Erciyes Üniversitesi
Fen Bilimleri Enstitüsü Dergisi
Cilt 38, Sayı 2 , 2022

Erciyes University
Journal of Institue Of Science and Technology
Volume 38, Issue 2, 2022

Predicting Liver Disease Using Decision Tree Ensemble Methods

Fırat ORHAN BULUCU[1,2], İrem ACER[2],Fatma LATİFOĞLU[2,*] ,Semra İÇER[2]
1 Inonu University, Department of Biomedical Engineering, Malatya, Turkey
*2 Erciyes University, Department of Biomedical Engineering, Kayseri, Turkey

(Alınış / Received: 28.01.2021, Kabul / Accepted: 31.05.2022, Online Yayınlanma / Published Online: 23.08.2022)

Keywords
Liver Disease,
Classification,
Ensemble learning,
Decision Trees,
SMOTE

Abstract: Damages that may occur in the liver, which has an important task for the human body, can cause fatal consequences. For this reason, early diagnosis of liver disease is important. In this study, liver disease was tried to be predicted using Ensemble learning methods based on various clinical values obtained from liver patients and healthy blood donors. The clinical values used in the study were obtained from the University of California, Irvine (UCI), which is shared as open access. Since the data used in the study showed an uneven distribution, the data were balanced by using the Synthetic Minority Sampling method (SMOTE) before the classification stage. The stabilized data were classified using the bagging and boosting models Random Forest (RF), J48, AdaBoost, Gradient Boosting Classifiers (GBC), and Light Gradient Boosting Machine (Light GBM) algorithms. As a result of the algorithms used, the Light GBM algorithm gave the most successful classification results using 10-fold cross-validation. Classification results were obtained as 98.8% accuracy, 98.1% precision, 99.4% recall and 0.98% kappa statistics rate.

Karar Ağacı Topluluğu Yöntemleri Kullanılarak Karaciğer Hastalığının Tahmin Edilmesi

Anahtar Kelimeler
Karaciğer Hastalığı,
Sınıflandırma,
Topluluk öğrenme,
Karar ağaçları,
SMOTE

Öz: İnsan vücudu için önemli bir görevi olan karaciğerde meydana gelebilecek hasarlar ölümcül sonuçlara neden olabilir. Bu nedenle karaciğer hastalığının erken teşhisi önemlidir. Bu çalışmada, karaciğer hastalarından ve sağlıklı kan bağışçılarından elde edilen çeşitli klinik değerlere bağlı olarak topluluk öğrenme yöntemleri kullanılarak karaciğer hastalığı tahmin edilmeye çalışılmıştır. Çalışmada kullanılan klinik değerler açık erişim olarak paylaşılan University of California, Irvine (UCI)'den elde edilmiştir. Yapılan çalışmada kullanılan veriler dengesiz dağılım gösterdiği için sınıflandırma aşamasından önce Sentetik Azınlık Örneklem Arttırma yöntemi (SMOTE) kullanılarak veriler dengeli hale getirilmiştir. Dengeli hale getirilen veriler torbalama ve artırma modellerinden Random Forest (RF), J48, AdaBoost, Gradient Boosting Classifiers (GBC) ve Light Gradient Boosting Machine (Light GBM) algoritmaları kullanılarak sınıflandırılmıştır. Kullanılan algoritmalar sonucunda en başarılı sınıflandırma sonuçlarını 10 kat çapraz doğrulama kullanılarak Light GBM algoritması vermiştir. Sınıflandırma sonuçları, %98,8 doğruluk, %98,1 kesinlik, %99,4 geri çağırma ve %0,98 kappa istatistik oranı olarak elde edilmiştir.

