Case-Controlled Study

ANNALS OF MEDICINE & SURGERY

OPEN

Artificial intelligence-based prediction of molecular and genetic markers for hepatitis C–related hepatocellular carcinoma

Cemil Colak, PhD[a], Zeynep Kucukakcali, PhD[a], Sami Akbulut, MD, PhD[a,b,*]

Background: Hepatocellular carcinoma (HCC) is the main cause of mortality from cancer globally. This paper intends to classify public gene expression data of patients with Hepatitis C virus-related HCC (HCV + HCC) and chronic HCV without HCC (HCV alone) through the XGboost approach and to identify key genes that may be responsible for HCC.

Methods: The current research is a retrospective case–control study. Public data from 17 patients with HCV + HCC and 35 patients with HCV-alone samples were used in this study. An XGboost model was established for the classification by 10-fold cross-validation. Accuracy (AC), balanced accuracy (BAC), sensitivity, specificity, positive predictive value, negative predictive value, and F1 score were utilized for performance assessment.

Results: AC, BAC, sensitivity, specificity, positive predictive value, negative predictive value, and F1 scores from the XGboost model were 98.1, 97.1, 100, 94.1, 97.2, 100, and 98.6%, respectively. According to the variable importance values from the XGboost, the HAO2, TOMM20, GPC3, and PSMB4 genes can be considered potential biomarkers for HCV-related HCC.

Conclusion: A machine learning-based prediction method discovered genes that potentially serve as biomarkers for HCV-related HCC. After clinical confirmation of the acquired genes in the following medical study, their therapeutic use can be established. Additionally, more detailed clinical works are needed to substantiate the significant conclusions in the current study.

Keywords: artificial Intelligence, chronic liver disease, genetic markers, hepatitis C infection, hepatocellular carcinoma

Introduction

According to the most recent epidemiological and clinical data, primary liver cancer is the sixth most prevalent type of cancer and the third leading cause of death worldwide. About 906 000 new cases and 830 000 deaths are reported each year. Primary liver cancer includes hepatocellular carcinoma (HCC) and intrahepatic cholangiocarcinoma, as well as other rare types[1,2]. Hepatitis B virus (HBV), alcoholism, Hepatitis C virus (HCV), and nonalcoholic fatty liver disease are among the most critical risk factors for HCC[3].

Hepatotropic RNA viruses, or HCV, are blood-borne infections that exclusively infect the liver. Most people infected with HCV will never be able to cure themselves of the virus, making their condition chronic and lifelong. Fibrosis, cirrhosis, and HCC are all serious liver diseases that occur due to HCV-induced chronic inflammation. Cirrhosis affects 20–30% of HCV-infected people, and yearly, 1–4% of cirrhotic individuals develop HCC[4,5]. In most cases, HCC develops alongside cirrhosis[6]. It has been reported that HCV infection can directly induce HCC without cirrhosis in patients who develop HCC[7]. Approximately 34% of all instances of HCC in the United States may be attributed to chronic HCV infection, making it the main cause of HCC in Western countries[7]. The risk of HCC is increased by 15 to 20-fold in patients infected with HCV[8,9]. Deaths from HCV-attributable HCC increased by 21.1 percent over the last decade when deaths from HCC secondary to causes other than HCV and alcohol remained stable[9]. Both geographic location and ethnicity influence the incidence of HCC caused by HCV. HCV is the most common cause of HCC in the United States, Europe, Japan, and South America, whereas HBV is the most common cause of HCC in Asia and Africa[10]. Concurrent liver disease, viral genotype, lifestyle factors, obesity, and diabetes mellitus are the most critical risk factors for developing HCC in chronic HCV infection[11].

HIGHLIGHTS

- In this study, we identified differentially expressed genes for hepatitis C virus (HCV)-associated hepatocellular carcinoma (HCC).
- HCV-related HCC and chronic HCV patients without HCC were classified using the proposed machine learning method.
- Possible genetic biomarkers for HCV-associated HCC were identified at the end of the classification model.

[a]Department of Biostatistics and Medical Informatics and [b]Department of Surgery, Inonu University Faculty of Medicine, Malatya, Turkey

Sponsorships or competing interests that may be relevant to content are disclosed at the end of this article.

*Corresponding author. Address: Department of Surgery and Liver Transplant Institute, Department Biostatistics and Medical Informatics Inonu University, Faculty of Medicine, Elazig Yolu 10. Km Malatya 44280, Turkey. Tel.: +90 422 3410660, fax: +90 422 3410036. E-mail: akbulutsami@gmail.com (S. Akbulut).

Annals of Medicine & Surgery (2023) 85:4674–4682

Received 27 April 2023; Accepted 12 August 2023

Published online 7 September 2023

http://dx.doi.org/10.1097/MS9.0000000000001210

The spread of HCV is becoming an international health crisis. In many regions, HCV is widespread and poses a growing challenge for healthcare systems. Long-term effects, including cirrhosis and HCC, are rising daily[12]. Chronic HCV infection typically progresses slowly, with few cases of severe liver disease appearing in the first 10 to 15 years following infection (even in individuals with co-factors for fibrosis progression). Therefore, key factors determining morbidity and mortality include the patient's age and the length of time they have had a chronic HCV infection[13]. Immune tolerance deteriorates due to chronic HCV infection, resulting in a protracted inflammatory response in the liver. Numerous nonviral agents induce cellular stress and liver damage, culminating in persistent sterile inflammation. The functional and morphogenetic changes in the liver are reversible if the substances or stressors that cause the damage are eliminated within the early stages of injury. As the liver shifts from normal to pathological adaptability, the continuation of chronic damage might result in irreparable impairment. Cirrhosis of the liver is a pathological response to cellular stress that results in structural and functional alterations in the liver parenchyma to avoid harm. Virus-induced cellular adaptation may be reversible or irreversible, depending on the type and degree of the stressor[4].

Noncoding RNAs, such as microRNAs (noncoding RNAs that are single-stranded and regulate gene expression primarily at the post-transcriptional level) and long noncoding RNAs (noncoding RNAs with lengths ranging from 200 nucleotides to 100 kilobases), are increasingly being recognized as crucial players in the regulation of liver function and hepatocarcinogenesis[14,15]. MiRNAs and lncRNAs have been reported in many studies of HCC and HCV-related HCC[16,17]. If the data is sufficient, artificial intelligence (AI) can classify relevant patterns without requiring a protracted time of learning[18].

