DOI: 10.4274/MMJ.galenos.2022.39049
Medeni Med J 2022;37:255-263

Original Article

Machine Learning-based Prediction of HBV-related Hepatocellular Carcinoma and Detection of Key Candidate Biomarkers

HBV ile İlişkili Hepatosellüler Karsinomun Makine Öğrenimi Tabanlı Tahmini ve Aday Biyobelirteçlerin Tespiti

Zeynep KUCUKAKCALI[1], Sami AKBULUT[1,2,3], Cemil COLAK[1]

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

ABSTRACT

Objective: This study aimed to classify open-access gene expression data of patients with hepatitis B virus-related hepatocellular carcinoma (HBV + HCC) and chronic HBV without HCC (HBV alone) using the XGBoost method, one of the machine learning methods, and reveal important genes that may cause HCC.

Methods: This case-control study used the open-access gene expression data of patients with HBV + HCC and HBV alone. Data from 17 patients with HBV + HCC and 36 patients with HBV were included. XGBoost was constructed for the classification via 10-fold cross-validation. Accuracy, balanced accuracy, sensitivity, selectivity, positive-predictive value, and negative-predictive value performance metrics were evaluated for model performance.

Results: According to the feature-selection method, 18 genes were selected, and modeling was performed with these input variables. Accuracy, balanced accuracy, sensitivity, specificity, positive-predictive value, negative-predictive value, and F1 score obtained from XGBoost model were 98.1%, 98.6%, 100%, 97.2%, 94.4%, 100%, and 97.1%, respectively. Based on the predictor importance findings acquired from XGBoost, the *RNF26, FLJ10233, ACBD6, RBM12, PFAS, H3C11,* and *GKP5* can be employed as potential biomarkers of HBV-related HCC.

Conclusions: In this study, genes that may be possible biomarkers of HBV-related HCC were determined using a machine learning-based prediction approach. After the reliability of the obtained genes are clinically verified in subsequent research, therapeutic procedures can be established based on these genes, and their usefulness in clinical practice may be documented.

Keywords: Hepatocellular carcinoma, hepatitis B infection, chronic liver disease, gene expression

ÖZ

Amaç: Bu çalışma, makine öğrenmesi yöntemlerinden XGBoost yöntemi kullanılarak hepatit B virüsü ilişkili hepatosellüler karsinom (HBV + HCC) ve HCC'siz kronik HBV (tek başına HBV) olan hastaların açık erişimli gen ekspresyon verilerini sınıflandırmayı ve HCC'ye neden olabilecek önemli genleri ortaya çıkarmayı amaçlamaktadır.

Yöntemler: Bu olgu-kontrol çalışmasında, yalnızca HBV + HCC ve HBV'li hastaların açık erişimli gen ekspresyonu verileri kullanılmıştır. Bu amaçla, çalışmaya HBV + HCC'li 17 hastadan ve tek başına HBV'li 36 hastadan alınan veriler dahil edildi. Sınıflandırma için on katlı çapraz geçerlilik yoluyla XGBoost modeli oluşturuldu. Model performansı için doğruluk, dengelenmiş doğruluk, duyarlılık, seçicilik, pozitif tahmin değeri, negatif tahmin değeri ve F1 skor performans metrikleri değerlendirildi.

Bulgular: Özellik seçim yöntemine göre 18 gen seçilmiş ve bu girdi değişkenleri ile modelleme yapılmıştır. XGBoost modelinden elde edilen doğruluk, dengelenmiş doğruluk, duyarlılık, özgüllük, pozitif prediktif değer, negatif prediktif değer ve F1 skoru sırasıyla %98,1, %98,6, %100, %97,2, %94,4, %100 ve %97,1 idi. XGBoost'tan elde edilen değişken önemliliği değerlerine göre, *RNF26, FLJ10233, ACBD6, RBM12, PFAS, H3C11* ve *GKP5* genleri, HBV ile ilişkili HCC için potansiyel biyobelirteçler olarak kullanılabilir.

Sonuçlar: Araştırma sonucunda, makine öğrenmesi temelli tahmin yaklaşımı ile HBV ile ilişkili HCC için olası biyobelirteç olabilecek genler belirlendi. Elde edilen genlerin güvenilirliği sonraki araştırmalarda klinik olarak doğrulandıktan sonra, bu genlere dayalı olarak terapötik prosedürler oluşturulabilir ve bunların klinik pratikteki yararları belgelenebilir.