*Corresponding Author, email: flatifoglu@erciyes.edu.tr

1. Introduction

The liver is an organ that has an important role in the human body. Liver diseases are one of the most common diseases in the world. According to the World Health Organization (WHO) data, more than 70 million people in the world suffer from liver-related diseases such as hepatitis C, fibrosis, and cirrhosis, and deaths occur due to these diseases [1]. Hepatitis C virus (HCV) is a disease that is carried to the human body through blood, can lead to serious conditions over time, and is accepted as the main health problem in society [2]. Fibrosis is a serious liver injury that occurs as a result of the progression of HCV, which occurs due to the accumulation of more than normal

fat in the liver [3, 4]. Cirrhosis, on the other hand, is a fatal disease that occurs as a result of deterioration of liver tissue, which is the last stage of fibrosis and is also known as liver failure [4]. As a result of a few laboratory tests to be taken from the human body, liver disease can be diagnosed early and its progression can be controlled. It has been stated in studies that clinical values such as Albumin (ALB), Alkaline phosphatase (ALP), Alanine amino-transferase (ALT), Aspartate amino-transferase (AST) are widely used in the diagnosis of liver disease and in research [5, 6].

In the field of health, the diagnosis of diseases can be made in traditional ways. However, it has been stated that with the clinical values obtained, machine learning methods have been used in recent studies for the purposes of disease prevention, and diagnosis. [7, 8]. With these methods, diseases can be diagnosed faster and can support physicians. There are a number of studies in the literature on predicting liver disease using machine learning methods. Oladimeji et al. predicted HCV with a success rate of 98.97% with RF [8]. In their study, Orooji and Kermani tried to detect HCV by using Multi-layer perceptron, Bayesian network, Decision tree algorithms and they obtained the most successful result in the RF algorithm with 99.9% [9]. Mostafa and Hasan tried to predict liver disease using many machine learning methods in their study and compared the results they obtained in the methods they used. The most successful result was obtained in the Support Vector Machine (SVM) algorithm with 98.23% [10]. In another study for HCV prediction, Stacked Ensemble models were used and a success rate of 99.8% was achieved [11].

In this study, it was aimed to predicted liver disease from machine learning using Bagging Ensemble and Boosting Ensemble methods using clinical data obtained from liver patients and control group. Since the data set used in the study showed an uneven distribution, it was turned into a balanced distribution by preprocessing. Accuracy, precision, recall, kappa statistic metric values were obtained with the classifiers preferred in the study from the data showing a balanced distribution, and the algorithms were compared. This study, it was aimed to contribute to the literature for the diagnosis of liver disease by determining the ensemble model that gave the most successful result. In the second part of the study, information about the data and methods used is given, while in the third part, the results of the research are given. In the last part, the results obtained in the study with similar studies in the literature are compared and discussed.

2. Material and Method

2.1 Dataset

The data set used in the study was shared by Hoffmann et al. [12] as open access in the University of California, Irvine (UCI) [13] data repository in 2020. The data set used consisted of a total of 615 people and was divided into two classes, 75 of whom were liver patients (HCV, fibrosis, and cirrhosis), and 540 people were healthy blood donors. The data show an uneven distribution between these two classes (Figure 1). Information on the clinical values obtained from these individuals and used in the study are given in Table 1, and more detailed information is given in [12].

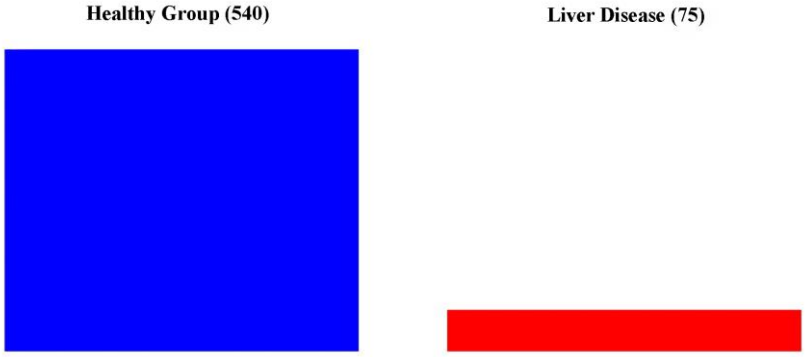

Figure 1. Distribution of data before balancing method

Table 1. Information on clinical values used in the study

Features used in this study
ALB
ALP
ALT
AST
Bilirubin
Choline esterase

Choline
Creatinine
γ-glutamyl-transferase
Total protein
Category: Healthy group (Negative)- Liver Disease (Positive)

