Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 5679837, 17 pages
https://doi.org/10.1155/2022/5679837

Research Article

Establishment and Analysis of a Combined Diagnostic Model of Liver Cancer with Random Forest and Artificial Neural Network

Runzhi Yu, Ziyi Cao, Yiqin Huang, Xuechun Zhang, and Jie Chen

Department of Gastroenterology, Clinical Research Center of Geriatric Frailty, Huadong Hospital Affiliated to Fudan University, Shanghai, China

Correspondence should be addressed to Jie Chen; laughchen@126.com

Received 27 February 2022; Revised 30 March 2022; Accepted 2 April 2022; Published 10 May 2022

Academic Editor: Yuchen Li

The incidence of liver cancer (hepatocellular carcinoma; HCC) is rising and with poor clinical outcome expected, a more accurate judgment of tumor tissues and adjacent nontumor tissues is necessary. The aim of this study was to construct a diagnostic model based on random forest (RF) and artificial neural network (ANN). It can be used to aid in the identification of diseased tissue such as cancerous tissue, for HCC clinical diagnosis and surgical guidance. GSE36376 and GSE121248 from Gene Expression Omnibus (GEO) were used as training sets in this investigation. R package "limma" and WGCNA were used to filter the training set for statistically significant ($p < 0.05$) differential genes. To better understand the biological function and characteristics, R software was used to perform GO and KEGG enrichment analyses. To pick out and further understand the key genes, we performed PPI analysis and random forest tree analysis. Next, we built the ANN to predict training sets and validation set (GSE84402), and ROC curve was plotted to calculate area under curve (AUC). Then immune cell infiltration indicated difference of immune cell subsets between control and case groups. Finally, the survival analysis of key genes was also carried out based on data in TCGA database. Based on the expression of these 9 genes, we built the artificial neural network (ANN) and the accuracy of the final models was assessed with an ROC curve. The areas under the ROC curve were 0.984 (95% CI 0.972–0.993) in training sets. Its predictive capability was further assessed using the validation set. And the areas under the ROC curve were 0.929 (95% CI 0.786–1.000). In summary, this method effectively classifies hepatocellular carcinoma tissues and the corresponding noncancerous tissues and provides reasonable new ideas for the early diagnosis of liver cancer in the future.

1. Introduction

Every year, more than 850 000 new cancer cases are diagnosed in the world's livers, with hepatocellular carcinoma accounting for over 90% of them [1]. The burden of liver cancer is projected to be over 1 million cases by 2030 [2]. Chronic hepatitis B and C virus infection, dietary toxin exposure (such as aflatoxin and aristolochic acid), metabolic illnesses (such as fatty liver disease and diabetes), and alcohol addiction are all key risk factors for HCC [3]. The neoplastic genesis of HCC is a multistep histological process. Hepatocellular necrosis is followed by hepatocyte growth after a hepatic injury. Chronic liver disease develops as a result of continuous destructive-regenerative cycles, resulting in liver cirrhosis, which is characterized by fibrosis and aberrant nodule development. Then hyperplastic and dysplastic nodules appear, leading to the development of HCC. HCC is further divided into three types: well-differentiated, moderately differentiated, and poorly differentiated tumors, with the last being the most dangerous type of primary HCC [4].

Hepatocellular carcinoma monitoring, diagnosis, and therapy have all improved significantly over the last decade [3]. Despite the disease's declining incidence rates, disease-specific death rates remain high [5], and early diagnosis is important to improving outcomes [6]. Biannual ultrasound (US) with or without alpha-fetoprotein (AFP) testing is recommended by international hepatic associations for screening for HCC in at-risk individuals [7–9]. The purpose of these guidelines is to enhance the possibility of detecting early-stage HCC that could be treated successfully [10]. However, alpha-fetoprotein (AFP), dynamic magnetic resonance imaging, and computed tomography have low sensitivity and specificity: the combination of US and AFP has a sensitivity of just 63 percent for early-stage HCC identification [11], and more than 5% of MRI-diagnosed HCC may be false positive or non-HCC lesions [12], calling into question their value as primary screening tools for hepatocellular carcinoma [9]. Furthermore, early diagnosis is more challenging due to confounding variables such as the presence of inflammation and cirrhosis [9].

Therefore, novel diagnostic models for HCC with higher diagnostic accuracy are still urgently required.

Recently, the availability of omics data for illnesses, such as tumors and HCC, is rapidly increasing [13, 14]. However, because of the high dimensionality of this data, detecting biologically relevant patterns might be difficult. This circumstance needs the creation of new analytical methods, such as using artificial intelligence (AI) and random forest to analyze data. Now the most popular example of artificial intelligence methods is the artificial neural network (ANN) [15], which enables computer systems to improve forecast accuracy by creating a probabilistic or statistical model based on current data [16]. For data-driven precision medicine, a hypothesis-free strategy to integrating huge data is essential. In the field of liver diseases, which is associated with multifactorial and complex characteristics, approaches based on ANN to combine multiple factors using available data appear to improve performance in diagnostic tasks and decision-making tasks based on treatment response or prognostic prediction [17].

ANNs, with all of their variants, are now the main tools in machine learning tasks, such as disease diagnosis and classification [18]. Allahverdi et al. [19] created a heart disease classification system that used an artificial neural network and attained an accuracy of 82.4%. Dongfang Jia et al. [20] proposed a methodology based on artificial neuronal networks that can accurately classify cancer tissues and normal tissues and provides reasonable new directions for the early diagnosis of cancer. In this context, we suggested a novel diagnostic model that overcomes some of the disadvantages of standard diagnostic approaches, fully utilizing the advances in omics technologies and achieving a more comprehensive optimization of diagnosis of HCC, allowing for a faster, more sensitive, and radiation-free HCC detection.

2. Materials and Methods

2.1. Materials

2.1.1. Data Collection. The study's workflow is depicted in Figure 1. The keywords "hepatocellular carcinoma/liver cancer", "normal", "Expression profiling by array", "Series", and "*Homo sapiens*" were used to search the Gene Expression Omnibus (GEO) databases for liver cancer gene expression profiles. Datasets collected from human case-control studies, with hepatocellular carcinoma tissues as the case group and corresponding noncancerous tissues as the control group, were included in our training set. Therefore, three datasets (GSE36376, GSE121248, and GSE84402) met the screening criteria. GSE36376 contained 240 tumor and 193 adjacent nontumor liver samples, and GSE121248 contained 70 tumor and 37 adjacent nontumor liver samples. GSE36376 and GSE121248 were combined as the training set to screen for the key genes to build the ANN model. GSE84402 (contained 14 tumor and 14 corresponding noncancerous samples) served as an independent validation set to verify the accuracy of the ANN model.

Flow diagram of the study. Data collection, analysis, key gene selection and validation.

2.1.2. Differentially Expressed Genes (DEGs) Screening. The differentially expressed genes (DEGs) were screened by limma [21]. The R software package's limma contains a solution for DEA of microarray data. The DEGs between tumor and neighboring nonneoplastic liver were screened using limma in the GSE36376 and GSE121248 datasets, respectively. Both $|\log FC| \geq 2.0$ and adjusted $P < 0.05$ were used as the thresholds for DEGs. All DEGs were visualized by a volcano plot.

2.1.3. Weighted Gene Coexpression Networks Construction and Module Selection. The gene modules were screened by WGCNA. After obtaining the gene expression profile, the WGCNA software tool in R [22] was used to create a gene coexpression network using the gene expression data of DEGs.

First, the appropriate soft-thresholding power (β) was selected by using the "pickSoftThreshold" function with the default parameters (herein, $\beta = 7$). Subsequently, Pearson's correlation matrix was calculated to evaluate the similarity among all the pairwise genes by using the "cor" function with the default parameters. Then, the adjacency was calculated based on β and Pearson's correlation matrix by using the "TOMsimilarity" function with the default parameters, and the corresponding dissimilarity (dissTOM) was also calculated. Finally, average linkage hierarchical clustering was conducted according to the dissTOM value with a minimum size of 48 for each gene dendrogram.

Module eigengenes (MEs), considered the first principal component (PC) of gene expression patterns of a corresponding module, were obtained for each module. To further strengthen the reliability of the modules, a cut line was set at 0.25 so that modules bearing <0.25 would be merged [23]. The module with the highest MS was considered as the key module related with liver cancer.

2.1.4. Candidate Gene Selection from the Most Significant Module. The intensity of intramodular interconnectivity (also known as module membership (MM)) was computed using the absolute value of Pearson's correlation coefficient between module eigengene and expression values to define candidate genes. Candidate genes with an MM of $r \geq 0.80$ and a substantial gene connection with liver cancer (at $p \leq 0.05$) were prioritized for further investigation. A

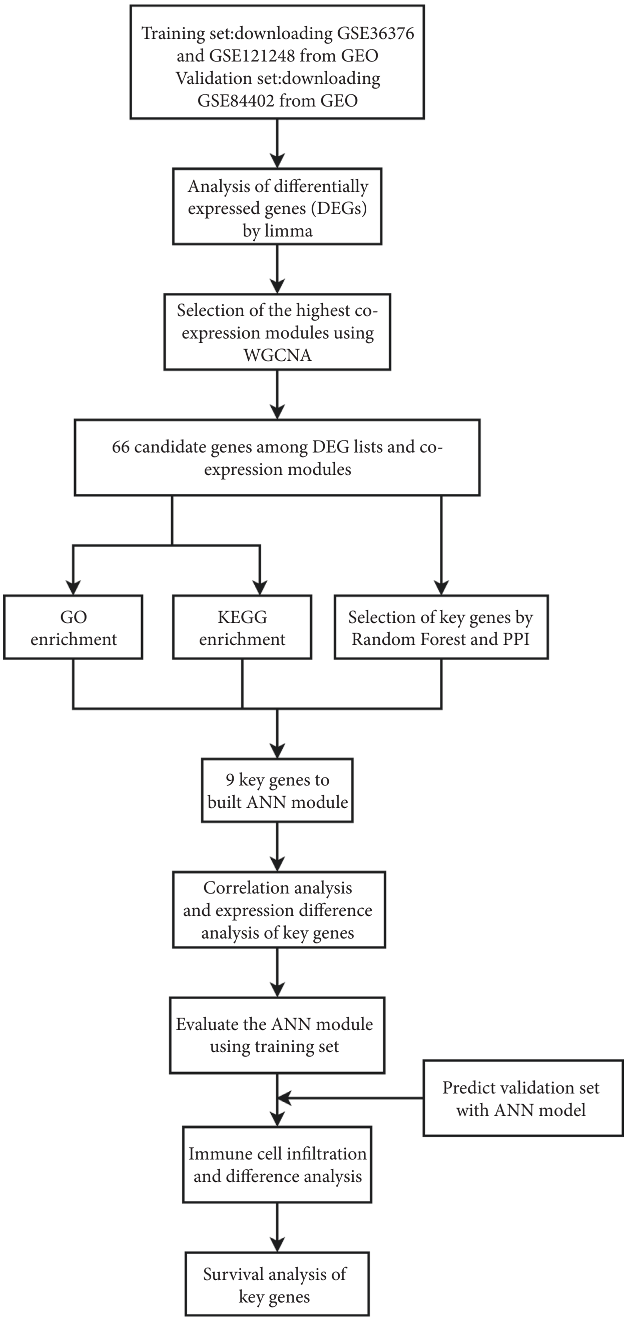

Training set:downloading GSE36376 and GSE121248 from GEO Validation set:downloading GSE84402 from GEO
Analysis of differentially expressed genes (DEGs) by limma
Selection of the highest co-expression modules using WGCNA
66 candidate genes among DEG lists and co-expression modules
GO enrichment
KEGG enrichment
Selection of key genes by Random Forest and PPI
9 key genes to built ANN module
Correlation analysis and expression difference analysis of key genes
Evaluate the ANN module using training set
Predict validation set with ANN model
Immune cell infiltration and difference analysis
Survival analysis of key genes

