ISTANBUL UNIVERSITY PRESS

RESEARCH / ARAŞTIRMA
DOI: 10.26650/IUITFD.1130442
İst Tıp Fak Derg 2022 / J Ist Faculty Med 2022

MACHINE LEARNING-BASED CLASSIFICATION OF HBV AND HCV-RELATED HEPATOCELLULAR CARCINOMA USING GENOMIC BIOMARKERS

GENOMİK BİYOBELİRTEÇLER KULLANILARAK HBV VE HCV İLE İLİŞKİLİ HEPATOSELLÜLER KARSİNOMUN MAKİNE ÖĞRENİMİ TABANLI SINIFLANDIRILMASI

Sami AKBULUT[1,2,3], Zeynep KÜÇÜKAKÇALI[2], Cemil ÇOLAK[2]

[1]Inonu University, Faculty of Medicine, Department of General Surgery, Malatya, Turkiye
[2]Inonu University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Turkiye
[3]Innonu University, Faculty of Medicine, Department of Public Health, Malatya, Turkiye

ORCID IDs of the authors: S.A. 0000-0002-6864-7711; Z.K. 0000-0001-7956-9272; C.Ç. 0000-0001-5406-098X

Cite this article as: Akbulut S, Kucukakcali Z, Colak C. Machine learning-based classification of HBV and HCV-related hepatocellular carcinoma using genomic biomarkers. J Ist Faculty Med 2022;85(4):532-40. doi: 10.26650/IUITFD.1130442

ABSTRACT

Objective: It is crucial to know the underlying causes of hepatocellular carcinoma (HCC) for optimal management. This study aims to classify open access gene expression data of HCC patients who have an HBV or HCV infection using the XGboost method.

Material and Methods: This case-control study considered the open-access gene expression data of patients with HBV-related HCC and HCV-related HCC. For this purpose, data from 17 patients with HBV+HCC and 17 patients with HCV+HCC were included. XGboost was constructed for the classification via ten-fold cross-validation. Accuracy, balanced accuracy, sensitivity, specificity, the positive predictive value, the negative predictive value, and F1 score performance metrics were evaluated for a model performance.

Results: With the feature selection approach, 17 genes were chosen, and modeling was done using these input variables. Accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and the F1 score obtained from the XGboost model were 97.1%, 97.1%, 94.1%, 100%, 100%, 94.4%, and 97%, respectively. Based on the variable importance findings from the XGboost, the *ALDOC, GLUD2, TRAPPC10, FLJ12998, RPL39, KDELR2,* and *KIAA0446* genes can be employed as potential biomarkers for HBV-related HCC.

Conclusion: As a result of the study, two different etiological factors (HBV and HCV) causing HCC were classified using a machine learning-based prediction approach, and genes that could be biomarkers for HBV-related HCC were identified. After the

ÖZET

Amaç: Hepatoselüler karsinomun (HCC) optimal yönetimi için altında yatan nedenleri bilmek çok önemlidir. Bu çalışma, HBV veya HCV enfeksiyonu olan HCC hastalarının açık erişim gen ekspresyon verilerini XGboost yöntemini kullanarak sınıflandırmayı amaçlamaktadır.

Gereç ve Yöntem: Bu vaka-kontrol çalışmasında, HBV ve HCV ile ilişkili HCC'li hastaların açık erişimli gen ekspresyonu verileri dikkate alınmıştır. Bu amaçla, 17 HBV+HCC ve 17 HCV+HCC hastadan elde edilen veriler çalışmaya dahil edildi. Sınıflandırma için on katlı çapraz geçerlilik kullanılarak XGboost modeli oluşturuldu. Model performansı için doğruluk, dengeli doğruluk, duyarlılık, özgüllük, pozitif tahmin değeri ve negatif tahmin değeri ve F1 skor performans metrikleri değerlendirildi.

Bulgular: Özellik seçimi yaklaşımı ile 17 gen seçilmiş ve bu girdi değişkenleri kullanılarak modelleme yapılmıştır. XGboost modelinden elde edilen doğruluk, dengeli doğruluk, duyarlılık, özgüllük, pozitif tahmin değeri, negatif tahmin değeri ve F1 skor sırasıyla %97,1, %97,1, %94,1, %100, %100, %94,4 ve %97 idi. XGboost'tan elde edilen değişken önemliliği bulgularına dayanarak, *ALDOC, GLUD2, TRAPPC10, FLJ12998, RPL39, KDELR2* ve *KIAA0446* genleri, HBV ile ilişkili HCC için potansiyel biyobelirteçler olarak kullanılabilir.

Sonuç: Çalışma sonucunda, HCC'ye neden olan iki farklı etiyolojik faktör (HBV ve HCV), makine öğrenimi tabanlı bir tahmin yaklaşımı kullanılarak sınıflandırıldı ve HBV ile ilişkili HCC için biyobelirteç olabilecek genler tanımlandı. Ortaya çıkan genler sonraki araştırmalarda klinik olarak doğrulandıktan sonra, bu

Corresponding author/İletişim kurulacak yazar: Sami AKBULUT – akbulutsami@gmail.com

Submitted/Başvuru: 14.06.2022 • **Revision Requested/Revizyon Talebi:** 15.06.2022 •
Last Revision Received/Son Revizyon: 20.06.2022 • **Accepted/Kabul:** 07.07.2022 • **Published Online/Online Yayın:** 23.09.2022

resulting genes have been clinically validated in subsequent research, therapeutic procedures based on these genes can be established and their utility in clinical practice documented.

