Original
ISSN 2413-0516

Potential biomarker detection for liver cancer stem cell by machine learning approach

Ali Farzane[1], Maryam Akbarzadeh[2], Reza Ferdousi[3], Mohammadreza Rashidi[4], Reza Safdari[1]

[1]Department of Health Information Management, School of Allied Medical Sciences, Tehran University of Medical Sciences, Tehran, Iran.
[2]Department of Biochemistry, Erasmus University Medical Center, Rotterdam, Netherlands.
[3]Department of Health Information Technology, School of Management and Medical Informatics, Tabriz University of Medical Sciences, Daneshgah St, Tabriz, Iran.
[4]Stem Cell and Regenerative Medicine Institute, Tabriz University of Medical Sciences, Tabriz, Iran.
Corresponding Author: Reza Safdari (E-mail: rsafdari@tums.ac.ir)
(Submitted: 05 October 2020 – Revised version received: 19 October 2020 – Accepted: 27 October 2020 – Published online: 30 December 2020)

Abstract
Objectives: In this study, we aimed to identify putative biomarkers for identification and characterization of these cells in liver cancer.
Methods: We employed a supervised machine learning method, XGBoost, to data from 13 GEO data series to classify samples using gene expression data.
Results: Across the 376 samples [129 cancer stem cell, CSCs and 247 non-CSCs cases], XGBoost displayed high performance in the classification of data. XGBoost feature importance scores and SHAP (Shapley Additive explanation) values were used for the interpretation of results and analysis of individual gene importance. We confirmed that expression levels of a 10-gene set (PTGER3, AURKB, C15orf40, IDI2, OR8D1, NACA2, SERPINB6, L1CAM, SMC1A, and RASGRF1) were predictive. The results showed that these 10 genes can detect CSCs robustly with accuracy, sensitivity, and specificity of 97%, 100%, and 95%, respectively.
Conclusions. We suggest that the 10-gene set may be used as a biomarker set for detecting and characterizing CSCs using gene expression data.
Key words: Liver cancer, cancer stem cell, machine learning, biomarker, gene expression

Introduction

Liver cancer is the sixth most common cancer in the worldwide.[1] The mechanisms of hepatocarcinogenesis are not fully understood; however, the theory of cancer stem cells (CSCs) has recently gained traction as a potential contributor to hepatic cancer. CSCs display great plasticity and self-renewal potential and play a decisive role in tumor formation and growth.[2] These cells are highly drug-resistant and metastatic, which may underlie the recurrence and drug resistance of liver cancer.[3, 4] Therefore, the identification of liver CSCs markers and therapeutic targets associated with them is necessary for improving treatment outcomes.[5, 6]
Biomarker is defined as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention." Biomarkers play an important role in the early diagnosis of diseases.[7, 8]

Microarray technology has revolutionized gene expression analysis and has been used widely for the identification of cancer biomarkers. The use of high-throughput techniques has resulted in an exponential growth in the amount of information available in biomedical databases, which then can be exploited to integrate gene expression data for applications such as biomarker discovery, disease classification, or phenotype comparisons, among others.[9]

However, these types of data are characterized by high dimensionality as the number of genes is far bigger than the number of samples. One of the biggest challenges of high-dimensional data are the curse of dimensionality, which describes the exponential increase in volume associated with adding extra dimensions in the Euclidean space. It is responsible for the breakdown of the optimal statistical model fitting.[10] To address the problem, researchers have applied various machine learning methods to reduce both cardinality and redundancy of gene expression data during the classification process, and most of these methods were developed to facilitate the analysis of microarray data to identify the best discriminative genes or biomarkers.[11]

Extreme gradient boosting (XGBoost) is a machine learning algorithm that assigns an importance score to each feature in the training phase, and these scores can then be used as the basis for identification of importance features. Since it uses multiple subsets of features to predict outcomes in the area of dimension cursed problem ensemble models may have better performance. Moreover, XGBoost is an optimized distributed gradient boosting that achieves state-of-the-art prediction performances.[12]

Feature importance ranking using the common tree ensemble models such as XGBoost and gbm R packages may provide inconsistent results. These methods only consider the effect of splits along the decision path. Therefore, when the model relies more on a given feature, the importance assigned to that feature changes incorrectly.[13] Regarding model interpretation, which is especially important when using machine learning models that are often difficult to interpret, several studies have used Shapley Additive exPlanations (SHAP).[14-16] Proposed by Lundberg and Lee, [17] SHAP is based on game theory[18] and local explanations[19] that offers a means to estimate the contribution of each feature. SHAP provides a consistent importance value, which is an alternative to permutation feature importance.[20]

In the present study, we applied machine learning to gene expression data from previous studies on liver CSCs to identify putative biomarkers for identification and characterization of these cells.

Materials & Methods

Data Preparation

The study was approved by the ethics committee at Tehran University of Medical Sciences, Tehran, Iran (ethics code: IR.TUMS.REC.1394.1589). Since we used non-identifiable information from a publicly available data set, no specific consent was required.