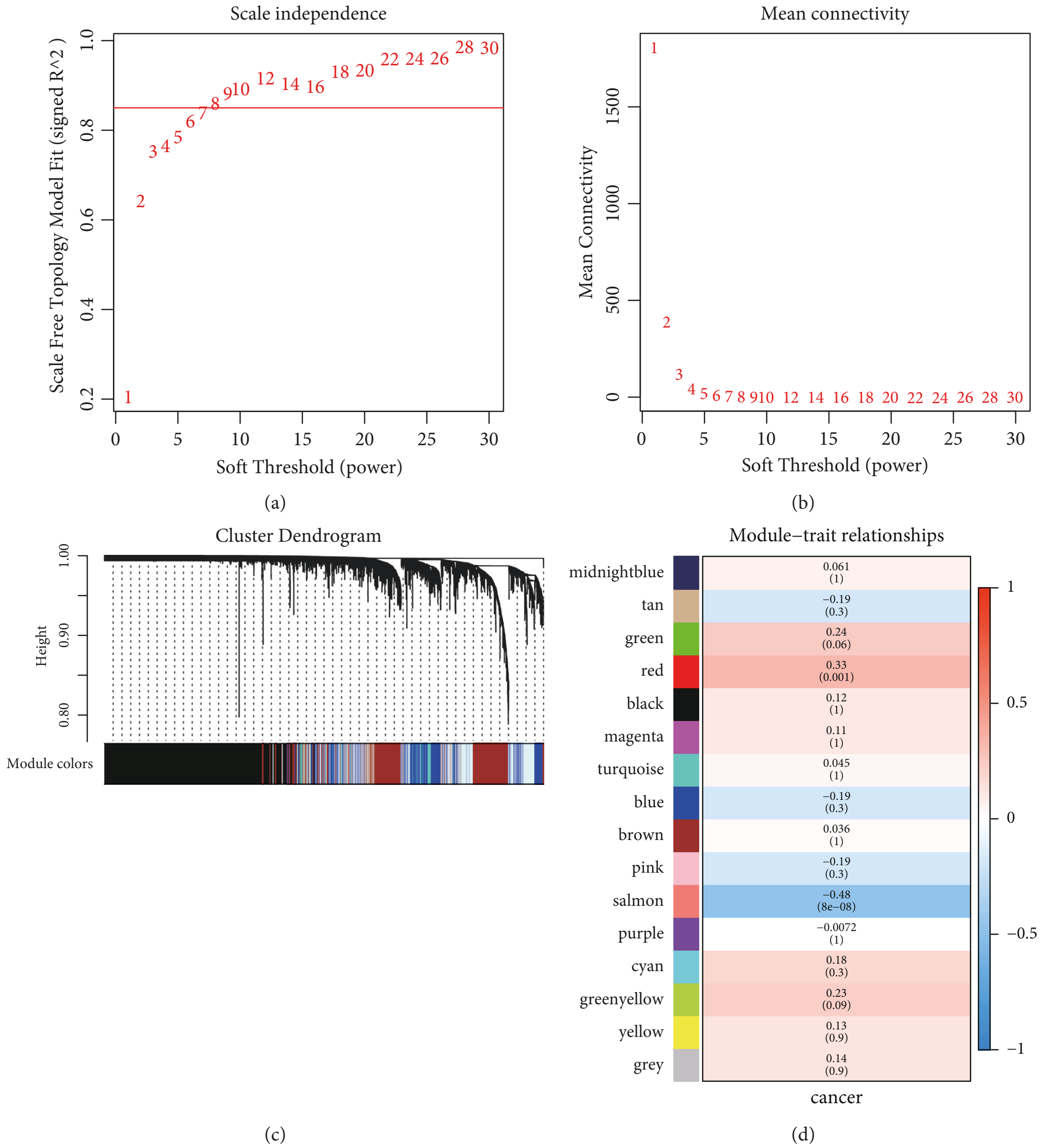
Figure 1: Flow diagram of the study.

volcano plot and heatmap were used to illustrate all of the candidate genes.

2.1.5. Functional Enrichment Analysis. The functional analysis was performed by Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) analyses. The KEGG and GO analyses were conducted by R package "clusterProfiler." [24].

2.1.6. Protein-Protein Interaction (PPI) Network Analysis. The information of the interaction of proteins and neighborhood, gene fusions were provided using the Search Tool for the Retrieval of Interacting Genes (STRING) database (a publicly available database; https://string-db.org/) [25]. In the present study, the input gene sets were 66 candidate genes and the species was *Homo sapiens*. To further explore the potential relevance of the candidate genes, the minimum required interaction score was 0.4.

2.1.7. Random Forest for Key Genes Screening. The random forest package in R was used with 500 trees and default parameters to do the random forest analysis [26]. To minimize overfitting, we used a decision tree-based method that included internal cross-validation and took into account



(a)

(b)

(c)

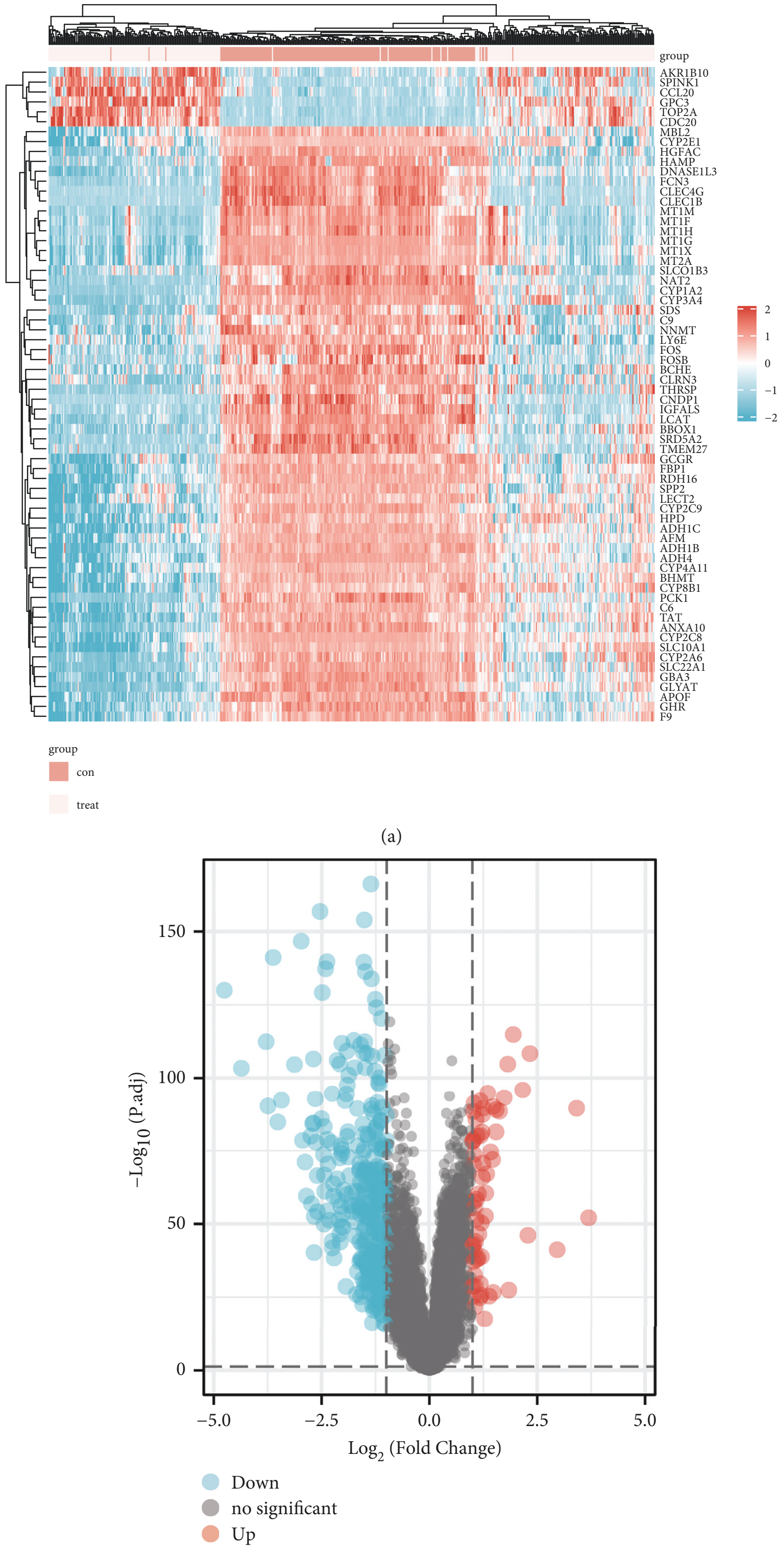
(d)

Figure 2: The results of weighted coexpression gene network analysis. An overview of the coexpressed genes in the current study, demonstrating the relevance of gene modules and phenotypes. (a) Screening soft-thresholding powers. (b) Mean network connectivity of soft-thresholding powers used in WGCNA. A soft threshold of 6 is the most suitable value. (c) Cluster dendrogram of the identified coexpression modules. In this figure, each gene is represented as a leaf and corresponds to a color module. Each color indicates that each gene in its corresponding cluster dendrogram belongs to the same module. If some genes have similar changes in expression, then these genes may be functionally related. Moreover, all these genes can further be included into a single module. The gray block represents the genes that do not coexpress with genes of any other color module. (d) Module-trait weighted correlations and corresponding P-values for the identified gene module and pathologic type (tumor tissues). The label of color on the right represents the strength of correlation, from 1 (red) to –1 (blue).

nonlinear data. We then obtained the best trees with the fewest cross-validated errors for crucial gene screening.

