Construction and validation of a joint diagnosis model based on random forest and artificial intelligence network for hepatitis B-related hepatocellular carcinoma

Ji liang  
the First People's Hospital of Changde City

Jie zhang  
Haikou People's Hospital /Affiliated Haikou Hospital of Xiangya Medical College, Central South University

Xili Jiang  
The Second People's Hospital of Hunan Province/Brain Hospital of Hunan Province

Shucai Xie  (✉️ 282791444@qq.com)  
Central South University

Research Article

Keywords: Hepatocellular carcinoma, Hepatitis B virus, random forest, artificial intelligence network, diagnostic model

Posted Date: August 26th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1960824/v1

License: ☋ ① This work is licensed under a Creative Commons Attribution 4.0 International License.  Read Full License
Abstract

Background: Hepatitis B virus (HBV) is the dominant pathogenic factor of HCC in Asia and Africa. This study aims to identify significant biomarkers and develop a novel genetic model for the efficient diagnosis of HBV-related HCC.

Methods: GSE19665, GSE55092, and GSE121248 were merged and used to identify significant differentially expressed genes (DEGs). The enrichment analysis was performed on Metascape and Database for Annotation, Visualisation and Integrated Discovery (DAVID) online tool. The random forest (RF) algorithm and artificial neural network (ANN) were used to select the potential predictive gene panels and construct an HBV-related HCC diagnostic model. Subsequently, GSE17548, GSE104310, GSE44074, and GSE136247 were used to test the accuracy of the ANN model. Finally, the CIBERSORT algorithm was used to assess the abundance of immune infiltrates in all samples.

Results: First, 116 genes were identified as DEGs from the merged dataset, and the enrichment analysis showed that DEGs were particularly enriched in cellular hormone metabolic process, monocarboxylic acid metabolic process, NABA ECM AFFILIATED steroid metabolic process and metabolism of bile acid and bile salt. **TOP2A, CLEC1B, BUB1B, FCN2, CXCL14, CAP2**, FCN3, KMO and CDHR2 were available to develop an HBV-related HCC diagnostic model. After validation, the diagnostic model has high sensitivity and specificity in four datasets, and the areas under the ROC curves efficiency was excellent. Finally, the percentage of infiltrating immune cell types for hepatitis B-related HCC were significantly different from that of non-cancerous liver tissue with HBV.

Conclusion: A novel early diagnostic model of HBV-related HCC was established, and the model has better efficiency in distinguishing HBV-related HCC from other non-cancerous with HBV individuals. This study will provide a promising theoretical basis for early diagnosis and immunity therapy of HBV-related HCC.

Background

Hepatocellular carcinoma (HCC) is the most prevalent primary liver cancer (90%) and the fourth leading cause of cancer-related death worldwide[1]. By 2025, it is estimated to threaten the health of more than 0.8 million people annually, with Chinese patients accounting for more than half of the global HCC burden[2]. Variations in the incidence rate for HCC globally are attributed to diversity in risk factors. The prevalence of hepatitis B and C virus infections, especially the hepatitis B virus (HBV), is responsible for the highest incidence of HCC in East Asia and sub-Saharan Africa, with HBV-induced HCCs accounting for ~ 60% of cases in Asia and Africa[1, 3]. Chronic HBV infection leads to persistent liver damage and impaired regeneration, a well-known driving force of liver fibrogenesis and carcinogenesis[4]. Therefore, there is an urgent need to identify reliable diagnostic markers to distinguish HBV-related HCC from other non-cancerous individuals with HBV.