Datasets

Gene expression data for samples of liver CSC and non-CSC were downloaded from GEO through accession numbers GSE56771, GSE59713, GSE66529, GSE84223, GSE68778, GSE126121, GSE112788, GSE66515, GSE42318, GSE131680, GSE103866, and GSE62905, and a data series including only non-CSC samples were downloaded via accession number GSE112790. The first nine of the above data series were expression profiling by array, while the remaining four were high-throughput sequencing (HiSeq). We excluded samples that had undergone intervention. Finally, a total of 129 CSC and 247 non-CSC samples were included in the study.

Preprocessing

R language was used for all processing steps, including preprocessing, modeling, and pathway analysis. The dropout effect was eliminated in the HiSeq series. We performed log transformation on data series, followed by quartile normalization where it was needed. Next, the Ensembl database (https://www.ensembl.org/index.html) was used to convert IDs from different data series to Ensembl IDs. Data series were integrated using the Merge function, and then the ComBat function from the sva R package was applied to remove batch effects. Finally, we had 8409 common transcript genes across all data series.

XGBoost model for biomarker signature identification

Data classification was performed using XGBoost, which is an efficient implementation of the gradient boosting framework proposed by Chen and Guestrin.[21] Gradient boosted decision tree is an ensemble learning method based on sequential decision trees whereby each decision tree learns from the previous tree to improve the model and build a strong learner.

Model tuning

In XGBoost, several parameters need to be selected to maximize model performance. However, the multiplicity of parameters may result in a model learning noises and random fluctuations and considering them meaningful, a phenomenon referred to as overfitting. Overfitting is a modeling error that occurs when a function is too closely fit to a limited set of data points.[22] Parameter tuning is an essential step in avoiding overfitting or undue complexity. The hyperparameters adopted in this study were "nrounds," "eta," "min_child_weight," "max_depth," "gamma," "colsample_bytree," "subsample," "lambda," and "alpha."[20]

The parameter "nrounds" is the number of trees that are fitted in the model. The "eta" parameter refers to the learning rate, which is used to make the model more robust. The "min_child_weight" is the minimum sum of instance weight needed in a child. If the tree partition step results in a leaf node with the sum of instance weight less than "min_child_weight," the building process will stop further partitioning.[21]

The "max_depth" parameter defines the maximum number of partitions, with greater maximum depth increasing the risk of overfitting. The minimum loss reduction required to make a further partition on a leaf node of the tree is defined as "gamma," with a larger "gamma" resulting in a more conservative algorithm. The "subsample" parameter refers to the fraction of observations randomly selected for the training instances, which is inversely related to overfitting. Another parameter useful in avoiding overfitting is "colsample_bytree." Finally, the parameters "lambda" and "alpha" are L2 and L1 regularization terms, respectively, that keep the weights small, thus preventing overfitting.

The random search method was used for model tuning. Random search means that hyperparameters are randomly picked from the predefined searching domain uniformly and the searching does not depend on the previous boosting result. It has been shown to be efficient for problems with high dimensions in some studies.[23]

Model evaluation

In this study, accuracy, sensitivity, and specificity were assessed to evaluate model performance. They are defined in Equations (1)–(3). The goal is to develop a model with high accuracy, sensitivity, and specificity.

$$accuracy = \frac{\sum True\ positive + \sum True\ negative}{\sum Total\ sample} \quad (1)$$

$$sensitivity = \frac{\sum True\ positive}{\sum Predicted\ positive} \quad (2)$$

$$specificity = \frac{\sum True\ negative}{\sum Predictednegative} \quad (3)$$

Gene selection

Two gene importance ranking lists obtained from XGBoost and SHAP values were considered for the selection of candidate genes. We selected the 10 highest-ranking genes common to both lists as the marker genes.

In XGBoost, feature importance is measured using three metrics, namely, gain, cover, and frequency. Gain is the contribution of a feature to the accuracy of the branches on them it is located. Cover measures the relative quantity of observations concerned by a feature. Frequency is a simpler way to measure the Gain. It just counts the number of times a feature is used in all generated trees. We used the gain score to create a ranked list of genes.[24]

The second ranking list was created with SHAP, which explains the prediction of an instance x by computing the contribution of each feature to the prediction. The Shapley value for a feature j is the feature's contribution to the prediction, weighted and summed over all possible feature value combinations that determined through Formula 4:

$$\varnothing_j = \sum_{S \subseteq \{x_1,\ldots,x_p\} \setminus \{x_i\}} \frac{|S|!\left(p-|S|-1\right)!}{p!}\left(v\left(S \bigcup \{x_i\}\right) - v(S)\right) \quad (4)$$

SHAP values can be used as feature importance scores: features with large absolute Shapley values are important. Since we wanted global importance, we averaged the absolute Shapley values per feature across the data (Formula 5):

$$I_j = \sum_{i=1}^{n} |\varnothing_j^{(i)}| \quad (5)$$

SHAP feature importance is an alternative to permutation feature importance. However, while permutation feature importance is based on the decrease in model performance, SHAP is based on the magnitude of feature attributions.[25]

Enrichment analysis

Pathway analysis for the top 10 genes was carried out using REACTOME, an open-source, open access, manually curated, and peer-reviewed pathway database. REACTOME provides intuitive bioinformatics tools for the visualization, interpretation, and analysis of pathway knowledge to support basic and clinical research, genome analysis, modeling, systems biology, and education. Gene ontology analysis was performed with Enrichr web-based tools and services.