Anahtar kelimeler: Hepatosellüler karsinom, hepatit B enfeksiyonu, kronik karaciğer hastalığı, gen ifadesi

Address for Correspondence: S. Akbulut, Inonu University Faculty of Medicine, Department of Biostatistics and Medical Informatics, General Surgery and Public Health, Malatya, Turkey
E-mail: akbulutsami@gmail.com **ORCID ID:** orcid.org/0000-0002-6864-7711

Received: 17 June 2022
Accepted: 07 August 2022
Online First: 23 August 2022

Cite as: Kucukakcali Z, Akbulut S, Colak C. Machine Learning-based Prediction of HBV-related Hepatocellular Carcinoma and Detection of Key Candidate Biomarkers. Medeni Med J 2022;37:255-263

INTRODUCTION

Current epidemiological and clinical data indicate that primary liver cancer is the sixth most frequently diagnosed cancer and the fourth among cancer-related deaths worldwide[1]. Approximately 841,000 people are diagnosed with primary liver cancer each year, and 782,400 people died from it. Hepatocellular carcinoma (HCC) accounts the majority of primary liver cancer cases. HCC is the world's fifth most common malignant tumor, with the second-highest mortality rate among malignant tumors[2,3]. The most important risk factors associated with HCC are hepatitis B virus (HBV), hepatitis C virus, alcohol abuse, and non-alcoholic fatty liver disease[4].

HBV infection is a global public health problem that causes significant morbidity and mortality. HBV is responsible for more than half of all HCC cases worldwide. The proportion of HCC attributable to HBV reflects the geographic distribution of HBV infection and varies significantly, accounting for <20% of all HCC cases in the United States and up to 65% in China and the Far East. Chronic HBV carriers have a 10- to 25-fold higher lifetime risk of developing HCC than non-infected ones[5].

Epidemiological studies have shown that many risk factors, especially hepatotropic viruses such as HBV, affect HCC development. Three basic mechanisms are suggested for HCC development from the background of HBV infection: (1) development of chronic inflammation and hepatocyte regeneration during the HBV infection process, (2) activation of the host genes responsible for proliferation as a result of the integration of the HBV DNA into the host genome, and (3) HBV-related proteins (HBx, etc.) support cell proliferation[6]. These results show that HBV-related HCC is considered not only a clinical disease but also a disease with a genetic basis. The biologically different behavioral patterns of the tumor indicate that genetic and epigenetic aberrations may be important in the HCC development and course[7,8]. With the detection of genetic and epigenetic anomalies in the pathogenesis of HCC, studies on the molecular pathogenesis of HCC have gained tremendous momentum in the last two decades. In these studies, thousands of genes, transcription, and translation pathways associated with these genes are analyzed, which is a complex and challenging process. Therefore, artificial intelligence (AI) models are needed to analyze thousands of data and interpret the analyses.

Machine learning (ML) is a subfield of AI that make predictions about new data by performing data-driven learning when exposed to new data. AI/ML methods are one of the technologies widely used in diagnosing diseases and clinical decision support systems in recent years and have a wide application area. ML has a wide application area in health and constitutes the basic infrastructure of applications in determining genetic diseases, early diagnosis of cancer, and identifying patterns in medical imaging. In the last decade, with the availability of large datasets and greater computing power, ML methods have achieved high performance in various situations[9,10]. At present, it is essential to diagnose HCC, determine or predict the genes that cause HCC as biomarkers, and use them concerning the HCC stage. Thus, many studies have used ML methods to identify genes that may be biomarkers related to HCC[11]. A study used gene expression profiling and supervised ML to predict HBV-positive metastatic HCCs[12]. In another study, genes that could be biomarkers were identified by ML methods using genome-wide data to predict relapse in patients with HCC[13]. This study aimed to classify open-access gene expression data of patients with HBV-related HCC (HBV + HCC) and chronic HBV without HCC (HBV alone) using XGBoost and reveal important genes that may cause HCC.

MATERIALS and METHODS

Study Design and Data

This is a retrospective case-control study, and XGBoost, one of the ML methods, was applied to open-access gene expression data of patients with HBV-related HCC and chronic HBV without HCC. Data from 17 HBV-related HCC and 36 chronic HBV samples were analyzed. Complementary DNA (cDNA) microarrays obtained from liver samples were used[14]. cDNA refers to a piece of DNA synthesized from a mature mRNA used as a template in a reaction catalyzed by the enzyme reverse transcriptase. cDNA is the double-stranded DNA version of the mRNA molecule. mRNA is more helpful in determining polypeptide sequence than the genomic sequence in eukaryotes. Since introns are cut out, researchers prefer to work with cDNA rather than mRNA. Therefore, RNA is inherently more unstable than DNA. In addition, no amplification and purification technique can be applied to the RNA molecule. mRNA is used as a template, and reverse transcriptase synthesizes single-stranded DNA molecules. This molecule is then utilized to synthesize double-stranded DNA[15].