Conventionally, early clinical diagnosis of hepatocellular carcinoma is dependent on clinical symptoms, serum AFP and imaging findings in patients with chronic hepatitis or cirrhosis. Despite significant improvement in the prevention, monitoring, early screening, diagnosis and therapy of HCC over the past decade, the prognosis for the vast majority of HCC patients is typically poor. In addition, most patients lack obvious clinical symptoms in the early stage of HCC, and the deep position of the liver within the body makes early diagnosis even more challenging. This highlights the importance of cancer research in identifying effective biomarkers, which are an attractive alternative for surveillance and early diagnosis of HCC since its objectivity and reproducibility. Previous studies have identified a few potential diagnostic markers for HBV-related HCC, including AFP, DCP, GPC 73, OPN, cell-free/circulating
tumour DNA, tumour-associated MicroRNAs and extracellular vesicles\textsuperscript{[5]}. Machine learning, unlike traditional statistical methods, is not rule-based programming but rather learning from examples. Machine learning is an emerging discipline based on the intersection of statistics and mathematical sciences. It builds a statistical model from learning large massive datasets data to achieve accurate prediction and to guide future research efforts\textsuperscript{[6, 7]}. The random forest (RF), an innovative and highly effective supervised ML algorithm, uses different prediction features in the training samples to effectively classify unknown samples by constructing a series of decision trees\textsuperscript{[7]}. As random forests overcome the common problem of over-fitting through the use of bootstrap aggregation, it appears to be more accurate in prediction than other algorithms\textsuperscript{[8]}. Another supervised ML algorithm is artificial neural network(ANN), which is based on the functioning of biological neural networks. The ANN is used to build a model of the complex relationship between input and output data and reveal patterns\textsuperscript{[9]}. Compared to conventional programming, neural networks are available to deal with problems that algorithms could not solve, or the available solutions are too complex\textsuperscript{[9]}. ANN models are widely used in disease diagnosis, classification, prediction, and survival analysis because their ability to handle linear and nonlinear relationship of data\textsuperscript{[10]}. It is well acknowledged that carcinogenesis and progression of HCC are closely related to mutation of genes, overexpression of various oncogenes and inactivation of tumour suppressor genes\textsuperscript{[11]}. With the rapid development in sequencing technology, huge volumes of gene expression profiling data related to cancer were generated for the identification of novel differential genes and diagnostic and prognostic biomarkers. In the previous studies, differentially expressed genes (DEGs) and the association pathways involved in HBV-induced HCCs were identified through integrated bioinformatics analysis using multiple datasets. In this study, three datasets were merged. The RF algorithm was then used to identify the key genes expressed in HBV-related HCC, and artificial neural networks constructed a genetic diagnostic model of HBV-related HCC. Finally, immune cell infiltration between HBV-related HCC samples and non-cancerous samples with HBV was evaluated.

**Materials And Methods**

Figure 1 shows the research framework of this study.

Gene expression profiles of GSE19665(https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE19665), GSE55092(https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE55092), GSE121248(https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE121248), GSE17548(https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE17548), GSE104310(https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE104310), GSE44074(https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE44074), and GSE136247(https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE136247) were obtained from the Gene Expression Omnibus (GEO) database of the National Centre for Biotechnology Information (https://www.ncbi.nlm.nih.gov/geo/)(Supplementary 1 raw data, Table 1). The seven datasets were divided into two groups: the training set, which including GSE19665, GSE55092, and GSE121248, and the remaining datasets were classified into the test set to verify the performance of the model. As previously described, the process of converting gene probe IDs to gene symbols was done using A Perl language command. The normalisation between arrays function was used to normalise the gene expression data, and the gene expression data were averaged when multiple probes correspond to a gene. Subsequently, the expression data of the three datasets in training sets were merged and used for the following analysis, the batch effect from the different datasets was removed, and the common genes were finally obtained. The gene expression data with a larger value was subjected to log\textsubscript{2} transformation in the limma R package.
Table 1
Details of the GEO dataset

