Frontiers in Health Informatics
2021; 10: 57

www.ijmi.ir
Open Access

Machine learning based methods for handling imbalanced data in hepatitis diagnosis

Azam Orooji[1], Farzaneh Kermani[2*]

[1]Assistant Professor of Medical Informatics, Department of Advanced Technologies, School of Medicine, North Khorasan University of Medical Sciences (NKUMS), Bojnurd, Iran
[2]Assistant Professor of Medical Informatics, Department of Health Information Technology, School of Allied Medical Sciences, Semnan University of Medical Sciences, Semnan, Iran

Article Info

Article type:
Research

Article History:
Received: 2020-11-23
Accepted: 2020-12-16
Published: 2021-01-23

**** Corresponding author:***
Farzaneh Kermani

Assistant Professor of Medical Informatics, Department of Health Information Technology, School of Allied Medical Sciences, Semnan University of Medical Sciences, Semnan, Iran

Email: f-kermani@semums.ac.ir

Keywords:
Hepatitis C Virus (HCV)
Prediction
Data Mining
Machine Learning
Imbalanced Data

ABSTRACT

Introduction: Hepatitis C virus is the leading cause of mortality from liver disease. Also, diagnosis systems are usable tools for better disease control and management. The aim of this study was to design an HCV disease prediction system and classify its severity based on data mining methods.

Material and Methods: This is an applied research that uses the hepatitis C dataset in the UCI library. The study was conducted in four steps including data preprocessing, data mining, evaluation and system design. In data pre-processing, data balancing techniques were performed. Then, three data mining algorithms (multi-layer perceptron, Bayesian network, and decision tree) were implemented and 10-fold cross-validation method was used to evaluate data mining algorithms. Finally, user interface was designed in MATLAB programming language (version 2016) based on the best algorithm.

Results: The results showed that the over-sampling method improved the performance measures of data mining algorithms in disease prediction, so that in the O-dataset the accuracy of the best method (random forest) was 99.9%. Also, the random forest for the O-dataset had the best performance measures in term of sensitivity, accuracy and f-measure (99.9%) and the 100% specificity amount.

Conclusion: Considering that the presented approach has performed better than all suggested methods in previous studies, the proposed system in this study can be used well in HCV diagnosing and determining its severity.

How to cite this paper
Orooji A, Kermani F. Machine Learning Based Methods for Handling Imbalanced Data in Hepatitis Diagnosis. Front Health Inform. 2021; 10: 57. DOI: 10.30699/fhi.v10i1.259

INTRODUCTION

Hepatitis C virus (HCV) is a single-stranded RNA virus [1]. This virus is one of the most important causes of chronic liver disease that due to its long-term treatment can cause cirrhosis and liver cell cancer [2]. The virus was first identified as the leading cause of non-A and non-B hepatitis in April 1989 [3], and there are now about 71 million people living with chronic hepatitis C worldwide. About 30% of people (15% to 45%) recover after six months without treatment. Chronic HCV spreads among the remaining 70% (55% to 85%). The risk of cirrhosis for this group will be between 15 to 30% in the next 20 years [4]. HCV is the leading cause of mortality from liver disease, with 333,000 deaths in 1990, 499,000 in 2010, and 704,000 in 2013 [4-6].

The incidence of this disease has been expressed in various studies from 0.5% to 2.8%. In high-income countries, the prevalence of chronic hepatitis C is less than 2%. [7, 8]. Countries with a high prevalence of HCV (more than 5%) include Egypt, Gabon, Uzbekistan, Cameroon, Mongolia, Pakistan, Nigeria and Georgia, which are low-middle income countries [9].

Getting HCV does not provide long-term immunity, and many cases of re-infection have been reported. As a result, in areas with a high prevalence of HCV, hybrid HCV genotypes are observed that result from more than one HCV infection. This is a major obstacle to developing a vaccine for this disease. Other challenges related to the control of the disease are [9]:

- Inadequate surveillance data
- Coverage of prevention programs is limited
- Few people know their hepatitis status and have access to treatment
- Medicines and diagnostics are unaffordable for most
- A public health approach to hepatitis is lacking
- Leadership and commitment are uneven

Hence, the WHO has recently developed the first global health sector strategy for the hepatitis virus. This strategy covers all 5 types of hepatitis virus (Hepatitis A, B, C, D and E) but focuses more on hepatitis C. According to the document, the number of infected people and the HCV death decreased 70% and 60% respectively by 2030 compared to 2010 years [10].

