JOURNAL OF INONU LIVER TRANSPLANTATION INSTITUTE

DOI: 10.14744/jilti.2022.21931
J Inonu Liver Transpl Inst 2023;1(1):5–9

Original Research

Risk Prediction of Liver Cancer based on the Proposed Artificial Intelligence Approach

Zeynep Kucukakcali,[1] Ipek Balikci Cicek,[1] Fatma Hilal Yagin,[1] Sami Akbulut,[2] Cemil Colak[1]

[1]Department of Biostatistics and Medical Informatics, Inonu University, Malatya, Türkiye
[2]Department of Surgery and Liver Transplant Institute, Inonu University, Malatya, Türkiye

Abstract

Objectives: Liver cancer is a primary worldwide public health concern, and it is critical to understand the disease's physiology and create therapies. The aim of this study is to classify open access liver cancer data and identify important risk factors with the Random Forest method.

Methods: The open-access liver cancer dataset was used to construct a predive model in the study. Random Forest was used to classify the disease. Balanced accuracy, accuracy, sensitivity, specificity, positive/negative predictive values were evaluated for model performance. In addition, risk factors were assessed with the logistic regression model.

Results: The accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1 score metrics obtained with the Random Forest model were 98.9%, 97.9%, 95.8%, 100%, 100%, and 98.3%, and 95.7% respectively. Also, the importance of the variables obtained, the most important risk factors for liver cancer were total proteins, albumin and globulin ratio, albumin, age, total bilirubin, aspartate aminotransferase, direct bilirubin, alamine aminotransferase, alkaline phosphatase, respectively. According to the logistic regression model results, age, direct bilirubin, and albumin variables were statistically significant.

Conclusion: According to the study results, with the machine learning model Random forest used, patients with and without liver cancer were classified with high accuracy, and the importance of the variables related to cancer status was determined. Factors with high variable importance can be considered potential risk factors associated with cancer status and can play an essential role in disease diagnosis.

Keywords: Classification, Liver cancer, Machine learning, Random Forest

Please cite this article as "Kucukakcali Z, Balikci Cicek I, Yagin FH, Akbulut S, Colak C. Risk Prediction of Liver Cancer based on the Proposed Artificial Intelligence Approach. J Inonu Liver Transpl Inst 2023;1(1):5–9".

Primary liver cancer is the sixth most common illness and the third leading reason of death from cancer with 906.000 new cases and 830.000 deaths in the last years. The incidence of liver cancer is ranked fifth on a global scale, but it has the second highest mortality rate for males. In most regions, men have two to three times higher rates of both incidence and mortality than women, and liver cancer ranks fifth in terms of global incidence. Cirrhosis is the underlying condition that leads to the majority (90%) of all cases of liver cancer. Infection with hepatitis B virus is the most common risk factor for developing liver cancer in our country. Cirrhosis brought on by alcohol, hepatitis C, and obesity is the three main causes of fatty liver disease. On the other hand, due to the fact that viral infections are un-

Address for correspondence: Zeynep Kucukakcali, MD. Department of Biostatistics and Medical Informatics, Inonu University, Malatya, Türkiye
Phone: +90 536 424 32 06 **E-mail:** zeynep.tunc@inonu.edu.tr

Submitted Date: 20.05.2022 **Revised Date:** 23.05.2022 **Accepted Date:** 26.05.2022 **Available Online Date:** 27.04.2023

der control and obesity rates are gradually rising, it is anticipated that liver cancers caused by cirrhosis that is brought on by a fatty liver will take the lead in the coming years.[1]

Data mining is a set of methods used to reveal hidden patterns in databases. Data mining is the process of using computer programs to discover relationships and rules that will allow us to forecast the future from enormous amounts of data. The primary goal of data mining, according to the definition, is to keep much of data in the data warehouse and extract meaningful information from it.[2] Machine learning, which is one of these techniques, is a sub-field of data mining that aims to make predictions about new data when exposed to new data by performing data-based learning. Machine learning includes the design and development processes of algorithms aimed at realizing data-driven learning. From the input and output sets given by machine learning, the outputs of the previously unlearned inputs can be predicted.[3]

In 2001, Breiman proposed the Random Forest (RF) method, which is one of the machine learning methods, by developing the bagging method, which envisages combining the decisions of many variables trees, each of which is trained with several training sets, instead of producing a single decision tree. This method uses bootstrapping technique to create different sub-training sets and random feature selection in the development of trees. The difference from the bagging method is that instead of using all the variables in the data set during the tree development phase, as in the bagging method, it branches each node by using the best among the randomly chosen factors at each node. The trees are built according to randomly selected variables.[4]

The aim of this study is to classify patients with and without liver cancer using the RF method. In addition, it is to determine the risk factors related with liver cancer and to find the variable importance of cancer-related factors.