Machine learning (ML) is a branch of AI that seeks to anticipate new data through data-driven learning when exposed to fresh data. AI/ML algorithms are widely employed in illness diagnosis and clinical decision support systems in recent years and have a wide range of applications. The fundamental infrastructure of apps for diagnosing genetic illnesses, early cancer diagnosis, and recognizing patterns in medical imaging relies on ML, which has a wide range of applications in the health sector. With the availability of massive datasets and increased processing power over the past decade, ML approaches have attained excellent performance in a variety of contexts[19,20]. Today, it is essential to diagnose HCC, to identify/predict the genes that cause the occurrence of HCC as biomarkers, and to use these biomarkers concerning the HCC stage. Consequently, several clinical articles have used ML techniques to discover genes that may serve as biomarkers for HCC[21]. One study studied Noncoding RNAs for HCV-associated HCC[17]. Another study employed ML to diagnose HCC with HCV-associated chronic liver disease[22]. This work intends to classify public gene expression data of patients with HCV-related HCC (HCV + HCC) and chronic HCV without HCC (HCV alone) using the XGboost approach and to identify significant genes that may contribute to the development of HCC.

Methods

Study design and data

The current research is a retrospective case–control study, and the XGboost method, one of the ML methods, was applied to open-access gene expression data of HCV-related HCC and chronic HCV patients without HCC. For this purpose, public data from 17 HCV-related HCC and 35 chronic HCV samples were encompassed in the present study. Complementary DNA (cDNA) microarrays achieved from liver samples were utilized for the targeted analysis[23]. In genetics, cDNA refers to a DNA fragment generated from mature mRNA using reverse transcriptase as a catalyst. cDNA is the double-stranded DNA counterpart to mRNA. In eukaryotes, mRNA is more valuable than genomic sequence for determining polypeptide sequence. Since introns are removed, scientists prefer to deal with cDNA to mRNA. Consequently, RNA is intrinsically more unstable than DNA. In addition, no technology for RNA molecule amplification and purification exists. Reverse transcriptase synthesizes single-stranded DNA molecules using mRNA as a template. This molecule is then used to generate DNA with two strands[24]. Based on the design of the current study, the primary outcome measure was the HAO_2 gene in the HCV-related HCC and chronic HCV patients without HCC.

Study protocol and ethics committee approval

This study, which utilized the open-access Gene Expression Omnibus dataset from the National Center for Biotechnology Information, involved human participants adhering to the ethical standards of the institutional and national research committees, the 1964 Helsinki Declaration and its later amendments, or other ethical standards. Ethical approval for this study (Ethical Committee No: 3647) was provided by the Ethical Committee of Inonu University, Malatya, Turkey on 07 June 2022. This retrospective case–control study is reported in line with the Strengthening the Reporting Of Cohort Studies in Surgery (STROCSS) guidelines[25].

Feature selection

Choosing which variables to include in a model is a crucial part of any predictive modeling process, and data selection is a crucial part of any statistical modeling process. Determining the most valuable elements of the dataset to be utilized in the study before dealing with massive datasets and models with high computing costs will lead to significant efficiency in terms of outcomes. Finding which aspects of a dataset affects the dependent variable is the goal of feature selection. There is a risk of over-learning the data and producing biased findings if there are too many explanatory factors and the computation time needed to process them is too great. Moreover, it is challenging to understand models that contain a large number of variables. Important influencing factors should be chosen before statistical modeling[26]. Large datasets can overwhelm the effectiveness of most ML and data mining techniques, leading to poor outcomes. As a result, reducing the dimensionality yields better outcomes using these approaches[27].

Gene expression huge datasets are massive. Modeling analyses require a long time due to the high amount of gene expression datasets, and these datasets may lead to computational inefficiencies in the performed studies. Because of the

high dimensionality, the model's performance may suffer. A classification method may overfit the training instances and undergeneralize novel samples in gene expression datasets with many genes. Several regularization strategies have been developed for model fitting and variable selection in poorly specified multiple regressions. These strategies include least absolute shrinkage and selection operator (LASSO), ridge, and elastic net. Compared to LASSO regularization, ridge regularization causes predictors to decrease, leading to more stable parameter estimation. However, LASSO regularization causes many regression coefficients to become precisely zero, which allows for automatic variable selection in which a single predictor is chosen from a set of correlated predictors. To get the best of both worlds, elastic-net regularization combines ridge and LASSO penalties. As a result, it offers shrinkage and automatic variable selection, as well as the ability to handle more effectively the severe multicollinearity common in genome-wide association study analysis[28–31].

XGBoost algorithm

In ML, Gradient Boost is a potent tool for regression and classification issues where ensemble versions of decision trees are typically the result of poor predictive models. The boosting-based Gradient Boost technique aims to build numerous sequentially weak learners and merge them into an elaborate model[32].

One of the most powerful supervised learning techniques is gradient boosting machines, and one of its applications is Extreme Gradient Boosting (XGBoost). It is established on gradient boosting and decision tree algorithms, which form its basic structure. Its speed and efficiency are much beyond those of competing algorithms. In addition to its strong predictiveness, XGBoost is 10 times quicker than competing algorithms and has many regularizations that boost overall performance while mitigating overfitting or over-learning. To produce a robust classifier, gradient boosting uses a collection of weak classifiers and the boosting technique to combine them. The powerful learner is educated in an iterative process, commencing with a primary learner. XGBoost works on the same fundamentals as gradient boosting. The main distinctions lie in the specifics of their use. It is possible to improve XGBoost's performance by using a variety of regularization approaches to the trees' complexity[33,34].