Predicting chronic diseases play a vital role in health informatics. Chronic disease diagnosis is very important because these diseases affect a person for a long time. The most common chronic diseases are diabetes, stroke, cardiovascular disease, cancer, hepatitis C and osteoarthritis. Early detection of chronic diseases improves prevention and increases the effectiveness of the treatment process [11].

Classification is a data mining technique that uses the train data to develop a model and the resulting model is applied to the test data for determining its predictive power. Various classification algorithms have been used to predict chronic diseases and their results have been very promising [12]. One of the challenges in using of data mining techniques is the unbalanced dataset. A dataset is called unbalanced if samples of one class (called the minority class) are much smaller than samples in other class (es) (called the majority class) [13]. This causes the algorithms have good accuracy in majority class and low accuracy for minority class [14]. There are problems with unbalanced datasets in many areas of research [15-17]. In medical science with disease diagnosis goal, the dataset is usually unbalanced because the number of patients is less than of healthy people which become more severe in rare disease. To solve this problem, they increase the number of minority class samples or decrease the number of majority class samples [18]. One solution is over-sampling of minority class samples and under-sampling of majority class samples [19].

Chronic disease diagnosis systems are valuable tools for better disease control and management [20]. Therefore, the aim of this study was to design an HCV disease prediction system and classify its severity based on data mining methods. To design this system, an unbalanced dataset was used and the results of three different data mining algorithms were compared, and the best result is the basis for designing the system.

MATERIAL AND METHODS

This is an applied research that uses the hepatitis C dataset in the UCI library (https://archive.ics.uci.edu/ml/datasets/HCV data). This dataset contains 615 records, of which 75 records are for people with HCV (Hepatitis: 24; Fibrosis: 21; Cirrhosis: 30). Out of the remaining 540 records, 7 suspected cases were recorded that were removed from the analysis (blood donor: 533 records). Descriptive statistics for the 608 records in this dataset (237: female and 371: male) are shown in Table 1.

Table 1: Descriptive statistic of records in dataset

Variable	Range	Mean (SD)	Missing values
Age	19-77	47.29 (9.993)	0
ALB	20-82.2	41.819 (5.4067)	1
ALP	11.3-416.6	67.821 (25.2744)	18
ALT	0.9-258	27.601 (21.2275)	1
AST	12-324	34.369 (32.6224)	0
BIL	1.8-254	11.474 (19.7706)	0
CHE	1.42-16.41	8.2049 (2.1684)	0
CHOL	1.43-9.67	5.3788 (1.1194)	10
CREA	8-1079	81.51 (49.721)	0
GGT	4.5-650.9	38.244 (51.9532)	0
PROT	51-90	72.253 (4.9323)	1

This study was conducted in four steps, which are explained below.

Step 1: Data pre-processing

Because of missing data in the dataset, the attribute values were replaced in unregistered records. According to the one to 7 ratios of the sum of the three minority classes to the majority class, this dataset needs to be balanced. For this purpose, random over-sampling and random under-sampling methods [19, 21] were used for balancing.

Step 2: Data Mining

After preparing the dataset, three data mining algorithms were implemented. These algorithms include multi-layer perceptron (MLP), Bayesian network, and decision tree. Each of which was implemented in the original and balanced dataset and their evaluation results were compared. All methods were implemented in MATLAB programming language (version 2016).

Step 3: data mining algorithms Evaluation

10-fold cross-validation method was used to evaluate data mining algorithms. For each algorithm, the performance measures included accuracy, precision, sensitivity, specificity, and F-Measure were calculated which is presented in Table 2.

Table 2: The performance evaluation measures

Accuracy= (TP+TN)/(TP+TN+FP+FN)
Precision= TP/(TP+FP)
Sensitivity= TP/(TP+FN)
Specificity= TN/(TN+FP)
F-measure=(2*Precision*Recall)/(Precision + Recall)

* True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN)

Step 4: System designing

After selecting the best algorithm in diagnosing the HCV and its severity, the system user interface was designed based on the best algorithm.

RESULTS

There were 31 unregistered data that were replaced using mean and median for continuous and discrete attributes, respectively. Then, two datasets were created with an equal minority to majority ratio by using two methods of random over-sampling and random under-sampling. The over-sampling dataset is called O-Dataset and the under-sampling dataset is called U-Dataset. In O-Dataset, the number of minority class samples increased and was equal to the number of majority class samples, and in U-Dataset, the number of majority class decreased and was equal to the number of minority class samples.