2.2 Data pre-processing

The data set consists of 540 healthy groups and 75 liver patients, showing an uneven distribution. To avoid this imbalance, the SMOTE proposed by Chawla et al. [14] was applied as a preprocessing step. Since it is stated in the literature that in case of uneven distribution of the data, the classification results cannot be interpreted properly and there may be a loss in success rates [8], the SMOTE method, which is frequently preferred in the literature, was preferred to balance the data before the classification stage. It has been observed that the SMOTE method improves the performance in predictive models [8] and therefore SMOTE was preferred in the study. With this method, the data in the minority class are randomly multiplied according to their nearest neighbors, making them close to or equal to the data in the majority class [15]. With this method, the data, which were divided into two groups as the healthy group and the liver patient, were stabilized as 540 individuals without any data loss (Figure 2). Balancing was performed using the Waikato Environment for Knowledge Analysis (WEKA) Version 3.8.3 program [16].

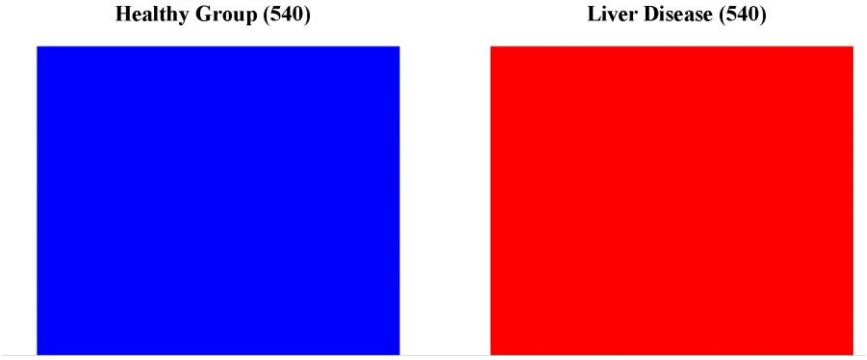

Figure 2. Distribution of data after balancing method

2.3.1 Ensemble learning methods

Ensemble learning is a machine learning method that can make some predictions using multiple similar or different classifier models. It is used to prevent prediction errors and negativities that may occur in machine learning by using different basic models [11, 17]. In studies using ensemble learning, bagging and boosting models are widely used. In this study, a comparison was made using bagging and boosting models.

A. Bagging method

In the bagging method, communities in multiple subsets are created by applying predictions to Bootstrap samples obtained from the original data set. A model is created in each subset and the models run independently and randomly. In the final stage, the predictions from all models are combined and the results are obtained [11, 17, 19]. In the bagging method, it is also known as Bootstrap Aggregating because the training data is divided into many sub-samples [11]. In this study, RF and J48 algorithms were used by using the bagging model approach. The RF algorithm is a typical bagging ensemble learning method in which multiple decision trees are applied to subsets of data and each decision tree produces results [18]. The J48 algorithm was developed by J. Ross Quinlan, and it is a C4.5 decision tree that is widely used in classification studies [20].

B. Boosting method

Boosting method, developed by Freund and Schapire [21], is a technique that combines weak classifications to achieve more successful performance. Unlike the bagging method, each tree structure created in the boosting method is connected with the previous tree structure. That is the models (training datasets and classifiers) created in each subset work sequentially with each other [21, 22]. In this study, Light GBM, AdaBoost, and GBC were preferred as boosting methods. The Light GBM algorithm, developed by Ke et al. [23] in 2017, is a gradient-based

decision tree technique that can be used in studies such as regression and classification. In the AdaBoost algorithm, the training set is trained with a weak learner. After the training, the wrongly estimated data is strengthened and retrained. In this way, it can produce strong classifiers from weak classifiers [18, 24]. GBC, proposed by Friedman [25], is a powerful machine learning method that can predict complex data quickly and accurately. The model created by this method can be combined with the previous model, thereby minimizing the estimation error. For more detailed information about the GBC algorithm, see [11, 25].