Keywords: Hepatocellular carcinoma, Hepatitis B infection, Hepatitis C infection, machine learning, classification

genlere dayalı terapötik prosedürler oluşturulabilir ve klinik uygulamada kullanımları belgelenebilir.

Anahtar Kelimeler: Hepatosellüler kanser, Hepatit B enfeksiyonu, Hepatit C enfeksiyonu, makine öğrenimi, sınıflandırma

INTRODUCTION

Primary liver cancer ranks as the sixth most prevalent kind of cancer that is diagnosed and the fourth most common cause of death from cancer globally (1). The great majority of instances of primary liver cancer, which account for roughly 75-85 percent of all cases, are caused by hepatocellular carcinoma (HCC) (1). The most important risk factors associated with HCC are Hepatitis B virus (HBV), Hepatitis C virus (HCV), alcohol abuse, non-alcoholic steatohepatitis (NASH), and non-alcoholic fatty liver disease (NAFLD) (1-3).

Hepatitis B virus contributes to the development of HCC via both direct and indirect mechanisms (4). Recent estimates demonstrate HBV is responsible for more than half of all HCC cases globally, ranking it second only to cigarettes as the most frequent carcinogen (5, 6). Chronic HBV carriers are 10- to 25-fold more likely to develop HCC throughout their lifetime than people without HBV (7). Alcohol consumption has a synergistic effect, increasing the carcinogenic risk of HBV by more than twofold. Tobacco use is also linked to an increased risk of HCC in patients with HBV-related cirrhosis, indicating a quantitative link between smoking and a cancer risk. In some subtropical areas of Asia and Africa, aflatoxin B1 exposure combined with HBV infection results in an exceptionally high HCC frequency (5, 8). Additionally, HBV replication, genotype, and HBV genomic mutations contribute to an increased likelihood of developing HCC. In the clinical environment, elevated levels of HBV DNA in the serum are linked to liver damage, the progression to cirrhosis, and the development of HCC (9, 10).

Hepatitis C virus is a hepatotropic RNA virus that only infects the liver and is spread through the bloodstream. HCV infects around 71 million individuals worldwide, yet only 20–30% of those infected develop liver cirrhosis, and only 1–4% of cirrhotic patients develop HCC each year (11, 12). The HCC risk is raised 15 to 20-fold in HCV-infected individuals, with the yearly incidence of HCC in cirrhotics estimated to be 1% to 4% over a 30-year period (13, 14). Over the last decade, mortality from HCV-related HCC has increased by 21.1%, whereas deaths from HCC caused by sources other than HCV and alcohol remained unchanged (14).

The role of many demographic, socioeconomic and clinical variables in the development of HCC has been studied in detail. However, the underlying molecular pathogenesis of HCC development such as genetic mutations and expression of gene products, has not been sufficiently clarified (15). The most important reasons for this are the popularity of genetic analyses in recent years, the lack of access to genetic tests, and the economic burden of these analyses. It is known that the genes or gene products play a vital role in the development of HCC. However, the comprehensive understanding molecular mechanism of HCC carcinogenesis and tumor prognosis remains unclear (15).

In recent years, in parallel with the development of next-generation sequencing (NGS) technology, important developments have been made regarding the molecular pathogenesis of HCC. In this context, the molecular mechanisms that play a role in the pathogenesis of HCC are roughly genomic, transcriptomic and, epigenetic alteration viral integration, tumor microenvironment, cancer stem cell, and cancer metabolism (16). Thanks to NGS, large-scale mutation screening and gene expression detection in HCCs has paved the way. However, instead of classical statistical analysis methods, it has become necessary to use artificial intelligence (AI) technology for their analysis and interpretation.

Machine learning (ML) is a subfield of AI that aims to make predictions about new data by performing data-driven learning when exposed to new data. AI/ML methods are one of the most commonly utilized technologies in illness detection and clinical decision support systems in recent years, with a wide range of applications. In the last decade, with the availability of large datasets and greater computing power, ML methods have achieved high performance in various situations (17, 18). Today, it is crucial to diagnose HCC and determine/predict the genes that cause the presence of HCC as biomarkers and use them concerning the HCC stage. For this reason, many studies have used ML methods to identify genes that may be biomarkers related to HCC (19). A study studied Non-Coding RNAs for HCV-associated HCC (20). Another study used ML to diagnose HCC with HCV (21). One study used gene expression profiling and supervised ML to predict HBV-related metastatic HCC (22). This study aims to classify open-access gene expression data of patients with HBV-related HCC and HCV-related HCC using the XGboost method and reveal important genes that may cause HCC.

MATERIAL AND METHODS

Data collection and variables

The present research originated from a case-control study published by Ueda et al. (23). The XGboost approach, one of the ML methods, was used to open-access gene expression data of HBV-related HCC and HCV-related HCC in the current investigation. For this purpose, data from 17 patients with HBV+HCC and 17 patients with HCV+HCC were included in the study. In the dataset, complementary DNA (cDNA) microarrays obtained from liver samples were used (23). cDNA is the double-stranded DNA version of the mRNA molecule. Since introns are cut out, researchers prefer to work with cDNA rather than mRNA. RNA is inherently more unstable than DNA. In addition, no amplification and purification technique can be applied to the RNA molecule (24). The primary output of the study is to classify HBV and HCV-associated HCC using machine learning methods and identify genes that may be biomarkers for HBV-related HCC.

Feature selection

Variable selection is an essential step in predictive modeling processes, and one of the most critical steps in developing a statistical model is deciding which data to include in the modeling. Feature selection identifies the most prominent features affecting a data set's dependent variable. Too many explanatory variables can lead to long computation times and the risk of over-learning the data and obtaining biased results (25). Most ML and data mining methods can produce ineffective results when working with extensive data. Therefore, these methods give more effective results when the dimensionality is reduced (26).