<table>
<thead>
<tr>
<th>Dataset ID</th>
<th>Sample</th>
<th>Platform</th>
<th>Non-cancerous samples with HBV</th>
<th>HBV-related HCC</th>
<th>Classification</th>
<th>Country</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE19665</td>
<td>hepatocellular carcinoma (HBV)</td>
<td>GPL570</td>
<td>5</td>
<td>5</td>
<td>training sets</td>
<td>Japan</td>
<td>Deng et al. (2010) [12]</td>
</tr>
<tr>
<td>GSE55092</td>
<td>hepatocellular carcinoma (HBV)</td>
<td>GPL570</td>
<td>91</td>
<td>49</td>
<td>training sets</td>
<td>USA</td>
<td>Melis et al. (2014) [13]</td>
</tr>
<tr>
<td>GSE121248</td>
<td>hepatocellular carcinoma (HBV)</td>
<td>GPL570</td>
<td>37</td>
<td>70</td>
<td>training sets</td>
<td>Singapore</td>
<td>Wang, Ooi &amp; Hui (2007) [14]</td>
</tr>
<tr>
<td>GSE17548</td>
<td>hepatocellular carcinoma (HBV)</td>
<td>GPL570</td>
<td>11</td>
<td>10</td>
<td>test sets</td>
<td>Turkey</td>
<td>Yildiz G et al. (2013) [15]</td>
</tr>
<tr>
<td>GSE104310</td>
<td>hepatocellular carcinoma (HBV)</td>
<td>GPL16791</td>
<td>7</td>
<td>9</td>
<td>test sets</td>
<td>China</td>
<td>Yun J et al. (2021, no published)</td>
</tr>
<tr>
<td>GSE44074</td>
<td>hepatocellular carcinoma (HBV)</td>
<td>GPL13536</td>
<td>36</td>
<td>34</td>
<td>test sets</td>
<td>Japan</td>
<td>Ueda T et al. (2013) [16]</td>
</tr>
<tr>
<td>GSE136247</td>
<td>hepatocellular carcinoma (HBV)</td>
<td>GPL17586</td>
<td>19</td>
<td>26</td>
<td>test sets</td>
<td>France</td>
<td>Cerapio JP et al. (2019) [17]</td>
</tr>
</tbody>
</table>

2.2 Identification of DEGs and enrichment analyses

The limma R package v.3.5.2 in R software was used to identify DEGs. The DEGs were selected based on the cut-off criterion that adjusted $P$-value $< 0.05$ and $|\log_2 \text{FC}| > 2$. Metascape (http://metascape.org/gp/#/main/step1), a common integrated portal, contains functional enrichment, interactome analysis, gene annotation and membership search to provide a comprehensive gene list annotation and analysis resource for users to grasp biological characteristics [18]. In present study, the DEGs enrichment analysis was performed in the “Express Analysis” module of Metascape. In addition, GO and KEGG pathway enrichment analysis were executed using DAVID (http://david.ncifcrf.gov) online tool. $P < 0.05$ was considered statistically significant.

2.3 RF screening for important genes

The RF software package (v.4.1.3) was used to filter out important variables and create a RF model that contributed most to the prediction of HBV-related HCC. First, the average model miscalculation rate of all genes based on out-of-band data was calculated. The best variable number for the binary tree at the node was set to 6, and 2000 was chosen as the best number of trees contained in the random forest [19]. Based on the point with the smallest error, the best RF model was then built, and the candidate genes for HBV-related HCC diagnosis were determined using the
mean decrease Gini. Finally, for the subsequent model construction, the genes with a significance score greater than 4 were chosen as disease-specific genes.

Subsequently, scores were assigned to the expression data of the selected DEGs using the following rules: If an upregulated gene's log FC value for a sample is greater than the gene's median expression value across all samples, its score was automatically assigned 1; otherwise, it was set to 0. If the log FC of the downregulated gene is greater than the mean expression value, its score was automatically assigned 0; otherwise, it was set to 1. The heatmap of the selected DEGs was drawn to show their expression in the merged dataset.

2.4 Neural network to build the disease classification model

The R software package neural net (v.1.44.2) was available to develop an ANN model of the important variables. The weight of each gene was obtained and five hidden layers were set as the model parameters to build a classification model of HBV-related HCC through the obtained gene score. The model accuracy results were obtained for HBV-related HCC samples and non-cancerous samples with HBV in the training set, and the pROC software package was used to calculate the areas under the ROC curves (AUC) classification performance verification results.

2.5 Validation of the predictive model

Four independent datasets (GSE17548, GSE104310, GSE44074, GSE136247) were used to verify the accuracy of the ANN model for classifying samples (HBV-related HCC or non-cancerous samples with HBV), and the ROC curves for each dataset were drawn using the pROC software package separately. At the same time, the optimal threshold in the ROC curve and the sensitivity and specificity in classifying cancer and normal samples under this threshold were calculated.

2.6 Evaluation of immune cell infiltration

The normalised gene expression data from the merged dataset was available to evaluate the abundance of immune infiltrates in all samples through the CIBERSORT algorithm. The percentages of 22 infiltrating immune cell types were calculated and output with the cutoff criterion that $P$-value < 0.05, and their correlations were displayed in a correlation heatmap drawn by the “corrplot” package[20]. The ratios of infiltrating immune cells in non-cancerous liver tissues from HBV patients and HBV-related HCC tissues were visualised by a histogram, and the difference was shown by violin diagrams.