Bioinformatics analysis

Gene expression patterns were analyzed for samples of HCV-related HCC and chronic hepatitis C patients using differential expression analysis performed via the limma package of the R programming language[35]. The statistical examination of normalized read count data to discover quantifiable variations in expression activity between treatment arms is known as differential expression analysis. A pipeline was built for the critical analysis using the R software environment. A table describing the relative importance of the genes and a graph showing the genes with various expression levels are provided as output. The most reliable genes are those with the fewest *P* values in the table of results, which also includes corrected P and log2-fold-change (log2FC) values. Genes with a log2FC greater than 1 were considered up-regulated, whereas those with a log2FC – 1 were considered down-regulated[36]. We used a volcano plot to visually emphasize readily noticeable high values concerning the key genes.

Biostatistical and power analyses

A posteriori/retrospective power analysis generated roughly 100% power, considering the type I error (alpha) of 0.05, the sample size of 35 in the HCV alone group, the effect size of 2.43 for the HAO_2 gene, and the two-sided alternative hypothesis (H_1). It was determined using the Shapiro–Wilk test if the values followed a normal distribution. The summarized information for that variable was shown as the median (minimum–maximum) or the mean + SD. Mann–Whitney *U* test was used to compare data that did not follow a normal distribution. At the same time, the independent-samples *t*-test was employed to evaluate the data that follow a normal distribution. Odds ratios for each gene were estimated using logistic regression (a measure of effect size). For logistic regression, we used Hosmer and Lemeshow's goodness of fit test and the omnibus test of model coefficients. Statistical significance was assumed at a *P*-value of less than 0.05. The study was conducted using IBM's SPSS Statistics 25.0.

AI/ML modeling process

One of the ML techniques utilized in the modeling was termed XGBoost. The *n*-fold cross-validation strategy was used for the analyses. The *n*-fold cross-validation technique divides the data into *n* subsets and then applies the model to each subset. The *n*-part dataset is divided as follows: one part is utilized for testing, while the remaining *n*-1 parts are used for model training. The cross-validation approach is assessed by looking at the median of the results. The modeling in this research used 10-fold cross-validation. The performance metrics employed were accuracy (AC), balanced accuracy (BAC), sensitivity (Sens), specificity (Spec), positive predictive value (PPV), negative predictive value (NPV), and F1 score. Moreover, variable importance scores were determined, which revealed how much each input variable contributed to the overall explanation of the outcome. Figure 1 shows a flowchart of all the research procedures that were conducted.

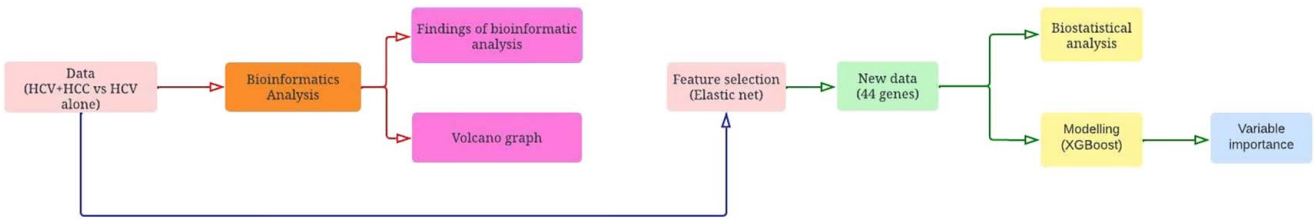

Figure 1. The flowchart of all the methods used in the study.

Table 1

The results of the bioinformatics analysis

ID (Gene name)	Symbol	Adjusted *P*	*P*	t	B	Log2FC	Diff-expressed
8479	HIRIP3	9.59E-11	1.84E-14	– 10.295045	22.4779	– 1.697632	Down
3456	GPC3	2.39E-10	9.19E-14	9.847219	20.9528	2.948391	Up
8733	SPINK1	8.32E-10	4.80E-13	9.392099	19.378	3.250553	Up
6164	GLYAT	1.18E-09	9.09E-13	– 9.218003	18.7695	– 1.821402	Down
3602	VIPR1	4.20E-09	4.04E-12	– 8.813517	17.3437	– 1.526099	Down
7647	HAO2	5.95E-09	7.36E-12	– 8.652349	16.7713	– 1.385486	Down
8938	TOMM20	5.95E-09	8.01E-12	8.629476	16.6899	0.852574	No
8871	SLC22A1	6.32E-09	9.73E-12	– 8.577299	16.504	– 2.118282	Down
4496	FBP1	1.93E-08	3.34E-11	– 8.247317	15.3235	– 1.8152	Down
1474	IGFBP3	2.06E-08	3.96E-11	– 8.202082	15.1611	– 1.978404	Down
5497	ASS	2.73E-08	5.77E-11	– 8.101588	14.7997	– 1.557922	Down
7542	CYP2C8	7.17E-08	1.66E-10	– 7.82152	13.7897	– 1.700409	Down
3439	ECHS1	1.34E-07	3.36E-10	– 7.634093	13.1117	– 0.930418	No
7208	CLEC2	2.38E-07	6.42E-10	– 7.462451	12.4898	– 2.580425	Down
1418	THBD	2.99E-07	8.64E-10	– 7.383885	12.2048	– 1.054523	Down

Results

Baseline characteristics and bioinformatics analysis

In the current study, 52 patients (HCV + HCC = 17; HCV alone = 35) were examined, of which 39 were male and 13 were female. The mean age of the patients was 61.78 ± 11.11 years. While 13 of the HCV + HCC patients were male and four were female, 26 of the patients with HCV alone were male, and nine were female. The mean age of HCV + HCC patients is 62.88 ± 9.96 years, and the mean age of patients with HCV alone is 61.25 ± 11.73 years. The dataset used contains 8516 expressions. According to the findings of the bioinformatics analysis, 'Table 1' contains a summary of the first fifteen results concerning the minimum adjusted-*P*-values. Based on the statistics from Table 1, seven genes (id: 8479, 6164, 3602, 7647, 8871, 4496, 1474, 5497, 7542, 7208, 1418) were down-regulation, two genes (id: 3456, 8733) were up-regulation, and two genes (id: 8938, 3439,) was unregulated. Table 1 presents descriptive statistics for the selected genes concerning the groups. According to Table 1, Log2FC values for the id = 8479, GPC3, SPINK1, GLYAT, VIPR1, HAO2, TOMM20, SLC22A1, FBP1, IGFBP3, ASS, CYP2C8, ECHS1, CLEC2, and THBD genes were – 1.697632, 2.948391, 3.250553, – 1.821402, – 1.526099, – 1.385486, 0.852574, – 2.118282, – 1.8152, – 1.978404, – 1,557922, – 1,700409, – 0,930418, – 2,580425, and – 1,054523, respectively.

