A systematic method for breast cancer classification using RFE feature selection. In: 2nd International conference on advance computing and innovative technologies in engineering (ICACITE), pp 1673–1676. <https://doi.org/10.1109/ICACITE53722.2022.9823464>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.