Feature Selection

Variable selection is an essential step in predictive modeling processes. One of the most critical steps in developing a statistical model is deciding which data to include in the model. Before working with large datasets and models with high computational costs,

determining the most valuable features of the dataset to be used in the study will lead to highly efficient results. Feature selection identifies the most prominent features that affect a data set's dependent variable. The use of numerous explanatory variables can lead to long computation times and risk of overlearning the data and obtaining biased results. In addition, models created with numerous variables are challenging to interpret. Before statistical modeling, selecting important variables that affect the dependent variable is recommended[16]. 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[17].

Gene expression datasets are large and complex and include raw data for the analyses. Modeling analyses take a long time because gene expression datasets are large, and these datasets can cause computational inefficiency in the analysis. As a result of the high-dimensionality issue, the model's performance may suffer. A classification algorithm can also overfit the training samples and under generalize new samples if there are numerous genes in gene expression datasets. In this study, LASSO, one of the feature-selection methods, was used to solve these problems. 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. Besides, the dataset should have many variables and few observations. Furthermore, by removing irrelevant variables unrelated to the response variable, LASSO improves model interpretability and eliminates overlearning[18].

XGBoost Algorithm

Gradient boost is defined as a powerful ML technique for regression and classification problems where weak predictive models often produce ensemble forms of decision trees. Gradient boost aims to construct many weak learners in sequence and incorporate them into a complex model because it is based on the boosting method[19].

XGBoost, the abbreviation for extreme gradient boosting, is one of the applications of gradient boosting machines, which is one of the most effective supervised learning algorithms. Its basic structure is established on gradient boosting and decision-tree algorithms. Compared with other algorithms, it is in a very advantageous position regarding speed and performance. Additionally, XGBoost is highly predictive, 10 times faster than other algorithms, and includes several regularizations that improve overall performance and reduce overfitting or overlearning. Gradient boosting is an ensemble method that combines weak classifiers with boosting to create a robust classifier. The strong learner is trained iteratively, starting with a basic learner. Both gradient boosting and XGBoost follow the same principle. They mainly differ in the implementation. By using different regularization techniques, XGBoost can achieve better performance by controlling the complexity of the trees[19].

Bioinformatics Analysis

For patients (HBV + HCC and HBV alone) whose gene expression profiles were examined, differential expression analyses were performed using the limma package in the R programming language[20]. 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 (Log_2FC) values, and genes with the smallest p values will be most reliable. Log_2FC >1 was used to identify upregulated genes, and Log_2FC <-1 was used to identify downregulated genes[21]. A volcano plot was graphed to highlight quickly large values regarding the relevant genes.

Study Protocol and Ethics Committee Approval

This study, which used the National Center for Biotechnology Information Gene Expression Omnibus open-access dataset involving human participants, was conducted in accordance with 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 for Non-Interventional Clinical Research (decision no: 2022/3646, date: 07.06.2022). Strengthening the Reporting of Observational Studies in Epidemiology guideline was utilized to assess the likelihood of bias and overall quality of this study.

Statistical Analysis

The Shapiro-Wilk test of normality was used to determine whether the variables followed 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 independent-sample t-tests were utilized to compare non-normally distributed data, where

appropriate. Logistic regression analysis was performed to estimate each gene's odds ratio (OR) (a measure of effect size). Hosmer and Lemeshow's test for the goodness of fit and omnibus test of model coefficients were calculated for logistic regression. P-value <0.05 was considered significant. IBM SPSS Statistics, version 25.0, was used in the analysis.

Modeling Process

XGBoost, one of the ML methods, was used in the modeling. Analyses were conducted using the n-fold cross-validation method. In the n-fold cross-validation method, data were first divided into n parts, and the model used was applied to n parts. One of the n parts is used for testing, whereas 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, positive-predictive value, negative-predictive value, and F1-score were used as performance evaluation criteria. In addition, variable importances were calculated, which gives information about how much the input variables explain the output variables.

RESULTS

In this study, 53 patients (HBV + HCC =17; HBV alone =36) were used, of which 42 were male and 11 were female. The mean age of the patients was 54.91±13.76 years. While 15 of the HBV + HCC group were male and two were female, 27 patients in the HBV alone group were male and nine were female. The mean age of patients with HBV + HCC was 60.47±9.01 years, and the mean age of patients with HBV alone is 52.28±14.90 years. The dataset used contains 8516 expressions. According to the bioinformatics analysis, the first 10 results are summarized concerning minimum adjusted p values in Table 1. As shown in Table 1, five genes (ID: 1474, 1817, 6277, 4496, and 7165) were downregulated, and the other five genes were unregulated.