Figure 2 shows the volcano plot displaying the differentially expressed genes. The volcano graph compares significance against fold-change in log2 based on the *y*-axes and *x*-axes, respectively, to identify rapid genes with significant expression differences.

Findings of AI modeling stage

Applying the Elastic Net feature selection approach to the 8516 expression results yielded 44 expression results. Table 2 gives some descriptive statistics for the chosen genes in terms of the categories and the odds ratio per gene for the output variable with respect to the groups. Table 2 shows that significant differences were found across groups for all save the

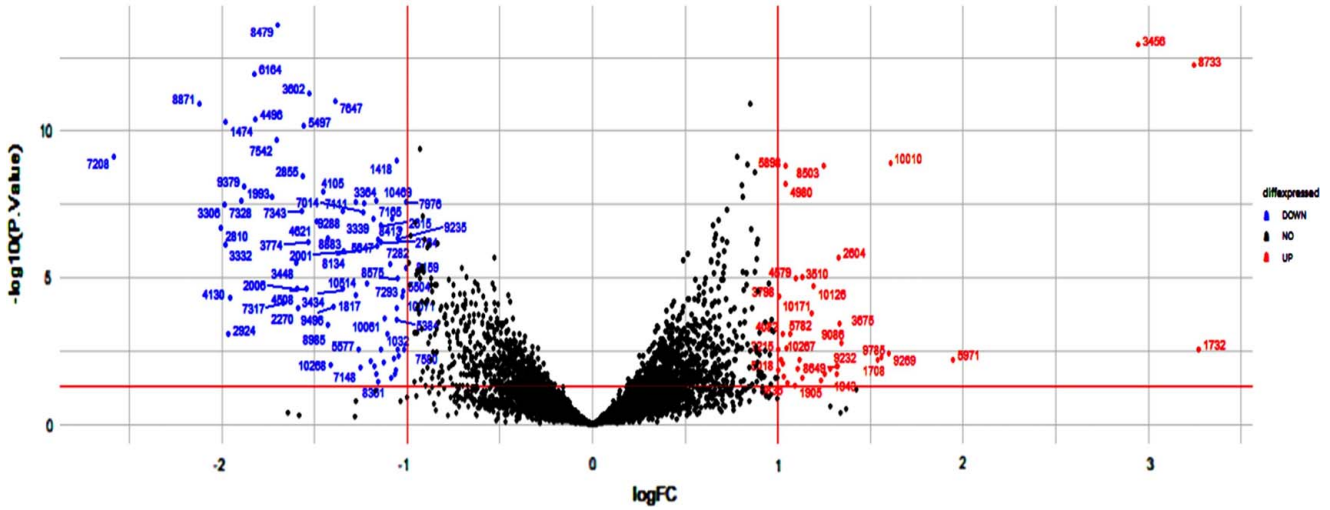

Figure 2. The volcano plot.

Table 2

Descriptive statistics for the input variables with respect to the study groups

		Groups					
		HCV + HCC ($n = 17$)		HCV Alone ($n = 35$)			
Gene name	Prop number	Mean ± SD	Median	Mean ± SD	Median	OR	*P*
THBD	1418	−1.22 ± 0.52	−1.19	−0.17 ± 0.45	−0.16	0.03	< 0.001**
IGFBP3	1474	−2.46 ± 1.18	−2.38	−0.48 ± 0.62	−0.38	0.08	< 0.001*
EPHX1	2604	+0.49 ± 0.98	+0.80	−0.85 ± 0.81	−0.82	5	< 0.001*
ANGPTL3	2764	−1.47 ± 0.82	−1.64	−0.33 ± 0.62	−0.28	0.12	< 0.001*
ALDOC	2765	+0.53 ± 0.39	+0.52	−0.17 ± 0.40	−0.24	88	< 0.001*
SFN	2810	−2.18 ± 1.19	−2.63	−0.18 ± 1.18	+0.07	0.32	< 0.001**
AFM	2855	−1.96 ± 0.92	−1.75	−0.39 ± 0.69	−0.39	0.08	< 0.001*
HNRNPF	3185	+0.19 ± 0.41	0.14	−0.34 ± 0.37	−0.30	87	< 0.001*
PCK1	3306	−2.28 ± 1.31	−2.07	−0.30 ± 0.95	−0.02	0.23	< 0.001**
UGT2B7	3364	−1.89 ± 0.77	−1.77	−0.66 ± 0.60	−0.57	0.09	< 0.001*
TCIRG1	3392	+0.87 ± 0.34	+0.83	+0.49 ± 0.36	+0.45	21	0.001*
GPC3	3456	+3.75 ± 1.31	+4.02	+0.80 ± 0.90	+0.89	63	< 0.001**
TIMP3	3510	+1.49 ± 0.85	+1.31	+0.38 ± 0.77	+0.45	16	< 0.001**
VIPR1	3602	−1.68 ± 0.83	−1.92	−0.15 ± 0.43	−0.19	0.02	< 0.001*
GOLGA5	3728	+0.68 ± 0.51	+0.57	−0.14 ± 0.34	−0.10	943	< 0.001*
RPL8	3781	+0.26 ± 0.49	+0.41	−0.39 ± 0.34	−0.44	34.8	< 0.001*
HFE2	3905	+0.86 ± 0.66	+0.74	+0.01 ± 0.37	0.05	132	< 0.001*
Homo sapiens mitochondrion complete genome	4011	−0.72 ± 0.68	−0.87	+0.14 ± 0.58	+0.04	0.08	< 0.001**
ACTA2	4579	+1.04 ± 1.13	+1.13	−0.06 ± 0.55	−0.09	–	< 0.001*
NT5E	4907	+0.55 ± 0.85	+0.48	+0.31 ± 0.47	+0.37	–	0.290
FDPS	4980	+0.83 ± 0.63	+0.96	−0.22 ± 0.44	−0.15	70	< 0.001*
VWF	5258	+1.62 ± 0.81	+1.52	+0.69 ± 0.86	+0.68	6.5	< 0.001**
PME-1	5445	+0.49 ± 0.47	+0.47	−0.17 ± 0.42	−0.21	26	< 0.001**
POR	5447	+0.82 ± 0.39	+0.92	+0.17 ± 0.36	+0.20	156	< 0.001*
ASS	5497	−1.66 ± 0.72	−1.69	−0.10 ± 0.63	−0.13	0.03	< 0.001*
KRTCAP2	5898	+0.90 ± 0.67	+0.79	−0.15 ± 0.36	−0.20	–	< 0.001*
GLYAT	6164	−2.09 ± 0.82	−2.17	−0.27 ± 0.60	−0.35	0.02	< 0.001*
DNCL2A	7013	+0.55 ± 0.30	+0.59	0.04 ± 0.28	+0.02	396	< 0.001*
C8A	7165	−1.63 ± 0.76	−1.67	−0.55 ± 0.51	−0.57	0.03	< 0.001*
YWHAE	7531	+0.57 ± 0.40	+0.44	−0.19 ± 0.28	−0.20	321650	< 0.001*
HAO2	7647	−1.84 ± 0.65	−1.78	−0.45 ± 0.48	−0.51	0.01	< 0.001*
GTF2A1	7807	−0.81 ± 0.27	−0.88	−0.28 ± 0.32	−0.33	0.001	< 0.001**
ARID5B	7812	−0.32 ± 0.35	−0.38	+0.22 ± 0.37	+0.23	0.01	< 0.001*
COL15A1	7856	+0.85 ± 0.77	+0.81	+0.129 ± 0.43	+0.12	7.77	0.002*
HIRIP3	8479	−1.71 ± 0.74	−1.61	−0.01 ± 0.45	+0.03	0.01	< 0.001*
S100A10	8503	+1.31 ± 0.66	+1.30	+0.055 ± 0.55	+0.13	92	< 0.001**
SPINK1	8733	+3.46 ± 2.06	+3.87	+0.21 ± 0.42	+0.18	–	< 0.001**
TOMM20	8938	+0.59 ± 0.39	+0.56	−0.27 ± 0.26	−0.31	25538	< 0.001*
BHMT	9379	−1.86 ± 1.25	−2.01	+0.02 ± 0.80	+0.24	0.14	< 0.001*
SIAHBP1	9402	+0.79 ± 0.48	+0.86	+0.12 ± 0.30	+0.12	158	< 0.001*
TANK	10010	+1.19 ± 0.66	+1.11	−0.38 ± 0.78	−0.29	195	< 0.001**
COL4A2	10344	+1.21 ± 0.73	+1.12	+0.39 ± 0.54	+0.46	9.5	< 0.001*
PLG	10469	−1.87 ± 0.69	−1.75	−0.71 ± 0.57	−0.64	0.02	< 0.001*
PSMB4	10537	+0.60 ± 0.49	+0.52	−0.28 ± 0.35	−0.23	1469	< 0.001**