2.1.8. Gene Scores for Removing Batch Effects. For each key genes screening by random forest, we calculated the median value of that gene across all arrays. Among the upregulated genes, whose expression is greater than the median value, we marked 1; otherwise mark 0. Among the downregulated genes, whose expression is greater than the median value, we marked 0; otherwise mark 1. Then we get the scores for these key genes in each sample to remove batch effects of training and validation set.

2.1.9. Correlation Analysis and Expression Difference Analysis of Key Genes. Spearman correlation analysis was used to analyze the correlation among key genes. Then we used the



(a)

(b)

Figure 3: Visualization of the 66 candidate genes. Each column of the heatmap represents the sample, the row represents the gene, and each grid represents the degree of gene expression in the sample. The row of the volcano graph represents log |FC|, and the column represents −log10 (adjusted P-value), and each point is the degree of gene expression. (a) Heat plot of candidate genes. (b) Volcano plot of candidate genes.

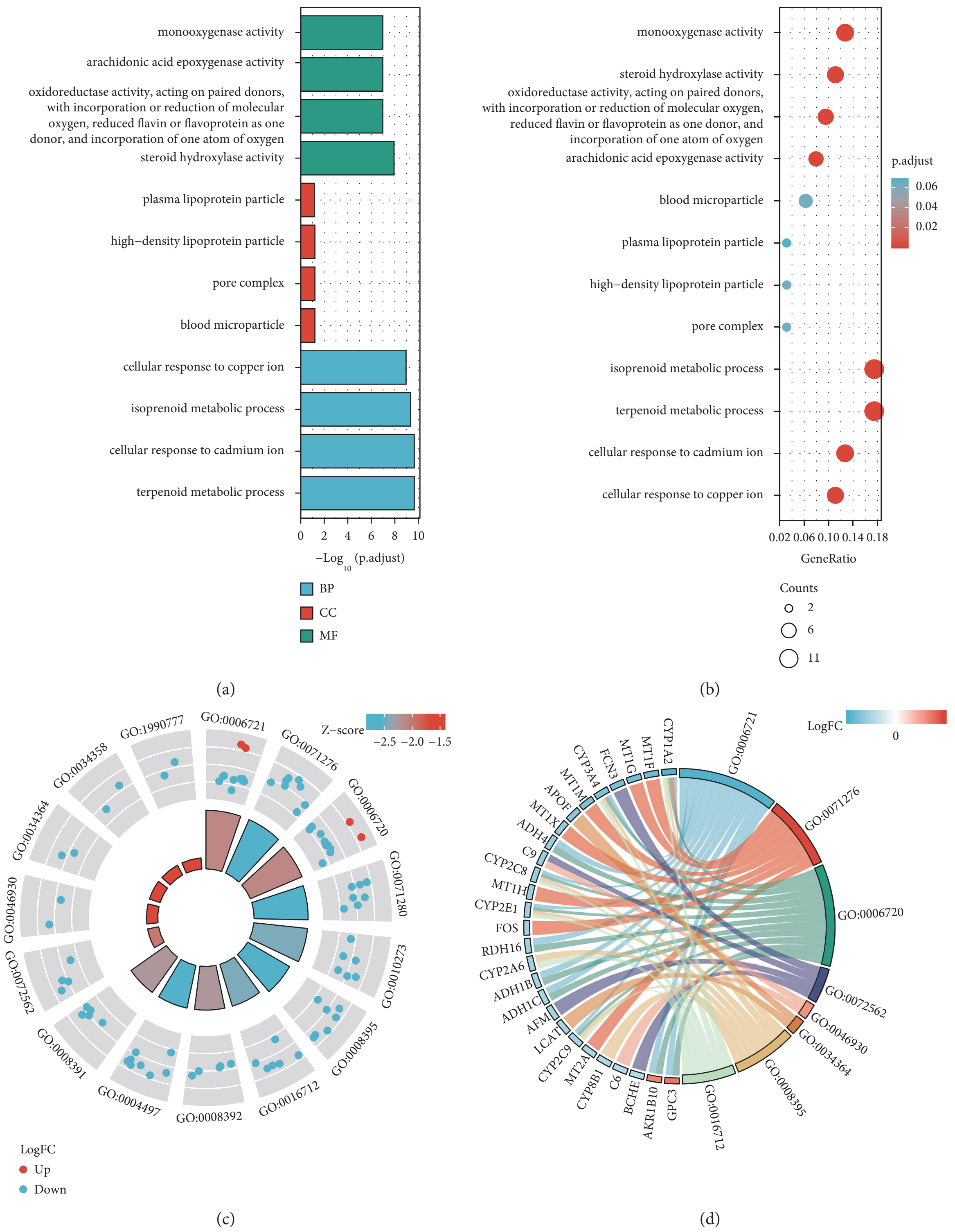


Figure 4: GO pathway enrichment analyses of candidate genes. (a) GO analysis indicated enrichment of the differentially expressed genes in biological processes, cellular components, and molecular functions. (b) Functional bubble map of gene enrichment. The size of the bubble represents the number of genes in the signaling pathway or the number of genes involved in the function. Color represents P-value; the darker the color the more significant the result. (c) Nodes in the concentric circle graph represent coexpressed genes clustered in specific biological process terms. The inner sectors with larger size and darker color represented more significant enrichment. (d) Ribbons with different colors corresponded to different enriched pathways terms from Metascape. GO, gene ontology.

Table 1: GO analysis of candidate genes.

Ontology	ID	Description
BP	GO: 0006721	Terpenoid metabolic process
BP	GO: 0071276	Cellular response to cadmium ion
BP	GO: 0006720	Isoprenoid metabolic process
BP	GO: 0071280	Cellular response to copper ion
BP	GO: 0010273	Detoxification of copper ion
CC	GO: 0072562	Blood microparticle
CC	GO: 0046930	Pore complex
CC	GO: 0034364	High-density lipoprotein particle
CC	GO: 0034358	Plasma lipoprotein particle
CC	GO: 1990777	Lipoprotein particle
MF	GO: 0008395	Steroid hydroxylase activity
MF	GO: 0016712	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
MF	GO: 0008392	Arachidonic acid epoxygenase activity
MF	GO: 0004497	Monooxygenase activity
MF	GO: 0008391	Arachidonic acid monooxygenase activity

Wilcox test to perform a differential analysis of the key genes between control and treat groups and ggplot2 (version 2.2.1) were used to construct boxplots of gene expression.

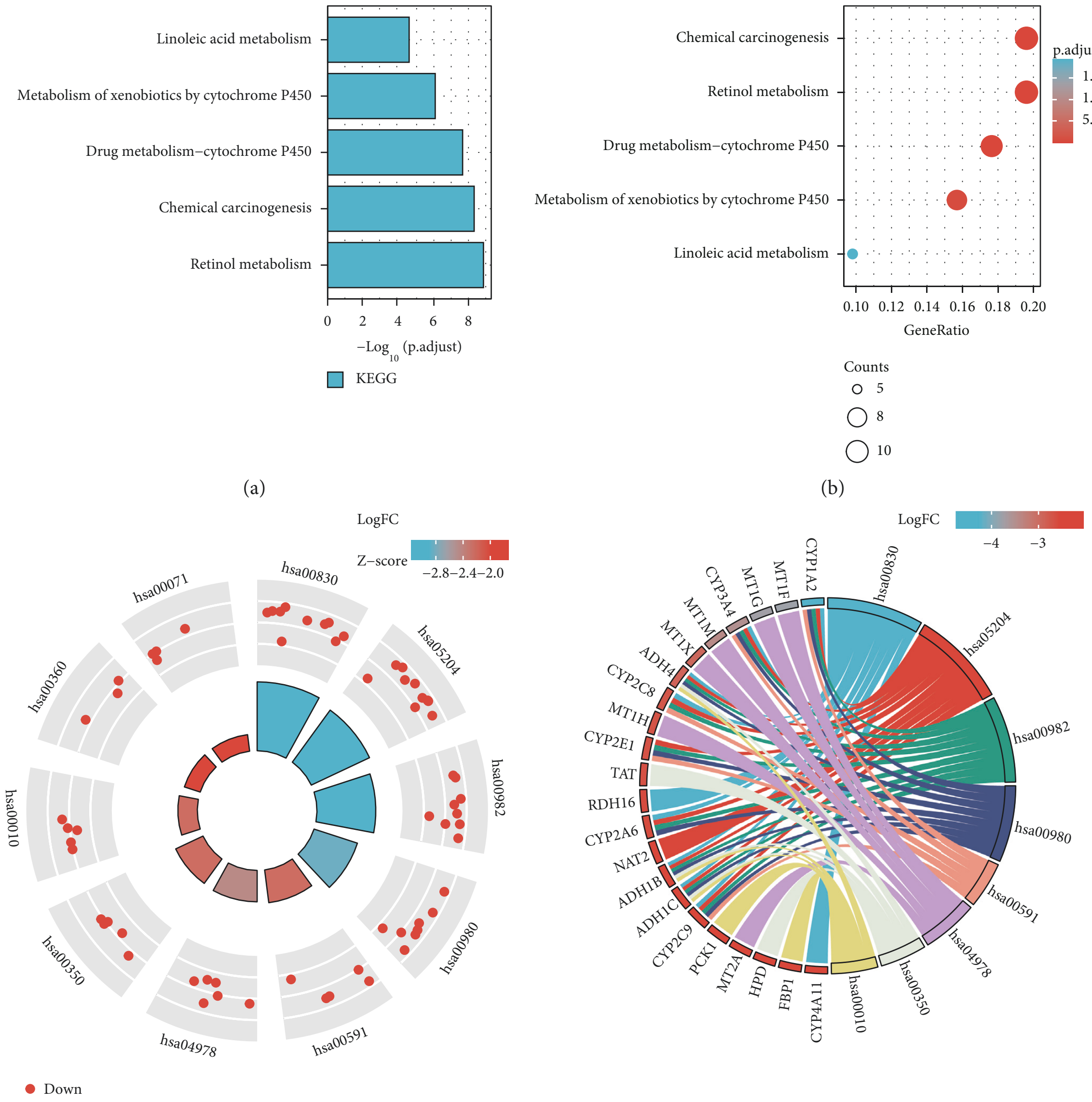
2.1.10. The Establishment of ANN Models and Test. In general, artificial neural networks (ANN) consist of three layers, namely, input, hidden, and output layers. We used key genes scores as input layers. And the hidden layer's main job is to extract categorized data from existing data. The output layer displays the network's ultimate output. The outputs of one layer's nodes are a weighted linear combination that has been altered by a nonlinear function. This nonlinear function enables the neural network to understand complex relationships between independent variables, hence improving the effectiveness of data-driven machine learning techniques [27–29]. In validation set, among the upregulated genes, whose expression is greater than the median value, we marked 1; otherwise mark 0. Among the downregulated genes, whose expression is greater than the median value, we marked 0; otherwise mark 1. Then we get the genes scores so that we can examine the accuracy of the ANN model in the validation set. At last, we plotted ROC curve to calculate area under curve (AUC) to show model accuracy. An area under the curve (AUC) value between 0.8 and 0.9 is considered an excellent classification, while greater than 0.9 is considered as outstanding discrimination [30].

