Biochemistry and Biophysics Reports 37 (2024) 101587

Contents lists available at ScienceDirect

Biochemistry and Biophysics Reports

journal homepage: www.elsevier.com/locate/bbrep

Bioinformatics and machine learning driven key genes screening for hepatocellular carcinoma

Ye Shen [a], Juanjie Huang [c], Lei Jia [d], Chi Zhang [e,*], Jianxing Xu [a,b,**]

[a] Department of Radiology, Wujin Hospital Affiliated with Jiangsu University, Changzhou, 213002, China
[b] Department of Radiology, The Wujin Clinical College of Xuzhou Medical University, Changzhou, 213002, China
[c] Department of General Surgery, Dongguan Qingxi Hospital, Dongguan, 523660, China
[d] International Health Medicine Innovation Center, Shenzhen University, ShenZhen, 518060, China
[e] Huaxia Eye Hospital of Foshan, Huaxia Eye Hospital Group, Foshan, Guangdong, 528000, China

ARTICLE INFO

Keywords:
Hepatocellular carcinoma cells
Differentially expressed genes
Biomarkers
Machine learning

ABSTRACT

Liver cancer, a global menace, ranked as the sixth most prevalent and third deadliest cancer in 2020. The challenge of early diagnosis and treatment, especially for hepatocellular carcinoma (HCC), persists due to late-stage detections. Understanding HCC's complex pathogenesis is vital for advancing diagnostics and therapies. This study combines bioinformatics and machine learning, examining HCC comprehensively. Three datasets underwent meticulous scrutiny, employing various analytical tools such as Gene Ontology (GO) function and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis, protein interaction assessment, and survival analysis. These rigorous investigations uncovered twelve pivotal genes intricately linked with HCC's pathophysiological intricacies. Among them, CYP2C8, CYP2C9, EPHX2, and ESR1 were significantly positively correlated with overall patient survival, while AKR1B10 and NQO1 displayed a negative correlation. Moreover, the Adaboost prediction model yielded an 86.8 % accuracy, showcasing machine learning's potential in deciphering complex dataset patterns for clinically relevant predictions. These findings promise to contribute valuable insights into the elusive mechanisms driving liver cancer (HCC). They hold the potential to guide the development of more precise diagnostic methods and treatment strategies in the future. In the fight against this global health challenge, unraveling HCC's intricacies is of paramount importance.

1. Introduction

Liver cancer and is becoming a widespread malignancy globally, which ranks as the sixth most prevalent type of cancer and the third deadliest worldwide in 2020 [1,2]. A significant number of patients with liver cancer are diagnosed at an advanced stage, limiting the treatment options and causing a high rate of recurrence. Only a small portion of these patients, less than 20 %, are eligible for surgical interventions such as resection or transplantation [3,4].

Liver malignancies are divided into two categories: primary and secondary. The most frequent form of primary liver cancer is hepatocellular carcinoma (HCC), which originates from hepatocytes [5]. Another type of primary liver malignancy is intrahepatic cholangiocarcinoma (ICC) and HCC that arises from bile duct epithelial cells [6]. Contributors to the development of HCC include common viral hepatitis infections, alcohol consumption, the fungal metabolite aflatoxin B1, liver flukes, autoimmune liver disease, non-alcoholic fatty liver, and metabolic syndrome [6]. The treatment options for HCC are limited, including surgical removal, interventional therapy, and liver transplantation. Despite being the most commonly used treatment, surgery has not significantly improved treatment outcomes. The pathogenesis of HCC is still not well understood due to its complexity and high variability, making exploration of its pathogenesis and the discovery of specific biomarkers and therapeutic targets crucial for improving HCC diagnosis and treatment. In this context, the utilization of nanomaterials in biosensor and medicine has garnered substantial attention due to their unique physicochemical properties and versatile applications. Recent advances in nanotechnology have led to the development of nanomaterial-based biosensors that enable the sensitive and specific detection of biomolecules, contributing to the early

* Corresponding author.
** Corresponding author. Department of Radiology, Wujin Hospital Affiliated with Jiangsu University, Changzhou, 213002, China.
E-mail addresses: mikezhang1980@outlook.com (C. Zhang), wjyycgz@126.com (J. Xu).

https://doi.org/10.1016/j.bbrep.2023.101587
Received 23 July 2023; Received in revised form 1 November 2023; Accepted 17 November 2023
Available online 25 November 2023

diagnosis of diseases including cancer [7–9].

Furthermore, the continued advancement of the Human Genome Project, the creation of the Cancer Genome Atlas, and the emergence of high-throughput technology and big data platforms have fueled the utilization of bioinformatics in the realm of tumor gene expression profiling, particularly in the fields of carcinogenesis and mechanism of progression. This allows for rapid analysis of large data sets and a more comprehensive understanding of differentially expressed genes in HCC progression. While various studies have explored the genetic landscape of HCC, the integration of bioinformatics techniques and machine learning holds the promise of unraveling novel insights into the molecular mechanisms driving HCC progression. Recent studies have shown that the potential for hepato-carcinogenesis can be predicted by evaluating genes with distinct expression profiles [10–14]. In light of the need for precise tools to combat HCC, cutting-edge approaches in bioinformatics and machine learning were introduced to advance our understanding of HCC at the molecular level [15–19]. Through the integration of multidisciplinary techniques, studies were strived to contribute to the development of effective strategies for the early detection and management of HCC. By analyzing gene expression profiles from diverse datasets, key differentially expressed genes associated with HCC could be identified, shedding light on potential therapeutic targets and diagnostic markers. Jia et al., for example, trained non-tumor liver tissue next to a tumor cancer as a normal sample and used it to derive a marker [20]. Ao et al. obtained an early diagnostic marker for HCC consisting of 19 gene pairs based on relative expression orderings (REO)-based strategy [21,22].With the development of sequencing techniques, next generation sequencing (NGS) technology [23,24] and single-cell RNA sequencing (scRNA-seq) were applied in the prediction hub-genes of HCC [22]. However, the high cost is one of the factors affecting the widely application of these strategies. Hence, mRNA expression profiles combined with machine learning approaches is still an effective strategies.

In this article, we present an exploration into the intricate landscape of HCC through the integration of bioinformatics techniques and machine learning. We introduced three comprehensive datasets of HCC and subjected them to meticulous screening using advanced computational methodologies. Concurrently, prediction models for HCC were meticulously crafted using machine learning algorithms, offering a novel approach to unravel the complex mechanisms underlying this formidable disease. Beginning with the systematic analysis of gene expression profiles across the datasets, we employed a series of analytical tools to decode the underlying biology. The culmination of this comprehensive analysis included GO function enrichment analysis, KEGG pathway enrichment analysis, protein interaction analysis, and survival analysis. These endeavors collectively unveiled an ensemble of twelve key genes intricately linked to the intricate tapestry of HCC. Central to our study was the implementation of the Adaboost prediction model, which demonstrated an exceptional accuracy rate of 86.8 %. This accomplishment underscored the potential of machine learning algorithms to decipher intricate patterns within complex datasets, thereby facilitating accurate predictions in a clinically significant context.

2. Materials and methods

2.1. Data collection

In this study, the datasets are selected by considering factors including tissue type, disease stage, sample size, clinical information, data quality, and platform compatibility. Relevant datasets with larger sample sizes, high-quality data, and detailed clinical information are prioritized based on research objectives. Hence, we selected high quality data sets with samples size are greater than 100, and the datasets with not too significant difference in positive and negative sample numbers. As a result, the mRNA expression profiles of three data sets GSE54236, GSE121248 and GSE164760 were downloaded from the GEO database. Totally, 428 patients were collected. GSE54236 belongs to the platform GPL6480, GSE121248 belongs to the platform GPL570, and GSE164760 belongs to the platform GPL13667. The linear regression model software package Limma was used to calculate the differences of different groups of chips and normalize them.

2.2. Differential expressed genes

Using R language to screen out differentially expressed genes with $|\log2FC|>1$ and P.Value < 0.05 as the standard. Through Wayne analysis, cross differentially expressed genes were screened out from these two data sets [25,26].

2.3. Gene ontology and KEGG analysis

Gene ontology(GO) function analysis generally includes biological process (Biological Process, BP), molecular function (Molecular Function, MF), and cellular component (Cellular Component, CC) [27]. In this study, the differentially expressed genes were imported into the online tool DAVID database (https://david.ncifcrf.gov/). Then, Official Gene Symbol was selected as gene identifier. Third, GO and KEGG enrichment analysis was performed respectively, and the obtained data were used to draw bubble charts [28,29].

2.4. Protein-protein interaction network

Use the online database String to construct a network diagram of protein-protein interactions between differentially expressed genes and analyze the interactions between protein functions. Then use Cytoscape software (3.8.0) online tool to visualize the PPI network module [30,31].

2.5. Survival analysis

After screening the target genes using PPI network analysis, the online tool GEPIA database was used to analyze the relationship between the expression level of target genes and OS in HCC patients with HCC patients with detailed clinical data in the TCGA database. Verify its participation in the process of HCC tumorigenesis.

2.6. Model prediction

In this study, six machine learning algorithms including AdaBoost [32–36], C4.5 [37–40], Random Tree [41–43], Random Forest [44–51], K-Nearest Neighbors (KNN) [52–54], and Bayesian Net [49,55–59] widely employed in the field of medicine science [60–62], life science [63] and food science [64–67]were applied to build prediction models for HCC. All the computation were performed by Software WEKA 3.7.