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

NT5E gene ($P < 0.05$). The results from the XGboost model's performance measures are detailed in Table 3.

AC, BAC, Sens, Spec, PPV, NPV, and F1 scores from the XGboost model were 98.1, 97.1, 100, 94.1, 97.2, 100, and 98.6%, respectively. The performance criteria values are plotted for the XGboost model in Figure 3. All of the identified genes' expression levels that contribute significantly to explaining the output variable are displayed in Figure 4. The HAO_2 gene was the most important predictor, with a value of 100%, followed by TOMM20 (68.047%), GPC3 (43.85%), and PSMB4 (34.907%).

Discussion

One of the most frequent forms of cancer, HCC, has a dismal survival rate for its victims. Although gene expression profiling in HCC and normal liver has been studied in-depth[2,3], the structure

Table 3

Performance metrics of the XGboost model

Metric	Value (%) (95% CI)
Accuracy	98.1 (94.3–100)
Balanced accuracy	97.1 (92.5–100)
Sensitivity	100 (90–100)
Specificity	94.1 (71.3–99.9)
Positive predictive value	97.2 (85.5–99.9)
Negative predictive value	100 (79.4–100)
F1 score	98.6 (0.954–100)

of ML-based prediction of HCV-related HCC and identification of essential candidate biomarkers using an AI strategy remains unclear. Therefore, this study aims to utilize the XGboost method to categorize gene expression data related to HCC caused by HCV and HCC not caused by HCV and to identify candidate genes for HCC.

HCC is among the most common causes of death from cancer worldwide and is responsible for a significant disease burden. It is the third leading cause of cancer-related mortality in many parts of the world[2,37,38]. The death rate and frequency of HCC demonstrate the large regional and national variability. The timing and severity of exposure to environmental and viral risk factors, access to medical care, early detection of HCC, and the availability of curative therapy all play a role in these discrepancies. Eastern Asia and Sub-Saharan Africa, two regions with exceptionally high poverty rates, account for a disproportionate share (85%) of the world's new cases of HCC[8,39]. Some possible risk factors for HCC are HBV, HCV, alcoholism, NASH, nonalcoholic fatty liver disease, and exposure to dietary toxins like aflatoxins and aristolochic acid[38]. Eighty percent of all instances of HCC are caused by chronic HBV and HCV infection[10]. One of the most widespread viral blood illnesses and a major killer globally is HCV[40,41]. There are around 71 million persons with chronic HCV infection or about 1% of the global population. The number of chronic HCV patients in Europe is 10.2 million, with the vast majority located in Eastern Europe (6.7 million), followed by Western Europe (2.3 million) and Central Europe (1.2 million)[42,43]. Nearly 1.75 million new HCV infections and 400 000 HCV-related fatalities occur annually despite high cure rates with direct-acting antiviral treatments[44]. The chance of developing HCC is increased 10-fold to 20-fold in people who have a chronic infection with the HCV[45]. There was a 21.1% rise in HCV-related HCC fatalities over the preceding decade, but deaths from other causes of HCC (e.g. alcohol) remained steady throughout this time[9].