Results

Data Preparation

Datasets. In this study, our aim was to screen and mine for specific biomarkers for liver CSCs using the online available data. At the first stage, we obtained the gene expression profile of liver CSCs cell line and tissue including 376 samples (129 liver CSCs and 247 non-CSCs) from published data in the GEO database.

Preprocessing. The ID for each data series, which was captured from its own platforms annotation file, was mapped to an Ensembl ID via the biomaRt package in R to unify the datasets. Then, the data series were merged, followed by a batch correction accomplished using the ComBat function from the sva package. The distributions of expression values before and after the batch effect correction for the combined datasets are shown in Fig. 1a–f. The difference in gene expression distribution between CSCs (Fig. 1a) and non-CSC samples was also magnified after batch effect correction (Fig. 1b). The quantile–quantile (Q–Q) plots (Fig. 1c–d) revealed a decrease in the distance between dots and the normal distribution line after removal of the batch effect. Finally, the PCA plots display the batch effect due to the integration of data from various studies (Fig. 1e), which has been resolved after applying the batch effect correction (Fig. 1f). Therefore, the gene expression data sets would be reliable for the subsequent analysis after batch effect correction.

XGBoost model for a biomarker signature

Model tuning. The hyperparameters adopted in this study were "eta," "nrounds," "max_depth," "gamma," "lambda," "alpha," "min_child_weight," "subsample," and "colsample_bytree." Table 1 presents the search domains and optimal values for the hyperparameters.

Gene selection and models evaluation. We randomly selected 65% of the data to train XGBoost and used the remaining 35% to test the model. Then, we employed a sevenfold cross-validation process to assess model performance stability. Fig. 2a displays the results of cross-validation for the three performance measures, i.e., accuracy, sensitivity, and specificity. The obtained accuracy was 88.68–94.45, sensitivity 86.68–94.11, and specificity 87.89–94.87. Overall, these indicators are significantly high and suggest that XGBoost can be used to model cell classification. XGBoost was finally retrained on the 75% of the training set and tested on the 25% of the testing set. The final performance indicators achieved were as follows: accuracy: 90%, sensitivity: 94%, and specificity: 89%, which is again indicative of significantly high performance.

For gene ranking, we created 1000 models using the same hyperparameters, with each model assigning an importance score to each gene. Then, the median of the 1000 scores for each gene was computed to obtain the average score for the gene. We re-ranked these genes based on SHAP values and the gain scores. The 10 top-ranking genes remained the same in both rankings. We select these genes as a potential biomarker set. The biomarker genes were PTGER3, AURKB, C15orf40, IDI2, OR8D1, NACA2, SERPINB6, L1CAM, SMC1A, and RASGRF1.

To see if the selected marker genes could serve as "universal" markers for cell classification using gene expression data, we trained the XGBoost model using only the selected marker genes. Interestingly, the selected genes did reasonably better than all-gene modeling. The performance indicators achieved were as follows: accuracy: 97%, sensitivity: 100%, and specificity: 95% (Fig. 2b). We also calculated the SHAP value for each marker gene (Fig. 3).

Enrichment analysis

We further tested the top 10 genes for enriched gene ontology terms by the Enrichr web service and analysis tools of the Reactome website for pathways (Fig. 4).

Discussion

Liver cancer is a leading global health issue associated with high morbidity and mortality rate.[26] In recent years, CSCs have been reported to make important contributions to tumor recurrence, progression, and therapeutic resistance. Therefore, therapeutic targeting of liver stem cells is necessary.[5]

In this study, we integrated data series from GEO to identify potential biomarkers for liver CSCs via machine learning classification-based gene selection. One application of integrated gene expression is biomarker discovery. The integration of data from multiple studies increases the sample size by incorporating samples from different cohorts, increasing the statistical power and the robustness of the results. However, it should be mentioned that increasing the sample size reduces the gene count in the integrated data set, resulting in information loss.[9, 27, 28]

Thirteen (13) data series from GEO were integrated producing a total of 385 samples in two groups (CSCs and non-CSCs). Batch correction was conducted with ComBat of the sva R package, which is frequently used in this area.[10, 29-32] We used XGBoost for cell type prediction, as it offers high prediction accuracy and has stronger interpretability owing to its state-of-the-art algorithms. Because of these advantages,