2.6.1. AdaBoost

Adaboost is a collective term for bagging and adaboosting, which was proposed by Freund and Schapire [68]. The main idea is that in machine learning, weak learning algorithms are equivalent to strong learning algorithms as long as you can find a weak learning algorithm that is slightly better than random guessing. Then, it could be boosted into a strong learning algorithm without the need to directly find strong learning algorithms, which are usually hard to obtain in most cases [69,70].

2.6.2. C4.5

C4.5 is a classic decision tree learning algorithm for building classification models [71,72].C4.5 can divide the training set into different subsets based on different features and feature values. It then selects the best features for node splitting by calculating the information gain. Next, C4.5 recursively builds a decision tree, starting from the root node and selecting features with the highest information gain for splitting,

creating child nodes until the stopping condition is satisfied. Finally, the decision tree generated by C4.5 can be used to categorize new data samples by traversing the branches and nodes of the tree to determine the class to which the sample belongs [73].

2.6.3. Random Tree

Random tree is a machine learning algorithm which is a variant of decision tree [74]. In Random Tree, a decision tree is constructed by introducing randomness to improve the generalization performance of the model and resist overfitting. The main feature of Random Tree is the possibility of random feature selection and random sample selection. Instead of using the entire training dataset in constructing each decision tree, a portion of samples from the training data is randomly selected for training [75]. This random sampling helps to introduce diversity and improves the generalization performance of the model. During each node split, instead of using all available features to select the best split feature, a subset of features from the feature set is randomly selected for the split. This helps reduce the risk of overfitting as there is no over-reliance on certain features [75].

2.6.4. Random forest

Random Forest is a powerful integrated machine learning algorithm that improves prediction performance by constructing multiple decision trees [76].Each decision tree is trained on a different subset of data, which is generated by random sampling. In addition, when splitting the nodes of each decision tree, instead of considering all available features, Random Forest randomly selects a subset from the feature set for splitting. This introduction of randomness helps to increase the diversity among decision trees, reduce the risk of overfitting, and improve the generalization performance of the overall model. Generally, the predictions of the random forest are derived by voting (for classification problems) or averaging (for regression problems) the individual predictions of each decision tree [77]. This integrated approach makes Random Forests excel at handling a wide range of complex tasks because it reduces the error of individual decision trees and improves the stability and accuracy of the overall model [78].

2.6.5. K-Nearest Neighbors (KNN)

The k-Nearest Neighbors (KNN) method uses a metric to measure the distance between two instances. After receiving an unknown sample, it selects k samples from the known samples that are closest to it and then determines the most common label among these k labeled samples to make a prediction for the unknown sample. The commonly used metric is the Euclidean distance, but other metrics are also possible. KNN is a lazy learning method that stores samples and classifies them only when needed. If the sample set is complex, it can lead to significant computational overhead [79].

2.6.6. Bayesian Net

The Bayesian Net (BN) is a probabilistic graphical model based on Bayes' theorem and graph theory. This model uses a directed acyclic graph to represent causal dependencies between random variables, where nodes represent random variables and edges represent dependencies between variables. Each node is associated with a conditional probability distribution that describes the conditional probability of the node given the value of its parent. Using Bayes' theorem, probabilistic inference can be performed in Bayesian net to compute a posteriori probability distributions for other nodes from known evidence for tasks such as prediction, diagnosis, and classification [80].

2.6.7. Prediction measurement

The predictive performance of the prediction models was evaluated by ten-folds cross-validation. Sensitivity (Sn), specificity (Sp), and Accuracy (ACC) were employed to measure the prediction ability of model. The SN, SP and ACC can be represented as:

$$SN = TP/[TP + FN]$$

$$SP = TN/[TN + FP]$$

$$ACC = [TP + TN]/[TP + TN + FP + FN],$$

Where TP, TN, FP and FN are true positives, true negatives, false positives and false negatives, respectively [81–83].

3. Results

3.1. Selection of differential expression genes

Following the examination of datasets (GSE54236, GSE121248, and GSE164760), we identified 12015, 25705, and 13503 genes with differential expression, respectively. The volcano map (Fig. 1) shows the differentially expressed genes in the three data sets. Red is high-expressed genes (logFC>1, P.Value < 0.05), and blue is low-expressed genes (logFC < −1, P.Value < 0.05). Hence,955, 1235, and 113 genes with differential expression were obtained after removing those differential expression genes in the range of [−1,1].

3.2. Biological function analysis

The GO function enrichment analysis is comprised of three distinct sections: Biological Process (BP), Cellular Component (CC), and Molecular Function (MF). Results from the BP analysis reveal that the differentially expressed genes contribute to processes such as oxidation-reduction process, cell adhesion, immune response and cell surface receptor signaling pathway (as seen in Fig. 2A). The MF analysis indicated that they play a role in calcium ion binding, iron ion binding and serine-type endopeptidase activity (as demonstrated in Fig. 2B), while CC analysis determined that these genes are located in extracellular region and extracellular exosomes (as depicted in Fig. 2C). Furthermore, the KEGG pathway enrichment analysis showed that these genes are predominantly involved in Chemical carcinogenesis and Linoleic acid metabolism (as illustrated in Fig. 2D).

3.3. Protein-protein interaction network analysis

PPI network maps of differentially expressed genes were generated using the STRING online database and Cytoscape software. Confidence level >0.4 and Degree >80 were set as the cut-off criteria to obtain the protein-protein interaction network DEG-A consisting of the differentially expressed genes with the highest clustering coefficients (Fig. 3 A).

Then, the DEG-A network was further analyzed by MCODE algorithms to refine the DEGs. As a result, 12 DEGs were obtained. Two densely connected network components were identified by the MCODE algorithm (Fig. 3B and C). As the score of MCODE 2 is higher thang MCODE 1, the genes in MCODE 2 are considered important for HCC, which include AKR1B10, CYP1A2, CYP2C8,CYP2C9, CYP2C19, EPHX2, ESR1, MAF, NAT2, NQO1, SRXN1 and TXNRD1 (Fig. 3C).

3.4. Survival analysis

To further investigate the target gene expression and its correlation with the overall survival(OS) of HCC patients, a Kaplan-Meier survival analysis was carried out using the TCGA database. The target genes were screened using PPI, and the median gene expression was used as the baseline for dividing into high and low expression groups. Results showed that the expression of CYP2C8, CYP2C9, EPHX2, and ESR1 were positively correlated with the patient's OS, meaning the higher the expression, the shorter the patient's survival. On the other hand, AKR1B10 and NQO1 showed a negative correlation with the patient's OS. However, no significant correlation was found between the expression of CYP1A2, CYP2C19, MAF, NAT2, SRXN1, and TXNRD1 and

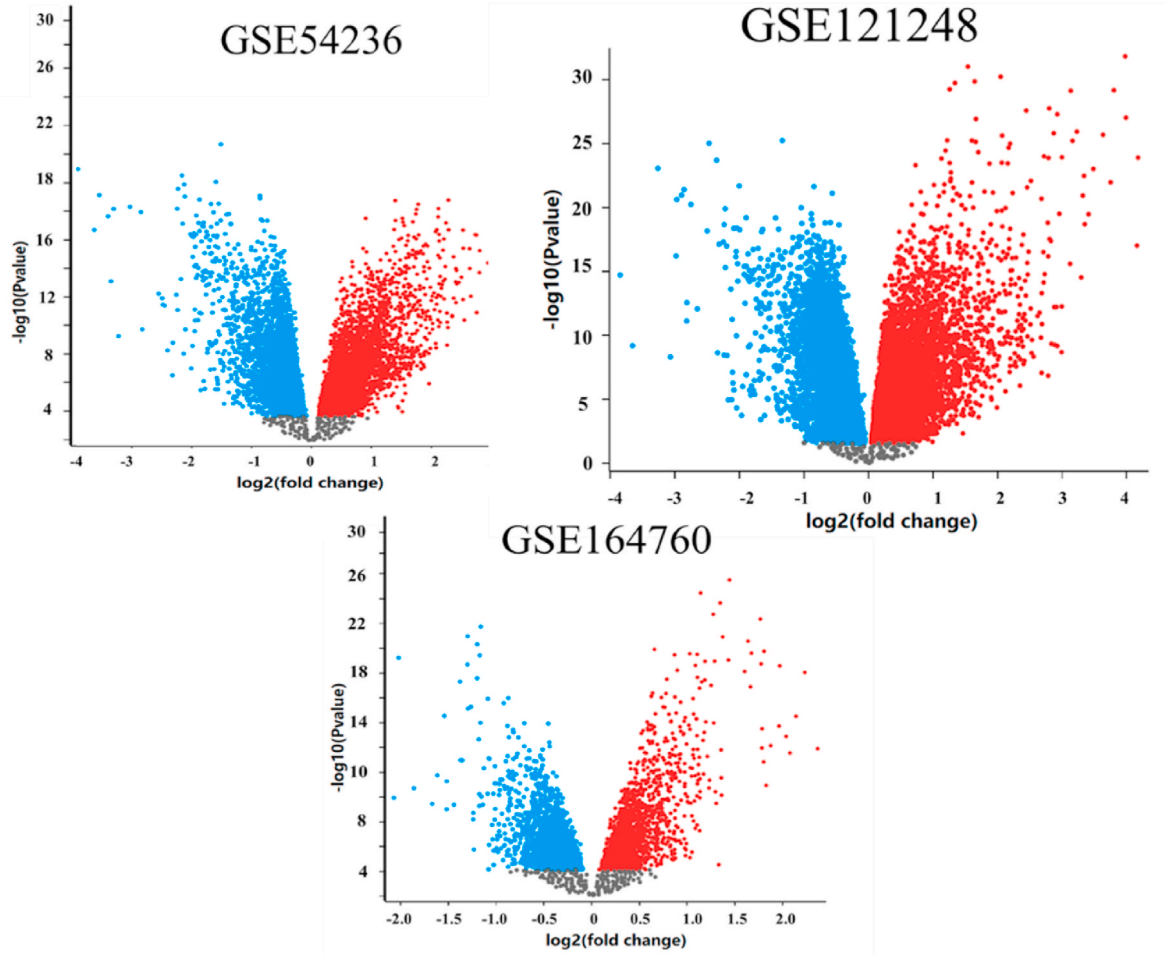

Fig. 1. Differential expression genes of GSE54236、GSE121248 and GSE164760(Red: significantly up-regulated genes with logFC>1, Blue: significantly down-regulated genes with logFC<1, Black: significantly down-regulated or up-regulated with logFC between −1 and 1, P < 0.05).

the survival of HCC patients (Fig. 4).