2.4 Evaluation of classifiers

The performances of the classification algorithms used in the study were examined in terms of accuracy, recall, precision, and kappa statistics. Accuracy, recall, precision rates were calculated using true positive (tp), true negative (tn), false positive (fp), and false negative (fn) values according to the confusion matrix given in Table 2. The accuracy rate is the rate of correctly predicted data. The recall value is also known as sensitivity and represents correctly classified positive rates. The precision value indicates how many of the samples classified as positive are actually positive. Kappa statistical value is a statistical criterion used to test the reliability of algorithms. The formulas for the calculated values are given in equations 1-4, respectively [26]. In the classification processes, cross-validation (CV) was set as 5-10 and the results were compared. Modeling processes of machine learning methods were carried out in the Google Colaboratory environment using the Python programming language

$$Accuracy = \frac{tp+tn}{fp+fn+fn+fp} \quad (1)$$

$$Recall = \frac{tp}{tp+fn} \times 100 \quad (2)$$

$$Precision = \frac{tp}{tp+fp} \times 100 \quad (3)$$

$$Kappa\ Statistic = \frac{p_0-p_e}{1-p_e} \quad (4)$$

Table 2. Confusion matrix

		Predicted Class	
		Negative (Healthy)	**Positive (Liver Disease)**
Actual Class	**Negative**	**TN**	**FP**
	Positive	**FN**	**TP**

3. Results

In this study, the clinical values obtained from liver patients with symptoms such as hepatitis C, fibrosis, cirrhosis, and healthy individuals were evaluated by bagging and boosting methods from ensemble models. Since the data set was initially unevenly distributed, balancing was applied to the data of liver patients and healthy individuals before proceeding to the classification stage. As a balancing method, the SMOTE method, which is widely used in studies, was preferred. While RF and J48 algorithms were preferred as bagging methods in the classification stage after the balancing process, Light GBM, AdaBoost, and GBC algorithms were used as boosting methods. In the classification process, 5-10 fold cross-validation technique was used and the performances of the algorithms were compared by obtaining the accuracy, precision, recall, kappa statistic metric values. Detailed information about the results obtained in this study is given in Table 3. Looking at Table 3, the most successful results were obtained with the Light GBM algorithm in terms of both 5-fold cross-validation and 10-fold cross-validation.

Table 3. Comparison of the performance of classification methods

Classifiers	Cross Validation	Accuracy (%)	Precision (%)	Recall (%)	Kappa Statistic
RF	5	98.2	97.4	99.1	0.96
	10	98.3	97.5	99.3	0.96
J48	5	96.9	95.8	98.1	0.94
	10	97.6	96.4	98.9	0.95
GBC	5	98.4	97.9	98.9	0.96
	10	98.5	98.1	98.8	0.97
Light GBM	5	98.7	97.9	99.5	0.97
	10	**98.8**	98.1	99.4	**0.98**
AdaBoost	5	98.1	97.5	98.6	0.96
	10	98.3	97.6	98.9	0.97

4. Discussion and Conclusion

In this study, liver disease was tried to be determined from clinical data by using bagging and boosting classification algorithms. Accuracy, precision, recall, kappa statistic values were obtained and the classifiers used were compared in terms of 5-10 fold cross-validation. Oversampling was applied as a pre-processing and the data set was balanced so that the classification could give a healthier result. As a result of the examination, the most successful performances were obtained with the Light GBM algorithm. The results of this study, which was carried out by examining the success rates obtained in studies on the predicted of similar data sets and liver disease, are compared in Table 4.