Methods

Dataset

The public dataset "ILPD (Indian Liver Patient Dataset) Data Set" was obtained from "https://www.kaggle.com/jeevannagaraj/indian-liver-patient-dataset" to classify the presence or absence of liver cancer via the RF method in the study. Explanations of the variables in the data set and their properties are given in Table 1.

Random Forest

RF is a classification/regression method proposed by Leo Breiman and Adele Cutler and includes the voting method. It consists of many decision trees coming together, and the individual trees are voted to determine the winning class. The decision trees in the forest are independent of one another and are built using the bootstrap technique from samples drawn from the data set.[4] The RF method is a forest classifier composed of several decision trees, and it can be used to establish classification or regression trees.[5] In the RF method, determining branching criteria and selecting a suitable pruning method are critical issues. The random forest classifier's branching criteria are determined using the Gini index method. The Gini index assesses the degree of weakness of class characteristics.[6] As in other classification methods, the RF method has parameters that the practitioner must determine. These parameters are the number of instances to be used at each node and the number of trees to be created, which are required in establishing the tree structure. In other words, during a classification process, the decision forest is created from K trees determined by the user.[7]

Table 1. Explanations of the variables in the dataset and their properties

Variable	Variable Description	Variable Type	Variable Role
Age	Patient's age	Quantitative	Predictor
Gender	Woman man	Qualitative	Predictor
tot_bilirubin	Total Bilirubin	Quantitative	Predictor
direct_bilirubin	Direct Bilirubin	Quantitative	Predictor
tot_proteins	Total Proteins	Quantitative	Predictor
albumin	Albumin	Quantitative	Predictor
ag_ratio	Albumin and Globulin Ratio	Quantitative	Predictor
sgpt	Alamine Aminotransferase	Quantitative	Predictor
sgot	Aspartate Aminotransferase	Quantitative	Predictor
alkphos	Alkaline Phosphatase	Quantitative	Predictor
is_patient	Sick/Not sick (the presence or absence of liver cancer)	Qualitative	Target

Data Analysis

To see if the variables had a normal distribution, the Kolmogorov-Smirnov test was used. The median (minimum-maximum) was used to summarize quantitative data, and the numbers were used to summarize qualitative variables (percentages). The Mann-Whitney U test was utilized to see if significant difference in the target exits. The logistic regression model was utilized by using a stepwise variable selection approach for target variable estimation. The model's fit was checked with Likelihood Ratio Test. P-value <0.05 was regarded significant. IBM SPSS Statistics 26.0 program was employed in the analysis.

Modeling

RF, one of the machine learning methods, was used in the modeling. Analyzes were carried out using the 10000 repeated bootstrap method. Balanced accuracy, accuracy, sensitivity, specificity, positive/negative predictive values, and F1-score were used as performance evaluation criteria.

Results

In the data set used in the study, there are 416 (71.4) liver cancer patients and 167 (28.6) without liver cancer patients, a total of 583 patients. Of the patients, 142 (24.4) were female, and 441 (75.6) were male.

Descriptive statistics for the target variable examined in this study are presented in Table 2. There is a significant difference between the diagnosis groups regarding other variables apart from the sgpt variable.

The results of the logistic regression model are given in Table 3. Odds ratios, their 95% confidence intervals (CI), and significance levels were also reported for convenience.

The results of the performance metrics obtained according to the results of the Random Forest model are given in Table 4. The model's fit was checked with Likelihood Ratio Tests (Chi-Square=110.048, df=3, p-value<0.001).

Accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1 score from the Random Forest model were 98.9%, 97.9%, 95.8%, 100%, 100%, and 98.3%, and 95.7% respectively.

In Figure 1, the values of performance criteria obtained from the RF model are plotted for visualization.

Variable importances obtained as a result of RF modeling are given in Table 5.

Figure 2 shows the importance levels of genes that are important for the Random Forest model.

Table 2. Descriptive statistics for target variables

Variables	is_patient		p*
	Patient (416) Median (Maks-Max)	Non Patient (167) Median (Maks-Max)	
Age	46 (7-90)	40 (4-85)	0.002
tot_bilirubin	1.40 (0.40-75)	0.80 (0.50-7.30)	<0.001
direct_bilirubin	0.50 (0.10-19.70)	0.20 (0.10-3.60)	<0.001
tot_proteins	229 (63-2110)	186 (90-1580)	<0.001
albumin	41 (12-2000)	27 (10-181)	<0.001
ag_ratio	53 (11-4929)	29 (10-285)	<0.001
sgpt	6.55 (2.70-9.60)	6.60 (3.70-9.20)	0.437
sgot	3.00 (0.90-5.50)	3.40 (1.40-5)	<0.001
alkphos	0.90 (0.30-2.80)	1.00 (0.37-1.90)	<0.001

*: Mann Whitney U test.