Table 2 presents descriptive statistics for the selected genes concerning the groups. According to Table 2, Log_2FC values for the IGFBP3, HGFAC, SLC39A14, CXCL12, PLG, FBP1, RNF26, ACBD6, C8A, and CCT3 were -1.54, -1.79, -1.06, -0.96, -0.92, -1.27, 0.46, 0.65, 1.02, and 0.66, respectively. Significant differences were determined in PFAS, FRA16B, GCNT2, GKP5, MEN1, MUC4, RBM12, RNF26, TIMP3, MCM3, VPS28, CRY1, SF3B2, H3C11, ACBD6, and FLJ10233 between the groups ($p<0.05$). CYP24A1 and homo sapiens chromosome 5 clone RP11-998B18 complete sequence genes were not significantly different between the groups ($p>0.05$).

The volcano plot used to visualize differentially expressed genes is given in Figure 1. On the y- and x-axes, the significance of the volcano graph plots versus the

Table 1. Top 10 results of the bioinformatics analysis.

ID	Adj p-value	p-value	t	B	Log_2FC	Gene name	Symbol	Diff. expressed
1474	0.000194	3.19E-08	-6.3807897	8.583	-1.54416328	Insulin-like growth factor binding protein 3	IGFBP3	Down
1817	0.000586	2.58E-07	-5.8319577	6.6719	-1.79464629	HGF activator	HGFAC	Down
6277	0.000586	2.89E-07	-5.8012278	6.5658	-1.06482155	Solute carrier family 39 (zinc transporter) member 14	SLC39A14	Down
4712	0.000737	4.85E-07	-5.66363	6.0922	-0.96830095	Chemokine (C-X-C motif) ligand 12 (stromal cell-derived factor 1)	CXCL12	No
10469	0.001358	1.26E-06	-5.4088551	5.2226	-0.92433824	Plasminogen	PLG	No
4496	0.001358	1.34E-06	-5.3910676	5.1623	-1.27427529	Fructose-1.6-bisphosphatase 1	FBP1	Down
5608	0.001514	1.75E-06	5.3196011	4.9206	0.4605353	Ring finger protein 26	RNF26	No
10010	0.001514	2.07E-06	5.2740142	4.7669	0.65932271	Acyl-Coenzyme A binding domain containing 6	ACBD6	No
7165	0.001514	2.24E-06	-5.2517915	4.6921	-1.0231292	Complement component 8. alpha polypeptide	C8A	Down
9041	0.002071	3.41E-06	5.1374571	4.309	0.66934959	Chaperonin containing TCP1. subunit 3 (gamma)	CCT3	No

fold change in log2 base show differentially expressed genes quickly.

Eighteen expression results were obtained by applying the LASSO feature-selection method to 8516 expression results. The explanations of the dataset with the selected expressions, examined target variable, and OR per gene for the target variable are presented in Table 2. The findings of the performance metrics from the XGBoost model are provided in Table 3.

Accuracy, balanced accuracy, sensitivity, specificity, positive-predictive value, negative-predictive value, and F1 score obtained from the XGBoost model were 98.1%, 98.6%, 100%, 97.2%, 94.4%, 100%, and 97.1%, 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. RNF26 had the highest predictor importance of 100.0%, followed by FLJ10233 at 66.21% and ACBD6 at 51.47%.

DISCUSSION

Although the gene expression profiling structure of HCC and the background liver has been widely examined[14], ML-based prediction of HBV-related HCC and detection of crucial candidate biomarkers have not

Table 2. Descriptive statistics for the selected genes concerning the groups.

Gene name	Prop number	Groups				OR	p-value
		HBV + HCC (n=17)		HBV alone (n=36)			
		Mean ± SD	Median (min-max)	Mean ± SD	Median (min-max)		
CYP24A1	1591	-0.11±0.35	-0.10 (-0.65-0.80)	0.06±0.53	0.20 (-0.94-1.41)	-	0.153*
PFAS	2390	+0.82±0.81	+0.59 (-0.42-2.52)	0.12±0.45	0.14 (-1.28-1.14)	10	<0.001**
FRA16B	2461	-0.16±0.30	-0.12 (-0.72-0.45)	0.20±0.39	0.25 (-0.54-1.09)	0.06	0.002
GCNT2	2651	-1.35±0.98	-1.23 (-2.99-0.39)	-0.56±0.53	-0.56 (-1.73-0.95)	0.21	<0.001*
GKP5	2715	-0.93±0.93	-0.59 (-3.52-0.10)	0.04±0.81	0.20 (-3.02-1.61)	0.24	<0.001**
MEN1	3785	+0.21±0.44	+0.30 (-0.53-0.90)	0.02±0.24	+0.02 (-0.60-0.57)	6.89	0.042*
Homo sapiens chromosome 5 clone RP11-998B18 complete sequence	4219	+0.20±0.58	+0.29 (-0.64-1.21)	-0.08±0.41	-0.07 (-1.08-0.86)	3.52	0.090*
MUC4	4585	+1.08±0.91	+1.28 (-0.37-2.97)	0.22±0.77	+0.35 (-1.84-2.46)	3.83	0.001*
RBM12	5520	+0.38±0.39	+0.30 (-0.246-1.11)	-0.14±0.31	-0.19 (-1.03-0.47)	170	<0.001*
RNF26	5608	+0.51±0.31	+0.45 (-0.01-1.20)	0.05±0.23	+0.02 (-0.37-0.55)	722	<0.001*
TIMP3	5815	+0.56±0.55	+0.57 (-0.75-1.26)	-0.02±0.46	+0.02 (-0.85-1.08)	10	<0.001*
MCM3	5906	+1.20±0.91	+1.05 (0.09-3.69)	0.20±0.65	+0.27 (-1.97-1.39)	9.99	<0.001**
VPS28	6292	+0.66±0.38	+0.64 (-0.24-1.49)	0.23±0.31	+0.27 (-0.63-0.69)	225	<0.001**
CRY1	6377	+0.15±0.30	+0.09 (-0.27-0.80)	0.56±0.53	+0.53 (-0.53-1.89)	0.1	0.001
SF3B2	8061	+0.09±0.50	+0.23 (-1.26-0.58)	-0.19±0.33	-0.18 (-0.88-0.50)	7.38	0.007
H3C11	8354	+0.35±0.51	+0.38 (-0.77-1.29)	-0.26±0.45	-0.19 (–1.53-0.70)	22	<0.001*
ACBD6	10010	+0.56±0.61	+0.43 (-0.52-1.90)	-0.10±0.29	-0.15 (-0.57-0.61)	57	<0.001*
FLJ10233	10333	+0.08±0.50	+0.14 (-1.44-0.68)	-0.31±0.29	-0.26 (-0.78-0.20)	26	<0.001**