Table 4. Comparison of this study with similar literature studies.

Study	Dataset	Classifier model	Accuracy (%)
Oladimeji et al. [8]	HCV	RF	98.97
Orooji and Kermani [9]	HCV	RF	99.9
Mostafa and Hasan [10]	Liver Disease	SVM	98.23
Gupta and Gupta [11]	HCV	Stacked Ensemble	99.8
Suwardika [27]	HCV	SVM/ Regresi Logistik	80/79.4
Chicco and Jurman [28]	Liver Disease	RF	97.1
Hashem et al. [29]	Liver Disease	Alternating Decision Tree (ADTree)	95.6
This Study	Liver Disease	Light GBM	**98.8**

Looking at Table 4, it is seen that liver disease or HCV are predicted with high accuracy using different algorithms such as RF, decision trees, and SVM. In this study, unlike the literature, liver disease was predicted with an accuracy rate of 98.8% by using Light GBM, one of the boosting methods. The performance criteria obtained by using RF and J48 algorithms as bagging methods and Light GBM, one of the boosting methods, as well as GBC and AdaBoost algorithms, which support the literature, were compared. As a result of the comparison, the most successful performance measures were obtained with the Light GBM algorithm (Figure 3).

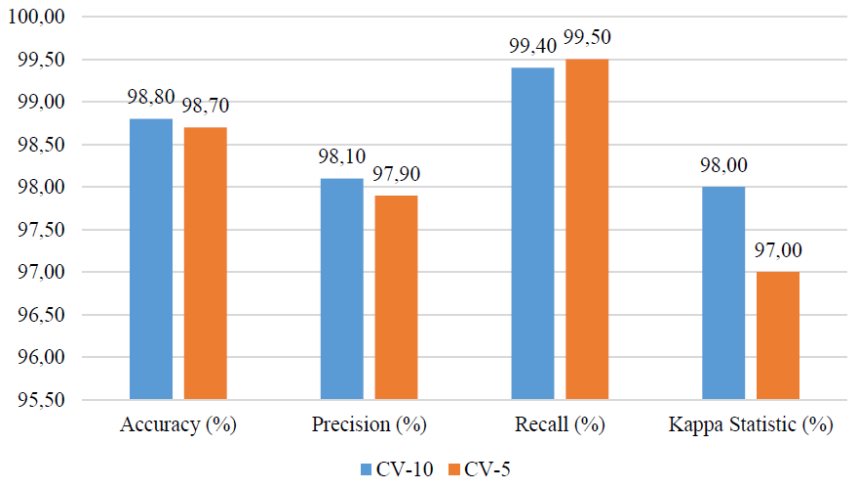

Figure 3. Light GBM Algorithm Performance Metrics

There are studies in the literature on predicting liver disease using clinical data using different algorithms [8-11, 27-29]. The results obtained in the study show that liver disease can be predicted using bagging and boosting algorithms using clinical data.

References

[1] World Health Organization (WHO), 2020. Hepatitis C Key Facts. https://www.who.int/newsroom/fact-sheets/detail/hepatitis-c (Accessed: September. 10, 2021).

[2] Hauri, A. M., Armstrong, G. L., & Hutin, Y. J. 2004. The global burden of disease attributable to contaminated injections given in health care settings. International journal of STD & AIDS, 15(1),pp.7-16.

[3] Khatun, M., & Ray, R. B. (2019). Mechanisms underlying hepatitis C virus-associated hepatic fibrosis. Cells, 8(10), 1249.

[4] Suk, K. T., & Kim, D. J. 2015. Staging of liver fibrosis or cirrhosis: The role of hepatic venous pressure gradient measurement. World journal of hepatology, 7(3), 607.

[5] Akkaya, O., Kiyici, M., Yilmaz, Y., Ulukaya, E., & Yerci, O. 2007. Clinical significance of activity of ALT enzyme in patients with hepatitis C virus. World journal of gastroenterology: WJG, 13(41), 5481.