Results

3.1 Identification of DEGs in HCC

GSE19665, GSE55092, and GSE121248 gene expression data were merged as a training dataset for subsequent analysis. A total of 133 non-cancerous liver tissues with HBV and 124 HBV-related HCC tissues were included in present analysis. As shown in the volcano graph (Fig. 2A), 116 genes were identified as DEGs according to the cut-off criterion that adjusted $P$-value < 0.05 and $|\log_2 FC| > 2$ (Table S1). Fig. 2B shows a heatmap of the top 10 up- and downregulated genes.

3.2 Functional enrichment analysis of DEGs in the training dataset
The enrichment analysis was performed on the “Express Analysis” module of Metascape and DAVID online tool to further investigate the biological functions of 116 DEGs. As shown in Fig. 3A, the top 20 clusters of enrichment analysis on Metascape showed that DEGs were particularly enriched in cellular hormone metabolic process, monocarboxylic acid metabolic process, NABA ECM AFFIALIATED, steroid metabolic process and bile acid and salt metabolism. Protein-protein interaction enrichment analysis was shown in Fig. 3B. As previously described [21], the GO analysis consisted of three functional groups, namely, the biological process group (BP), the cellular component group (CC) and the molecular function group (MF). The GO analysis results showed that the DEGs were particularly enriched in the CC (Fig. 3C, Table S2), including the extracellular region, extracellular space, membrane attack complex, spindle and condensed nuclear chromosome kinetochore. Positive regulation of the Wnt signalling pathway, planar cell polarity pathway, cellular response to cadmium ion, complement activation, xenobiotic metabolic process and retinol metabolic process were all enriched in DEGs for the BP. Iron ion binding, heme binding and oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, monooxygenase activity and oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, reduced flavin or flavoprotein as one donor, and incorporation of one atom of oxygen were principally enriched in the MF DEGs. The KEGG results revealed that the DEGs were particularly enriched in metabolic pathways, bile secretion, caffeine metabolism, retinol metabolism and chemical carcinogenesis-DNA adducts, as shown in Fig. 3D.

3.3 RF screening for DEGs

All 116 DEGs were included in the RF classifier. Fig. 4A shows the relationship between the model error and the number of decision trees. The final model showed a stable error when the number of decision trees is 2000. Therefore, the RF model was built with 2000 trees as the parameter of the final model. Genes with an importance score were shown in Fig. 4B; genes with an importance score greater than 4 were selected as the candidate genes for subsequent analysis. Finally, nine genes were selected, TOP2A, CLEC1B, BUB1B, FCN2, CXCL14 and CAP2 being the most important, followed by FCN3, KMO and CDHR2. As shown in Fig. 4C, CAP2/TOP2A/BUB1B were upregulated in HBV-related HCC samples, while KMO/CDHR2/CXCL14/FCN2/CLEC1B were upregulated in non-cancerous liver tissue with HBV.

3.4 Construction of the ANN model

Expression data for these nine genes in each sample were assigned a score of 1 or 0. Based on these nine important variables, an ANN model was constructed and used to distinguish HBV-related HCC tissues and non-cancerous liver tissue with HBV in 257 samples of the merge datasets (Fig. 5A). As a result, the model could correctly predict 132 cases in the HBV-related HCC group with 99.2% (132/133) accuracy and 120 cases in the non-cancerous with HBV group with 96.8% (120/124) accuracy. The AUC of the model in the training dataset were close to 1 (average AUC > 0.99), showing the highly stable of the model in diagnosing HBV-related HCC (Fig. 5B).

3.5 Validation of the ANN model

Four independent datasets (GSE17548, GSE104310, GSE44074, GSE136247) were used to verify the performance of the ANN model to classify samples (HBV-related HCC tissues or non-cancerous liver tissues with HBV). As a result, the model could correctly predict 23 cases in the HBV-related HCC group with 88.5% (23/26) accuracy and 19 cases in the non-cancerous with HBV group with 100% (19/19) accuracy in GSE136247, 9 cases in the HBV-related HCC group with 90% (9/10) accuracy and 9 cases in the non-cancerous with HBV group with 81.8% (9/11) accuracy in GSE17548, 23 cases in the HBV-related HCC group with 88.5% (23/26) accuracy and 17 cases in the non-
cancerous with HBV group with 89.5%(17/19) accuracy in GSE104310, and 26 cases in the HBV-related HCC group with 76.5%(26/34) accuracy and 26 cases in the non-cancerous with HBV group with 72.2%(26/36) accuracy in GSE44074. The areas under the ROC curves (AUC) of the model in the test dataset were 1 (95% CI 1–1), 0.927 (95% CI 0.791–1), 0.921 (95% CI 0.738–1) and 0.833 (95% CI 0.725–0.918), respectively (Fig. 6).