Indeed, validation of the identified key genes in independent HCC datasets is an important aspect that strengthens the credibility and applicability of our findings. To address the issue of validation, three additional independent HCC datasets(GSE101685, GSE62232 and GSE 45267) were collected. These datasets will cover diverse patient populations, including different demographics and disease stages, to ensure a comprehensive validation process. The identified key differentially expressed genes, including CYP2C8, CYP2C9, EPHX2, ESR1, AKR1B10 and NQO1, were critically evaluated in these datasets. The results show that aforementioned six genes displayed significant differences in the three data sets(See Table 1).

3.5. Model prediction

The 12 key genes chosen were linked to HCC, making them potential predictors for the disease. Six predictive models were created using a variety of algorithms, each undergoing 10-folds cross-validation. As demonstrated in Table 2, all six models demonstrated excellent prediction capabilities. Among these models, Adaboost demonstrated the best performance when compared to the other algorithms.

4. Discussion

In this study, the differential expression genes were identified and screened. The results of the GO functional enrichment analysis indicated that these genes were associated with various processes including oxidation-reduction process, cell adhesion, immune response, proteolysis, cell surface receptor signaling pathway, calcium ion binding, enzyme binding, iron ion binding, serine-type endopeptidase, plasma membrane, extracellular region, and extracellular exosomes. Furthermore, the KEGG pathway enrichment analysis showed that these genes are predominantly involved in metabolic pathways, linoleic acid metabolism, chemical carcinogenesis, mineral absorption, arachidonic acid metabolism, retinol metabolism, and drug metabolism - cytochrome P450. The cytochromes P450 are part of a family of hemoglobin molecules that play a role in the metabolism of both endogenous and exogenous substances, drugs, and compounds [84]. These CYP450 proteins not only participate in the metabolism of various drugs in the liver but are also closely related to various liver diseases, including HCC. When the liver experiences pathological changes, CYP activity and expression decreases, which is linked not only to the severity of liver disease but also to the cause of cirrhosis [85].

The survival analysis revealed interesting patterns of gene expression and its correlation with patient survival. Specifically, the expression levels of CYP2C8, CYP2C9, EPHX2 and ESR1 showed a positive correlation with patients' OS, indicating that higher gene expression was associated with shorter survival. This unexpected finding suggests that these genes may play a role in promoting HCC progression or signaling a dysregulated state within the tumor microenvironment. Conversely, AKR1B10 and NQO1 showed a negative correlation with patients' OS, indicating that increased expression of these genes was associated with longer survival in HCC patients. This observation suggests that AKR1B10 and NQO1 may have tumor-suppressive properties or may indicate a

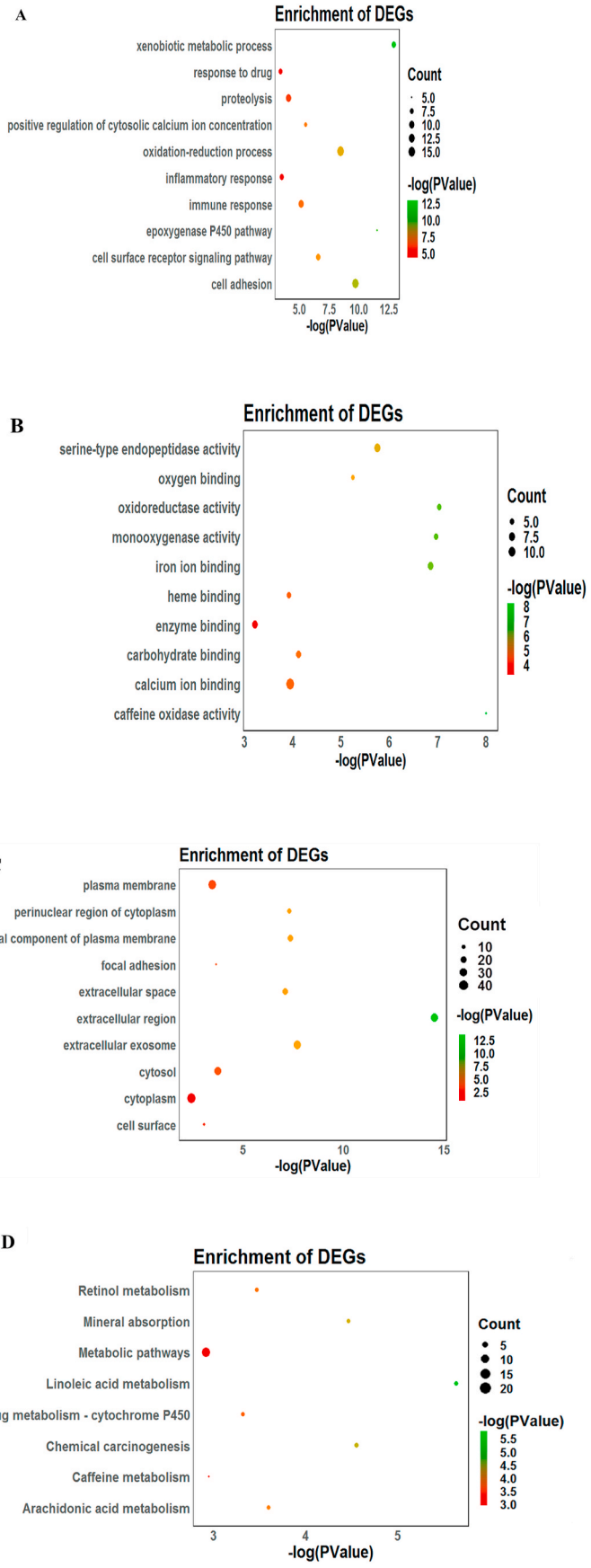

Fig. 2. Enrichment of biological function of differentially expressed genes in GSE54236、GSE121248 and GSE164760(A: Differentially expressed genes in Biological Process of Gene Ontology, B: Differentially expressed genes in Molecular Function of Gene Ontology, C: Differentially expressed genes in Cellular Component of Gene Ontology, D: differentially expressed genes in KEGG).