The 10-fold cross-validation method was used for evaluation. The acquired accuracy for data mining algorithms on the three datasets: original, O-Dataset, and U-Dataset are shown in Fig 1.

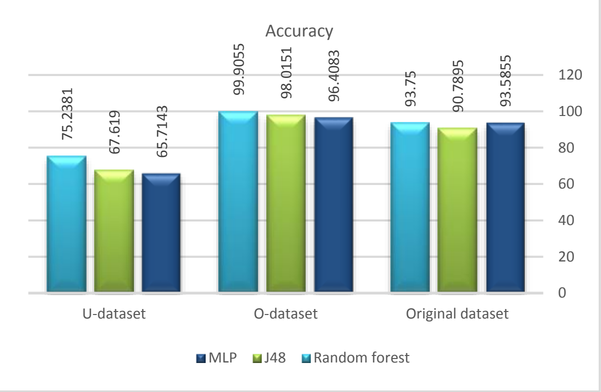

Fig 1: Comparison of data mining algorithms on three datasets: original, O-Dataset and U-Dataset based on accuracy measure

According to results, all three algorithms in the O-Dataset have higher accuracy measure than the original dataset and U-Dataset. However, due to the unbalanced data and display the accuracy in general, accuracy measure cannot indicate the superiority of one method over another. Fig 2 to 4 show other performance measures for the three data mining algorithms and the three datasets. In these figures, the performance measures are reported for each class separately. Table 3 also shows the average of these measures for each dataset.

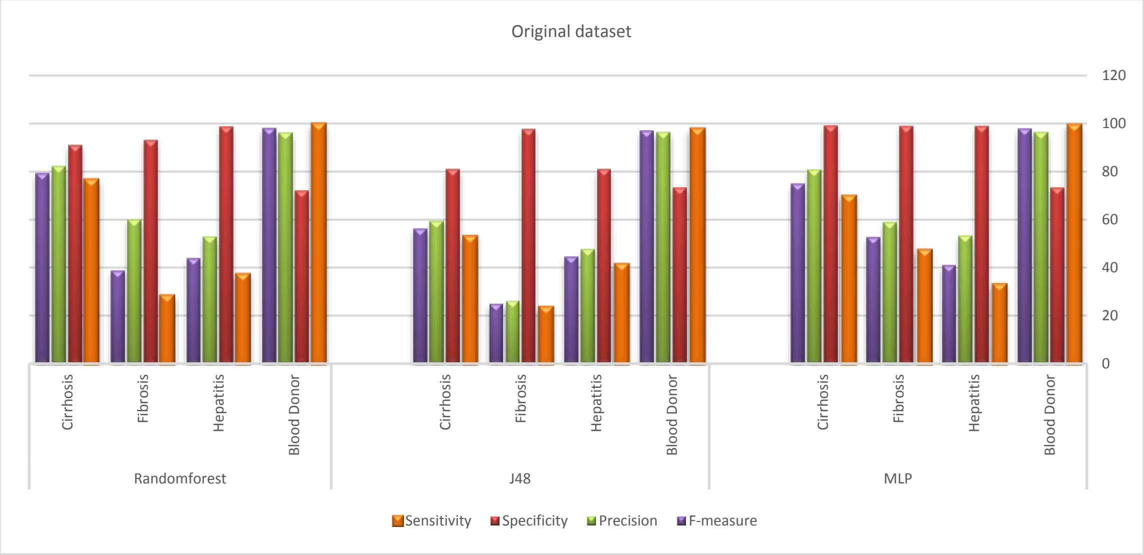

Fig 2: Comparison of the performance measures of data mining algorithms in the original dataset