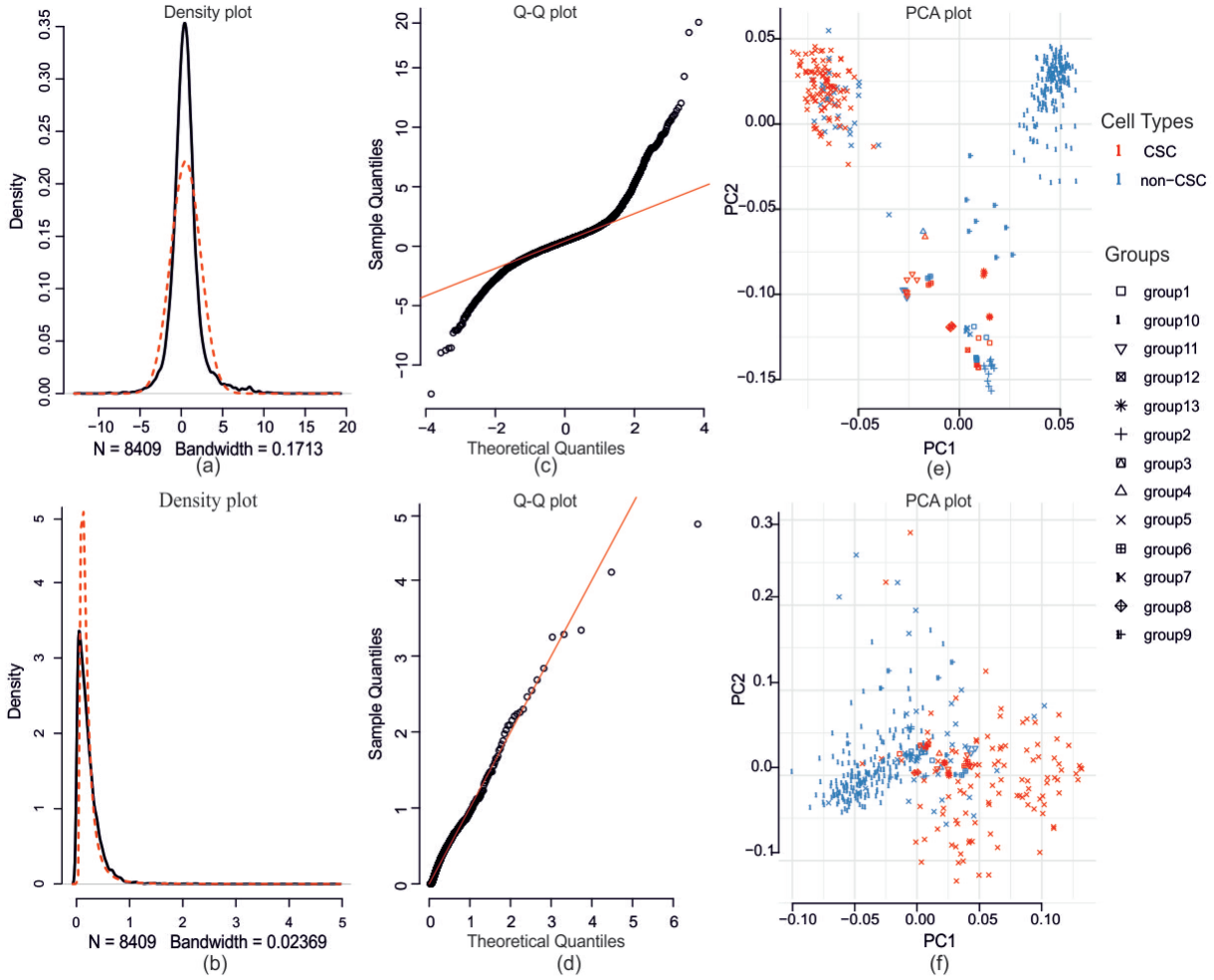

Fig 1. **The density, Q–Q, and PCA plots for evaluating the effect of the batch removal method on overall data.** (a) The density plot before batch effect removal. (b) The density plot after batch effect removal (c) The Q–Q plot before batch effect removal. (d) The Q–Q plot after batch effect removal. (e) The PCA plot before batch effect removal. (f) The PCA plot after batch effect removal. The dashed line in the density plot represents CSCs samples, and the continuous line represents non-CSCs samples. N, the gene number in the combined data set.

Table 1. **Tuned hyperparameter and Searching domain in XGBoost.**

Name	Domain	Transformation function	Optimal Hyperparameter Value
eta	[0.01 , 0.10]	-	0.088
nrounds	[100 , 1000]	-	475
max_depth	[4 , 10]	-	5
gamma	[-1 , 0]	$F(x) = 10^x$	0.62
lambda	[-1 , 1]	$F(x) = 10^x$	1.59
alpha	[-1 , 1]	$F(x) = 10^x$	1.08
min_child_weight	[1 , 12]	-	3.28
subsample	[0.5 , 1]	-	0.61
colsample_bytree	[0.5 , 1]	-	0.85

researchers are increasingly using XGBoost in biomarker discovery.[12, 33, 34] To improve the prediction performance, we tuned hyperparameters using random search, which is more efficient than either a traditional manual or grid search and evaluates more of the search space, especially when the search space has more than three dimensions.[35] As we did not have an external data set to evaluate our model, we performed a sevenfold cross-validation with accuracy as the overall metric and sensitivity and specificity as the class-specific metrics. In our model, the values for the three metrics indicate high performance at both training and testing stages, suggesting that XGBoost can effectively distinguish the two classes.[12, 20]