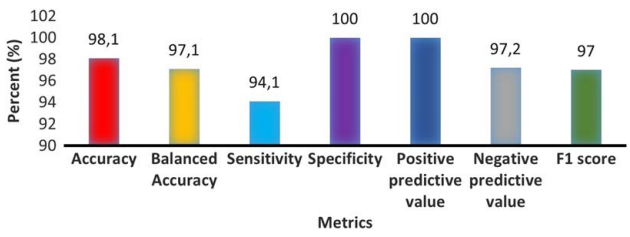

Figure 3. Graph of values for performance criteria obtained from XGboost models.

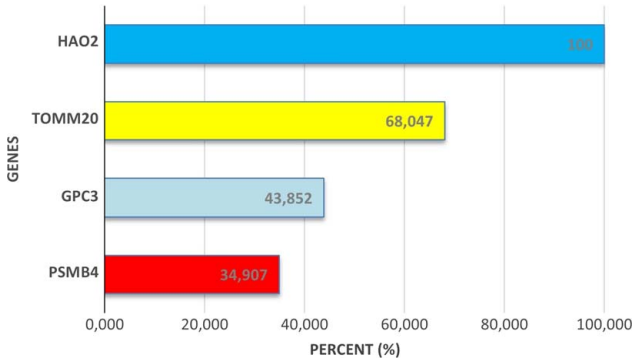

Figure 4. The graphic of gene importance values for predicting the output variable.

Patients with HCC have a dismal 5-year survival rate; thus, it is imperative that we rationally increase our efforts to minimize HCC risk factors so that we may lessen the disease's worldwide impact. To better understand the causes of liver carcinogenesis and enhance the clinical care of HCC patients, there is an increasing interest in genomics and molecular biology investigations to uncover early diagnostic and prognostic indicators and new therapeutic targets. Building on these findings, advancements in HCC surveillance promise to significantly reduce the global burden of HCC over the next few decades[15,38].

Genomic data from liver tissue samples collected from 17 patients with HCV-related HCC and 35 individuals with chronic HCV but no HCC were analyzed in this study. The samples were needed to produce cDNA microarrays, and the dataset used included 8516 expressions. The bioinformatics analysis results (Table 2) show that the HIRIP3 gene is expressed 3.22-fold less in HCV-related HCC patients than in chronic HCV patients, as measured by Log2FC values. Similarly, the GLYAT gene had 3.53-fold lower gene expression, the VIPR1 gene 2.86-fold, HAO2 had 2.60-fold, the SLC22A1 had 4.31-fold, the FBP1 gene had 3.50-fold, the IGFBP3 gene had 3.91-fold, the ASS gene had 2.92-fold, the CYP2C8 gene had 3.24-fold, the CLEC2 gene had 5.97-fold, the THBD gene had 2.07-fold lower gene expression. The GPC3 gene had 7.67-fold, the SPINK1 gene 9.51-fold upper gene expression HCV-related HCC patients than chronic HCV patients. Finally, there was no difference in the expression of the TOMM20 gene or the ECHS1 gene between the two groups. Because of their sheer quantity, gene expression data provide unique challenges for modeling. Therefore, the most crucial genes linked with the output variable were chosen using the Elastic Net variable selection approach before modeling using the current data set. To construct XGboost, forty-four genes were chosen using the Elastic Net technique. AC, BAC, Sens, Spec, PPV, NPV, and F1 scores from the XGboost model were 98.1, 97.1, 100, 94.1, 97.2, 100, and 98.6%, respectively. Measures of assessment suggested that the suggested XGboost, which relies on AI to make its determinations, successfully distinguished the two types of patients. Among the genes whose OR values were calculated, YWHAE, TOMM20, PSMB4, GOLGA5, DNCL2A, TANK, SIAHBP1, POR, HFE2, and S100A10 genes were found to have the highest top 10 OR values, respectively. According to the variable importance obtained from the XGboost method, HAO2, TOMM20, GPC3, and PSMB4 genes can be used as candidate predictive biomarkers for HCV-related HCC. In addition, the

calculated OR values and the variable importance values in the study support each other. According to variable significance results, genes with tremendous OR values were determined as contributing genes to HCV-related HCC. The suggested pipeline also generated a volcano graphic to illustrate the up-and-down-regulation of genes. Thousands of duplicate data points between two conditions are commonly included in omics research like genomics, proteomics, and metabolomics, making these charts more prevalent[46].

In a medical study, HAO_2 was shown to be down-regulated in HCC, predicting metastasis and poor survival[47]. In another study, HAO_2 was found to be down-regulated in HCC[48]. In an experimental study in rats, HAO_2 was found to be down-regulated in HCC, and it was stated to be involved in the mechanism that contributes to the development of liver cancer[49]. A study has shown the relationship between TOM M20 with HCC[50]. Another study reported that GPC3 is effective in HCC progression[51]. Another study found GPC3 as one of the potential biomarkers of human HCC[52]. In one study, PSMB4 was associated with HCC[53]. In another study, HAO_2 was found to be down-regulated in HCC, while GPC3 was found to be up-regulated[54].

All diseases that induce chronic liver cell (hepatocyte) damage are predisposing factors for developing HCC. As a result, following up on such patients following international guidelines is critical for detecting possible HCC or detecting it at an early stage[55]. Societies of the European Association for the Study of Liver Diseases (EASL)[56], American Association for the Study of Liver Diseases (AASLD)[57], and Asian Pacific Association for the Study of the Liver (APASL)[58] publish the most authoritative guidelines on monitoring chronic liver patients regularly[55]. HCC tumor doubling time is known to range between 4 and 6 months. As a result, the guidelines mentioned above recommend that patients with chronic liver disease who do not have suspected HCC be followed up with ultrasonography and AFP at 6-month intervals[55]. Patients with suspected HCC (nodule diameter 10 mm) should be evaluated with AFP and US every 3 or 6 months. Patients with a strong suspicion of HCC should be evaluated by the US and AFP. Further radiological examinations are recommended for patients with nodule diameters greater than 10 mm and/or AFP levels greater than 20 ng/ml[55].