2.1.11. Immune Infiltration by CIBERSORT Analysis. To forecast the infiltration of 22 different kinds of immune cells in each tissue sample, the CIBERSORT algorithm is often utilized [31]. Seven types of T cells [CD8 + T cells, naïve CD4 + T cells, resting memory CD4 + T cells, activated memory CD4 + T cells, follicular helper T cells, regulatory T cells (Tregs), and gamma delta T cells], three types of macrophages (M0, M1, and M2), naïve B cells, memory B cells, plasma cells, resting natural killer (NK) cells, activated NK cells, monocytes, resting dendritic cells, activated dendritic cells, resting mast cells, activated mast cells, eosinophils, and neutrophils are among the 22 immune cells identified. The CIBERSORT method was used to transform a normalized gene expression matrix into 22 different types of immune cell matrix. The immune cell matrix was filtered using P0.05 criteria, and the relative expression of 22 categories of immune cells was determined using R packages between tumor and neighboring nontumor samples. The difference between tumor and neighboring nontumor samples was also determined using principal component analysis (PCA).

2.1.12. Survival Analysis. All the expressions of key genes were calculated, and patients were separated by the median expression level of each gene (highly expressed group and lowly expressed group). The Kaplan–Meier (KM) survival analyses were used to compare the survival difference between lowly and highly expressed groups based on each key gene group, with log-rank test.

3. Results

3.1. Screening the Candidate Genes. Weighted gene coexpression network analysis can be used to screen out the gene modules related to cancer tissues. First, we checked outliers in the sample, which were found and deleted from all samples (Figure 2(a)). The proper power value was then determined. Scale independence reached 0.8 and mean connection was more than zero when the soft threshold power value was equal to 7 (Figure 2(b)). As a result, the soft threshold power value for further analysis was set at 7. And 16 coexpression modules were discovered, with the gray module representing a gene that was not allocated to any module (Figure 2(c)). The genes of each part had been matched with different colors. The eigengene adjacency heatmap was used to identify correlations between different modules (Figure 2(d)). Modules that were grouped together into a single branch may have functionalities that are comparable. As shown in Figure 2(d), the highest connection with liver cancer was seen in the salmon module, which



(a)

(b)

(c)

(d)

Figure 5: KEGG pathway enrichment analyses of candidate genes. (a) KEGG pathway enrichment analysis. (b) Functional bubble map of gene enrichment. The size of the bubble represents the number of genes in the signaling pathway or the number of genes involved in the function. Color represents P-value; the darker the color the more significant the result. (c) Nodes in the concentric circle graph represent coexpressed genes clustered in specific biological process terms. The inner sectors with larger size and darker color represented more significant enrichment. (d) Ribbons with different colors corresponded to different enriched pathways terms from Metascape.

Table 2: KEGG analysis of candidate genes.

Ontology	ID	Description
KEGG	hsa00830	Retinol metabolism
KEGG	hsa05204	Chemical carcinogenesis
KEGG	hsa00982	Drug metabolism, cytochrome P450
KEGG	hsa00980	Metabolism of xenobiotics by cytochrome P450
KEGG	hsa00591	Linoleic acid metabolism
KEGG	hsa04978	Mineral absorption
KEGG	hsa00350	Tyrosine metabolism
KEGG	hsa00010	Glycolysis/gluconeogenesis
KEGG	hsa00360	Phenylalanine metabolism
KEGG	hsa00071	Fatty acid degradation

contained 66 genes. The heatmap of these 66 candidate genes was shown in Figure 3(a), indicating that these 66 candidate genes are differentially expressed between tumor and neighboring nontumor samples. In the heat plot, each cell represents the degree of gene expression, red represents upregulation, and green represents downregulation. We take log $|FC|$ as the horizontal axis and −log10 (adj. P-value) as the vertical axis to make volcano plots (Figure 3(b)), where the red and green dots represent the upregulated and downregulated genes, respectively.

3.2. Functional Enrichment of Candidate Genes.

To explore the functions of candidate genes, the GO and KEGG

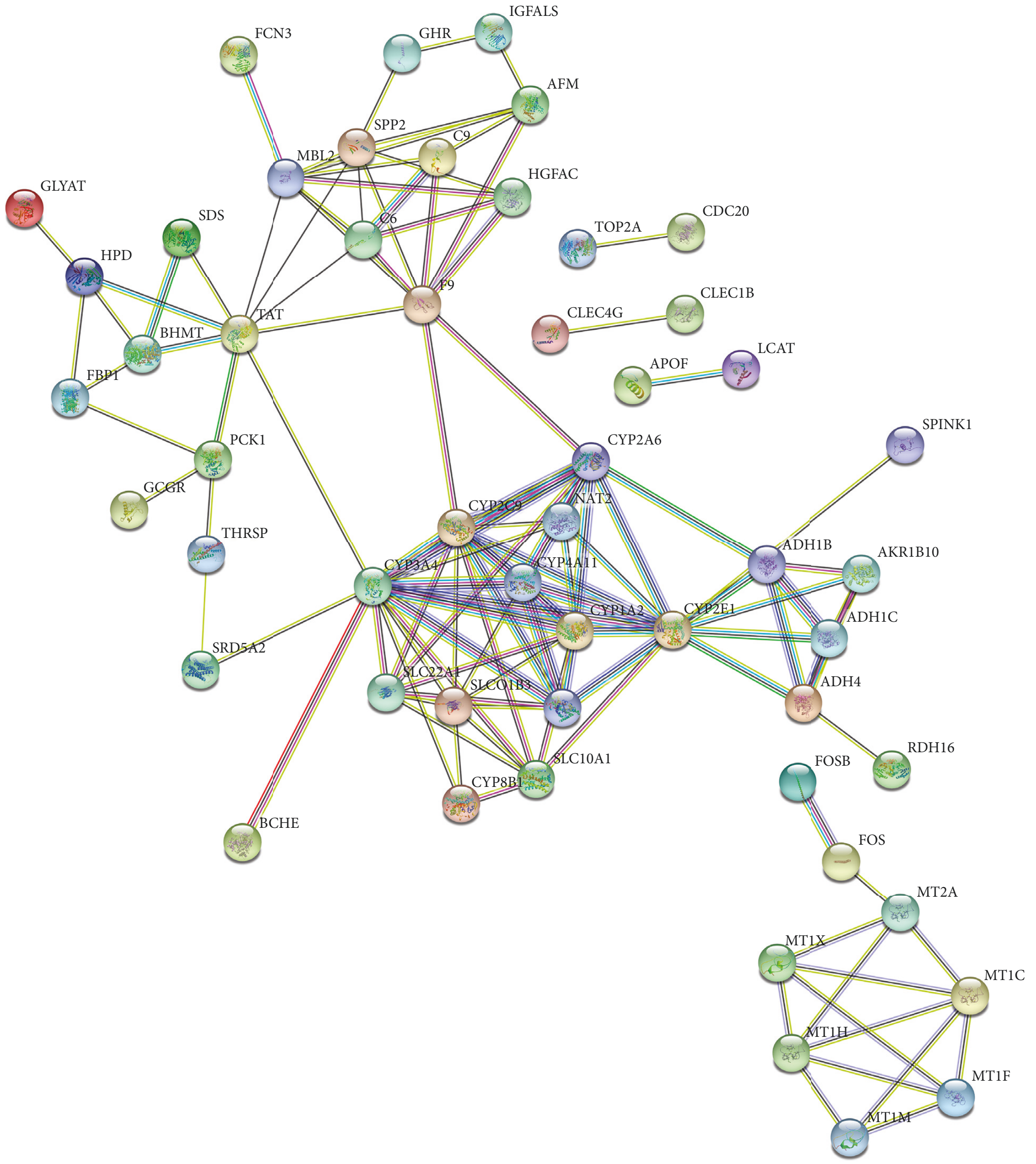


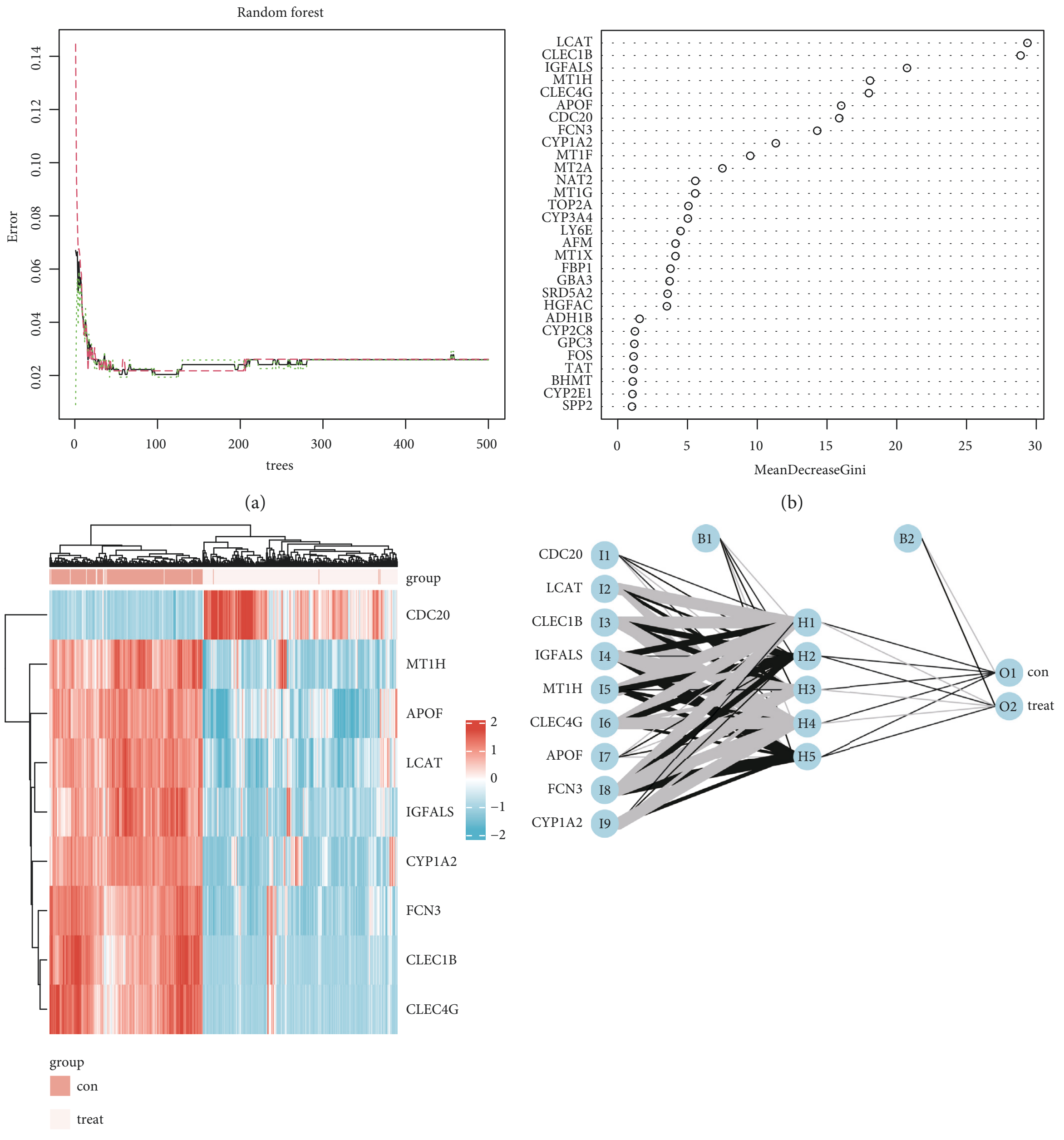
Figure 6: PPI network of the candidate genes.