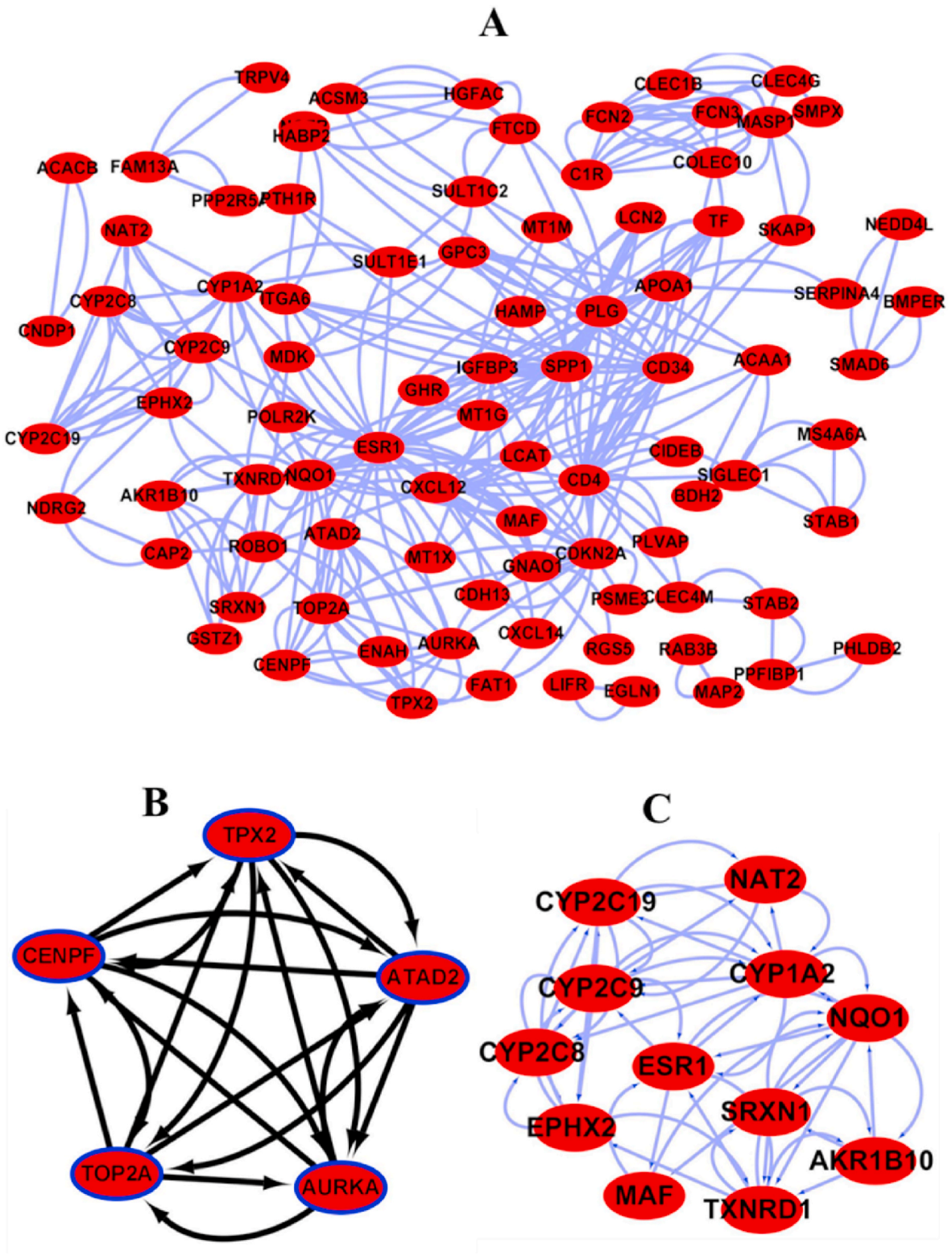

Fig. 3. Protein-protein interaction(PPI) network of differentially expressed genes (A: PPI network with Confidence level >0.4 and Degree >80; B: MCODE 1 network; C: MCODE 2 network).

more favorable prognosis. Notably, for certain genes, such as CYP1A2, CYP2C19, MAF, NAT2, SRXN1 and TXNRD1, no significant correlation was found between their expression levels and survival of HCC patients. This underscores the complexity of the molecular landscape in HCC, where specific genes may have unique influences on patient outcomes.

High expression of CYP2C8 and CYP2C9 associated with shorter overall survival (OS) in HCC patients may be counterintuitive given that these genes are members of the cytochrome P450 family, known for their role in drug metabolism [86]. However, it is important to consider that these enzymes have multiple functions outside of drug metabolism, including involvement in inflammatory and oxidative stress pathways. It is possible that increased expression of these genes could indicate a

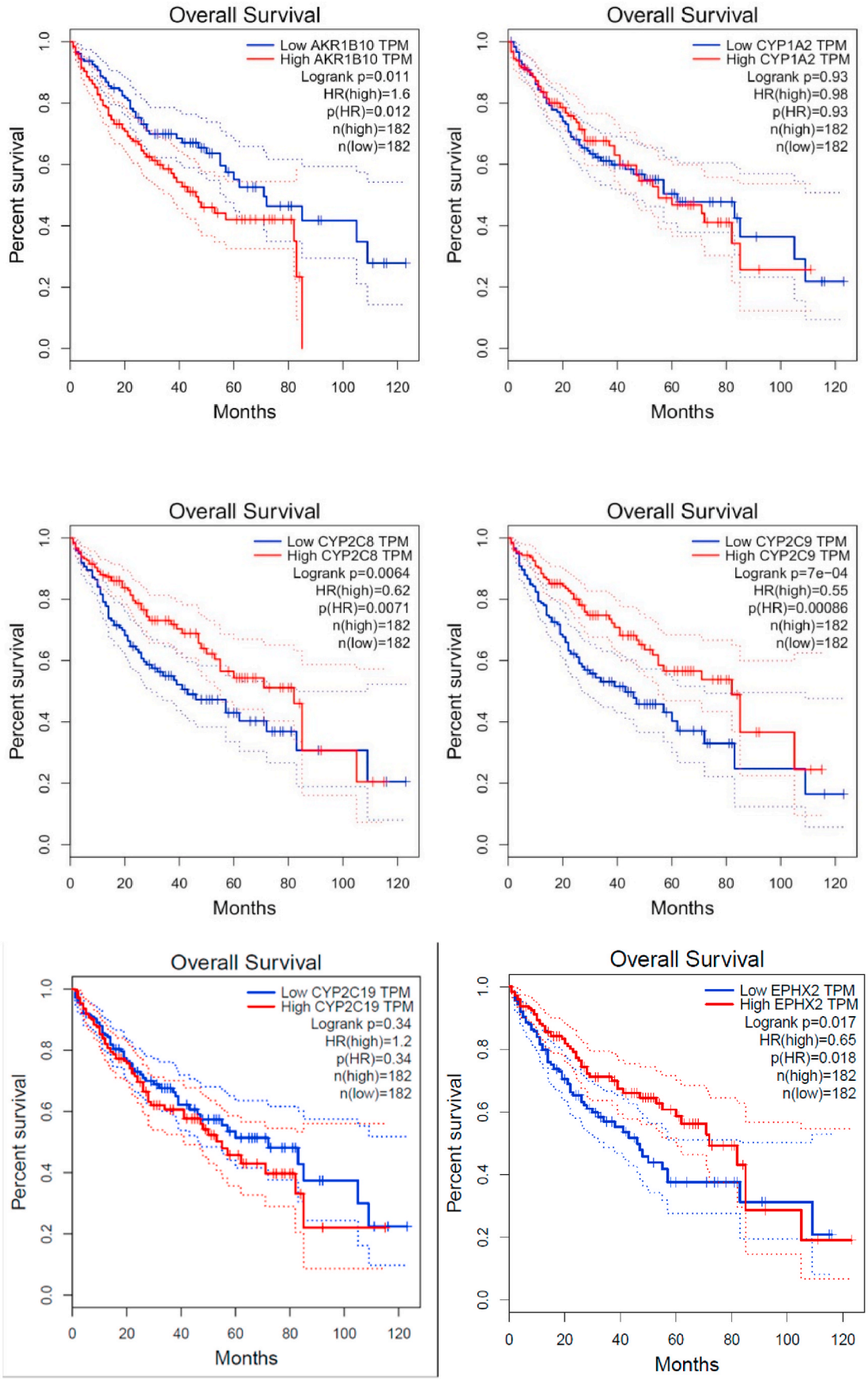

Fig. 4. Survival analysis of twelve differentially expressed genes. (A: AKR1B10, B:CYP1A2, C: CYP2C8, D: CYP2C9, E: CYP2C19, F:EPHX2,G:ESR1, H:MAF, I:NAT2, J:NQO1, K:SRXN1, L:TXNRD1.).

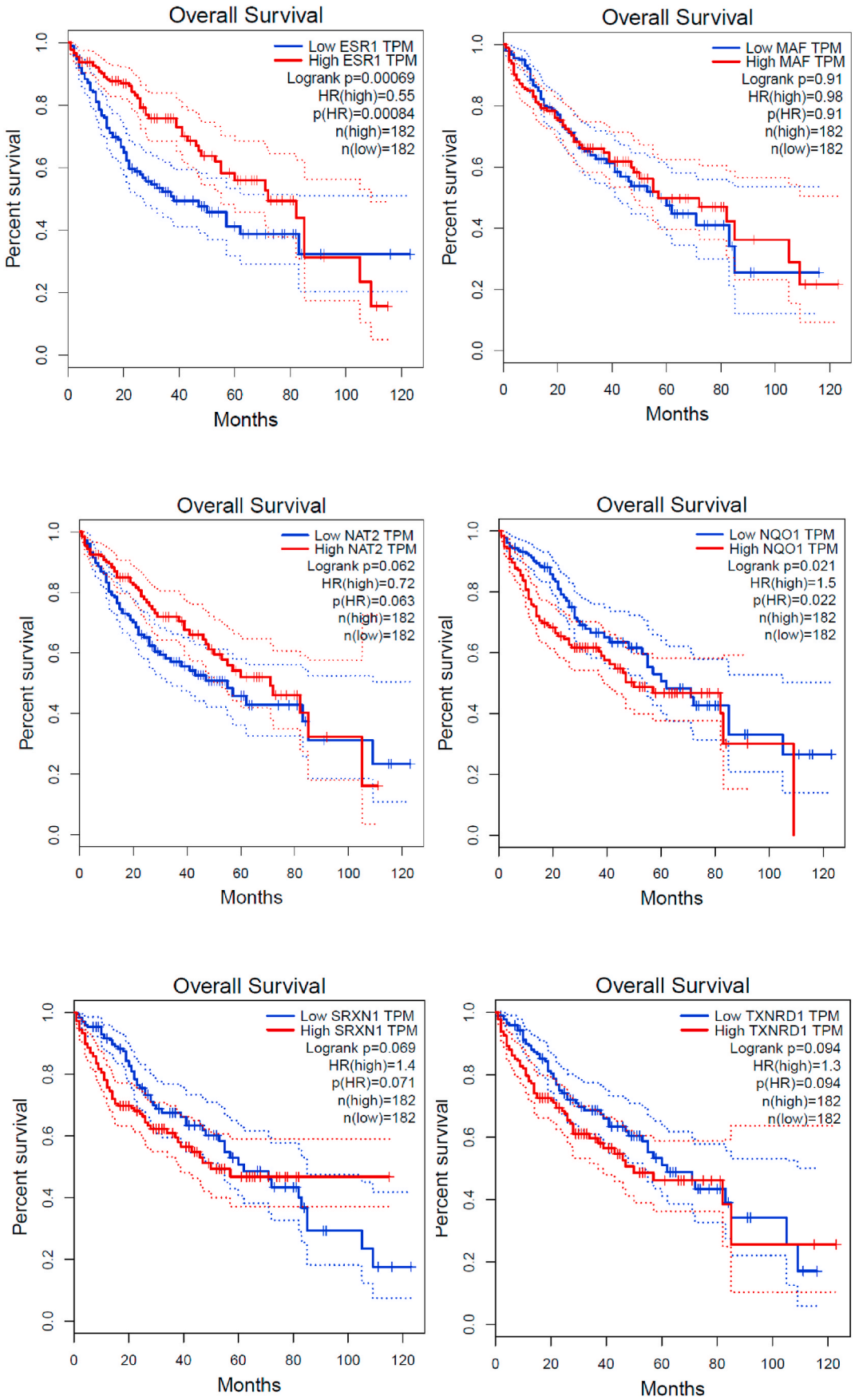

Fig. 4. (*continued*).