[6] Pradat, P., Alberti, A., Poynard, T., Esteban, J. I., Weiland, O. et al. 2002. Predictive value of ALT levels for histologic findings in chronic hepatitis C: a European collaborative study. Hepatology, 36(4), pp.973-977.

[7] Awan, S. E., Bennamoun, M., Sohel, F., Sanfilippo, F. M., & Dwivedi, G. 2019. Machine learning-based prediction of heart failure readmission or death: implications of choosing the right model and the right metrics. ESC heart failure, 6(2), pp.428-435..

[8] Oladimeji, O. O., Oladimeji, A., Olayanju, O. 2021. Machine Learning Models for Diagnostic Classification of Hepatitis C Tests. Frontiers in Health Informatics, 10(1), 70.

[9] Orooji, A., Kermani, F. 2021. Machine learning based methods for handling imbalanced data in hepatitis diagnosis. Frontiers in Health Informatics, 10(1), 57.

[10] Mostafa, F. B., Hasan, E. 2021. Machine Learning Approaches for Binary Classification to Discover Liver Diseases using Clinical Data. medRxiv.

[11] Gupta, S., Gupta, M. K. 2021. A comprehensive data-level investigation of cancer diagnosis on imbalanced data. Computational Intelligence.

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

[13] Dua, D., Graff, C. 2019. UCI Machine Learning Repository Irvine, CA: University of California, School of Information and Computer Science. http://archive.ics.uci.edu/ml. (Accessed: April, 26, 2021).
[14] Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, pp.321-357.
[15] Yıldırım, P. 2016. Pattern classification with imbalanced and multiclass data for the prediction of albendazole adverse event outcomes. Procedia Computer Science, 83, pp.1013-1018.
[16] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. 2009. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), pp.10-18.
[17] Types of Ensemble methods in Machine learning. Accessed: September. 11, 2021. [Online]. Available: https://towardsdatascience.com/types-of-ensemble-methods-in-machine-learning-4ddaf73879db. (Accessed: September. 11, 2021).
[18] Pal, M. 2005. Random forest classifier for remote sensing classification. International journal of remote sensing, 26(1), pp.217-222.
[19] Breiman, L. 1996. Bagging predictors. Machine learning, 24(2), pp.123-140.
[20] Quinlan, J. R. 2014. C4. 5: programs for machine learning. Elsevier.
[21] Freund, Y., Schapire, R. E. 1996. Experiments with a new boosting algorithm. In icml Vol. 96, pp. 148-156.
[22] Skurichina, M., Duin, R. P. 2002. Bagging, boosting and the random subspace method for linear classifiers. Pattern Analysis & Applications, 5(2), pp.121-135.
[23] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., et al. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30, pp.3146-3154.
[24] An, T. K., Kim, M. H. 2010. A new diverse AdaBoost classifier. In 2010 International conference on artificial intelligence and computational intelligence. IEEE, (Vol. 1, pp. 359-363).
[25] Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. Annals of statistics, pp.1189-1232.
[26] Hossin, M., Sulaiman, M. N. 2015. A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process, 5(2), 1.
[27] Suwardika, G. 2017. Pengelompokan Dan Klasifikasi Pada Data Hepatitis Dengan Menggunakan Support Vector Machine (SVM), Classification And Regression Tree (Cart) Dan Regresi Logistik Biner. Journal of Education Research and Evaluation, 1(3), pp.183-191.
[28] Chicco, D., Jurman, G. 2021. An ensemble learning approach for enhanced classification of patients with hepatitis and cirrhosis. IEEE Access, 9, pp.24485-24498.
[29] Hashem, S., ElHefnawi, M., Habashy, S., El-Adawy, M., Esmat, G., et al. 2020. Machine Learning Prediction Models for Diagnosing Hepatocellular Carcinoma with HCV-related Chronic Liver Disease. Computer methods and programs in biomedicine, 196, 105551.