Gene expression data sets are pretty large. Modeling analyses take a long time because gene expression datasets are large, and these datasets can cause computational inefficiency in the analysis. LASSO, one of the feature selection methods, was used to solve these problems in this study. The LASSO method requires that the sum of the model parameters' absolute values be less than a fixed value (upper limit). The method achieves this by penalizing the coefficients of the regression variables, causing some of them to drop to zero. It is beneficial when the data set has a lot of variables and few observations. Furthermore, by removing irrelevant variables unrelated to the response variable, LASSO improves model interpretability and eliminates the problem of over-learning (27).

XGBoost

Gradient Boost is a powerful ML technique used for regression and classification problems where weak predictive models often produce ensemble forms of decision trees. Gradient Boost is based on boosting techniques (28, 29).

XGBoost, the abbreviation for Extreme Gradient Boosting, is one of the applications of gradient boosting machines (GBM), one of the most effective supervised learning algorithms. Its basic structure is based on gradient boosting and decision tree algorithms. Compared to other algorithms, it is in a very advantageous position regarding speed and performance. Gradient boosting is an ensemble method combining weak classifiers with boosting to create a strong classifier. The strong learner is trained iteratively, starting with a basic learner (29, 30).

Bioinformatics analysis

For the samples of HBV-related HCC and HCV-related HCC patients whose gene expression profiles were examined, differential expression analyses were performed using the limma package in the R programming language (31). Differential expression analysis is the statistical analysis of normalized read count data to find quantitative differences in expression activities between treatment arms. A pipeline is designed for the relevant analyses via the R software environment. The achieved results are presented from a table of genes in order of importance and a graph to visualize differentially expressed genes. The result table contains adjusted P and log2-fold change (log2FC) values, with genes with the smallest p values will be the most reliable. Log2FC>1 was used to identify up-regulated genes, and log2FC<-1 was used to identify down-regulated genes (32). A volcano plot was graphed to highlight quickly large values regarding the relevant genes.

Study protocol and ethics committee approval

This study, which was prepared using the National Center for Biotechnology Information Gene Expression Omnibus open-access dataset involving human participants, followed the ethical standards of the institutional and national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. Ethical approval was obtained from the Inonu University Institutional Review Board (IRB) for Non-Interventional Clinical Research (Date: 07.06.2022, No: 2022/3648). The STROBE (Strengthening the reporting of observational studies in epidemiology) guideline was utilized to assess the likelihood of bias and overall quality of this study (33).

Biostatistical analysis

The Shapiro Wilk test of normality was used to determine whether the variables had a normal distribution. Data were given as median (minimum-maximum) or mean± standard deviation. The Mann-Whitney U test was employed to compare non-normally distributed data, and an independent-sample t-test was utilized to compare non-normally distributed data where appropriate. Logistic regression analysis was performed to estimate each gene's odds ratio (a measure of effect size). Hosmer & Lemeshow's test for the goodness of fit and an omnibus

test of model coefficients were calculated for logistic regression. A P-value <0.05 was considered statistically significant. The IBM SPSS Statistics 25.0 program was used in the analysis.

Modeling process

The XGBoost, one of the ML methods, was used in the modeling. Analyses were carried out using the n-fold cross-validation method. In the n-fold cross-validation method, the data is first divided into n parts, and the model is applied to n parts. One of the n parts is used for testing, while the other n-1 parts are used for training the model. The mean of the obtained values is evaluated for the cross-validation method. In this study, 10-fold cross-validation was employed for the modeling process. Accuracy, balanced accuracy, sensitivity, selectivity, a positive predictive value, a negative predictive value, and an F1-score were used as performance evaluation criteria. In addition, variable importances were calculated, which gives information about how much the input variables explain to the output variable.

RESULTS

In the study, 34 HCC patients were used, of which 28 were male and six were female. The mean age of the patients was 61.7±9.4 years. While 15 of the HBV+HCC patients were male and two were female, 13 of the HCV+HCC patients were male, and four were female. The mean age of HBV+HCC patients is 60.5±9.0 years, and the mean age of HCV+HCC patients is 62.9±9.9 years. The dataset used contains 8516 expressions. According to the bioinformatics analysis, the first ten results are summarized concerning minimum adjusted-p values in Table 1. Based on the statistics from Table 1, two genes (ID: 7109 and 9136) were down-regulation, one gene (ID: 6412) was up-regulation, and the other seven genes were unregulated. According to Table 1, Log2FC values for the ID=7179, *ALDOC, RPL39, IFITM3, FLJ12998, KIAA0446, GLUD1, TNIP1, FLJ30092,* and *MRPS21* genes were -1,6096623, -0,8756088, -1,1163435, -0,8729706, -0,7040085, -0,9362293, 1,0475908, -0,7960824, -0,9509129, and -0,7807535, respectively. The volcano plot used to visualize differentially expressed genes is given in Figure 1. On the y- and x-axes, the volcano graph plots significance versus fold-change in log 2 base to observe differentially expressed genes quickly.