enrichment analyses were conducted by R packages. GO analysis was applied from 3 aspects: biological process (BP), cellular component (CC), and molecular function (MF). In the BP part, the upregulated robust DEGs were mainly enriched in terpenoid metabolic process, cellular response to cadmium ion, and isoprenoid metabolic process. For CC, the upregulated genes were particularly enriched in blood microparticle, pore complex, and high-density lipoprotein particle. The top three significantly enriched terms were oxidoreductase activity, steroid hydroxylase activity, and aromatase activity in the MF group (Figure 4 and Table 1). The result of KEGG pathway enrichment analysis is shown in Figure 5 and Table 2. Retinol metabolism, chemical carcinogenesis, and drug metabolism-cytochrome P450 were highly associated with tumor progression.

3.3. PPI Network Analysis. To pick out and further understand the key genes, PPI network analysis was performed using STRING. The PPI network of candidate genes is shown in Figure 6. A total of 65 nodes and 118 interaction pairs were included in the network.

Using the STRING online database, a total of 65 nodes and 118 interaction pairs were included in the network.

3.4. Key Genes Screened by Random Forest Tree to Build ANN Model. Figure 7(a) shows the relationship between error rate and the number of classification trees. Genes with mean decrease Gini greater than 10 are identified as the final signature shown in Figure 7(b) (9 key genes). The heatmap of these 9 key genes was shown in Figure 7(c). Then we used these key genes scores to build ANN models, as shown in



(a)

(b)

(c)

(d)

Figure 7: Key genes screened by random forest tree and ANN model. (a) The relationship between error rate and the number of classification trees in random forest. (b) Genes with mean decrease Gini greater than 10 are identified as the final signature. (c) Heat plot of 9 key genes. (d) Artificial neural network structure diagram.

Figure 7(d), which consists of 9 input nodes, 5 hidden nodes, and 2 output nodes. The output layer can judge the properties of the samples and divide them into control groups and case groups.

3.5. Correlation Analysis. The major key genes were utilized to explore the relationships among these genes by a correlation analysis. Most of the genes had previously shown a strong positive correlation, while CDC20 was negatively correlated with the rest (Figure 8).

Blue represents negative correlation and red represents positive correlation. The depth of color indicates the intensity of the correlation between covariates. The darker the color, the higher the correlation.

3.6. Expression Difference Analysis of Key Genes. Among them, CDC20 showed significantly high expression in tumor tissues, while the other eight genes were low expressed in tumor tissues (Figure 9).

3.7. Model Accuracy on Training and Validation Sets. The ROC curve was displayed to validate the predictive accuracy of the model. The area under the receiver-operating

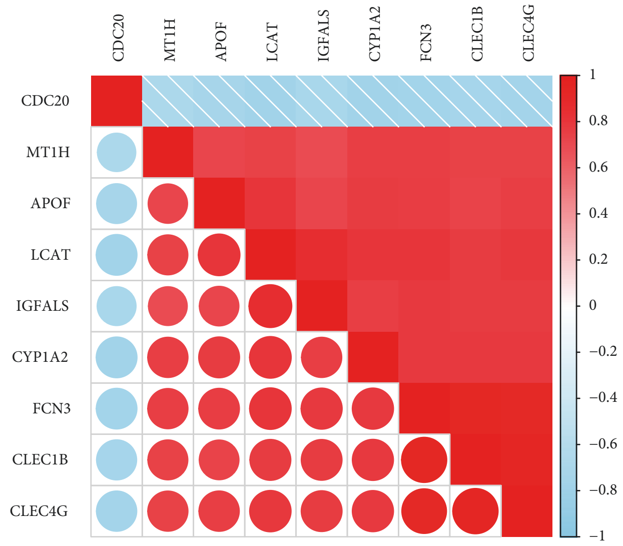
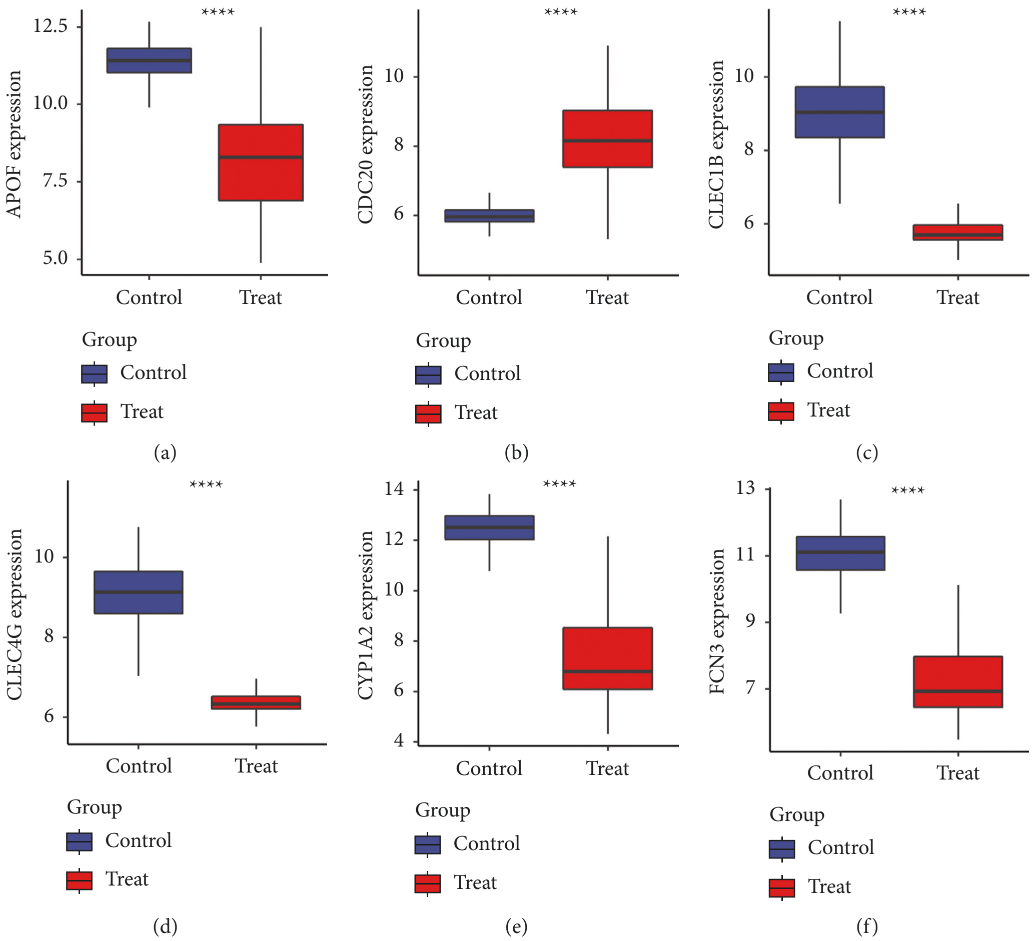


FIGURE 8: Correlation analysis of 9 key genes.



(a)

(b)

(c)

(d)

(e)

(f)

FIGURE 9: Continued.

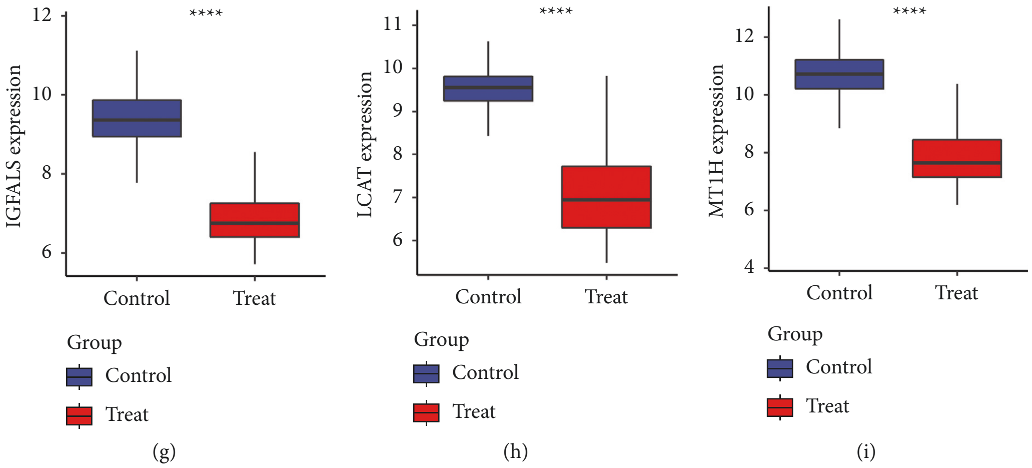


Figure 9: Expression difference analysis of key genes. The boxplot shows the key genes expression level between tumor tissues and adjacent nontumor tissues.