*Independent samples t-test, **Mann-Whitney U test, OR: Odds ratio, SD: Standard deviation, min-max: Minimum-maximum, HBV: Hepatitis B virus, HCC: Hepatocellular carcinoma, HBV + HCC: Hepatitis B virus-related hepatocellular carcinoma

been clarified using an AI approach. Thence, this study intends to classify HBV-related HCC and HBV without HCC gene expression data using the XGBoost method and identify important genes that may cause HCC.

Table 3. Performance metrics of the XGBoost model.

Metric	Value (%) (95% CI)
Accuracy	98.1 (94.5-1)
Balanced accuracy	98.6 (95.5-1)
Sensitivity	100 (80.5-1)
Specificity	97.2 (85.5-99.9)
Positive-predictive value	94.4 (72.7-99.9)
Negative-predictive value	100 (90-1)
F1 score	97.1 (95.4-1)

CI: Confidence interval

HBV is widespread worldwide, with varying levels of infection in different regions. According to the World Health Organization, approximately two billion people have been infected with HBV worldwide, with 240 million people infected with chronic HBV and approximately 650,000 people die annually from hepatic failure and liver cirrhosis and HCC caused by HBV infection. HBV infection is responsible for 30% and 45% of patients with liver cirrhosis and HCC worldwide[22,23].

The overall survival of patients with HCC is low, and the 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 research to identify diagnosis early, prognostic markers, and new therapeutic targets to uncover the mechanisms of liver carcinogenesis and thus improve the

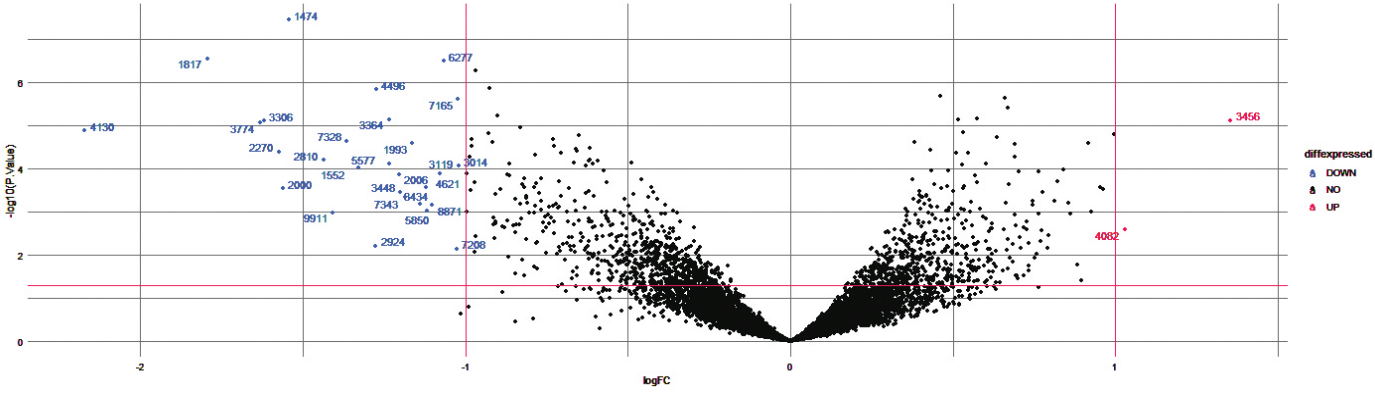

Figure 1. Volcano plot.