Seventeen expression results were obtained by applying the LASSO feature selection method to 8516 expression results. Table 2 presents descriptive statistics for the selected genes concerning the groups. The explanations of the data set with the selected expressions, the examined target variable, and the odds ratio per gene for the target variable are presented in Table 2. Based on the statistics in Table 2, significant differences were detected between groups in all genes (p<0.05). The findings of the performance metrics from the XGboost model are given in Ta-

Table 1: The results of the bioinformatics analysis

Gene ID	Gene	Gene product	Adj P value	p value	t	B	Log2FC	diffexpressed
7109	*GML*	glycosylphosphatidylinositol anchored molecule like	0.0000215	3.60E-09	-7.57217	10.3385	-1.60966	Down
2765	*ALDOC*	Aldolase C, fructose-bisphosphate	0.0006396	2.59E-07	-6.21883	6.59	-0.87561	No
9136	*RPL39*	Ribosomal protein L39	0.0006396	3.22E-07	-6.15147	6.3996	-1.11634	Down
4853	*IFITM3*	Interferon-induced transmembrane protein 3 (1-8U)	0.0021807	2.06E-06	-5.56952	4.7487	-0.87297	No
9176	*FLJ12998*	Hypothetical protein FLJ12998	0.0021807	2.13E-06	-5.55998	4.7216	-0.70401	No
7556	*KIAA0446*	KIAA0446 gene product	0.0021807	2.19E-06	-5.55042	4.6945	-0.93623	No
6412	*GLUD1*	Glutamate dehydrogenase 1	0.0022021	2.58E-06	5.498999	4.5485	1.047591	Up
3752	*TNIP1*	TNFAIP3 interacting protein 1	0.0040048	5.37E-06	-5.26902	3.8962	-0.79608	No
5909	*FLJ30092*	AF-1 specific protein phosphatase	0.0070976	1.07E-05	-5.05112	3.2804	-0.95091	No
7010	*MRPS21*	Mitochondrial ribosomal protein S21	0.0124601	2.09E-05	-4.83884	2.6838	-0.78075	No

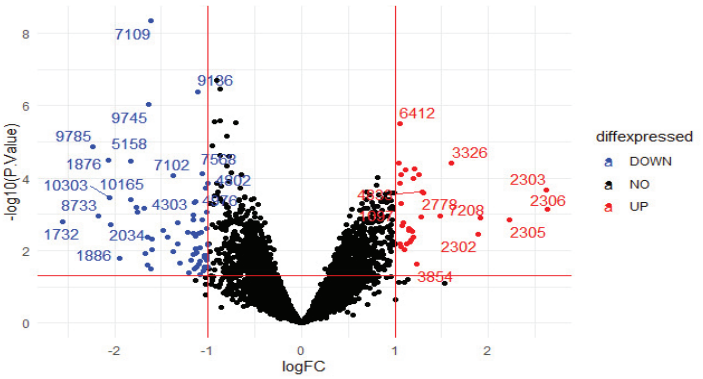

Figure 1: The volcano plot

Table 2: Descriptive statistics for Input variables

Gene	Prop number	Grups				OR	p
		HCV+HCC		HBV+HCC			
		Mean±SD	Median (min-max)	Mean±SD	Median (min-max)		
GLUD2	2747	0.87±0.31	0.89 (0.23-1.38)	-0.01±0.43	-0.02 (-0.64-0.72)	0.003	<0.001*
ALDOC	2765	0.53±0.39	0.52 (-0.33-1.17)	-0.34±0.35	-0.35 (-1.01-0.34)	0.003	<0.001*
TNIP1	3752	0.40±0.36	0.49 (-0.48-0.83)	-0.40±0.46	-0.46 (-1.24-0.61)	0.019	<0.001*
MX1	4303	0.75±1.07	0.43 (-1.32-2.98)	-0.39±0.72	-0.40 (-1.72-0.59)	0.187	0.001*
IFITM3	4853	0.55±0.40	0.56 (-0.17-1.14)	-0.32±0.46	-0.37 (-0.94-0.44)	0.011	<0.001*
C7orf30	4904	0.63±0.50	0.57 (-0.10-1.66)	-0.04±0.53	-0.01 (-1.56-0.72)	0.029	0.001*
RPL41	6171	-1.91±0.69	-1.83 (-3.47--0.77)	-0.98±0.82	-0.58 (-2.51-0.15)	4.779	0.004**
TRAPPC10	7109	1.74±0.69	1.69 (0.40-3.01)	0.13±0.57	0.01 (-0.78-1.94)	-	<0.001**
KIAA0446	7556	0.69±0.48	0.69 (-0.23-1.78)	-0.25±0.47	-0.11 (-1.72-0.35)	0.002	<0.001**
KDELR2	7919	0.33±0.50	0.22 (-0.57-1.38)	-0.34±0.49	-0.34 (-1.14-0.66)	0.050	<0.001*
OS-9	7949	0.17±0.38	0.22 (-0.45-1.10)	-0.36±0.38	-0.34 (-1.09-0.21)	0.014	<0.001*
ACP1	8178	0.16±0.25	0.12 (-0.24-0.69)	-0.23±0.28	-0.20 (-0.83-0.16)	0.001	<0.001*
RPL39	9136	0.99±0.52	0.90 (0.27-2.23)	-0.13±0.53	-0.26 (-1.02-0.59)	0.003	<0.001*
FLJ12998	9176	0.70±0.31	0.77 (0.09-1.13)	-0.01±0.32	-0.11 (-0.41-0.67)	0.001	<0.001*
WTAP	9589	0.55±0.33	0.58 (-0.31-1.05)	0.07±0.40	0.00 (-0.58-0.64)	0.028	0.001*
LMNA	9744	0.88±0.53	0.90 (-0.18-1.94)	0.09±0.76	0.27 (-2.30-1.60)	0.067	<0.001**
FKBP1A	10014	0.76±0.46	0.66 (0.13-1.84)	0.27±0.33	0.23 (-0.37-1.07)	0.022	0.001*