Table 1
Expressions of CYP2C8, CYP2C9, EPHX2, ESR1, AKR1B10 and NQO1 in Test set.

Data sets	CYP2C8	CYP2C9	EPHX2	ESR1	AKR1B10	NQO1
GSE101685	3.96	3.01	1.96	4.36	−4.07	−2.70
GSE62232	2.54	2.42	1.61	3.28	−4.92	−1.63
GSE 45267	3.11	3.26	2.00	3.41	−2.75	−1.74

Table 2
Perdition results base on different algorithm.

Algorithms	Sn (%)	Sp (%)	Acc (%)
AdaBoost	81.4	90.6	86.3
C4.5	76	91.5	84.2
RandomTree	80.9	82.1	81.5
Random Forest	83.8	83.3	83.6
KNN	77	82.9	80.1
Bayesian Net	80.4	80.4	80.3

dysregulated metabolic state or altered cellular processes leading to tumor progression.

Considering that ESR1 encodes estrogen receptor α, which is a well-known breast cancer target, less is known about the role of estrogen receptors in HCC, but estrogen signalling has been implicated in the progression of HCC. Elevated ESR1 expression may be associated with hormonal imbalances or other molecular pathways leading to a more aggressive phenotype [87,88]. The ESR1 gene codes for an estrogen receptor and a transcription factor activated by ligand. This factor regulates the transcription of genes inducible by estrogen that is linked to growth, metabolism, sexual development, pregnancy and other reproductive processes. Additionally, ESR1 is believed to have a contribution in the development of breast cancer, endometrial cancer and osteoporosis [89,90]. However, several recent studies have suggested that ESR1 may also act as master regulator for the expression in Cytochrome P450 enzymes in the human liver [91,92].

Expression of AKR1B10 and NQO1 was negatively correlated with patients' OS, suggesting their potential protective role in HCC. AKR1B10 and NQO1 are enzymes involved in detoxification processes and protection against oxidative stress. High expression of these genes may indicate that patients have a greater ability to counteract the damaging effects of reactive oxygen species and other toxic compounds, leading to improved survival outcomes. Recently, a number of research studies have indicated that AKR1B10 is expressed at an elevated level in hepatocellular carcinoma. According to a study by Wang et al., the expression of AKR1B10 is controlled by miR-383–5p and contributes to the advancement of HCC tumors [22]. The AKR1B10 protein belongs to the NAD (P) H-dependent oxidoreductase superfamily and plays a crucial role in the proliferation and development of tumors. It reduces carbonyl groups and regulates lipid metabolism, leading to the promotion of cell survival and modulation of the retinoic acid signaling pathway. Additionally, AKR1B10 stabilizes ACCA, resulting in the synthesis of long-chain fatty acids. These fatty acids are essential components of membrane phospholipids and lipid second messengers, which drive cell growth, proliferation and survival by mediating cell signaling. Research has demonstrated that AKR1B10 is mainly expressed in the normal human small intestine and colon, and is frequently overproduced in various types of cancer [93–96]. Recently, several studies have suggested that AKR1B10 may be overexpressed in cases of hepatocellular carcinoma. According to a study by Wang et al., AKR1B10 is regulated by miR-383–5p and has been linked to promoting tumor progression in HCC cases [97]. Another study by Shi et al. found that overexpression of AKR1B10 in tumors may be associated with reduced overall survival in hepatocellular carcinoma patients [98].

NQO1 is a flavoprotein homodimer that performs the elimination of quinone through a one-step two-electron reduction reaction, converting it into hydroquinone. In contrast, if not reduced by NQO1, quinone will undergo a one-electron reduction to form semi-hydroquinone, which creates reactive oxygen species through a redox cycle. This process helps prevent DNA damage due to environmental stressors. Moreover, NQO1 helps maintain the reduced forms of ubiquinone and alpha-biquinone and is crucial in safeguarding the body's endogenous antioxidants [99,100]. Research suggests that NQO1 is a significant gene linked to the metabolic patterns and apoptosis of tumor cells, making it a potential therapeutic target for hepatocellular carcinoma (HCC) [101–103].

A positive correlation between EPHX2 expression and shorter OS in HCC patients suggests that EPHX2 may be associated with HCC aggressiveness.EPHX2 encodes an enzyme involved in the metabolism of cyclic eicosatrienoic acids (EETs), which are lipid mediators with both pro- and anti-inflammatory properties [87,88].Dysregulation of EPHX2 may disrupt the homeostasis of these mediators, with potential effects on tumour growth, vasculature generation and immune responses. However, the exact mechanism of the association between EPHX2 expression and HCC survival requires further investigation.

Besides, identification of biomarkers that can serve as potential predictors for the early detection and management of hepatocellular carcinoma (HCC) is a major area of research. In this study, 12 key genes linked to the development and progression of HCC were then used to create six different predictive models using a variety of algorithms. To ensure the accuracy and reliability of the predictive models, each one underwent 10-folds cross-validation, a widely used technique in machine learning to evaluate the model's performance. This technique involves dividing the data into ten equal subsets, where each subset serves as a test dataset once, while the rest serve as training datasets. The results of this study were promising, with all six models showing excellent prediction capabilities. However, among these models, the Adaboost algorithm outperformed the other algorithms in terms of predictive accuracy. Adaboost is a machine learning algorithm that combines multiple weak classifiers to form a strong classifier. The algorithm's strength lies in its ability to adapt to the complexity of the data and avoid overfitting. The high performance of the Adaboost algorithm in predicting HCC based on the expression of the 12 key genes underscores its potential usefulness in clinical settings. Its ability to accurately predict HCC can aid clinicians in early detection and intervention, which can significantly improve patient outcomes.

This study uses bioinformatics techniques and machine learning to predict hepatocellular carcinoma (HCC)-related genes, a promising approach to identify potential biomarkers and elucidate disease mechanisms. However, an important aspect to consider is the lack of reliable validation. Although the key genes identified in this study are promising candidates, their role in HCC may be complex and context-specific. More in-depth functional studies, such as gene knockout or overexpression experiments, are needed to validate their specific contributions to HCC development. In addition, the computational methods used in this study may provide valuable insights into gene-gene interactions and potential HCC candidate genes, but we must recognise that the transition from computational prediction to actual clinical application is a multifaceted process. Clinical validation is a critical step in bridging the gap between predictions and their actual utility in diagnosing, treating or monitoring patients with HCC. Validation of predictive genetic profiles or biomarkers identified by computational analyses requires rigorous experimentation and evaluation in actual patient cohorts. This requires the collection and analysis of gene expression data from clinical samples, ideally across different patient populations and disease stages. The predictive power of the identified genes or traits must then be assessed against clinical endpoints such as disease progression, patient survival or response to therapy. In addition, the clinical validation process addresses potential confounders and sources of bias inherent in real-world patient data. These factors include patient demographics, comorbidities and treatment history, among others, which can affect the reliability and generalizability of calculated predictions. Rigorous statistical analysis and validation protocols are essential to ensure that calculated results can reliably inform clinical decisions.

In addition, machine learning algorithms were used in this study to construct predictive models. The accuracy and reliability of machine learning analyses are highly dependent on the quality and consistency of the data used. Datasets from different sources may differ in experimental techniques, sample sizes and data pre-processing methods, which can introduce noise and bias into the results. Discuss how data heterogeneity affects the robustness of study results and the likelihood of false positives or false negatives. Cancer, especially HCC, is a multifactorial disease influenced by various genetic, epigenetic and environmental factors. Bioinformatics analyses may oversimplify this complexity and miss important interactions and regulatory mechanisms. Exploring how the chosen method may not fully capture all contributing factors can lead to incomplete or inaccurate conclusions. In addition, overfitting or underfitting is a limitation of machine learning that can affect the predictive accuracy of a model. Therefore, it is necessary to regularly update the training and test sets to refine the model and improve its ability to generalize.