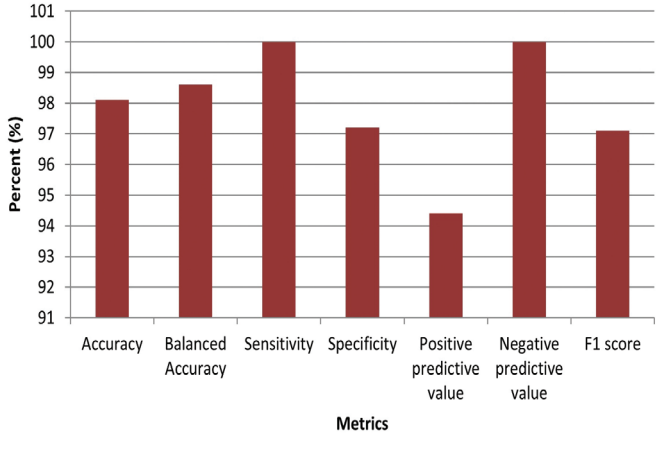

Figure 2. Values for the performance criteria obtained from XGBoost models.

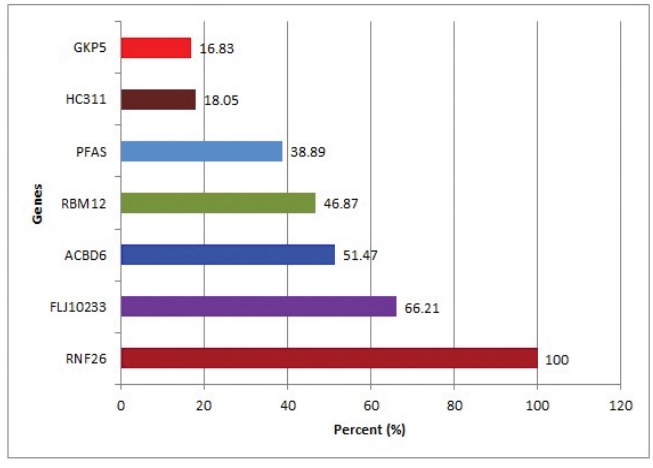

Figure 3. Gene importance values for predicting the output variable.

clinical management of patients with HCC. Building on these studies, advances in HCC surveillance promise to significantly reduce the worldwide burden of HCC over the next few decades[24,25].

In the dataset analyzed in this study, the genomic data of samples obtained from liver tissues of 17 patients with HBV-related HCC and 36 with chronic HBV without HCC were used for the relevant analyses. cDNA microarrays were obtained from the samples, and the dataset used contained 8516 expressions. According to the Log_2FC values used to determine the expression fold changes between the two groups from the findings of the bioinformatics analyses (Table 2), IGFBP3 has 2.90-fold lower gene expression in patients with HBV-related HCC than in patients with chronic HBV. Similarly, HGFAC had 3.45-fold lower gene expression, SLC39A14 had 2.08-fold, FBP1 had 2.41-fold, and C8A had 2.02-fold lower gene expression. CXCL12, PLG, RNF26, and ACBD6 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. Therefore, before modeling with the existing dataset, the most important genes associated with the output variable were selected with the LASSO variable-selection method. Eighteen genes selected by the LASSO method were used in building the XGBoost model. The accuracy, balanced accuracy, sensitivity, specificity, positive and negative-predictive values, and F1 score metrics obtained with the XGBoost model were 98.1%, 98.6%, 100%, 97.2%, 94.4%, 100%, and 97.1%, respectively. The performance metrics indicated that the proposed XGBoost model could correctly classify two groups of patients based on the AI approach. Among the genes whose OR values were calculated, *RNF26* (OR =722), *VPS28* (OR =225), *RBM12* (OR =170), *ACBD6* (OR =57), *FLJ10233* (OR =26), *H3C11* (OR =22), *PFAS* (OR =10), and *TIMP3* (OR =10) genes were found to have the highest OR values, respectively. According to the variable importance obtained from XGBoost, RNF26, FLJ10233, ACBD6, RBM12, PFAS, H3C11, and GKP5 can be used as candidate predictive biomarkers of HBV-related HCC. In addition, the calculated OR values and variable importance values in the study support each other. According to variable significance results, genes with huge OR values were determined as genes contributing to HBV-related HCC development. Additionally, the proposed pipeline produced a volcano plot, representing the up- and downregulation of the genes. These plots are becoming more common in omics experiments, such as genomics, proteomics, and metabolomics, where there are often thousands of replicate data points between two conditions[26].