*: Independent sample t-test; **: Mann Whitney U test; OR. Odds ratio; SD: Standard deviation

Table 3: Performance metrics of the XGboost model

Metric	Value (%) (95% CI)
Accuracy	97.1 (91.4-100)
Balanced accuracy	97.1 (91.4-100)
Sensitivity	94.1 (71.3-99.9)
Specificity	100 (80.5-100)
Positive predictive value	100 (79.4-100)
Negative predictive value	94.4 (72.7-99.9)
F1 score	97 (91.2-100)

ble 3. Accuracy, balanced accuracy, sensitivity, specificity, the positive predictive value, the negative predictive value, and the F1 score obtained from the XGboost model were 97.1%, 97.1%, 94.1%, 100%, 100%, 94.4%, and 97%, respectively. The performance criteria values are plotted for the XGboost model in Figure 2. Figure 3 shows the importance levels of expressions for the selected genes in explaining the output variable. The *ALDOC* gene had the highest predictor importance of 100%, followed by *GLUD2* at 77.2 %, *TRAPPC10* at 59.2%, *FLJ12998* at 51.0%, *RPL39* at 33.2%, *KDELR2* at 24.8%, and *KIAA0446* at 23.8%.

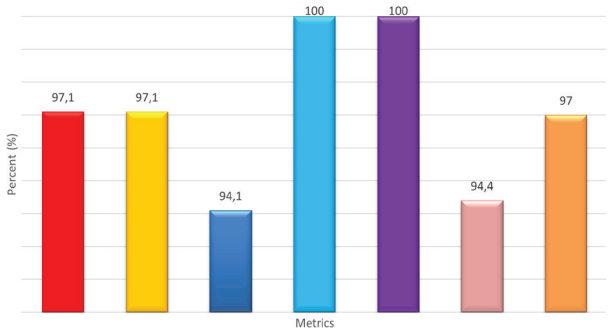

Figure 2: Graph of values for performance criteria obtained from XGboost models

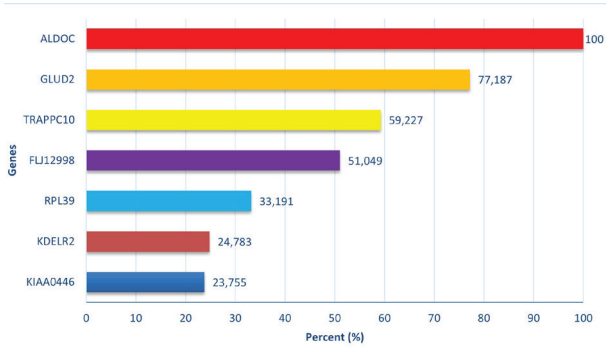

Figure 3: Graph of values for performance criteria obtained from XGboost models

DISCUSSION

Although the structure of gene expression profiling in HCC and the background liver tissue structure has been extensively studied, ML-based prediction of HBV-related HCC and HCV-related HCC and the detection of critical candidate biomarkers using an AI approach have not been clarified (23). The present study uses the XGboost method to classify HBV-related HCC and HCV-related HCC and identify important genes that may cause HBV-related HCC.

HCC is an aggressive type of cancer with well-defined epidemiological features. HCC continues to be an important public health problem worldwide, as it causes a significant economic and disease burden (1, 3, 34). The incidence and fatality rates of HCC vary significantly throughout the world. Discrepancies in the timing and quantity of exposure to environmental and infectious risk factors, the availability of healthcare resources, and the capacity to identify HCC at an earlier stage and administer possibly curative therapy are all variables that contribute to these differences (13, 35). HCC develops due to prolonged chronic hepatitis. In this case, patients have developed liver cirrhosis due to HBV or HCV infection. In patients with cirrhosis owing to chronic HBV or HCV infection, the annual incidence of HCC ranges from 2 to 5 percent overall. Chronic HBV and HCV infection are the major causes of HCC globally, accounting for 80% of all cases (34).

Except in northern Africa, where HCV incidence is most significant, chronic HBV infection is the primary cause of HCC throughout Eastern Asia and most African nations (36, 37). It is estimated that 257 million people worldwide have a chronic HBV infection. This situation leads to the high prevalence of chronic viral liver disease and HCC. It is also estimated that 20 million deaths can be attributed to acute hepatitis, chronic hepatitis, cirrhosis, and HCC caused by HBV between 2015 and 2030, with 5 million deaths from HCC alone (34).

HCV infection is still one of the most frequent bloodborne viral diseases and the leading cause of global infectious disease mortality (38, 39). HCV infection affects an estimated 71 million individuals worldwide, representing 1% of the population (40). Although direct-acting antiviral treatments have a high cure rate, 1.75 million new HCV infections and 400,000 HCV-related deaths occur yearly (41). HCV infection is a firmly established risk factor for HCC, increasing risk by 10- to 20-fold. Fatalities from HCV-related HCC grew by 21.1 percent during the last decade, but deaths from HCC caused by sources other than HCV and alcohol remained unchanged (14).

The overall survival of patients affected by HCC is low, and management of HCC risk factors needs to be rationally expanded to reduce the burden of HCC worldwide. There is a growing interest in genomics and molecular biology studies to identify early diagnosis and prognostic markers and new therapeutic targets to uncover the mechanisms of liver carcinogenesis and thus improve the clinical management of HCC patients (34, 42).