However, because it is not always simple for people in impoverished or developing nations to contact healthcare practitioners, these measures may not consistently deliver the intended outcomes. Since ultrasound is a radiological device, the rate of false-negative findings may be higher than predicted, depending on the operator's skill. The likelihood of developing HCC is believed to increase with the length of chronic liver disease. Furthermore, like with all cancers, gene mutation and mutation-related mRNA expression changes are to be expected in HCC. Therefore, after a fixed amount of time in the follow-up of patients with chronic liver disease, a fundamental genetic analysis may be carried out to ascertain whether there is a genetic mutation. This study shows that it is possible to closely monitor patients and start preventative therapy if changes in the expression of genes highly related to HCC are found. Nonetheless, there is a lack of evidence-based data on the optimal time for genetic analysis for chronic liver disease. Therefore, it is necessary to conduct prospective multicenter research to determine when genetic analysis should be carried out on individuals with chronic liver disease. Based on this important discovery, recruiting more patients might increase the study's power and expose more of the human genome.

This study has several limitations. First, the inability to reach the patients' necessary demographic and clinical data is an essential handicap of studies using available datasets. Second, there is no data to analyze the dynamic relationship between AFP and radiological instruments frequently used in HCC screening programs and the gene mutations used in this study.

Conclusions

Using gene expression data from HCV-related HCC and individuals with chronic HCV alone, this study discovered possible genetic biomarkers for HCV-related HCC. More in-depth investigations will evaluate the accuracy of these genes, permitting the development of targeted therapies and elucidating their clinical relevance. Thus, the ground will be prepared for personalized medicine and immunotherapy to find wider practical use. Additionally, more detailed clinical works are needed to substantiate the significant conclusions in the current study.

Ethical approval

Ethical approval for this study (Ethical Committee N° 3647) was provided by the Ethical Committee of Inonu University, Malatya, Turkey on 07 June 2022.

Consent

This study was prepared using the open-access dataset of a published paper. Therefore, patient consent is not required for this study.

Sources of funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for profit sectors.

Author contribution

S.A., Z.K., and C.C.: conceptualization, investigation, writing – original draft preparation, writing – review and editing, and project administration. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest disclosure

The authors declare that they have no conflicts of interest.

Research registration unique identifying number (UIN)

1. Name of the registry: not applicable.
2. Unique identifying number or registration ID: not applicable.
3. Hyperlink to your specific registration (must be publicly accessible and will be checked): not applicable.

Guarantor

Sami Akbulut.

Data availability statement

The datasets analyzed during the current study are available from the corresponding author upon reasonable request.

Provenance and peer review

Not commissioned, externally peer-reviewed.

References

[1] Sung H, Ferlay J, Siegel RL, *et al*. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021;71:209–49.

[2] Kim E, Viatour P. Hepatocellular carcinoma: old friends and new tricks. Experimental & Molecular Medicine 2020;52:1898–907.

[3] Llovet JM, Kelley RK, Augusto V, *et al*. Hepatocellular carcinoma. Nat Rev Dis Primers 2021;7:6.

[4] Dash S, Aydin Y, Widmer KE, *et al*. Hepatocellular carcinoma mechanisms associated with chronic HCV infection and the impact of direct-acting antiviral treatment. J Hepatocell Carcinoma 2020;7:45–76.

[5] Gower E, Estes C, Blach S, *et al*. Global epidemiology and genotype distribution of the hepatitis C virus infection. J Hepatol 2014;61(1 Suppl): S45–57.

[6] Desai A, Sandhu S, Lai J-P, *et al*. Hepatocellular carcinoma in non-cirrhotic liver: a comprehensive review. World J Hepatol 2019;11:1–18.

[7] Lok AS, Everhart JE, Wright EC, *et al*. Maintenance peginterferon therapy and other factors associated with hepatocellular carcinoma in patients with advanced hepatitis C. Gastroenterology 2011;140:840–9.

[8] El-Serag HB. Epidemiology of viral hepatitis and hepatocellular carcinoma. Gastroenterology 2012;142:1264–273. e1.

[9] Axley P, Ahmed Z, Ravi S, *et al*. Hepatitis C virus and hepatocellular carcinoma: a narrative review. J Clin Transl Hepatol 2018;6:79.

[10] Yang JD, Roberts LR. Hepatocellular carcinoma: a global view. Nat Rev Gastroenterol Hepatol 2010;7:448–58.

[11] Kruse RL, Kramer JR, Tyson GL, *et al*. Clinical outcomes of hepatitis B virus coinfection in a United States cohort of hepatitis C virus-infected patients. Hepatology 2014;60:1871–8.

[12] Hajarizadeh B, Grebely J, Dore GJ. Epidemiology and natural history of HCV infection. Nat Rev Gastroenterol Hepatol 2013;10:553–62.

[13] Seeff LB. Natural history of chronic hepatitis C. Hepatology 2002;36(5 Suppl 1):s35–46.

[14] Imbeaud S, Ladeiro Y, Zucman-Rossi J. Identification of novel oncogenes and tumor suppressors in hepatocellular carcinoma. Semin Liver Dis 2010;30:75–86.

[15] Ghidini M, Braconi C. Non-coding RNAs in primary liver cancer. Front Med (Lausanne) 2015;2:36.

[16] Pezzuto F, Buonaguro L, Buonaguro FM, *et al*. The role of circulating free DNA and microRNA in non-invasive diagnosis of HBV-and HCV-related hepatocellular carcinoma. Int J Mol Sci 2018;19:1007.

[17] Plissonnier M-L, Herzog K, Levrero M, *et al*. Non-coding RNAs and hepatitis C virus-induced hepatocellular carcinoma. Viruses 2018;10:591.

[18] Park S-H, Park H-M, Baek K-R, *et al*. Artificial intelligence based real-time microcirculation analysis system for laparoscopic colorectal surgery. World J Gastroenterol 2020;26:6945–62.

[19] Polikar R. Ensemble learning. In: Zhang C, Ma Y, eds. Ensemble Machine Learning: Methods and Applications. Springer New York, NY, 2012:1–34(Chapter No : 1)

[20] Akman M, Genç Y, Ankarali H. Random forests methods and an application in health science. Turkiye Klinikleri J Biostat 2011;3:36.

[21] Piñero F, Dirchwolf M, Pessôa MG. Biomarkers in hepatocellular carcinoma: diagnosis, prognosis and treatment response assessment. Cells 2020;9:1370.