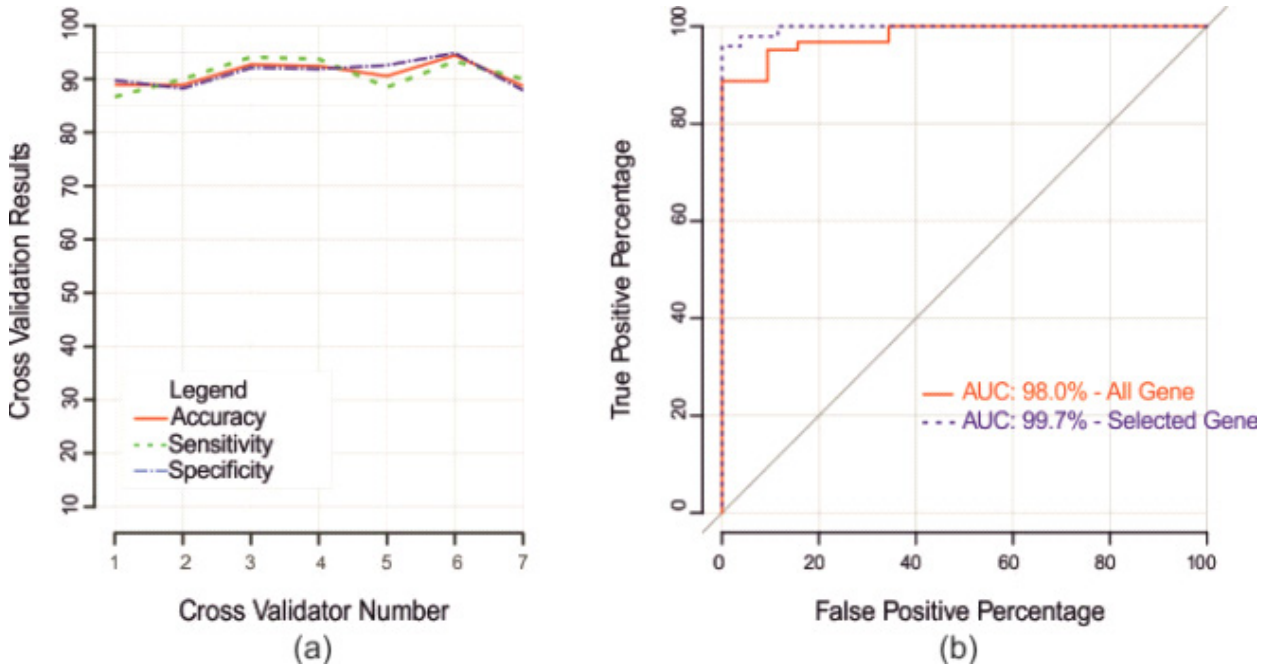

Fig. 2. **Models performance presentation.** (a) The XGBoost model's sevenfold cross-validation plot. (b) Selected gene XGBoost model ROC generates an AUC value greater than that achieved using all-gene XGBoost.

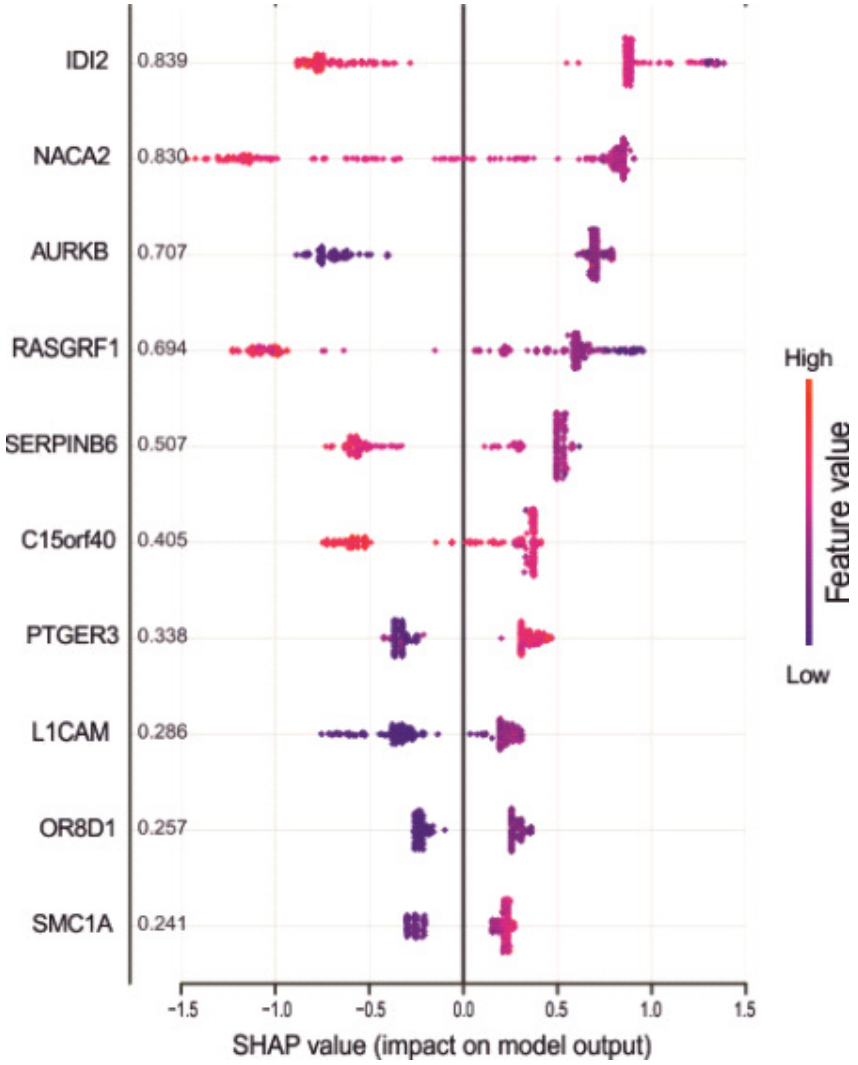

Fig. 3. **SHAP summary plot.** Contribution of each gene to model (XGBoost with top 10 genes) output.