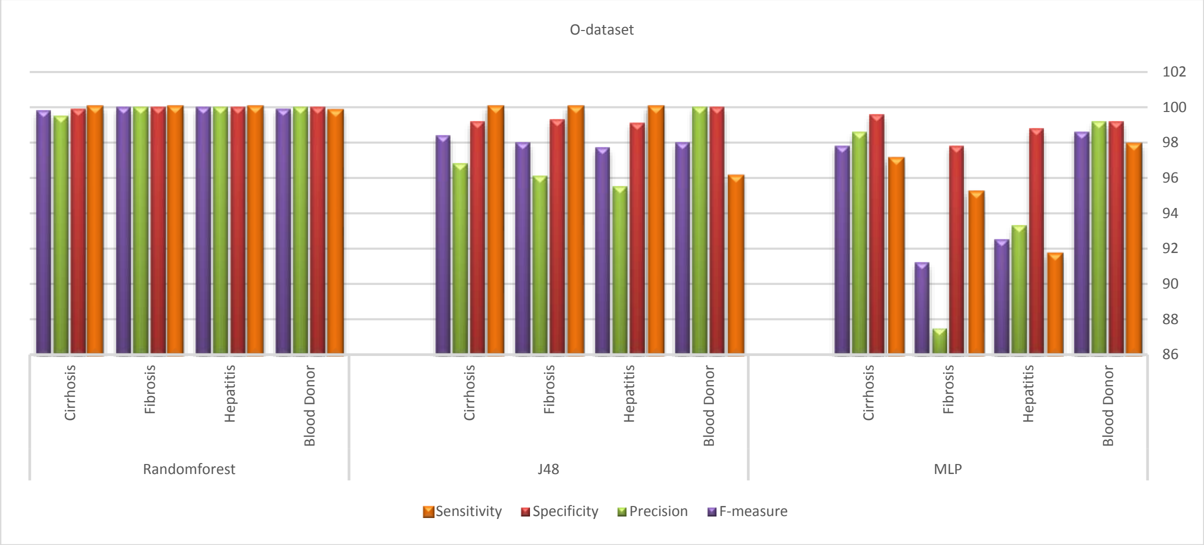

Fig 3: Comparison of the performance measures of data mining algorithms in the O-dataset

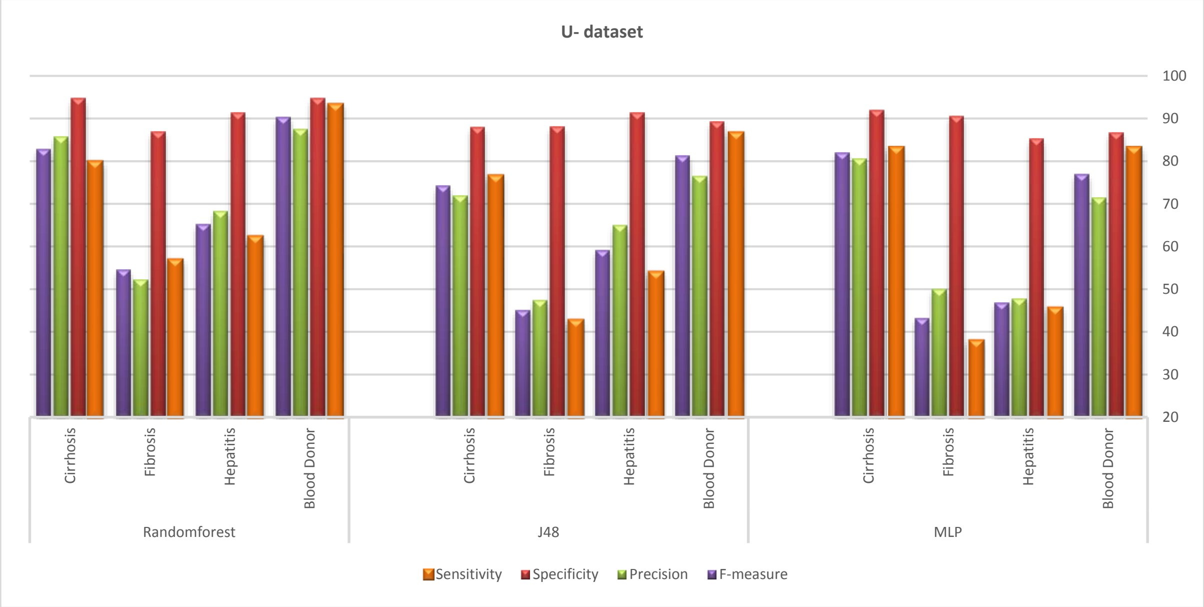

Fig 4: Comparison of the performance measures of data mining algorithms in the U-dataset

Table 3: Weighted average performance measures in the three datasets

Data set	Algorithm	Sensitivity	Specificity	Precision	F-measure
Original dataset	MLP	93.6	76.5	92.6	92.9
	J48	90.8	76.4	90.1	90.4
	Random forest	93.8	75.3	92.6	92.9
O-dataset	MLP	96.4	99	96.5	96.4
	J48	98	99.6	98.1	98
	Random forest	99.9	100	99.9	99.9
U-dataset	MLP	65.7	88.6	64.4	64.7
	J48	67.6	89.2	66.7	66.9
	Random forest	75.2	92.4	75.5	75.3

According to Fig 2, the sensitivity, specificity and accuracy for the original dataset in the three patients' group are much less than in the healthy group (majority class), and as a result, the F-measure, which is a result of sensitivity and specificity, are low. These results are also repeated in the U-Dataset (Fig 4) and show that under-sampling has not been effective in resolving imbalances. While, Fig 3 shows that the over-sampling method is very effective and significant progress has been made in all performance measures in all groups (including patients and healthy individuals). Among the data mining methods, the random forest method had better performance in the U-dataset and O-dataset, but the MLP method was better in the original dataset. Also, Table 3 confirms the findings in Fig 2 to 4. Considering that the best results were obtained for random forest in O-dataset, it became the basis for the development of the hepatitis prediction system. The system user interface simulated in the MATLB environment is shown in Fig 5.