[22] Hashem S, ElHefnawi M, Habashy S, *et al*. Machine learning prediction models for diagnosing hepatocellular carcinoma with HCV-related chronic liver disease. Comput Methods Programs Biomed 2020;196: 105551.

[23] 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.

[24] Chang HY, Thomson JA, Chen X. Microarray analysis of stem cells and differentiation. Methods Enzymol 2006;420:225–54.

[25] Mathew G, Agha R. STROCSS 2021: strengthening the reporting of cohort, cross-sectional and case–control studies in surgery. Ann Med Surg (Lond) 2021;72:103026.

[26] Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. bioinformatics 2007;23:2507–17.

[27] Fodor IK. A survey of dimension reduction techniques. Lawrence Livermore National Lab., CA (US). 2002.

[28] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol 1996;58:267–88.

[29] Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc 2006;101:1418–29.

[30] Le Cessie S, Van Houwelingen JC. Ridge estimators in logistic regression. J R Stat Soc Series C (Applied Statistics) 1992;41:191–201.

[31] Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol 2005;67:301–20.

[32] Wang J, Li P, Ran R, *et al*. A short-term photovoltaic power prediction model based on the gradient boost decision tree. Appl Sci 2018;8:689.

[33] Dikker J, Boosted tree learning for balanced item recommendation in online retail, 2017.

[34] Salam Patrous Z. Evaluating XGBoost for user classification by using behavioral features extracted from smartphone sensors. Computer Science. KTH Royal Institute of Technology; 2018.

[35] Smyth GK. Limma: linear models for microarray data. In: Gentleman R, Carey VJ, Huber W, Irizarry RA, Dudoit S, eds. Bioinformatics and computational biology solutions using R and Bioconductor. Springer New York; 2005:397–420 (Chapter: 23).

[36] 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.

[37] Fitzmaurice C, Allen C, Barber RM, *et al*. Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015: a systematic analysis for the global burden of disease study. JAMA Oncol 2017;3:524–48.

[38] Yang JD, Hainaut P, Gores GJ, *et al*. A global view of hepatocellular carcinoma: trends, risk, prevention and management. Nat Rev Gastroenterol Hepatol 2019;16:589–604.

[39] Tang A, Hallouch O, Chernyak V, *et al*. Epidemiology of hepatocellular carcinoma: target population for surveillance and diagnosis. Abdom Radiol (NY) 2018;43:13–25.

[40] Jefferies M, Rauff B, Rashid H, *et al*. Update on global epidemiology of viral hepatitis and preventive strategies. World J Clin Cases 2018;6:589–99.

[41] 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; 3:117–23.

[42] Blach S, Zeuzem S, Manns M, *et al*. Global prevalence and genotype distribution of hepatitis C virus infection in 2015: a modelling study, Lancet Gastroenterol Hepatol 2017;2:161–76.

[43] Marshall AD, Pawlotsky J-M, Lazarus JV, *et al*. The removal of DAA restrictions in Europe–one step closer to eliminating HCV as a major public health threat. J Hepatol 2018;69:1188–96.

[44] Page K, Melia MT, Veenhuis RT, *et al*. Randomized trial of a vaccine regimen to prevent chronic HCV infection. N Engl J Med 2021;384:541–9.

[45] El-Serag HB, Kanwal F. Epidemiology of hepatocellular carcinoma in the United States: where are we? Where do we go? Hepatology 2014;60:1767–75.

[46] Blum BC, Emili A. Omics notebook: robust, reproducible and flexible automated multiomics exploratory analysis and reporting. Bioinform Adv 2021;1:vbab024.

[47] Mattu S, Fornari F, Quagliata L, *et al*. The metabolic gene HAO_2 is downregulated in hepatocellular carcinoma and predicts metastasis and poor survival. J Hepatol 2016;64:891–8.

[48] Li Y, Zhang M, Li X, *et al*. Hydroxyacid Oxidase 2 (HAO2) inhibits the tumorigenicity of hepatocellular carcinoma and is negatively regulated by miR-615-5p. J Immunol Res 2022;2022:5003930.

[49] Fornari F, Gramantieri L, Callegari E, *et al*. MicroRNAs in animal models of HCC. Cancers (Basel) 2019;11:1906.

[50] Yang X, Song D, Zhang J, *et al*. PRR34-AS1 sponges miR-498 to facilitate TOMM20 and ITGA6 mediated tumor progression in HCC. Exp Mol Pathol 2021;120:104620.

[51] Zhu Xt, Yuan Jh, Zhu Tt, *et al*. Long noncoding RNA glypican 3 (GPC3) antisense transcript 1 promotes hepatocellular carcinoma progression via epigenetically activating GPC3. FEBS Journal 2016;283:3739–54.
[52] Jia H-L, Ye Q-H, Qin L-X, *et al*. Gene expression profiling reveals potential biomarkers of human hepatocellular carcinoma. Clin Cancer Res 2007;13:1133–9.
[53] Midorikawa Y, Tsutsumi S, Nishimura K, *et al*. Distinct chromosomal bias of gene expression signatures in the progression of hepatocellular carcinoma. Cancer Res 2004;64:7263–70.
[54] Zhou Z, Li Y, Hao H, *et al*. Screening hub genes as prognostic biomarkers of hepatocellular carcinoma by bioinformatics analysis. Cell Transplant 2019;28(1_suppl):76S–86S.
[55] Akbulut S, Garzali IU, Hargura AS, *et al*. Screening, surveillance, and management of hepatocellular carcinoma during the COVID-19 pandemic: a narrative review. J Gastrointest Cancer 2022:1–12.
[56] EASL-EORTC clinical practice guidelines: management of hepatocellular carcinoma. J Hepatol 2012;56:908–43.
[57] Marrero JA, Kulik LM, Sirlin CB, *et al*. Diagnosis, staging, and management of hepatocellular carcinoma: 2018 practice guidance by the American Association for the Study of Liver Diseases. Hepatology 2018; 68:723–50.
[58] Omata M, Cheng AL, Kokudo N, *et al*. Asia-Pacific clinical practice guidelines on the management of hepatocellular carcinoma: a 2017 update. Hepatol Int 2017;11:317–70.