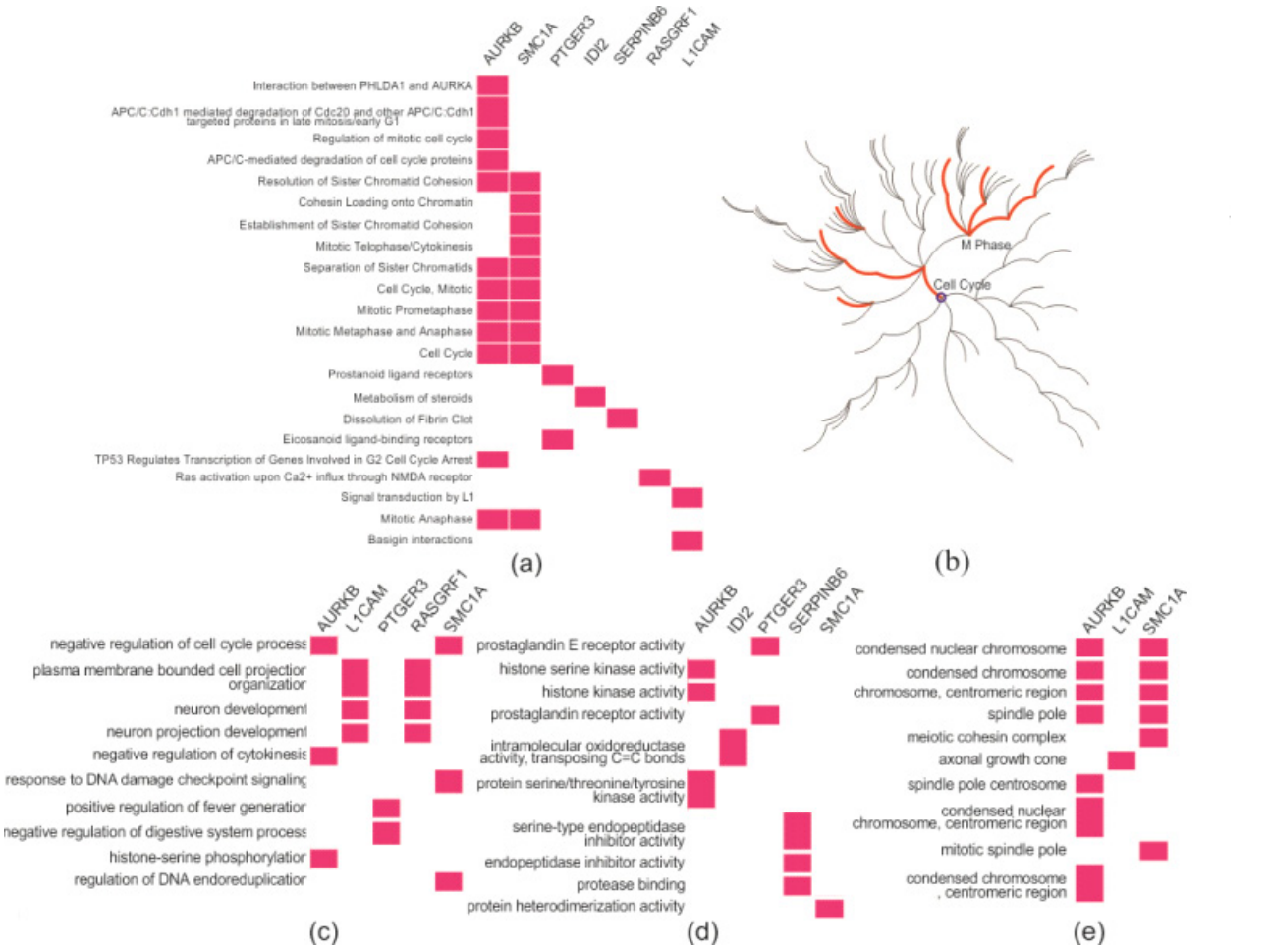

Fig. 4. **Pathway and gene ontology analysis of selected genes.** (a) Selected genes-involved pathways. (b) Overall view of cell cycle pathways in which the selected genes were involved. (c) Selected genes involved in biological processes. (d) Selected genes involved in molecular functions. (e) Selected genes involved in cellular component.