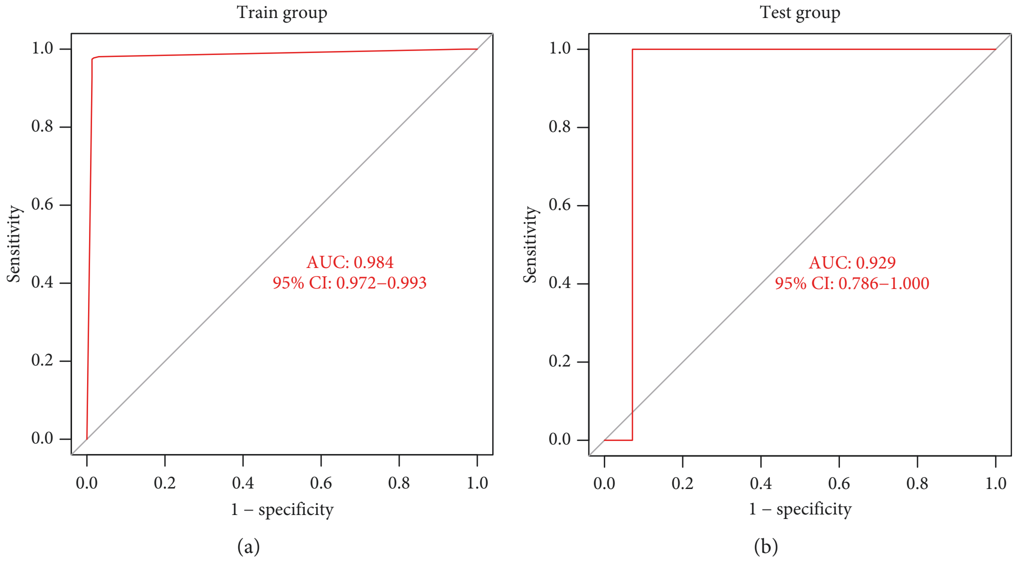


Figure 10: The ROC curve. (a) Train group (GSE36376 and GSE121248). (b) Test group (GSE84402).

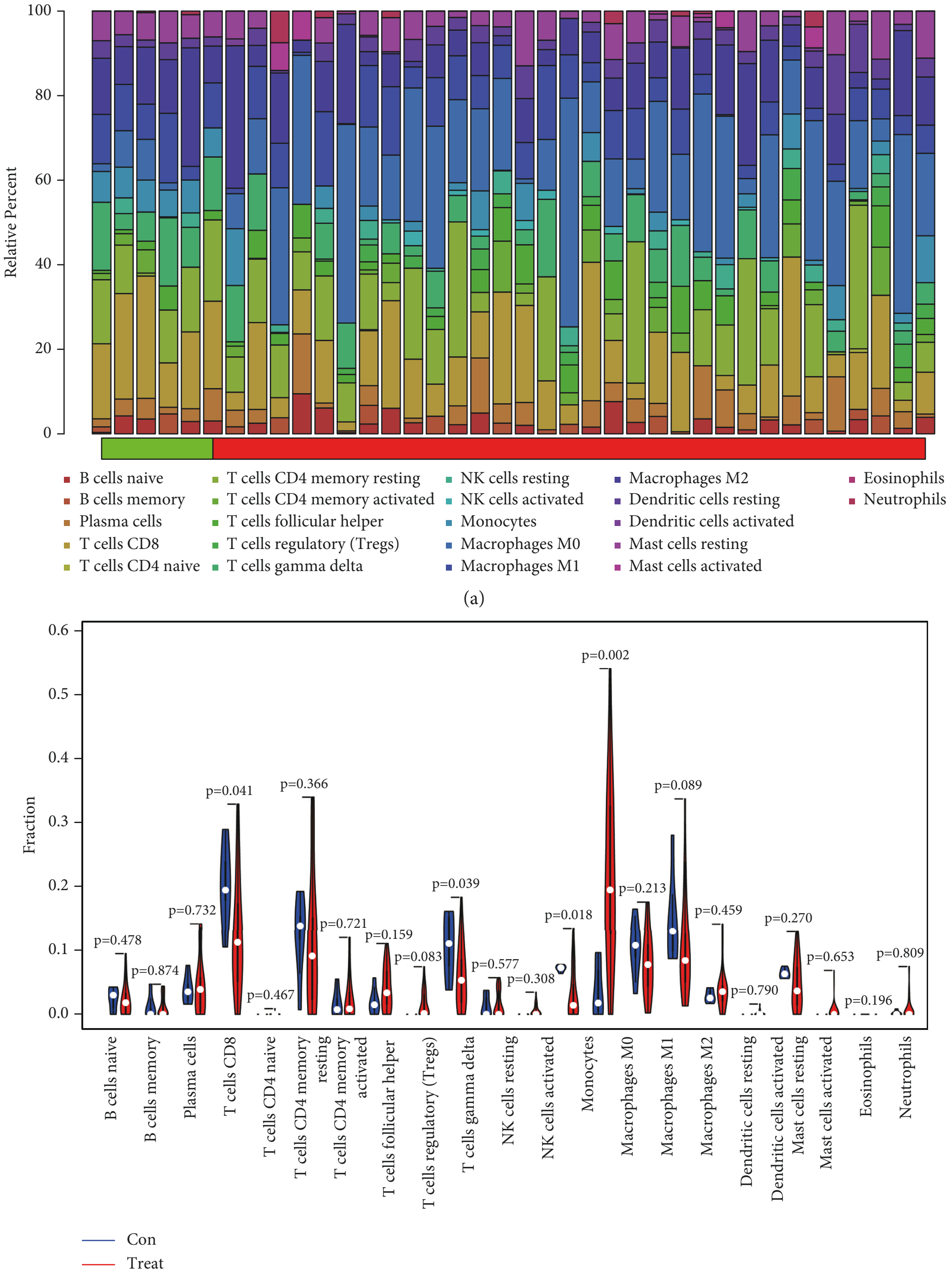
characteristic (ROC) curve (AUC) of the training and validation set was 0.984 and 0.929, respectively (Figure 10).

3.8. Difference of Immune Cell Subsets between Control and Case Groups. Figure 11(a) shows the infiltration of 22 types of immune cells in all samples in the training set using the CIBERSORT algorithm. Macrophages M0 account for a large proportion of case group immune cell infiltration. The changes in immune cells between control and case samples are further examined in Figure 11(b). T cells CD4 memory activated and T cells CD8 showed the strongest positive correlation (Pearson correlation = 0.65), while macrophages M0 and mast cells resting showed the strongest negative correlation (Pearson correlation = −0.63) in the target database at a CIBERSORT $p < 0.05$ (Figure 11(c)).

3.9. Survival Analysis. We explored the prognostic roles of each key gene by KM curves with the log-rank test. APOF (log-rank $p = 0.00063$), CDC20 (log-rank $p < 0.0001$), CLEC1B (log-rank $p = 0.0014$), CLEC4G (log-rank $p = 0.0095$), CYP1A2 (log-rank $p = 0.1$), FCN3 (log-rank $p = 0.0011$), IGFALS (log-rank $p = 0.00072$), LCAT (log-rank $p < 0.0001$), and MT1H (log-rank $p = 0.051$) were identified as prognostic genes in the TCGA database (Figure 12).

4. Discussion

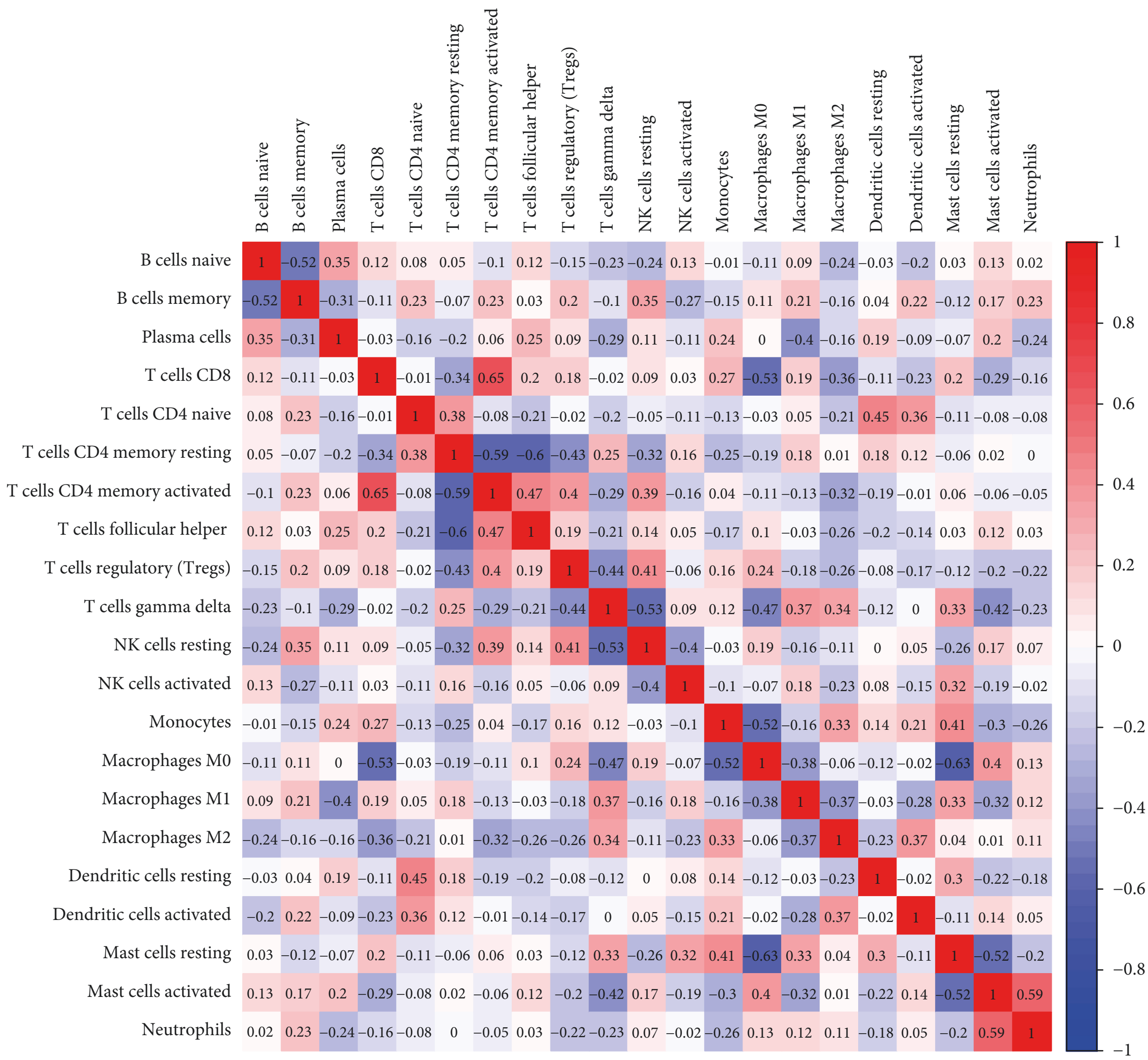
This work innovatively combined comprehensive biological information analysis and artificial neural network (ANN) to classify hepatocellular carcinoma tissues and the corresponding noncancerous tissues. In this study, WGCNA was performed to reveal key modules with clinical significance,



(a)

(b)

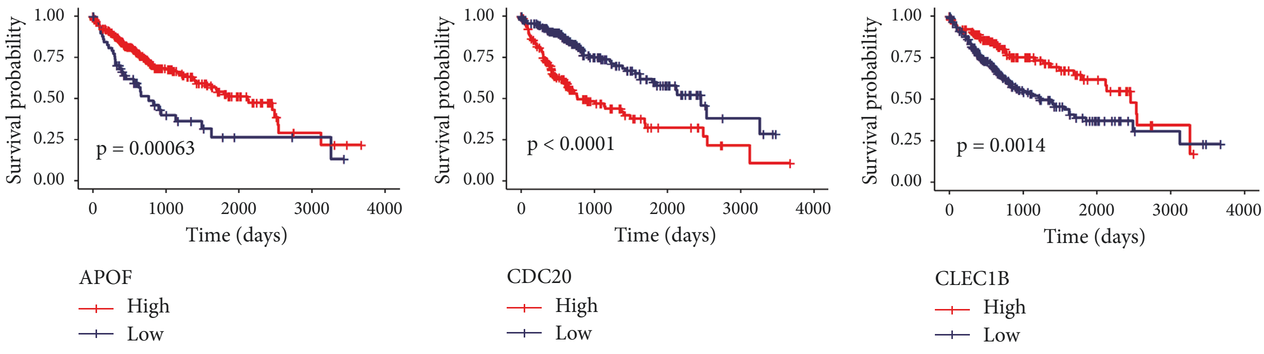
FIGURE 11: Continued.