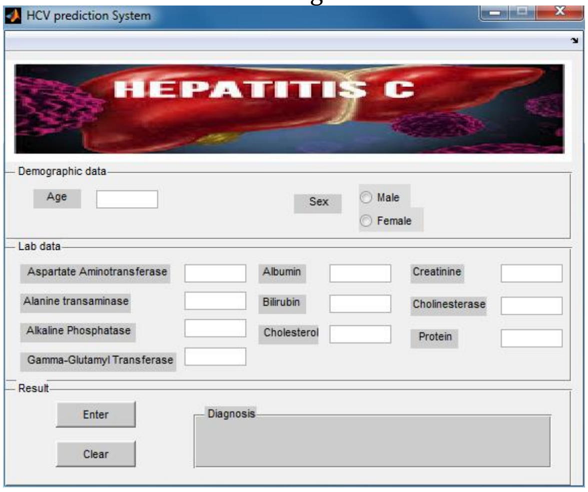

Fig 5: Hepatitis prediction system user interface

DISCUSSION

In this study, a hepatitis C prediction system was designed using demographic data and blood tests. Due to the imbalance of classes in the dataset, the random under-sampling and random over-sampling methods were used. The results showed that the over-sampling method improved the performance measures of data mining algorithms in disease prediction, so that in the O-dataset the accuracy of the best method (random forest) was 99.9% and about 6% higher than the original dataset. However, due to the data imbalance, algorithms were compared using other performance measures. According to results, the random forest for the O-dataset had the best performance measures in term of sensitivity, accuracy and f-measure (99.9%) and the 100% specificity amount.

Given that the dataset was recently published in the UCI, only two related studies have covered it. In the study of Hoffmann et al. [22], only three groups of patients were considered and the healthy individual's data were deleted and as a result, the problem of data imbalance has not occurred. By adding the enhanced liver fibrosis (ELF) score to the dataset, they examined two modes: the ELF dataset and the ELF-free dataset. The C Tree and rpart algorithms were simulated in R software and the evaluation was performed using the leave-one-out method. rpart had the best result for the ELF dataset with 73.33% accuracy. In Chawathe study [23], a large number of classification algorithms were compared in terms of accuracy, F-measure, AUC, classification time, Training time and model size. In addition to examining the performance measures of algorithms in original dataset classification, three important features have been identified using 7 feature selection methods. These three features include: ALT, AST, CHE. Then, all the algorithms were compared by these three features in the dataset. The results showed that, Bayes Net (an unlimited number of parents per node: BNt-u) methods based on accuracy and F-measure measures and random forest based on AUC measure performed better than other algorithms for the original dataset. Also, for the collection with three important features, random forest has been the best method. Although the exact values of the measures are not specified, but according to the graphs, the accuracy and F-measure of all algorithms in both datasets are less than 93% and 98%, respectively.

CONCLUSION

Considering that the presented approach has performed better than all suggested methods in previous studies, the proposed system in this study can be used well in HCV diagnosing and determining its severity.

AUTHOR'S CONTRIBUTION

The authors agree on this final form of the manuscript, and attested that all authors contributed in the final draft of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest regarding the publication of this study.

FINANCIAL DISCLOSURE

No financial interests related to the material of this manuscript have been declared.

REFERENCES

1. Yager EJ, Konan KV. Sphingolipids as potential therapeutic targets against enveloped human rna viruses. Viruses. 2019; 11(10): 912. PMID: 31581580 DOI: 10.3390/v11100912 [PubMed]

2. Ghadir M, Jafari E, Amiriani M, Rezvan H, Aminikafiabad S, Pourshams A. Hepatitis C in Golestan Province-Iran. Govaresh. 2006; 11(3): 158-62.