XGBoost assigns an importance score for each feature, which can be used for feature selection. SHAP values also provide a means of ranking features as well as providing a measure of prediction consistency. Therefore, for more certainty, we used both XGBoost importance and SHAP values.[13, 20] The SHAP values and gain scores for genes were imputed, as described by Riberio et al and Chen et al, respectively.[19, 21]

We created 1000 XGBoost models by the same tuning parameters and then obtained feature importance score (gain) and SHAP values for each of the 1000 models. Finally, the median of 1000 scores was calculated for each gene. These scores are reliable and can be used as gene ranks.[12] Genes were ordered separately by gain score and SHAP value. The top 10 genes from the SHAP list were selected, 9 of which were also among the top 10 genes in the gain score list and the remaining 1 corresponded to the 11th gene in this list. So, the 10 highest-ranking genes from the SHAP list were selected as the biomarker set, including PTGER3, AURKB, C15orf40, IDI2, OR8D1, NACA2, SERPINB6, L1CAM, SMC1A, and RASGRF1. To ensure that this gene set can be used as a biomarker set, we trained our model using only these 10 genes, which offered better prediction performance compared with all-gene models.

Many studies have shown that CSCs have one or more abnormalities in signaling pathways that regulate cell cycle and self-renewal. The cellular pathways in which the key genes are most involved are pathways associated with cell cycle regulation.[36] For example, the aurora kinase b (Aurkb)-protein phosphatase 1 (PP1) axis has been shown to mediate the resetting of Oct4 during the cell cycle in embryonic stem cells. Aurkb-PP1 axis also plays a critical role in cell cycle-dependent changes in kinetochore assembly by regulating the balance between phosphorylation and dephosphorylation of kinetochore substrates.[37]

SMC1A has also a key role in tumor metastasis and resistance to radiation therapy. This gene is associated with CSCs, epithelial-to-mesenchymal transition, and DNA–damage response pathways. Yadav et al demonstrated that suppression of SMC1A expression reduces the self-renewal capacity of prostate cancer cells.[38] PTGER3 induces tumorigenesis and drug resistance in ovarian cancer.[36] LMBR1 is a regulator of nuclear stemness marker BMI1 in gastrointestinal stromal tumors.[39] Another study has reported PFKL to play a vital role in the maintenance of CSC-like phenotype in hepatocellular carcinoma.[40] L1CAM has been implicated in maintaining the growth and survival of CD133+ glioma cells both in vitro and in vivo and has been suggested to be a CSC-specific therapeutic target for improving the treatment of malignant gliomas and other brain tumors.[41]

Conclusions

These 10 key genes were found to play important roles in liver CSC maintenance. It seems that AURKB is more important for controlling the stemness and may help in the treatment of liver cancer. This gene may be a therapeutic target for inhibiting liver cancer stemness characteristics. However, this conclusion is based on retrospective data, and validation of these findings warrants further biological studies.

Acknowledgments

This work supported by the Department of Health Information Management, School of Allied Medical Science, Tehran University of Medical Sciences.

Conflict of interest

The authors of this paper declare that they do not have any conflict of interest.