In the dataset investigated in this study, genomic data of samples obtained from liver tissues of 17 HBV-related HCC and 17 HCV-related HCC patients were used for the relevant analyses. cDNA microarrays were obtained from the samples, and the dataset used contained 8516 expressions. According to the Log2FC values used to determine the expression fold changes between the two groups from the bioinformatics analyses (detailed in Table 2), the GML gene has three-fold lower gene expression in HBV-related HCC patients than HCV-related HCC. Similarly, the RPL39 gene had a 2.15-fold lower gene expression. The *GLUD1* gene had two-fold upper gene expression in HBV-related HCC patients than in HCV-related HCC patients. Finally, the *ALDOC* gene, *IFITM3* gene, *FLJ12998* gene, *KIAA0446* gene, TNIP1 gene, *FLJ30092* gene, and *MRPS21* gene had the same expression between the two groups. In this instance, gene expression data are so large that modeling with these datasets can result in long analysis times and computational inefficiency in the analysis due to the size. Therefore, before mod-

eling with the existing data set, the most important genes associated with the output variable were selected with the Lasso variable selection method. Seventeen genes selected by the Lasso method were used in building Xgboost modeling. The accuracy, balanced accuracy, sensitivity, specificity, positive and negative predictive value, and F1 score metrics obtained with the XGboost model were 97.1%, 97.1%, 94.1%, 100%, 100%, 94.4%, and 97%, respectively. The performance metrics indicated that the proposed XGboost could correctly classify two groups of patients based on the AI approach. According to the variable importance obtained from the XGboost method, *ALDOC, GLUD2, TRAPPC10, FLJ12998, RPL39, KDELR2*, and *KIAA0446* genes can be used as candidates for predictive biomarkers for HBV-related HCC. According to the statistical analysis, 17 genes obtained by variable selection showed statistically significant differences for the two patient groups. Of the genes whose odds values were calculated, all genes, except *RPL41*, were downregulated in HBV-related HCC patients at significantly higher folds than in HCV-related HCC patients. The *RPL41* gene, on the other hand, was upregulated 4.779 fold in HBV-related HCC patients compared to HCV-related HCC patients. The OR values that were determined throughout the study and the Log2FC values support each other and support the values that were identified in the genes according to the variable significance. Additionally, the proposed pipeline produced a volcano plot, representing the up-and-down-regulation of the genes in this research. These plots are becoming more common in omics experiments such as genomics, proteomics, and metabolomics, where thousands of replicate data points between two conditions are often present.

One study reported that the *ALDOC* gene is associated with HBV-related HCC and is up-regulated by the MLX protein (43). In another study, it was reported that *ALDOC* was up-regulated in patients with HBV (44). In a study using matched tumor and adjacent liver tissues from 159 patients with HBV-related HCC, *GLUD2* showed high expression (45). Another study showed *GLUD2 d*own-regulation for the same condition (46). Another study found *GLUD2*, a potentially relevant gene for HCC (47). In one study, overexpression of RPL39 was reported to be associated with HCC (48). In one study, *KDELR2* was identified as a potential gene associated with HBV (49).

As it is known, all diseases that cause chronic liver damage are risk factors for the development of HCC. Therefore, international guidelines' follow-up of such patients is crucial for detecting possible HCC or its detection at an early stage (50). The most authoritative guidelines on monitoring chronic liver patients are published periodically by EASL, APASL, and AASLD (50). The above guidelines suggest that patients with chronic liver disease without suspected HCC should be followed up with ultrasonography and AFP at six-month intervals (50). Patients with suspected HCC should be followed up with ultrasound and AFP at three or six-month intervals. Patients with a strong suspicion of HCC should be followed up with ultrasound and AFP.

However, these approaches may not always give the expected results because it is not always easy for patients to reach healthcare providers in underdeveloped or developing countries. False-negative results may be higher than expected, especially since ultrasonography is an operator-dependent examination. It is a known fact that there is a correlation between the duration of chronic liver disease and the probability of developing HCC. In addition, as in all other cancer types, gene mutation and mutation-related mRNA expression changes are expected in HCC. Therefore, in the follow-up of patients with chronic liver disease, fundamental genetic analysis can be performed after a certain period to determine whether there is a genetic mutation. As seen in the results of this study, if changes in the expression of genes strongly associated with HCC are detected, and ideas are formed about the genetic mechanism underlying the different etiologies that cause HCC, patients can be followed more closely, and preventive treatments can be started when necessary. However, there is no evidence-based data on when genetic analysis should be performed on chronic liver disease. Therefore, a prospective multicenter study is needed on the timing of genetic analysis for patients with chronic liver disease. With this vital finding, increasing the number of patients may further increase the scope of genetic information and the power of the study.

CONCLUSION

In conclusion, this study identified potential genomic biomarkers for HBV-associated HCC using gene expression data from patients with HBV-associated HCC and HCV-associated HCC. The reliability of the genes discovered in the future, more thorough analyses may be evaluated, therapy techniques can be devised based on these genes, and their clinical utility can be detailed.

Ethics Committee Approval: This study was approved by Inonu University Institutional Review Board (IRB) for Non-Interventional Clinical Research (Date: 07.06.2022, No: 2022/3648).

Peer Review: Externally peer-reviewed.

Author Contributions: Conception/Design of Study- S.A., Z.K.; Data Acquisition- S.A., Z.K.; Data Analysis/Interpretation- Z.K., C.C.; Drafting Manuscript- S.A., Z.K.; Critical Revision of Manuscript- S.A., Z.K., C.C.; Approval and Accountability- S.A., Z.K., C.C.

Conflict of Interest: The authors have no conflict of interest to declare.

Financial Disclosure: The authors declared that this study has received no financial support.