3.6 Immune cell infiltration results
A total of 133 cases of non-cancerous liver tissues from HBV patients and 124 cases of HBV-related HCC tissues were selected for the immune cell infiltration analysis. Based on the cut-off criterion that \( p < 0.05 \), 38 cases of HBV-related HCC tissues and 31 cases of non-cancerous liver tissues from HBV patients were selected for CIBERSORT analysis. First, the percentages of 22 kinds of immune cells in each sample were visualised in a histogram (Fig. 7A). The correlations of 22 kinds of infiltrating immune cells between HBV-related HCC tissues and non-cancerous liver tissue with HBV were analysed (Fig. 7B). For example, T follicular helper cells were positively correlated with T cells CD8+ and macrophages M1. NK cells resting were positively associated with neutrophils and T cells CD4 naïve. The Wilcoxon test was used to detect significantly different immune cell infiltrates between HBV-related HCC tissues and non-cancerous liver tissue with HBV. The results that presented 12 types of immune cells with \( p < 0.05 \) are shown in a violin diagram in Fig. 7C.

Discussion
In present study, 116 DEGs were identified in the merged dataset formed from three HBV-related HCC datasets. Nine important candidate DEGs were acquired through the RF classifier, and a neural network model was created. Four independent datasets were used to verify the classification (HBV-related liver cancer or non-cancerous liver tissues with HBV) efficiency of the model, and the AUC efficiency was excellent. The immune cell infiltration result shows that the percentages of 12 types of immune cells were significantly different between HBV-related HCC tissues and non-cancerous liver tissue with HBV.

Random Forest (RF) and Neural Network (NN) are different types of algorithms. Random forests (RF) is an ensemble decision tree approach where each decision tree processes a sample and predicts an output label. Decision trees in an ensemble are independent. ANN is composed of many layers of nodes that carry the signal and process it to make the final decision\cite{22}. An ANN model for the diagnosis and screening of HBV-related HCC was constructed based on nine important genes from random forests. Of these nine genes, TOP2A and BUB1B have been extensively studied in HCC\cite{23-27}. KMO\cite{28, 29}, CDHR2\cite{30}, CLEC1B\cite{31}, CXCL14\cite{32} and FCN2\cite{33} were significantly decreased in HCC tissues (or) and cell lines\cite{30} overexpression of these genes exhibits tumour-inhibitory effects towards HCC\cite{28, 29}, including inhibits tumor formation and the growth of subcutaneous tumours, suppresses proliferation, migration and invasion of HCC cells, EMT and induced apoptosis. FCN3 expression was significantly lower in HCC tissues than in normal tissues\cite{34}. However, more in vitro and in vivo experiments are needed to further confirm its effect on HCC. KMO\cite{29}, CXCL14\cite{35}, CAP2\cite{36} and FCN3\cite{37} were prognostic markers in HCC, and the combination of PD-L1\textsubscript{high} and CLEC1B\textsubscript{low} expression has been shown to predict worse outcomes\cite{38}.

CAP2 was a valuable molecular marker in the histological diagnosis of early HCC\cite{39}, and its overexpression might be related to multistage hepatocarcinogenesis\cite{40}. In addition, CAP2 transcriptional levels were significantly suppressed in silibinin-treated HCC cells. Silibinin could be a potential therapeutic agent against HCC, particularly for HBV-related HCCs\cite{41}. These findings indicate that CAP2 may play a critical role in the carcinogenesis or
progression of HBV-related HCC. CXCL14 was markedly suppressed in HBV-related HCC tissues, and its polymorphisms were associated with advanced-stage chronic HBV infection\[^{42}\]. FCN2 is active in hepatitis B infection\[^{43}\], and ficolin-2 serum levels and FCN2 haplotypes contribute to the outcome of HBV infection in a Vietnamese cohort\[^{43}\]. Ficolin-2 was implied, which may play a crucial role in innate immunity against HBV infection.