Table 3. Results of Logistic regression analysis

Variables in the Equation	Odds Ratio	95% CI for Odds Ratio		p
		Lower	Upper	
Intercept				0.001
Age	1.019	1.007	1.031	0.002
Direct Bilirubin	1.941	1.362	2.770	<0.001
Albumin	1.015	1.007	1.022	<0.001

Table 4. Values for the metrics of the classification performance of the Random Forest model

Metric	Value (%)
Accuracy	98.9
Balanced Accuracy	97.9
Sensitivity	95.8
Specificity	100
Positive predictive value	100
Negative predictive value	98.3
F1 score	95.7

Discussion

Liver cancer is a significant cause of cancer death, and its incidence is increasing. Liver cancers have a poor prognosis, and the etiology of the disease includes metabolic syndrome, obesity, chronic hepatitis B and C infection, cirrhosis, non-alcoholic steatohepatitis (NASH), and aflatoxin B1 or other mycotoxins and alcohol consumption. Because of the poor prognosis for liver cancer, scientists and doctors are looking for new treatment options to help patients live longer.[8, 9]

This study aims to classify liver cancer and reveal the risk factors associated with liver cancer using the open-access liver cancer dataset. For this purpose, variable importance values were calculated due to modeling by using the Random Forest method, one of the machine learning methods. In addition, the factors associated with cancer were determined by the logistic regression model. The accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1 score metrics obtained with the Random Forest model were 98.9%, 97.9%, 95.8%, 100%, 100%, and 98.3%, and 95.7% respectively. According to these results, the disease was classified correctly. According to the importance of the variables obtained, the most important risk factors for liver cancer were total proteins, albumin and globulin ratio, albumin, age, total bilirubin, aspartate aminotransferase, direct bilirubin, alamine aminotransferase, alkaline phosphatase, respectively. According to the logistic regression model results, age, direct bilirubin, and albumin variables were statistically significant and included in the model ($p<0.05$). An increase of one unit in the age variable increases the status of liver cancer by 1.02 (OR) fold. An increase of one unit in the direct bilirubin increases the condition of having liver cancer by 1.94 (OR) fold. A 1 (one) unit increase in the albumin variable increases the status of liver cancer by 1.02 (OR) fold. Multiple logistic regression analysis suggested three significant factors (i.e., age, direct bilirubin, albumin) associated with the presence or absence of liver cancer. When the outcomes of the LR model were assessed, the most significant predictor was direct bilirubin (OR=1.96), followed by age (OR=1.019) and albumin (OR=1.015) factors.

A study has presented the NBTree algorithm, a combination of the Decision Tree and Naive Bayes algorithms. The accuracy of the NB Tree method was 67.01%, but the accuracy of the Decision Tree and Nave Bayes algorithms were 66.14% and 56.14%, respectively.[10] Another study used Decision Tree, K-Nearest neighbor, and Logistic Regression models on the same dataset. In conclusion, the accuracy of

Table 5. Variable importances obtained as a result of RF

Variables	Variable importance (%)
tot_proteins	100
ag_ratio	97.44
albumin	93.08
age	81.52
tot_bilirubin	76.21
sgot	70.67
direct_bilirubin	65.32
sgpt	65.08
alkphos	58.76
gender	0

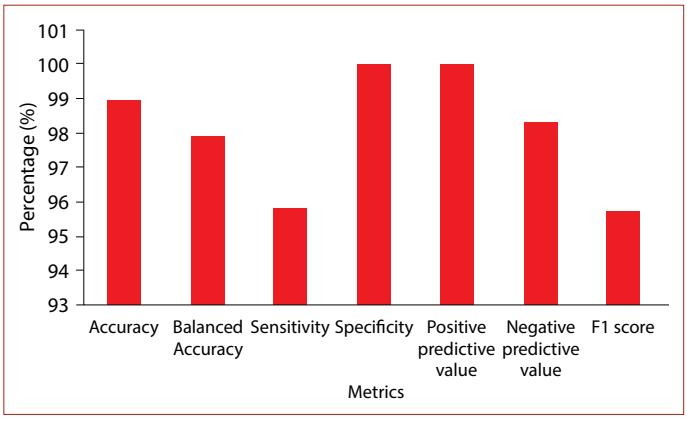

Figure 1. Graph of values for performance criteria obtained from Random Forest models.