5. Conclusions

HCC is a complexly-arisen tumor. The results of this study indicate that genes such as CYP2C8, CYP2C9, EPHX2, ESR1, AKR1B10, and NQO1 may hold a significant impact on the development of HCC. These genes have the potential to serve as a diagnostic tool or prognostic marker for HCC, as well as a new target for HCC treatment, potentially enhancing the therapeutic outcome. Overall, this study provides important insights into the potential of the 12 key genes and the Adaboost algorithm in predicting HCC. Further studies can build upon these findings and explore the clinical utility of these predictive models in large patient cohorts.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated and/or analyzed during the current study are available in the GEO repository, GSE54236, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE54236,

GSE121248,https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE121248, and GSE164760,https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE164760.

Funding

This work is supported by Natural Science Foundation of Fujian Province of China 2023J011581) and Huaxia Scientific Research Funding (HXKY202304D003/HXKY202305D004).

Authors' contributions

Y·S., C.Z. and J.X. designed the study; Y·S.,J.H. and L.J. drafted the manuscript; C.Z.,L.Z.,Y·S.,J.H.and H.L. conducted the experiments, and C.Z. and L.J.arranged the study plan. All authors read and approved the final manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

Not applicable.

References

[1] M. Wang, Y. Wang, X. Feng, R. Wang, Y. Wang, H. Zeng, J. Qi, H. Zhao, N. Li, J. Cai, et al., Contribution of hepatitis B virus and hepatitis C virus to liver cancer in China north areas: experience of the Chinese National Cancer Center, Int. J. Infect. Dis. 65 (2017) 15–21.
[2] J. Calderaro, M. Ziol, V. Paradis, J. Zucman-Rossi, Molecular and histological correlations in liver cancer, J. Hepatol. 71 (2019) 616–630.
[3] P.-P. Song, J.-F. Xia, Y. Inagaki, K. Hasegawa, Y. Sakamoto, N. Kokudo, W. Tang, Controversies regarding and perspectives on clinical utility of biomarkers in hepatocellular carcinoma, World J. Gastroenterol. 22 (2016) 262–274.
[4] E.U. Cidon, Systemic treatment of hepatocellular carcinoma: past, present and future, World J. Hepatol. 9 (2017) 797–807.
[5] C.E. DeSantis, J. Ma, M.M. Gaudet, L.A. Newman, K.D. Miller, A.G. Sauer, A. Jemal, R.L. Siegel, Breast cancer statistics, 2019, Ca-Cancer J. Clin. 69 (2019) 438–451.
[6] M.M. Center, A. Jemal, International Trends in Liver Cancer Incidence Rates, vol. 20, Cancer Epidemiology Biomarkers & Prevention, 2011, pp. 2362–2368.
[7] G. Dastgeer, Z.M. Shahzad, H. Chae, Y.H. Kim, B.M. Ko, J. Eom, Bipolar junction transistor exhibiting excellent output characteristics with a prompt response against the selective protein, Adv. Funct. Mater. 32 (2022).
[8] M. Shahzadi, S. Nisar, D.-K. Kim, N. Sarwar, A. Rasheed, W. Ahmad, A.M. Afzal, M. Imran, M.A. Assiri, Z.M.M. Shahzad, et al., Highly efficient, non-covalent functionalization of CVD-graphene via novel pyrene-based supporter construct, Chemosensors 11 (2023).
[9] G. Dastgeer, S. Nisar, Z.M. Shahzad, A. Rasheed, D.-K. Kim, S.H.A. Jaffery, L. Wang, M. Usman, J. Eom, Low-power negative-differential-resistance device for sensing the selective protein via supporter molecule engineering, Adv. Sci. 10 (2023).
[10] X. Chen, C.C. Yan, X. Zhang, Z.-H. You, Long non-coding RNAs and complex diseases: from experimental results to computational models, Briefings Bioinf. 18 (2017) 558–576.
[11] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, P.-A. Heng, H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes, IEEE Trans. Med. Imag. 37 (2018) 2663–2674.
[12] Y. Lu, Z. Li, C. Lin, J. Zhang, Z. Shen, Translation role of circRNAs in cancers, J. Clin. Lab. Anal. 35 (2021).
[13] S. Ucar, Analysis of hepatitis B disease with fractal-fractional Caputo derivative using real data from Turkey, J. Comput. Appl. Math. 419 (2023).
[14] Y.-J. Zhu, B. Zheng, G.-J. Luo, X.-K. Ma, X.-Y. Lu, X.-M. Lin, S. Yang, Q. Zhao, T. Wu, Z.-X. Li, et al., Circular RNAs negatively regulate cancer stem cells by physically binding FMRP against CCAR1 complex in hepatocellular carcinoma, Theranostics 9 (2019) 3526–3540.
[15] J. Calderaro, T.P. Seraphin, T. Luedde, T.G. Simon, Artificial intelligence for the prevention and clinical management of hepatocellular carcinoma, J. Hepatol. 76 (2022) 1348–1361.
[16] Z.H. Foda, A.V. Annapragada, K. Boyapati, D.C. Bruhm, N.A. Vulpescu, J. E. Medina, D. Mathios, S. Cristiano, N. Niknafs, H.T. Luu, et al., Detecting liver cancer using cell-free DNA fragmentomes, Cancer Discov. 13 (2023) 616–631.
[17] P. Johnson, Q. Zhou, D.Y. Dao, Y.M.D. Lo, Circulating biomarkers in the diagnosis and management of hepatocellular carcinoma, Nat. Rev. Gastroenterol. Hepatol. 19 (2022) 670–681.
[18] D. Nam, J. Chapiro, V. Paradis, T.P. Seraphin, J.N. Kather, Artificial intelligence in liver diseases: improving diagnostics, prognostics and response prediction, Jhep Reports 4 (2022).
[19] A. Tabari, S.M. Chan, O.M.F. Omar, S.I.I. Iqbal, M.S.S. Gee, D. Daye, Role of machine learning in precision oncology: applications in gastrointestinal cancers, Cancers 15 (2023).
[20] H.-L. Jia, Q.-H. Ye, L.-X. Qin, A. Budhu, M. Forgues, Y. Chen, Y.-K. Liu, H.-C. Sun, L. Wang, H.-Z. Lu, et al., Gene expression profiling reveals potential biomarkers of human hepatocellular carcinoma, Clin. Cancer Res. 13 (2007) 1133–1139.
[21] L. Ao, Z. Zhang, Q. Guan, Y. Guo, Y. Guo, J. Zhang, X. Lv, H. Huang, H. Zhang, X. Wang, et al., A qualitative signature for early diagnosis of hepatocellular carcinoma based on relative expression orderings, Liver Int. 38 (2018) 1812–1819.
[22] L.S. Mou, L. Liu, Y.M. Qiu, Y.Y. Liang, Z.H. Pu, Construction of a novel predictive model with seven metabolism-related genes for hepatocellular carcinoma by machine learning, Genes & Diseases 10 (2023) 1806–1808.
[23] J.J. Harding, S. Nandakumar, J. Armenia, D.N. Khalil, M. Albano, M. Ly, J. Shia, J.F. Hechtman, R. Kundra, I. El Dika, et al., Prospective genotyping of hepatocellular carcinoma: clinical implications of next-generation sequencing for matching patients to targeted and immune therapies, Clin. Cancer Res. 25 (2019) 2116–2126.
[24] J.S. Ross, K. Wang, L. Gay, R. Al-Rohil, J.V. Rand, D.M. Jones, H.J. Lee, C. E. Sheehan, G.A. Otto, G. Palmer, et al., New routes to targeted therapy of