A medical study reported that RNF26 was abnormally expressed in patients with HCC[27]. In another study, VPS28 was upregulated[28]. Another study showed that a high RBM12 level in HCC indicates a poor patient prognosis[29]. One study reported that ACBD6 was expressed differently in HCC and chronic hepatitis[30]. In a study, high-grade tumors exhibited progressively higher levels of PFAS, ATIC, IMPDH1, IMPDH2, GMPS, and ADSL than low-grade tumors or normal liver tissue[31]. In one study, *TIMP3* was found as a candidate gene in HBV-related HCC[32]. Another study determined that epigenetic methylation of TIMP3 is associated with HBV-associated HCC[33].

In a study, *SHCBP1, FOXM1, KIF4A, ANLN, KIF15, KIF18A, FANCI, NEK2, ECT2,* and *RAD51AP1* were found as the top 10 most important genes for HBV-related HCC[34]. In addition, patients with FOXM1, NEK2, RAD51AP1, ANLN, and KIF18A showed worse overall survival. In another study with HCC, the expression levels for *PER1, PER2, PER3,* and *CRY2* genes were lower[35]. Another study showed that high expression of FOXM1 causes a poor prognosis for HBV-related HCC and promotes tumor metastasis[36].

All diseases that cause chronic liver damage are risk factors for HCC development. Therefore, international guidelines' follow-up of such patients is crucial for detecting possible HCC or its detection at an early stage[37]. The most authoritative guidelines on monitoring patients with chronic liver are published periodically by European Association for the Study of the Liver, Asian-Pacific Association for the Study of the Liver, and American Association for the Study of Liver Diseases[37]. The tumor doubling time of HCC varies between 4 and 6 months. Therefore, the abovementioned guidelines suggest that patients with chronic liver disease without HCC should be followed up with ultrasonography (US) and alpha-fetoprotein (AFP) at 6-month intervals[37]. Patients with suspected HCC (nodule diameter <10 mm) should be followed up with US and AFP at 3 or 6-month intervals. Patients with a strong suspicion of HCC should be followed up with US and AFP. Patients with nodule diameter >10 mm and/or AFP >20 ng/mL should be evaluated further with radiological examinations[37].

However, these approaches may not always provide 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, because US is an operator-dependent examination. There is a correlation between

the duration of chronic liver disease and probability of HCC development. 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 shown in our results, if changes are detected in the expression of genes that are strongly associated with HCC, patients can be followed more closely, and preventive treatments can be initiated 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 to determine the timing of genetic analysis for patients with chronic liver disease. With this important finding, increasing the number of patients may further increase the scope of genetic information and power of the study.

CONCLUSION

In conclusion, this study revealed possible genomic biomarkers of HBV-related HCC using gene expression data from patients with HBV-related HCC and patients with chronic HBV alone. The reliability of the genes obtained with more comprehensive analyses to be made in the future can be tested, treatment approaches can be developed based on these genes, and their usability in clinical practice can be detailed. Thus, individual-based treatments and immunotherapy approaches more applicable to clinical practice are possible.

Ethics

Ethics Committee Approval: Ethical approval was obtained from the Inonu University Institutional Review Board for Non-Interventional Clinical Research (decision no: 2022/3646, date: 07.06.2022).

Informed Consent: Retrospective study.

Peer-review: Externally and internally peer-reviewed.