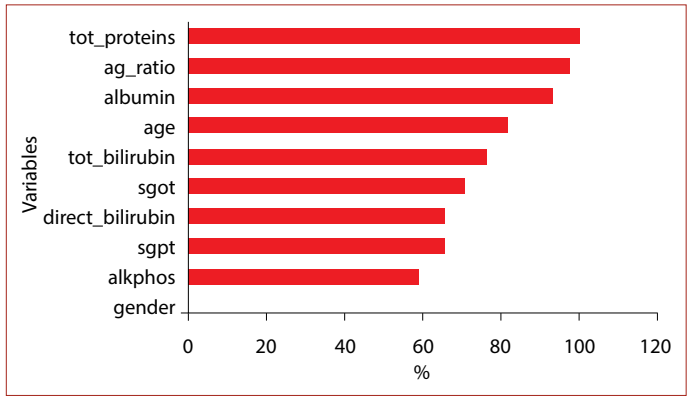

Figure 2. The graphic of variable importance values for the Random Forest model.

the Decision tree with the highest performance was 69.40.[11] In another study, Bayesian Network, Support Vector Machine, J48, Multi-Layer Perceptron, and Random Forest were performed on the same data. Thence, the Random Forest Algorithm produced the best performance with 71.87% accuracy.[12] In another study, they applied a support vector machine and Naive Bayes classification algorithms to the same data set. It was found that SVM outperformed Naive Bayes with 79.66% accuracy.[13] In a study using the same data set, logistic regression, support vector machines, random forest, AdaBoost, and bagging methods were employed for the classification task. The results obtained from the models used in the mentioned study were 73.5, 70.94, 66.66, 74.35, and 72.64, respectively.[14] In another study, the same data set was classified with Boosted C5.0 and CHAID, and the accuracies were obtained as 93.75% and 65%, respectively.[15] In a study that used the Indian Liver Patient Dataset, different classification algorithms such as Logistic Regression, K-NN and SVM were used for classification. The performances of these algorithms were evaluated for assessment metrics, and LR had the highest sensitivity.[16]

According to the study results, the proposed model (i.e., Random forest) can discriminate the patients with and without liver cancer with high performance. Factors with high variable importance can be considered possible risk factors associated with cancer status and can play an influential role in diagnosing the disease.

Disclosures

Ethics Committee Approval: Open-sourced data were used in the current study.

Peer-review: Externally peer-reviewed.

Conflict of Interest: None declared.

Authorship Contributions: Concept – Z.K., S.A.; Design – İ.B.Ç., F.H.Y., C.C.; Supervision – S.A.; Materials – Z.K., İ.B.Ç, F.H.Y.; Data collection &/or processing – Z.K.; Analysis and/or interpretation – Z.K.; Literature search – Z.K., İ.B.Ç., F.H.Y.; Writing – Z.K., S.A.; Critical review – S.A., C.C.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians 2021;71(3):209–49.
2. Akpınar H. Veri tabanlarında bilgi keşfi ve veri madenciliği. İstanbul Üniversitesi İşletme Fakültesi Dergisi. 2000;29(1):1-22.
3. Polikar R. Ensemble learning. Ensemble machine learning: Springer; 2012. p. 1–34.
4. Breiman L. Random forests. Machine learning 2001;45(1):5–32.
5. Akman M, Genç Y, Ankaralı H. Random forests methods and an application in health science. Turkiye Klinikleri J Biostat 2011;3:36–48.
6. Mather PM, Koch M. Computer processing of remotely-sensed images: an introduction: John Wiley & Sons; 2011.
7. Pal M. Random forest classifier for remote sensing classification. International Journal of Remote Sensing 2005;26:217–22.
8. Erstad DJ, Tanabe KK. Hepatocellular carcinoma: early-stage management challenges. Journal of hepatocellular carcinoma. 2017;4:81.
9. Anwanwan D, Singh SK, Singh S, Saikam V, Singh R. Challenges in liver cancer and possible treatment approaches. Biochimica et Biophysica Acta (BBA)-Reviews on Cancer 2020;1873(1):188314.
10. Alfisahrin SNN, Mantoro T, editors. Data mining techniques for optimization of liver disease classification. 2013 International Conference on Advanced Computer Science Applications and Technologies; 2013: IEEE.
11. Jin H, Kim S, Kim J. Decision factors on effective liver patient data prediction. International Journal of Bio-Science and Bio-Technology 2014;6(4):167–78.
12. Gulia A, Vohra R, Rani P. Liver patient classification using intelligent techniques. International Journal of Computer Science and Information Technologies 2014;5(4):5110–5.
13. Vijayarani S, Dhayanand S. Liver disease prediction using SVM and Naïve Bayes algorithms. International Journal of Science, Engineering and Technology Research (IJSETR) 2015;4(4):816–20.
14. Idris K, Bhoite S. Applications of machine learning for prediction of liver disease. Int J Comput Appl Technol Res 2019;8(9):394–6.
15. Abdar M, Zomorodi-Moghadam M, Das R, Ting I-H. Performance analysis of classification algorithms on early detection of liver disease. Expert Systems with Applications 2017;67:239–51.
16. Arbain AN, Balakrishnan BYP. A comparison of data mining algorithms for liver disease prediction on imbalanced data. International Journal of Data Science and Advanced Analytics (ISSN 2563-4429) 2019;1(1):1–11.