intrahepatic cholangiocarcinomas revealed by next-generation sequencing, Oncol. 19 (2014) 235–242.
[25] A. Radoaca, Simple venn diagrams for multisets, in: 2015 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, 2015.
[26] M.E. Ritchie, B. Phipson, D. Wu, Y. Hu, C.W. Law, W. Shi, G.K. Smyth, Limma powers differential expression analyses for RNA-sequencing and microarray studies, Nucleic Acids Res. 43 (2015) e47.
[27] J.A. Blake, M. Dolan, H. Drabkin, D.P. Hill, L. Ni, D. Sitnikov, S. Bridges, S. Burgess, T. Buza, F. McCarthy, et al., Gene ontology annotations and resources, Nucleic Acids Res. 41 (2013) D530–D535.
[28] D.W. Huang, B.T. Sherman, Q. Tan, J. Kir, D. Liu, D. Bryant, Y. Guo, R. Stephens, M.W. Baseler, H.C. Lane, et al., DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists, Nucleic Acids Res. 35 (2007) W169–W175.
[29] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, et al., KEGG for linking genomes to life and the environment, Nucleic Acids Res. 36 (2008) D480–D484.
[30] S. Tripathi, M.O. Pohl, Y. Zhou, A. Rodriguez-Frandsen, G. Wang, D.A. Stein, H. M. Moulton, P. DeJesus, J. Che, L.C. Mulder, et al., Meta- and orthogonal integration of influenza "OMICs" data defines a role for UBR4 in virus budding, Cell Host Microbe 18 (2015) 723–735.
[31] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K.P. Tsafou, et al., STRING v10: protein-protein interaction networks, integrated over the tree of life, Nucleic Acids Res. 43 (2015) D447–D452.
[32] M.T. Ramakrishna, V.K. Venkatesan, I. Izonin, M. Havryliuk, C.R. Bhat, Homogeneous Adaboost ensemble machine learning algorithms with reduced entropy on balanced data, Entropy 25 (2023).
[33] J. Sun, H. Li, H. Fujita, B. Fu, W. Ai, Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting, Inf. Fusion 54 (2020) 128–144.
[34] Y. Wu, Y. Ke, Z. Chen, S. Liang, H. Zhao, H. Hong, Application of alternating decision tree with AdaBoost and bagging ensembles for landslide susceptibility mapping, Catena 187 (2020).
[35] C. Xiao, N. Chen, C. Hu, K. Wang, J. Gong, Z. Chen, Short and mid-term sea surface temperature prediction using time-series satellite data and LSTM-AdaBoost combination approach, Rem. Sens. Environ. 233 (2019).
[36] B. Niu, Y. Lu, J.Y. Wang, Y. Hu, J.H. Chen, Q. Chen, G.W. He, L.F. Zheng, 2D-SAR, Topomer CoMFA and molecular docking studies on avian influenza neuraminidase inhibitors, Comput. Struct. Biotechnol. J. 17 (2019) 39–48.
[37] A. Jaiswal, R. Kumar, Breast cancer diagnosis using stochastic self-organizing map and enlarge C4.5, Multimed. Tool. Appl. 82 (2023) 18059–18076.
[38] M.A. Muslim, S.H. Rukmana, E. Sugiharti, B. Prasetiyo, S. Alimah, Iop, Optimization of C4.5 algorithm-based particle swarm optimization for breast cancer diagnosis, in: International Conference on Mathematics, Science and Education (ICMSE), Univ Negeri Semarang, Fac Math & Nat Sci, Semarang, INDONESIA, 2017, 983.
[39] M. Park, P. Summons, Diabetic retinopathy classification using C4.5, in: 15th Pacific Rim International Conference on Artificial Intelligence (PRICAI)/15th Pacific Rim Knowledge Acquisition Workshop (PKAW), 2018, pp. 90–101, 11016. (Nanjing, PEOPLES R CHINA.
[40] I.M. Wirawan, T. Widiyaningtyas, S.B. Nurwakiah, Ieee, Nutritional status of infants classification by calculating anthropometry through C4.5 algorithm, in: International Conference on Electrical, Electronics and Information Engineering (ICEEIE), 2019, pp. 216–219. Denpasar, INDONESIA.
[41] E. Abdulhay, M. Alafeef, A. Abdelhay, A. Al-Bashir, Classification of normal, ictal and inter-ictal EEG via direct quadrature and random forest tree, J. Med. Biol. Eng. 37 (2017) 843–857.
[42] M. Kretowska, Random forest of dipolar trees for survival prediction, in: L. Rutkowski, R. Tadeusiewicz, L.A. Zadeh, J. Zurada (Eds.), Artificial Intelligence and Soft Computing - Icaisc 2006, Proceedings, Volume 4029, 2006, pp. 909–918.
[43] R. Maree, P. Geurts, J. Piater, L. Wehenkel, Biomedical image classification with random subwindows and decision trees, in: Y. Liu, T. Jiang, C.S. Zhang (Eds.), Computer Vision for Biomedical Image Applications, Proceedings, ume 3765, 2005, pp. 220–229.
[44] M. Gu, C. Li, L. Chen, S. Li, N. Xiao, D. Zhang, X. Zheng, Insight from untargeted metabolomics: revealing the potential marker compounds changes in refrigerated pork based on random forests machine learning algorithm, Food Chem. 424 (2023).
[45] A.R. Khan, W.A. Wicaksono, N.J. Ott, A.T. Poret-Peterson, G.T. Browne, Random forest analysis reveals taxa predictive of Prunus replant disease in peach root microbiomes, PLoS One 17 (2022).
[46] J.Y. Ryu, W.D. Jang, J. Jang, K.-S. Oh, PredAOT: a computational framework for prediction of acute oral toxicity based on multiple random forest models, BMC Bioinf. 24 (2023).
[47] H. Zhang, M. Chi, D. Su, Y. Xiong, H. Wei, Y. Yu, Y. Zuo, L. Yang, A random forest-based metabolic risk model to assess the prognosis and metabolism-related drug targets in ovarian cancer, Comput. Biol. Med. 153 (2023).
[48] B. Niu, R.R. Liang, G.Y. Zhou, Q. Zhang, Q. Su, X.S. Qu, Q. Chen, Prediction for Global Peste des Petits Ruminants Outbreaks Based on a Combination of Random Forest Algorithms and Meteorological Data, Front. Vet. Sci. 7 (2021).
[49] B. Niu, H. Zhang, G.Y. Zhou, S.W. Zhang, Y.F. Yang, X.J. Deng, Q. Chen, Safety risk assessment and early warning of chemical contamination in vegetable oil, Food Control 125 (2021).
[50] R.R. Liang, Y. Lu, X.S. Qu, Q. Su, C.X. Li, S.J. Xia, Y.X. Liu, Q. Zhang, X. Cao, Q. Chen, et al., Prediction for global African swine fever outbreaks based on a combination of random forest algorithms and meteorological data, Transboundary and Emerging Diseases 67 (2020) 935–946.
[51] B. Niu, M.M. Zhao, Q. Su, M.Y. Zhang, W. Lv, Q. Chen, F.X. Chen, D.C. Chu, D. S. Du, Y.H. Zhang, 2D-SAR and 3D-QSAR analyses for acetylcholinesterase inhibitors, Mol. Divers. 21 (2017) 413–426.
[52] K.V. Archana, G. Komarasamy, A novel deep learning-based brain tumor detection using the Bagging ensemble with K-nearest neighbor, J. Intell. Syst. 32 (2023).
[53] N. Kour, S. Gupta, S. Arora, A vision-based clinical analysis for classification of knee osteoarthritis, Parkinson's disease and normal gait with severity based on k-nearest neighbour, Expet Syst. 39 (2022).
[54] Z.-R. Tang, Y. Chen, R. Hu, H. Wang, Predicting hematoma expansion in intracerebral hemorrhage from brain CT scans via K-nearest neighbors matting and deep residual network, Biomed. Signal Process Control 76 (2022).
[55] J. Atkinson, A. Rivas, Discovering novel causal patterns from biomedical natural-language texts using bayesian nets, IEEE Trans. Inf. Technol. Biomed. 12 (2008) 714–722.
[56] L. Casini, P.M. Illari, F. Russo, J. Williamson, RECURSIVE bayesian nets for prediction, explanation and control in cancer science A position paper, in: 1st International Conference on Bioinformatics (BIOINFORMATICS 2010), 2010, pp. 233–238. Valencia, SPAIN.
[57] K. Topuz, H. Uner, A. Oztekin, M.B. Yildirim, Predicting pediatric clinic no-shows: a decision analytic framework using elastic net and Bayesian belief network, Ann. Oper. Res. 263 (2018) 479–499.
[58] H. Xiong, S. Liu, E. Coiera, S. Berkovsky, R.V. Sharan, Weak label based Bayesian U-Net for optic disc segmentation in fundus images, Artif. Intell. Med. 126 (2022).
[59] H. Zhang, Q.D. Liu, X.R. Sun, Y.R. Xu, Y.L. Fang, S.L. Cao, B. Niu, C. Li, Integrated bioinformatics and machine learning algorithms analyses highlight related pathways and genes associated with alzheimer's disease, Curr. Bioinf. 17 (2022) 284–295.
[60] R.R. Liang, J.Y. Xie, C. Zhang, M.Y. Zhang, H. Huang, H.Z. Huo, X. Cao, B. Niu, Identifying cancer targets based on machine learning methods via chou's 5-steps rule and general pseudo components, Curr. Top. Med. Chem. 19 (2019) 2301–2317.
[61] Y. Hu, Y. Lu, S. Wang, M. Zhang, X. Qu, B. Niu, Application of machine learning approaches for the design and study of anticancer drugs, Curr. Drug Targets 20 (2019) 488–500.
[62] Y. Hu, G. Zhou, C. Zhang, M. Zhang, Q. Chen, L. Zheng, B. Niu, Identify compounds' target against alzheimer's disease based on in-silico approach, Curr. Alzheimer Res. 16 (2019) 193–208.
[63] M.Y. Zhang, Q. Su, Y. Lu, M.M. Zhao, B. Niu, Application of machine learning approaches for protein-protein interactions prediction, Med. Chem. 13 (2017) 506–514.
[64] J. Kuang, N. Luo, Z. Hao, J. Xu, X. He, J. Shi, NI-Raman spectroscopy combined with BP-Adaboost neural network for adulteration detection of soybean oil in camellia oil, J. Food Meas. Char. 16 (2022) 3208–3215.
[65] E. Mirzaee-Ghaleh, A. Taheri-Garavand, F. Ayari, J. Lozano, Identification of fresh-chilled and frozen-thawed chicken meat and estimation of their shelf life using an E-nose machine coupled fuzzy KNN, Food Anal. Methods 13 (2020) 678–689.
[66] N. Gerhardt, S. Schwolow, S. Rohn, P. Ruiz Perez-Cacho, H. Galan-Soldevilla, L. Arce, P. Weller, Quality assessment of olive oils based on temperature-ramped HS-GC-IMS and sensory evaluation: comparison of different processing approaches by LDA, kNN, and SVM, Food Chem. 278 (2019) 720–728.
[67] M. Keramat-Jahromi, S.S. Mohtasebi, H. Mousazadeh, M. Ghasemi-Varnamkhasti, M. Rahimi-Movassagh, Real-time moisture ratio study of drying date fruit chips based on on-line image attributes using kNN and random forest regression methods, Measurement 172 (2021).
[68] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. Syst. Sci. 55 (1997) 119–139.
[69] A. Lazarevic, Z. Obradovic, Boosting algorithms for parallel and distributed learning, Distributed Parallel Databases 11 (2002) 203–229.
[70] B. Niu, R.R. Liang, S.W. Zhang, H. Zhang, X.S. Qu, Q. Su, L.F. Zheng, Q. Chen, Epidemic analysis of COVID-19 in Italy based on spatiotemporal geographic information and Google Trends, Transboundary and Emerging Diseases 68 (2021) 2384–2400.
[71] W.X. Sun, J. Chen, J.Q. Li, Decision tree and PCA-based fault diagnosis of rotating machinery, Mech. Syst. Signal Process. 21 (2007) 1300–1317.
[72] J.R. Quinlan, C4.5: Program for Machine Learning, Morgan Kaufmann), San Mateo, CA, 1993.
[73] S. Ruggieri, Efficient C4.5, IEEE Trans. Knowl. Data Eng. 14 (2002) 438–444.
[74] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Wadsworth International Group, 1984.
[75] I. Witten, F. E, Data Mining: Practical Machine Learning Tools and Techniques 3Edition, Acm Sigmod Rec, 2011.
[76] M.T. Markus, P.J.F. Groenen, An introduction to the bootstrap, Psychometrika 63 (1998) 97–101.
[77] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[78] Y.-Q. Song, X. Yao, Z. Liu, X. Shen, J. Mao, An improved C4.5 algorithm in bagging integration model, IEEE Access 8 (2020) 206866–206875.
[79] P. Cunningham, S.J. Delany, K-Nearest neighbour classifiers - a tutorial, ACM Comput. Surv. 54 (2021).
[80] A. Antonucci, M. Zaffalon, Fast algorithms for robust classification with Bayesian nets, Int. J. Approx. Reason. 44 (2007) 200–223.
[81] J.J. Zhong, W.D. Xuan, S. Lu, S.H. Cui, Y.H. Zhou, M.T. Tang, X.S. Qu, W.C. Lu, H. Z. Huo, C. Zhang, et al., Discovery of ANO1 Inhibitors based on Machine learning