References

1. Dasgupta P, Henshaw C, Youlden DR, et al. Global trends in incidence rates of primary adult liver cancers: A systematic review and meta-analysis. Front Oncol 2020;10:171.
2. Bai KH, He SY, Shu LL, et al. Identification of cancer stem cell characteristics in liver hepatocellular carcinoma by WGCNA analysis of transcriptome stemness index. Cancer Med 2020.
3. Prasad S, Ramachandran S, Gupta N, et al. Cancer cells stemness: a doorstep to targeted therapy. Biochim Biophys Acta (BBA)-Mol Basis Dis 2020;1866:165424.
4. Shibata M and Hoque MO. Targeting cancer stem cells: a strategy for effective eradication of cancer. Cancers 2019; 11: 732.
5. Xiang Y, Yang T, Pang B-y, et al. The progress and prospects of putative biomarkers for liver cancer stem cells in hepatocellular carcinoma. Stem Cells Int 2016;2016.
6. Najafi M, Farhood B and Mortezaee K. Cancer stem cells (CSCs) in cancer progression and therapy. J Cell Physiol 2019;234:8381-8395.
7. Chang N-W, Dai H-J, Shih Y-Y, et al. Biomarker identification of hepatocellular carcinoma using a methodical literature mining strategy. Database 2017;2017.
8. Shahrjooihaghighi A, Frigui H, Zhang X, et al. An ensemble feature selection method for biomarker discovery. In: 2017 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT) 2017, pp. 416-421. IEEE.
9. Toro-Domínguez D, Martorell-Marugán J, López-Domínguez R, et al. ImaGEO: integrative gene expression meta-analysis from GEO database. Bioinformatics 2019;35:880-882.
10. Diegues I, Vinga S and Lopes MB. Identification of common gene signatures in microarray and RNA-sequencing data using network-based regularization. In: International Work-Conference on Bioinformatics and Biomedical Engineering 2020, pp.15–26. Springer.
11. Shukla AK, Singh P and Vardhan M. A two-stage gene selection method for biomarker discovery from microarray data for cancer classification. Chemomet Intell Lab Syst 2018;183:47-58.
12. Li Y, Umbach DM, Bingham A, et al. Putative biomarkers for predicting tumor sample purity based on gene expression data. BMC Genom 2019;20:1-12.
13. Lundberg SM and Lee S-I. Consistent feature attribution for tree ensembles. arXiv preprint arXiv:170606060 2017.
14. Guthrie NL, Carpenter J, Edwards KL, et al. Emergence of digital biomarkers to predict and modify treatment efficacy: machine learning study. BMJ Open 2019;9:e030710.
15. Caly H, Rabiei H, Coste-Mazeau P, et al. Pregnancy data enable identification of relevant biomarkers and a partial prognosis of autism at birth. bioRxiv 2020.
16. Hathaway QA, Roth SM, Pinti MV, et al. Machine-learning to stratify diabetic patients using novel cardiac biomarkers and integrative genomics. Cardiovasc Diabetol 2019;18:78.
17. Lundberg SM and Lee S-I. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 2017, pp.4765-4774.
18. Štrumbelj E and Kononenko I. Explaining prediction models and individual predictions with feature contributions. Knowl Inform Syst 2014;41:647-665.
19. Ribeiro MT, Singh S and Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining 2016, pp.1135-1144.
20. Parsa AB, Movahedi A, Taghipour H, et al. Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. Accident Anal Prev 2020;136:105405.
21. Chen T and Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining 2016, pp.785-794.
22. Werpachowski R, György A and Szepesvári C. Detecting overfitting via adversarial examples. In: Advances in Neural Information Processing Systems 2019, pp.7858-7868.
23. Wang Y and Ni XS. A XGBoost risk model via feature selection and Bayesian hyper-parameter optimization. arXiv preprint arXiv:190108433 2019.
24. developers x. Understand your dataset with XGBoost R-Project. https://xgboost.readthedocs.io/en/latest/R-package/discoverYourData.html (2020).
25. Molnar C. Interpretable machine learning. Lulu com 2019.
26. Wang W, Smits R, Hao H, et al. Wnt/β-catenin signaling in liver cancers. Cancers 2019;11:926.
27. Toro-Domínguez D, Villatoro-García JA, Martorell-Marugán J, et al. A survey of gene expression meta-analysis: Methods and applications. Brief Bioinform 2020.
28. Ghosheh N, Küppers-Munther B, Asplund A, et al. Human pluripotent stem cell-derived hepatocytes show higher transcriptional correlation with adult liver tissue than with fetal liver tissue. ACS Omega 2020;5:4816-4827.
29. Bai J, Zhang X, Kang X, et al. Screening of core genes and pathways in breast cancer development via comprehensive analysis of multi gene expression datasets. Oncol Lett 2019;18:5821-5830.
30. Xia L, Su X, Shen J, et al. ANLN functions as a key candidate gene in cervical cancer as determined by integrated bioinformatic analysis. Cancer Manage Res 2018;10:663.
31. Kuang Y, Wang Y, Zhai W, et al. Genome-wide analysis of methylation-driven genes and identification of an eight-gene panel for prognosis prediction in breast cancer. Front Genet 2020;11:301.
32. Guo T, Ma H and Zhou Y. Bioinformatics analysis of microarray data to identify the candidate biomarkers of lung adenocarcinoma. PeerJ 2019;7:e7313.
33. Zhang X, Li T, Wang J, et al. Identification of cancer-related long non-coding RNAs using XGBoost with high accuracy. Front Genet 2019;10:735.
34. Ding W, Chen G and Shi T. Integrative analysis identifies potential DNA methylation biomarkers for pan-cancer diagnosis and prognosis. Epigenetics 2019;14:67-80.
35. Si M, Xiong Y, Du S, et al. Evaluation and calibration of a low-cost particle sensor in ambient conditions using machine-learning methods. Atmos Measure Tech 2020;13.
36. Rodriguez-Aguayo C, Bayraktar E, Ivan C, et al. PTGER3 induces ovary tumorigenesis and confers resistance to cisplatin therapy through up-regulation Ras-MAPK/Erk-ETS1-ELK1/CFTR1 axis. EBioMedicine 2019;40:290-304.
37. Shin J, Kim TW, Kim H, et al. Aurkb/PP1-mediated resetting of Oct4 during the cell cycle determines the identity of embryonic stem cells. Elife 2016;5:e10877.
38. Yadav S, Kowolik CM, Lin M, et al. SMC1A is associated with radioresistance in prostate cancer and acts by regulating epithelial-mesenchymal transition and cancer stem-like properties. Mol Carcinogen 2019;58:113-125.
39. Bai C, Liu X, Xu J, et al. Expression profiles of stemness genes in gastrointestinal stromal tumor. Human Pathol 2018;76:76-84.
40. Lin S-H, Liu T, Ming X, et al. Regulatory role of hexosamine biosynthetic pathway on hepatic cancer stem cell marker CD133 under low glucose conditions. Scient Rep 2016;6:1-10.
41. Bao S, Wu Q, Li Z, et al. Targeting cancer stem cells through L1CAM suppresses glioma growth. Cancer Res 2008;68:6043-6048.

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported License which allows users to read, copy, distribute and make derivative works for non-commercial purposes from the material, as long as the author of the original work is cited properly.