	B cells naive	B cells memory	Plasma cells	T cells CD8	T cells CD4 naive	T cells CD4 memory resting	T cells CD4 memory activated	T cells follicular helper	T cells regulatory (Tregs)	T cells gamma delta	NK cells resting	NK cells activated	Monocytes	Macrophages M0	Macrophages M1	Macrophages M2	Dendritic cells resting	Dendritic cells activated	Mast cells resting	Mast cells activated	Neutrophils
B cells naive	1	−0.52	0.35	0.12	0.08	0.05	−0.1	0.12	−0.15	−0.23	−0.24	0.13	−0.01	−0.11	0.09	−0.24	−0.03	−0.2	0.03	0.13	0.02
B cells memory	−0.52	1	−0.31	−0.11	0.23	−0.07	0.23	0.03	0.2	−0.1	0.35	−0.27	−0.15	0.11	0.21	−0.16	0.04	0.22	−0.12	0.17	0.23
Plasma cells	0.35	−0.31	1	−0.03	−0.16	−0.2	0.06	0.25	0.09	−0.29	0.11	−0.11	0.24	0	−0.4	−0.16	0.19	−0.09	−0.07	0.2	−0.24
T cells CD8	0.12	−0.11	−0.03	1	−0.01	−0.34	0.65	0.2	0.18	−0.02	0.09	0.03	0.27	−0.53	0.19	−0.36	−0.11	−0.23	0.2	−0.29	−0.16
T cells CD4 naive	0.08	0.23	−0.16	−0.01	1	0.38	−0.08	−0.21	−0.02	−0.2	−0.05	−0.11	−0.13	−0.03	0.05	−0.21	0.45	0.36	−0.11	−0.08	−0.08
T cells CD4 memory resting	0.05	−0.07	−0.2	−0.34	0.38	1	−0.59	−0.6	−0.43	0.25	−0.32	0.16	−0.25	−0.19	0.18	0.01	0.18	0.12	−0.06	0.02	0
T cells CD4 memory activated	−0.1	0.23	0.06	0.65	−0.08	−0.59	1	0.47	0.4	−0.29	0.39	−0.16	0.04	−0.11	−0.13	−0.32	−0.19	−0.01	0.06	−0.06	−0.05
T cells follicular helper	0.12	0.03	0.25	0.2	−0.21	−0.6	0.47	1	0.19	−0.21	0.14	0.05	−0.17	0.1	−0.03	−0.26	−0.2	−0.14	0.03	0.12	0.03
T cells regulatory (Tregs)	−0.15	0.2	0.09	0.18	−0.02	−0.43	0.4	0.19	1	−0.44	0.41	−0.06	0.16	0.24	−0.18	−0.26	−0.08	−0.17	−0.12	−0.2	−0.22
T cells gamma delta	−0.23	−0.1	−0.29	−0.02	−0.2	0.25	−0.29	−0.21	−0.44	1	−0.53	0.09	0.12	−0.47	0.37	0.34	−0.12	0	0.33	−0.42	−0.23
NK cells resting	−0.24	0.35	0.11	0.09	−0.05	−0.32	0.39	0.14	0.41	−0.53	1	−0.4	−0.03	0.19	−0.16	−0.11	0	0.05	−0.26	0.17	0.07
NK cells activated	0.13	−0.27	−0.11	0.03	−0.11	0.16	−0.16	0.05	−0.06	0.09	−0.4	1	−0.1	−0.07	0.18	−0.23	0.08	−0.15	0.32	−0.19	−0.02
Monocytes	−0.01	−0.15	0.24	0.27	−0.13	−0.25	0.04	−0.17	0.16	0.12	−0.03	−0.1	1	−0.52	−0.16	0.33	0.14	0.21	0.41	−0.3	−0.26
Macrophages M0	−0.11	0.11	0	−0.53	−0.03	−0.19	−0.11	0.1	0.24	−0.47	0.19	−0.07	−0.52	1	−0.38	−0.06	−0.12	−0.02	−0.63	0.4	0.13
Macrophages M1	0.09	0.21	−0.4	0.19	0.05	0.18	−0.13	−0.03	−0.18	0.37	−0.16	0.18	−0.16	−0.38	1	−0.37	−0.03	−0.28	0.33	−0.32	0.12
Macrophages M2	−0.24	−0.16	−0.16	−0.36	−0.21	0.01	−0.32	−0.26	−0.26	0.34	−0.11	−0.23	0.33	−0.06	−0.37	1	−0.23	0.37	0.04	0.01	0.11
Dendritic cells resting	−0.03	0.04	0.19	−0.11	0.45	0.18	−0.19	−0.2	−0.08	−0.12	0	0.08	0.14	−0.12	−0.03	−0.23	1	−0.02	0.3	−0.22	−0.18
Dendritic cells activated	−0.2	0.22	−0.09	−0.23	0.36	0.12	−0.01	−0.14	−0.17	0	0.05	−0.15	0.21	−0.02	−0.28	0.37	−0.02	1	−0.11	0.14	0.05
Mast cells resting	0.03	−0.12	−0.07	0.2	−0.11	−0.06	0.06	0.03	−0.12	0.33	−0.26	0.32	0.41	−0.63	0.33	0.04	0.3	−0.11	1	−0.52	−0.2
Mast cells activated	0.13	0.17	0.2	−0.29	−0.08	0.02	−0.06	0.12	−0.2	−0.42	0.17	−0.19	−0.3	0.4	−0.32	0.01	−0.22	0.14	−0.52	1	0.59
Neutrophils	0.02	0.23	−0.24	−0.16	−0.08	0	−0.05	0.03	−0.22	−0.23	0.07	−0.02	−0.26	0.13	0.12	0.11	−0.18	0.05	−0.2	0.59	1

(c)

Figure 11: Immune cells infiltration analysis. (a) The distribution of 22 types of immune cells between tumor tissues and adjacent nontumor tissues. (b) Violin plot visualizing the differentially infiltrated immune cells ($P < 0.05$). (c) The difference of immune cells infiltration between tumor tissues and adjacent nontumor tissues visualized by heatmap.



(a) (b) (c)

Figure 12: Continued.

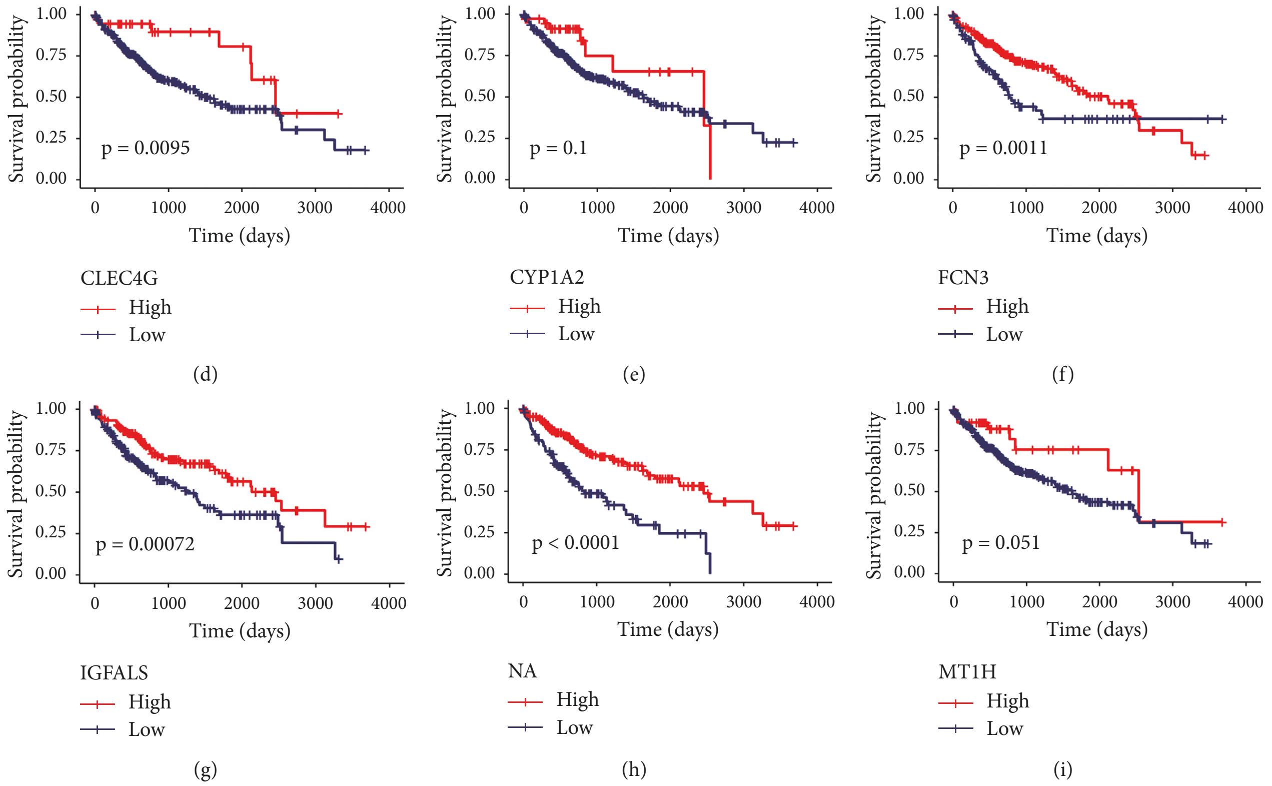


Figure 12: Survival analysis. Gene changes of APOF (a), CDC20 (b), CLEC1B (c), CLEC4G (d), CYP1A2 (e), FCN3 (f), IGFALS (g), LCAT (h), and MT1H (i) were significantly correlated with the overall survival of HCC patients.

and salmon module was screened out through preservation evaluation. The GO and KEGG analyses revealed that the genes in salmon module were significantly enriched in the biological processes of the cell cycle, cell division, and liver-related functions. All these biological functions are closely related to liver cancer.