This study aimed to establish an effective diagnostic model for HBV-related HCC based on gene expression data from GEO. The three datasets in the training group are from different countries, using the same sequencing platform, which minimised the effect of confounding factors to some extent. Four independent datasets from different countries and regions were used to assess the performance of this diagnostic model, increasing the stability, usefulness and credibility of this model. The results show that the diagnostic model has high sensitivity and specificity in four test datasets, and the AUC efficiency was excellent.

Other types of diagnostic and predictive models for HBV-related HCC have also been established previously. Integrated analysis of the microbiome and host transcriptome revealed that six important microbial markers associated with the tumour immune microenvironment or bile acid metabolism showed the potential to predict clinical outcomes\[^{44}\]. LncRNA was also a potential diagnostic biomarker for HBV-related HCC, and AL356056.2, AL445524.1, TRIM52-AS1, AC093642.1, EHMT2-AS1, AC003991.1, AC008040.1, LINC00844 and LINC01018 were screened out by machine learning\[^{45}\]. Based on the data from the hospital authority data collaboration lab, 124,006 patients with CVH with complete data were included to build the models, and HCC-RS from the ridge regression machine learning model accurately predicted HCC in patients with chronic viral hepatitis\[^{46}\]. In addition, another study identifies noninvasive biomarkers by applying a urinary proteomic strategy\[^{47}\].

Infiltrating immune cells, a component of the tumour microenvironment, are involved in many processes, including tumour growth, invasion and metastasis. Accumulating evidence has shown that HCC tumours harbour a significant level of immune cell infiltration, and the status of immune cell infiltration and its characteristics are usually associated with different prognostic outcomes\[^{48, 49}\]. In this study, the density of B cells memory, T cells CD8, T cells regulatory (Tregs), NK cells resting, macrophages M0, dendritic cells activated in tumour tissues significantly increased compared with non-cancerous liver tissues with HBV. In contrast, the density of B cells naïve, Plasma cells, T cells CD4 memory resting, T cells gamma delta, NK cells activated, mast cells activated in HBV-related HCC tissues significantly decreased. T cells, B cells, NK cells, macrophages and mast cells have been previously reported to be present in immune cell infiltrates of HCC and play essential roles in the development, prognosis and immunotherapy treatment of HCC. High densities of naïve B cells and plasma cells were associated with superior survival\[^{50}\]. The antitumor or tumour-promoting effects of tumour-infiltrating lymphocytes depend on the proportion of the lymphocyte subsets constituent in the tumour microenvironment, and T lymphocytes are the primary TIL cells in HCC\[^{51}\]. The mechanism of mast cell activation in HCC is unclear, but its activation facilitates immune escape and resultant tumor growth\[^{49}\]. More importantly, HBV-specific CD8+ T cells, HBV-non-specific CD8+, CD4+T, B and NK/NKT are all involved in the development of HBV-related HCC\[^{52}\].

This study also has some limitations. First, HCC exhibits high heterogeneity, which contains etiologic, geographic and molecular heterogeneity. Molecular heterogeneity can be further classified into interpatient, intertumor and intratumor heterogeneity\[^{53}\]. The HBV-related HCC diagnosis model using an ANN was solely based on gene expression data. Therefore, it is difficult to use a single model to accurately diagnose HCC at an early stage,
although the model performed satisfactorily on the training and validation datasets. Second, the number of samples used for the construction and validation of this model was relatively small. Third, subsequent confirmatory experiments and clinical practice are needed to further monitor the accuracy and stability of the diagnostic model.

**Conclusion**

In conclusion, a combination of three datasets’ expression data was used to select important variables through random forest. An ANN model was formulated for the early diagnosis and screening of HBV-related HCC. Finally, the ratio of infiltrating immune cells in non-cancerous liver tissues from HBV patients and HBV-related HCC tissues were assessed. The findings give a deeper and more comprehensive understanding of the occurrence and progression of HCC and its association with HBV and a valuable reference for the early screening and directions for improving the clinical efficacy of HBV-related HCC.