REFERENCES

1. Aly A, Ronnebaum S, Patel D, Doleh Y, Benavente F. Epidemiologic, humanistic and economic burden of hepatocellular carcinoma in the USA: a systematic literature review. Hepat Oncol 2020;7(3):HEP27. doi: 10.2217/hep-2020-0024. [CrossRef]
2. Llovet JM, Kelley RK, Augusto V, Singal AG, Eli P, Sasan R, et al. Hepatocellular carcinoma. Nat Rev Dis Primers 2021;7(1):6. [CrossRef]
3. Sayiner M, Golabi P, Younossi ZM. Disease burden of hepatocellular carcinoma: a global perspective. Dig Dis Sci 2019;64(4):910-7. [CrossRef]
4. Levrero M, Zucman-Rossi J. Mechanisms of HBV-induced hepatocellular carcinoma. J Hepatol 2016;64(1 Suppl):S84-101. [CrossRef]
5. El–Serag HB, Rudolph KL. Hepatocellular carcinoma: epidemiology and molecular carcinogenesis. Gastroenterology 2007;132(7):2557-76. [CrossRef]
6. Venook AP, Papandreou C, Furuse J, Ladrón de Guevara L. The incidence and epidemiology of hepatocellular carcinoma: a global and regional perspective. Oncologist 2010;15(S4):5-13. [CrossRef]
7. Fattovich G, Stroffolini T, Zagni I, Donato F. Hepatocellular carcinoma in cirrhosis: incidence and risk factors. Gastroenterology 2004;127(5):S35-50. [CrossRef]
8. Ming L, Thorgeirsson SS, Gail MH, Lu P, Harris CC, Wang N, et al. Dominant role of hepatitis B virus and cofactor role of aflatoxin in hepatocarcinogenesis in Qidong, China. Hepatology 2002;36(5):1214-20. [CrossRef]
9. Chen C-J, Yang H-I, Su J, Jen C-L, You S-L, Lu S-N, et al. Risk of hepatocellular carcinoma across a biological gradient of serum hepatitis B virus DNA level. JAMA 2006;295(1):65-73. [CrossRef]
10. Di Bisceglie AM. Hepatitis B and hepatocellular carcinoma. Hepatology 2009;49(S5): S56-60. [CrossRef]
11. Dash S, Aydin Y, Widmer KE, Nayak L. Hepatocellular carcinoma mechanisms associated with chronic HCV infection and the impact of direct-acting antiviral treatment. J Hepatocell Carcinoma 2020;7:45. [CrossRef]
12. 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):S45-57. [CrossRef]
13. El-Serag HB. Epidemiology of viral hepatitis and hepatocellular carcinoma. Gastroenterology 2012;142(6):1264-73. [CrossRef]
14. Axley P, Ahmed Z, Ravi S, Singal AK. Hepatitis C virus and hepatocellular carcinoma: a narrative review. J Clin Transl Hepatol 2018;6(1):79. [CrossRef]
15. Hu B, Ma X, Fu P, Sun Q, Tang W, Sun H, et al. miRNA-mRNA regulatory network and factors associated with prediction of prognosis in hepatocellular carcinoma. Genomics Proteomics Bioinformatics 2021;S1672-0229(21)00059-0. [Epub ahead of print] [CrossRef]
16. Ho DW, Lo RC, Chan LK, Ng IO. Molecular pathogenesis of hepatocellular carcinoma. Liver Cancer 2016;5(4):290-302. [CrossRef]
17. Polikar R. Ensemble learning. Ensemble machine learning. Springer; 2012: pp. 1-34. [CrossRef]
18. Akman M, Genç Y, Ankarali H. Random forests yöntemi ve saglık alanında bir uygulama/Random forests methods and an application in health science. Turkiye Klinikleri J Biostat 2011;3(1):36-48.
19. Piñero F, Dirchwolf M, Pessôa MG. Biomarkers in hepatocellular carcinoma: diagnosis, prognosis and treatment response assessment. Cells 2020;9(6):1370. [CrossRef]
20. Plissonnier M-L, Herzog K, Levrero M, Zeisel MB. Non-coding RNAs and hepatitis C virus-induced hepatocellular carcinoma. Viruses 2018;10(11):591. [CrossRef]
21. Hashem S, ElHefnawi M, Habashy S, El-Adawy M, Esmat G, Elakel W, et al. Machine learning prediction models for diagnosing hepatocellular carcinoma with HCV-related chronic liver disease. Comput Methods Programs Biomed 2020;196:105551. [CrossRef]
22. Ye Q-H, Qin L-X, Forgues M, He P, Kim JW, Peng AC, et al. Predicting hepatitis B virus–positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning. Nat Med 2003;9(4):416-23. [CrossRef]
23. Ueda T, Honda M, Horimoto K, Aburatani S, Saito S, Yamashita T, et al. Gene expression profiling of hepatitis B-and hepatitis C-related hepatocellular carcinoma using graphical Gaussian modeling. Genomics 2013;101(4):238-48. [CrossRef]
24. Chang HY, Thomson JA, Chen X. Microarray analysis of stem cells and differentiation. Methods Enzymol 2006;420:225-54. [CrossRef]
25. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics 2007;23(19):2507-17. [CrossRef]
26. Fodor IK. A Survey of Dimension Reduction Techniques. Lawrence Livermore National Lab (CA). Department of Energy (US). 2002 May. Report No: UCRL-ID-148494 TRN: US200408%%150. [CrossRef]
27. Fonti V. Research Paper in Business Analytics: Feature Selection with LASSO. Amsterdam: VU Amsterdam 2017.
28. Wang J, Li P, Ran R, Che Y, Zhou Y. A short-term photovoltaic power prediction model based on the gradient boost decision tree. Appl Sci 2018;8(5):689. [CrossRef]
29. Salam Patrous Z. Evaluating XGBoost for User Classification by using Behavioral Features Extracted from Smartphone Sensors (dissertation). Stockholm: KTH Royal Institute of Technology. 2018.
30. Dikker J. Boosted tree learning for balanced item recommendation in online retail (dissertation). Eindhoven: The Eindhoven University of Technology. 2017.
31. Smyth GK. Limma: linear models for microarray data. Bioinformatics and computational biology solutions using R and Bioconductor. In: Gail M, Samet JM. Statistics for Biology and Health. Springer; 2005: pp. 397-420. [CrossRef]
32. Yan H, Zheng G, Qu J, Liu Y, Huang X, Zhang E, et al. Identification of key candidate genes and pathways in multiple myeloma by integrated bioinformatics analysis. J Cell Physiol 2019; 234(12):23785-97. [CrossRef]
33. Cevallos M, Egger M, Moher D. STROBE (Strengthening the Reporting of Observational Studies in Epidemiology). In: Moher D, Altman DG, Schulz KF, Simera I, Wager E, editors. Guidelines for Reporting Health Research: A User's Manual. John Wiley & Sons, Ltd; 2014. p. 169-79. [CrossRef]
34. Yang JD, Hainaut P, Gores GJ, Amadou A, Plymoth A, Roberts LR. A global view of hepatocellular carcinoma: trends, risk, prevention and management. Nat Rev Gastroenterol Hepatol 2019;16(10):589-604. [CrossRef]