Author Contributions

Concept: Z.K., S.A., Design: Z.K., C.C., Data Collection and/or Processing: Z.K., C.C., Analysis and/or Interpretation: Z.K., S.A., C.C., Literature Search: Z.K., S.A., C.C., Writing: Z.K., S.A., 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. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68:394-424.
2. McGlynn KA, Petrick JL, London WT. Global epidemiology of hepatocellular carcinoma: an emphasis on demographic and regional variability. Clin Liver Dis. 2015;19:223-38.
3. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin. 2019;69:7-34.
4. Llovet JM, Kelley RK, Villanueva A, et al. Hepatocellular carcinoma. Nat Rev Dis Primers. 2021;7:6.
5. Fattovich G, Stroffolini T, Zagni I, Donato F. Hepatocellular carcinoma in cirrhosis: incidence and risk factors. Gastroenterology. 2004;127(5 Suppl)35-50.
6. An P, Xu J, Yu Y, Winkler CA. Host and Viral Genetic Variation in HBV-Related Hepatocellular Carcinoma. Front Genet. 2018;9:261.
7. Andrisani O. Epigenetic mechanisms in hepatitis B virus-associated hepatocellular carcinoma. Hepatoma Res. 2021;7:12
8. Takeda H, Takai A, Inuzuka T, Marusawa H. Genetic basis of hepatitis virus-associated hepatocellular carcinoma: linkage between infection, inflammation, and tumorigenesis. J Gastroenterol. 2017;52:26-38.
9. Akman M, Genç Y, Ankarali H. Random Forests Methods and an Application in Health Science. Turkiye Klinikleri J Biostat. 2011;3:36-48.
10. Polikar R. Ensemble learning. Ensemble machine learning: Springer; 2012. p. 1-34.
11. Piñero F, Dirchwolf M, Pessôa MG. Biomarkers in hepatocellular carcinoma: diagnosis, prognosis and treatment response assessment. Cells. 2020;9:1370.
12. Ye QH, Qin LX, Forgues M, et al. Predicting hepatitis B virus-positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning. Nature Med. 2003;9:416-23.
13. Shen J, Qi L, Zou Z, et al. Identification of a novel gene signature for the prediction of recurrence in HCC patients by machine learning of genome-wide databases. Sci Rep. 2020;10:4435.
14. Ueda T, Honda M, Horimoto K, et al. Gene expression profiling of hepatitis B-and hepatitis C-related hepatocellular carcinoma using graphical Gaussian modeling. Genomics. 2013;101:238-48.
15. Chang HY, Thomson JA, Chen X. Microarray analysis of stem cells and differentiation. Methods Enzymol. 2006;420:225-54.
16. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23:2507-17.
17. Fodor IK. A Survey of Dimension Reduction Techniques. Lawrence Livermore National Lab., CA (US); 2002.
18. Fonti V. Feature Selection with LASSO. VU Amsterdam research paper in Business Analytics. 2017;30:1-25.
19. Salam Patrous Z. Evaluating XGBoost For User Classification By Using Behavioral Features Extracted From Smartphone Sensors. [Master Thesis]: KTH Royal Institute of Technology, School of Computer Science and Communication, Sweden; 2018.
20. Smyth GK. Limma: linear models for microarray data. Springer; 2005. p. 397-420.

21. Yan H, Zheng G, Qu J, et al. Identification of key candidate genes and pathways in multiple myeloma by integrated bioinformatics analysis. J Cell Physiol. 2019;234:23785-97.
22. Hou J, Wang G, Wang F, et al. Guideline of prevention and treatment for chronic hepatitis B (2015 update). J Clin Transl Hepatol. 2017;5:297-318.
23. Tang LS, Covert E, Wilson E, Kottilil S. Chronic hepatitis B infection: a review. JAMA. 2018;319:1802-13.
24. Ghidini M, Braconi C. Non-coding RNAs in primary liver cancer. Front Med (Lausanne). 2015;2:36.
25. 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;16:589-604.
26. Blum BC, Emili A. Omics Notebook: robust, reproducible and flexible automated multiomics exploratory analysis and reporting. Bioinformatics Advances. 2021:1-4.
27. Augello C, Colombo F, Terrasi A, et al. Expression of C19MC miRNAs in HCC associates with stem-cell features and the cancer-testis genes signature. Dig Liver Dis. 2018;50:583-93.
28. Azhar NA, Bakar SAA, Citartan M, Ahmad NH. mRNA Transcriptomic Profiling of Human Hepatocellular Carcinoma Cells HepG2 Treated with Catharanthus roseus-Silver Nanoparticles. Preprints. 2021.
29. Gao C, Shen J, Chen W, et al. Increased RBM12 expression predicts poor prognosis in hepatocellular carcinoma based on bioinformatics. J Gastrointest Oncol. 2021;12:1905-26.
30. Ura S, Honda M, Yamashita T, et al. Differential microRNA expression between hepatitis B and hepatitis C leading disease progression to hepatocellular carcinoma. Hepatology. 2009;49:1098-112.
31. Chong YC, Toh TB, Chan Z, et al. Targeted inhibition of purine metabolism is effective in suppressing hepatocellular carcinoma progression. Hepatol Commun. 2020;4:1362-81.
32. Dong X, Hou Q, Chen Y, Wang X. Diagnostic value of the methylation of multiple gene promoters in serum in hepatitis B virus-related hepatocellular carcinoma. Dis Markers. 2017;2017:2929381.
33. Lai H, Lo SJ. Epigenetic methylation of TIMP-3 may play a role in HBV-associated hepatocellular carcinoma. Chang Gung Med J. 2005;28:453-5.
34. Xie S, Jiang X, Zhang J, et al. Identification of significant gene and pathways involved in HBV-related hepatocellular carcinoma by bioinformatics analysis. Peer J. 2019;7:e7408.
35. Lin YM, Chang JH, Yeh KT, et al. Disturbance of circadian gene expression in hepatocellular carcinoma. Mol Carcinog. 2008;47:925-33.
36. Xia L, Huang W, Tian D, et al. Upregulated FoxM1 expression induced by hepatitis B virus X protein promotes tumor metastasis and indicates poor prognosis in hepatitis B virus-related hepatocellular carcinoma. J Hepatol. 2012;57:600-12.
37. 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.