**Abbreviations**

ANN: artificial neural network; BP: biological process group; CC: cellular component group; DAVID: Database for Annotation, Visualisation and Integrated Discovery; DEGs: differentially expressed genes; GEO: Gene Expression Omnibus; HCC: Hepatocellular carcinoma; HBV: Hepatitis B virus; MCODE: Molecular Complex Detection; MF: molecular function group; RF: random forest; RR: ribonucleotide reductase; SMDs: small-molecule drugs; STRING: The Search Tool for Retrieval of Interacting Genes; Tregs: T cells regulatory.

**Declarations**

**Author contribution**

This study was designed by SX, JZ, and JL. JZ, XJ, and SX performed the statistical analysis. SX and JL finished writing the article. XJ and SX provided supervision and final check. All the authors read the final version of this paper and approved it.

**Ethics approval and consent to participate**

Not applicable

**Availability of data and materials**

Gene expression profiles of GSE19665, GSE55092, GSE121248, GSE17548, GSE104310, GSE44074, and GSE136247 were downloaded from the GEO database of the NCBI. These six datasets used in the study are publicly available.

**Competing interests**

The authors declare that they have no conflict of interest.

**Funding**

This work was supported by the China Postdoctoral Science Foundation (No.2022M713535).

**Acknowledgment**
References

4. LI J, CHENG L, JIA H, et al. IFN-gamma facilitates liver fibrogenesis by CD161(+)CD4(+) T cells through a regenerative IL-23/IL-17 axis in chronic hepatitis B virus infection [J]. Clin Transl Immunology, 2021, 10(11): e1353
5. PANDYARAJAN V, GOVALAN R, YANG J D. Risk Factors and Biomarkers for Chronic Hepatitis B Associated Hepatocellular Carcinoma [J]. Int J Mol Sci, 2021, 22(2)
35. LIN T, ZHANG E, MAI P P, et al. CXCL2/10/12/14 are prognostic biomarkers and correlated with immune infiltration in hepatocellular carcinoma [J]. Biosci Rep, 2021, 41(6)
52. CHEN Y, TIAN Z. HBV-Induced Immune Imbalance in the Development of HCC [J]. Front Immunol, 2019, 10: 2048

**Figures**
Figure 1

Schematic illustration of the research design

Figure 2

(A) Volcano plot of differential expression analysis results. The abscissa is log2FC and the ordinate is -log10 adjusted P-value. The red dots represent the upregulated genes based on an adjusted P < 0.05 and log2FC > 2; the green dots represent the downregulated genes based on an adjusted P < 0.05 and log2FC < -2; the black dots represent the remaining stable genes. (B) Heatmap of the top 10 up- and downregulated genes. Colours on the graph from red to blue indicate high to low expression. On the upper part of the heatmap, the blue band indicates the non-cancerous HBV samples and the red band indicates HBV-related HCC samples.
Figure 3

(A) The top 20 clusters of enrichment analysis were performed on Metascape; (B) Protein-protein interaction enrichment analysis was performed on Metascape; GO analysis (C) and KEGG pathway enrichment analysis (D) of DEGs using the online database DAVID;

Figure 4

(A) The influence of the number of decision trees on the error rate. The x-axis represents the number of decision trees, and the y-axis indicates the error rate. (B) The importance of the top 30 genes ranked by mean accuracy decreases. (C) Heatmap of the nine important genes generated by RF. The red colour indicates high expression genes in the samples, the blue colour indicates low expression genes in the samples, the red band on the upper side of the heatmap represents HBV-related HCC samples, and the blue band indicates non-cancerous liver tissue with HBV.

Figure 5

(A) Construction of a neural network: the neural network topology of the dataset with five hidden layers. (B) The ROC curve of the predictive model of the training dataset.

Figure 6

The ROC curve of the ANN Model in the Validation dataset. (A) GSE136247. (B) GSE17548. (C) GSE104310. (D) GSE44074

Figure 7

Immune cell infiltration in HBV-related HCC tissues and non-cancerous liver tissue with HBV. (A) The compositions of 22 immune cell types in each sample were shown in a histogram. (B) The correlations of 22 types of immune cells in HBV-related HCC tissues were evaluated. Red: positive correlation; blue: negative correlation. (C) Wilcoxon test was conducted to analyse the different immune cell infiltrates in HBV-related HCC and HBV non-cancerous liver tissues.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS1diff.xls
- TableS2GOandKEGG.xlsx
• rawdata.zip