35. Tang A, Hallouch O, Chernyak V, Kamaya A, Sirlin CB. Epidemiology of hepatocellular carcinoma: target population for surveillance and diagnosis. Abdom Radiol (NY) 2018;43(1):13-25. [CrossRef]
36. Park JW, Chen M, Colombo M, Roberts LR, Schwartz M, Chen PJ, et al. Global patterns of hepatocellular carcinoma management from diagnosis to death: the BRIDGE Study. Liver Int 2015;35(9):2155-66. [CrossRef]
37. Yang JD, Gyedu A, Afihene MY, Duduyemi BM, Micah E, Kingham PT, et al. Hepatocellular carcinoma occurs at an earlier age in Africans, particularly in association with chronic hepatitis B. Am J Gastroenterol 2015;110(11):1629-31. [CrossRef]
38. Jefferies M, Rauff B, Rashid H, Lam T, Rafiq S. Update on global epidemiology of viral hepatitis and preventive strategies. World J Clin Cases 2018;6(13):589-99. [CrossRef]
39. Hill AM, Nath S, Simmons B. The road to elimination of hepatitis C: analysis of cures versus new infections in 91 countries. J Virus Erad 2017;3(3):117-23. [CrossRef]
40. Marshall AD, Pawlotsky J-M, Lazarus JV, Aghemo A, Dore GJ, Grebely J. The removal of DAA restrictions in Europe–one step closer to eliminating HCV as a major public health threat. J Hepatol 2018;69(5):1188-96. [CrossRef]
41. Page K, Melia MT, Veenhuis RT, Winter M, Rousseau KE, Massaccesi G, et al. Randomized trial of a vaccine regimen to prevent chronic HCV infection. N Engl J Med 2021;384(6):541-9. [CrossRef]
42. Ghidini M, Braconi C. Non-coding RNAs in primary liver cancer. Front Med (Lausanne) 2015;2:36. [CrossRef]
43. Xie Q, Fan F, Wei W, Liu Y, Xu Z, Zhai L, et al. Multi-omics analyses reveal metabolic alterations regulated by hepatitis B virus core protein in hepatocellular carcinoma cells. Sci Rep 2017;7(1):41089. [CrossRef]
44. Li H, Zhu W, Zhang L, Lei H, Wu X, Guo L, et al. The metabolic responses to hepatitis B virus infection shed new light on pathogenesis and targets for treatment. Sci Rep 2015;5(1):8421. [CrossRef]
45. Gao Q, Zhu H, Dong L, Shi W, Chen R, Song Z, et al. Integrated proteogenomic characterization of HBV-related hepatocellular carcinoma. Cell 2019;179(2):561-77. [CrossRef]
46. Wei X, Su R, Yang M, Pan B, Lu J, Lin H, et al. Quantitative proteomic profiling of hepatocellular carcinoma at different serum alpha-fetoprotein level. Translational oncology 2022;20:101422. [CrossRef]
47. Lu M, Kong X, Wang H, Huang G, Ye C, He Z. A novel microRNAs expression signature for hepatocellular carcinoma diagnosis and prognosis. Oncotarget 2017;8(5):8775-84. [CrossRef]
48. El Khoury W, Nasr Z. Deregulation of ribosomal proteins in human cancers. Biosci Rep 2021;41(12):BSR20211577. [CrossRef]
49. Li F, Deng Y, Zhang S, Zhu B, Wang J, Wang J, et al. Human hepatocyte-enriched miRNA-192-3p promotes HBV replication through inhibiting Akt/mTOR signalling by targeting ZNF143 in hepatic cell lines. Emerg Microbes Infect 2022;11(1):616-28. [CrossRef]
50. Akbulut S, Garzali IU, Hargura AS, Aloun A, Yilmaz S. Screening, Surveillance, and Management of Hepatocellular Carcinoma During the COVID-19 Pandemic: a Narrative Review. J Gastrointest Cancer 2022:1-12. [CrossRef]