The basic unit of the ANN is neuron. To get better performance, the weight and bias of each neuron were constantly updated during training. Classification results of ANN indicated that the average accuracy is 0.929 in validation set which showed that the model in this paper is highly accurate.

Through immune cell infiltration analysis, we compared the infiltration of immune cells in tumor and corresponding noncancerous samples. The results show that a high proportion of macrophages M0 infiltration existed in tumor samples. Finally, survival analysis of hub genes based on the TCGA database was also performed in our study. The results imply that these key genes are potentially associated with the prognosis of HCC as well.

This model can be applied to the early diagnosis of cancer. In this method, probes are firstly used to measure gene expression, and then deep learning methods are used to classify cancer samples. There is no instrument contact during the whole diagnosis process, so there is no risk of radiation compared with CT, ultrasonic imaging, X-ray examination, and PET. In addition, CT and MRI scans for individuals with HCC have a number of limitations. Contrast-related allergies, respiratory movements, renal impairment, and cumulative radiation doses are all problems to consider when getting a CT scan, especially in young patients. Ultrasonography (US) is frequently used for screening, sometimes almost entirely, and other times in combination with CT or MRI, depending on the particular patient's risk factors, doctor preference, and advanced imaging approved by health insurance. If a lesion is discovered on screening US, a contrast-enhanced scan is necessary to better characterize the lesion [32]. Despite the use of dynamic contrast-enhanced MRI (DCE-MRI), the imaging diagnosis of HCC is difficult due to atypical imaging presentations and the variety of liver tumors [33]. Other diagnostic approaches, such as α-fetoprotein, are expensive and lack sensitivity in HCC detection [34]. This article's categorization is computer-assisted, and while pathological diagnosis necessitates manual operation throughout the procedure, this technique is more suited for quick diagnosis.

In addition, tumor individualized medical care is becoming increasingly important, relying on accurate risk stratification systems. These systems aid in the selection of the most appropriate therapy and the evaluation of the treatment's effectiveness. Gene sequencing is useful for predicting therapy effects and assisting doctors in selecting the best customized treatment approach for patients.

At the same time, preventing early recurrence is an important issue in the management of HCC. Early recurrence, defined as recurrence within 1-2 years after resection or ablation, is a significant predictor of poor prognosis in HCC patients [33]. By detecting recurrent lesions early, the first-line treatment for recurrences can be initiated early. The diagnostic model based on ANN and random forest analysis can distinguish the tumor tissue from the normal adjacent tissue, so as to predict early recurrence after surgery or curative ablation in HCC patients.

A vast amount of single-cell sequencing data will need to be investigated in the future. Classification and clustering challenges are common research subjects. We can actually construct small-scale quick diagnosis equipment for cancer with the advancement of sequencing data and deep learning. It gives people the chance to avoid and treat cancer.

Data Availability

The expression profiling by array data used to support the findings of this study has been deposited in the GEO repository (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE36376, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE121248, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84402).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Ziyi Cao contributed to the supervision of this study.

References

[1] A. B. El-Khoueiry, B. Sangro, T. Yau et al., "Nivolumab in patients with advanced hepatocellular carcinoma (CheckMate 040): an open-label, non-comparative, phase 1/2 dose escalation and expansion trial," *The Lancet*, vol. 389, no. 1008, pp. 2492–2502, 2017.

[2] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-Tieulent, and A. Jemal, "Global cancer statistics, 2012," *CA: A Cancer Journal for Clinicians*, vol. 65, no. 2, pp. 87–108, 2012.

[3] J. D. Yang, P. Hainaut, G. J. Gores, A. Amadou, A. Plymoth, and L. R. Roberts, "A global view of hepatocellular carcinoma: trends, risk, prevention and management," *Nature Reviews Gastroenterology & Hepatology*, vol. 16, no. 10, pp. 589–604, 2019.

[4] A. Rossetto, V. De Re, A. Steffan et al., "Carcinogenesis and metastasis in liver: cell physiological basis," *Cancers*, vol. 11, no. 11, 1731 pages, 2019.

[5] C. Fitzmaurice, T. F. Akinyemiju, L. F. H. Al et al., J. J. Choi, D. J. Christopher, S. C. Chung et al., S. I. Hay, B. Heibati, M. K. Hiluf et al., Global burden of disease cancer collaboration. 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," vol. 3, no. 4, pp. 524–548, 2017.

[6] J. D. Yang, W. R. Kim, K. W. Park et al., "Model to estimate survival in ambulatory patients with hepatocellular carcinoma," *Hepatology*, vol. 56, no. 2, pp. 614–621, 2012.

[7] M. Omata, A. L. Cheng, N. Kokudo et al., "Asia-Pacific clinical practice guidelines on the management of hepatocellular carcinoma: a 2017 update," *Hepatology International*, vol. 11, no. 4, pp. 317–370, 2017.

[8] J. K. Heimbach, L. M. Kulik, R. S. Finn et al., "AASLD guidelines for the treatment of hepatocellular carcinoma," *Hepatology*, vol. 67, no. 1, pp. 358–380, 2018.

[9] P. R. Galle, A. Forner, J. M. Llovet et al., "EASL clinical practice guidelines: management of hepatocellular carcinoma," *Journal of Hepatology*, vol. 69, no. 1, pp. 182–236, 2018.

[10] V. L. Chen, A. G. Singal, E. B. Tapper, and N. D. Parikh, "Hepatocellular carcinoma surveillance, early detection and survival in a privately insured US cohort," *Liver International*, vol. 40, no. 4, pp. 947–955, 2020.

[11] K. Tzartzeva, J. Obi, N. E. Rich et al., "Surveillance imaging and alpha fetoprotein for early detection of hepatocellular carcinoma in patients with cirrhosis: a meta-analysis," *Gastroenterology*, vol. 154, no. 6, pp. 1706–1718, 2018.

[12] Y. J. Lee, J. M. Lee, J. S. Lee et al., "Hepatocellular carcinoma: diagnostic performance of multidetector CT and MR imaging-a systematic review and meta-analysis," *Radiology*, vol. 275, no. 1, pp. 97–109, 2015.

[13] C. Li, X. Zeng, H. Yu, Y. Gu, and W. Zhang, "Identification of hub genes with diagnostic values in pancreatic cancer by bioinformatics analyses and supervised learning methods," *World Journal of Surgical Oncology*, vol. 16, no. 1, 223 pages, 2018, https://doi.org/10.1186/s12957-018-1519-y.

[14] X. Zeng, C. Li, Y. Li et al., "A network-based variable selection approach for identification of modules and biomarker genes associated with end-stage kidney disease," *Nephrology*, vol. 25, no. 10, pp. 775–784, 2020.

[15] H. A. Afan, M. F. Allawi, A. El-Shafie et al., "Input attributes optimization using the feasibility of genetic nature inspired algorithm: application of river flow forecasting," *Scientific Reports*, vol. 10, no. 1, p. 4684, 2020.

[16] S. Wang and R. M. Summers, "Machine learning and radiology," *Medical Image Analysis*, vol. 16, no. 5, pp. 933–951, 2012, https://doi.org/10.1016/j.media.2012.02.005.

[17] M. Sato, R. Tateishi, Y. Yatomi, and K. Koike, "Artificial intelligence in the diagnosis and management of hepatocellular carcinoma," *Journal of Gastroenterology and Hepatology*, vol. 36, no. 3, pp. 551–560, 2021.

[18] D. Joksas, P. Freitas, Z. Chai et al., "Committee machines-a universal method to deal with non-idealities in memristor-based neural networks," *Nature Communications*, vol. 11, no. 1, p. 4273, 2020.

[19] H. Kahramanli and N. Allahverdi, "Design of a hybrid system for the diabetes and heart diseases," *Expert Systems with Applications*, vol. 35, no. 1-2, pp. 82–89, 2008.

[20] D. Jia, C. Chen, C. Chen et al., "Breast cancer case identification based on deep learning and bioinformatics analysis," *Frontiers in Genetics*, vol. 12, Article ID 628136, 2021.

[21] M. E. Ritchie, B. Phipson, D. Wu et al., "Limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Research*, vol. 43, no. 7, e47 pages, 2015.

[22] P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis," *BMC Bioinformatics*, vol. 9, no. 1, 559 pages, 2008.

[23] L. Chen, L. Yuan, Y. Wang et al., "Co-expression network analysis identified FCER1G in association with progression and prognosis in human clear cell renal cell carcinoma," *International Journal of Biological Sciences*, vol. 13, no. 11, pp. 1361–1372, 2017.

[24] G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He, "ClusterProfiler: an R package for comparing biological themes among gene

clusters," *OMICS: A Journal of Integrative Biology*, vol. 16, no. 5, pp. 284–287, 2012.

[25] D. Szklarczyk, A. Franceschini, S. Wyder et al., "STRING v10: protein-protein interaction networks, integrated over the tree of life," *Nucleic Acids Research*, vol. 43, pp. D447–52, 2015.

[26] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, pp. 18–22, 2002.

[27] D. Kim, S. You, S. So et al., "A data-driven artificial intelligence model for remote triage in the prehospital environment," *PLoS One*, vol. 13, no. 10, 2018.

[28] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[29] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.

[30] L. Stanley, W. David, J. R. Hosmer, and X. S. Rodney, *Applied Logistic Regression*, Wiley, Hoboken, NJ, USA, 2nd edition, 2000.

[31] B. Chen, M. S. Khodadoust, C. L. Liu, A. M. Newman, and A. A. Alizadeh, "Profiling tumor infiltrating immune cells with CIBERSORT," *Methods in Molecular Biology*, vol. 1711, pp. 243–259, 2018.

[32] K. L. McGillen, S. Zaidi, A. Ahmed, S. Harter, and N. S. Yee, "Contrast-enhanced ultrasonography for screening and diagnosis of hepatocellular carcinoma: a case series and review of the literature," *Medicine*, vol. 7, no. 9, 51 pages, 2020.

[33] B. Feng, X. H. Ma, S. Wang, W. Cai, X. B. Liu, and X. M. Zhao, "Application of artificialIntelligence in preoperative imaging of hepatocellular carcinoma: current status and future perspectives," *World Journal of Gastroenterology*, vol. 27, no. 32, pp. 5341–5350, 2021.

[34] X. Wang, A. Zhang, and H. Sun, "Power of metabolomics in diagnosis and biomarker discovery of hepatocellular carcinoma," *Hepatology*, vol. 57, no. 5, pp. 2072–2077, 2013.