and molecule docking simulation approaches, Eur. J. Pharmaceut. Sci. 184 (2023).
[82] L.F. Zheng, X.Y. Qin, J. Wang, M.Y. Zhang, Q.L. An, J.Z. Xu, X.S. Qu, X. Cao, B. Niu, Discovery of MAO-B inhibitor with machine learning, topomer CoMFA, molecular docking and multi-spectroscopy approaches, Biomolecules 12 (2022).
[83] B. Niu, C.F. Liang, Y. Lu, M.M. Zhao, Q. Chen, Y.H. Zhang, L.F. Zheng, K.C. Chou, Glioma stages prediction based on machine learning algorithm combined with protein-protein interaction networks, Genomics 112 (2020) 837–847.
[84] D.W. Nebert, K. Wikvall, W.L. Miller, Human Cytochromes P450 in Health and Disease, vol. 368, Philosophical Transactions of the Royal Society B-Biological Sciences, 2013.
[85] X.K. Wang, X.W. Liao, C.K. Yang, K.T. Huang, T.D. Yu, L. Yu, C.Y. Han, G.Z. Zhu, X.M. Zeng, Z.Q. Liu, et al., Identification of prognostic biomarkers for patients withhepatocellular carcinoma after hepatectomy, Oncol. Rep. 41 (2019) 1586–1602.
[86] X. Wang, T. Yu, X. Liao, C. Yang, C. Han, G. Zhu, K. Huang, L. Yu, W. Qin, H. Su, et al., The prognostic value of CYP2C subfamily genes in hepatocellular carcinoma, Cancer Med. 7 (2018) 966–980.
[87] M. Bhat, E. Pasini, C. Pastrello, M. Angeli, C. Baciu, M. Abovsky, A. Coffee, O. Adeyi, M. Kotlyar, I. Jurisica, Estrogen receptor 1 inhibition of wnt/beta-catenin signaling contributes to sex differences in hepatocarcinogenesis, Front. Oncol. 11 (2021).
[88] Y. Guo, G. Wu, J. Yi, Q. Yang, W. Jiang, S. Lin, X. Yang, X. Cai, L. Mao, Anti-hepatocellular carcinoma effect and molecular mechanism of the estrogen signaling pathway, Front. Oncol. 11 (2022).
[89] Y.-H. Huang, W.-H. Fang, D.-J. Tsai, Y.-H. Chen, Y.-C. Wang, W. Su, C.-C. Kao, K. Yi, C.-C. Wang, S.-L. Su, The decisive case-control study elaborates the null association between ESR1 XbaI and osteoarthritis in asians: a case-control study and meta-analysis, Genes 12 (2021).
[90] A. Juliansyah, S. Rahman, I. Indra, B. Nelwan, P. Prihantono, Association of ERalpha-36 expression with the denovo resistance of tamoxifen in ER-positive breast cancer, Breast Dis. 40 (2021) S123–S127.
[91] J.M. Collins, Z.G. Huo, D.X. Wang, ESR1 ChIP-seq identifies distinct ligand-free ESR1 genomic binding sites in human hepatocytes and liver tissue, Int. J. Mol. Sci. 22 (2021).
[92] J.M. Collins, D. Wang, Co-expression of drug metabolizing cytochrome P450 enzymes and estrogen receptor alpha (ESR1) in human liver: racial differences and the regulatory role of ESR1, Drug metabolism and personalized therapy 36 (2021) 205–214.
[93] M.W. Kang, E.S. Lee, S.Y. Yoon, J. Jo, J. Lee, H.K. Kim, Y.S. Choi, K. Kim, Y. M. Shim, J. Kim, et al., AKR1B10 is associated with smoking and smoking-related non-small-cell lung cancer, J. Int. Med. Res. 39 (2011) 78–85.
[94] Y.T. Chung, K.A. Matkowskyj, H.N. Li, H. Bai, W.Y. Zhang, M.S. Tsao, J. Liao, G. Y. Yang, Overexpression and oncogenic function of aldo-keto reductase family 1B10 (AKR1B10) in pancreatic carcinoma, Mod. Pathol. 25 (2012) 758–766.
[95] K.A. Reddy, P.U. Kumar, M. Srinivasulu, B. Triveni, K. Sharada, A. Ismail, G. B. Reddy, Overexpression and enhanced specific activity of aldoketo reductases (AKR1B1 & AKR1B10) in human breast cancers, Breast 31 (2017) 137–143.
[96] J. Ma, D.X. Luo, C.F. Huang, Y. Shen, Y.W. Bu, S. Markwell, J. Gao, J.H. Liu, X. Y. Zu, Z. Cao, et al., AKR1B10 overexpression in breast cancer: association with tumor size, lymph node metastasis and patient survival and its potential as a novel serum marker, Int. J. Cancer 131 (2012) E862–E871.
[97] J.Q. Wang, Y.Y. Zhou, X.C. Fei, X.H. Chen, Y.J. Chen, Biostatistics mining associated method identifies AKR1B10 enhancing hepatocellular carcinoma cell growth and degenerated by miR-383-5p, Sci. Rep. 8 (2018).
[98] J. Shi, L.X. Chen, Y. Chen, Y.F. Lu, X.R. Chen, Z.G. Yang, Aldo-Keto Reductase Family 1 Member B10 (AKR1B10) overexpression in tumors predicts worse overall survival in hepatocellular carcinoma, J. Cancer 10 (2019) 4892–4901.
[99] L.J. Lin, J. Sun, Y. Tan, Z.L. Li, F.Y. Kong, Y. Shen, C. Liu, L.T. Chen, Prognostic implication of NQO1 overexpression in hepatocellular carcinoma, Hum. Pathol. 69 (2017) 31–37.
[100] H.-Z. Zhou, H.-Q. Zeng, D. Yuan, J.-H. Ren, S.-T. Cheng, H.-B. Yu, F. Ren, Q. Wang, Y.-P. Qin, A.-L. Huang, et al., NQO1 potentiates apoptosis evasion and upregulates XIAP via inhibiting proteasome-mediated degradation SIRT6 in hepatocellular carcinoma, Cell Commun. Signal. 17 (2019).
[101] W. Zhao, L. Jiang, T. Fang, F. Fang, Y. Liu, Y. Zhao, Y. You, H. Zhou, X. Su, J. Wang, et al., Beta-lapachone selectively kills hepatocellular carcinoma cells by targeting NQO1 to induce extensive DNA damage and PARP1 hyperactivation, Front. Oncol. 11 (2021).
[102] N. Geng, Y.Y. Jin, Y.R. Li, S.X. Zhu, H. Bai, AKR1B10 inhibitor epalrestat facilitates sorafenib-induced apoptosis and autophagy via targeting the mTOR pathway in hepatocellular carcinoma, Int. J. Med. Sci. 17 (2020) 1246–1256.
[103] B.Y. Cheng, E.Y. Lau, H.-W. Leung, C.O.-N. Leung, N.P. Ho, S. Gurung, L.K. Cheng, C.H. Lin, R.C.-L. Lo, S. Ma, et al., IRAK1 augments cancer stemness and drug resistance via the AP-1/AKR1B10 signaling cascade in hepatocellular carcinoma, Cancer Res. 78 (2018) 2332–2342.