3. Choo Q-L, Kuo G, Weiner AJ, Overby LR, Bradley DW, Houghton M. Isolation of a cDNA clone derived from a blood-borne non-A, non-B viral hepatitis genome. Science. 1989; 244(4902): 359-62. PMID: 2523562 DOI: 10.1126/science.2523562 [PubMed]

4. World Health Organization. Hepatitis C: Fact sheet [Internet]. 2016 [cited: 15 Oct 2020; updated: 27 July 2020]. Available from: https://www.who.int/en/news-room/fact-

sheets/detail/hepatitis-c.

5. Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: A systematic analysis for the Global Burden of Disease Study 2010. The lancet. 2012; 380(9859): 2095-128. PMID: 23245604 DOI: 10.1016/S0140-6736(12)61728-0 [PubMed]

6. Vos T, Barber RM, Bell B, Bertozzi-Villa A, Biryukov S, Bolliger I, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: A systematic analysis for the Global Burden of Disease Study 2013. The lancet. 2015; 386(9995): 743-800. PMID: 26063472 DOI: 10.1016/S0140-6736(15)60692-4 [PubMed]

7. Mohd Hanafiah K, Groeger J, Flaxman AD, Wiersma ST. Global epidemiology of hepatitis C virus infection: new estimates of age-specific antibody to HCV seroprevalence. Hepatology. 2013; 57(4): 1333-42. PMID: 23172780 DOI: 10.1002/hep.26141 [PubMed]

8. Gower E, Estes C, Blach S, Razavi-Shearer K, Razavi H. Global epidemiology and genotype distribution of the hepatitis C virus infection. J Hepatol. 2014; 61(1 Suppl): S45-57. PMID: 25086286 DOI: 10.1016/j.jhep.2014.07.027 [PubMed]

9. Lanini S, Easterbrook PJ, Zumla A, Ippolito G. Hepatitis C: Global epidemiology and strategies for control. Clin Microbiol Infect. 2016; 22(10): 833-8. PMID: 27521803 DOI: 10.1016/j.cmi.2016.07.035 [PubMed]

10. World Health Organization. Global health sector strategy on viral hepatitis 2016-2021 [Internet]. 2016 [cited: 15 Oct 2020]. Available from: http://www.who.int/hepatitis/strategy2016-2021/ghss-hep/en/

11. Jain D, Singh V. Feature selection and classification systems for chronic disease prediction: A review. Egyptian Informatics Journal. 2018; 19(3): 179-89.

12. Han J, Pei J, Kamber M. Data mining: Concepts and techniques. Elsevier; 2011.

13. Beyan C, Fisher R. Classifying imbalanced data sets using similarity based hierarchical decomposition. Pattern Recognition. 2015; 48(5): 1653-72.

14. Devarriya D, Gulati C, Mansharamani V, Sakalle A, Bhardwaj A. Unbalanced breast cancer data classification using novel fitness functions in genetic programming. Expert Systems with Applications. 2020; 140: 112866.

15. Bhardwaj A, Tiwari A, RameshKrishna M, Vishaal Varma M. An innovative genetic programming framework in modelling a real time epileptic seizure detection system. ASE BigData/SocialInformatics/PASSAT/BioMedCom Conference. Harvard University; 2014.

16. Bhardwaj H, Sakalle A, Bhardwaj A, Tiwari A. Classification of electroencephalogram signal for the detection of epilepsy using Innovative Genetic Programming. Expert Systems. 2019; 36(1): e12338.

17. Mera D, Bolon-Canedo V, Cotos JM, Alonso-Betanzos A. On the use of feature selection to improve the detection of sea oil spills in SAR images. Computers & Geosciences. 2017; 100: 166-78.

18. More A. Survey of resampling techniques for improving classification performance in unbalanced datasets. arXiv preprint arXiv:160806048. 2016.

19. Shelke MS, Deshmukh PR, Shandilya VK. A review on imbalanced data handling using undersampling and oversampling technique. Int J Recent Trends in Eng & Res. 2017; 3: 444-9.

20. Hussein AS, Omar WM, Li X, Ati M. Efficient chronic disease diagnosis prediction and recommendation system. IEEE-EMBS Conference on Biomedical Engineering and Sciences. IEEE; 2012.

21. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. 2002; 16: 321-57.

22. Hoffmann G, Bietenbeck A, Lichtinghagen R, Klawonn F. Using machine learning techniques to generate laboratory diagnostic pathways—a case study. J Lab Precis Med. 2018; 3: 58.

23. Chawathe SS. Diagnostic classification using hepatitis C tests. International IOT, Electronics and Mechatronics Conference. IEEE; 2020.