Explainable Biomarker Discovery in Hepatocellular Carcinoma: an XGBoost and SHAP Framework

Guangrui Xu [1*], Xiaoyu Shen [2]

{202300172014@mail.edu.sdu.cn*}

Intelligent Medical Engineering, Shandong University, Jinan, Shandong, China [1], Central University of Finance and Economics - University of New South Wales, Beijing, China[2]

Abstract. Hepatocellular carcinoma (HCC) is one of the leading causes of cancer-related deaths worldwide. Although ensemble machine learning models have demonstrated high accuracy in HCC classification and prediction, their "black-box" nature impedes their reliable adoption in clinical practice. Therefore, this study proposes a two-stage research framework: first, constructing a high-accuracy HCC risk prediction model using the extreme gradient boosting algorithm; second, applying the SHAP method to provide global and local interpretations of the model's predictions, quantifying the contribution of each feature to the predictive outcome, thereby identifying key features. The research is based on the HCC clinical feature dataset, while also drawing on experiences from multi-center databases such as TCGA. This work aims to provide clinicians with transparent decision support, promote personalized treatment, and offer clear target maps for subsequent biological validation studies.

Keywords: XGBoost, SHAP, Hepatocellular Carcinoma.

1 Introduction

Hepatocellular carcinoma (HCC) is the sixth most prevalent malignant disease in the world and the third leading cause of cancer-related mortality. There is a huge difference between early-stage and late-stage patients, which is why it is highly important to observe patients at the earliest possible stage and enhance the effectiveness of the diagnostic process and patient stratification [1, 2]. However, in recent years, artificial intelligence has become very promising in high-dimensional clinical, imaging, and genomic data analysis to enhance the accuracy of HCC diagnosis and treatment [3, 4]. Among them, ensemble learning models (e.g., XGBoost) are often used to offer higher predictive efficiencies learning with multiple base learners. Nonetheless, sophisticated AI models in the form of deep neural networks and sophisticated ensemble models are literally black boxes and their decision-making processes are opaque [5-7]. The problem of interpretability becomes devastating in the case of high-risk clinical decision-making, such as HCC, as it hinders the trustworthiness of the models by the physician and reduces the application of physicians to clinical settings. Thus, the creation of explainable artificial intelligence, where model predictions are provided in the form comprehensible to

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365119

humans, became one of the priorities in ensuring the responsible use of AI in medical practice [8-10].

In an effort to overcome the above challenges, this paper outlines a research framework to be used in interpretable AI with the view of identifying liver cancer biomarkers. The main idea lies in the synthesis of the two notions, namely, the high-performance prediction and the post-hoc interpretability. In particular: 1) Model Construction: To select the XGBoost algorithm as the central classifier, this paper chose it because this algorithm can be used to analyze tabular data and avoid overfitting. 2) Interpretation and Discovery: this paper uses SHAP which is a model-agnostic post-hoc interpretation method. As a result of game theory SHAP values can reasonably estimate the value that particular features in the dataset make to single prediction. This can not only gain a global ranking of the most significant features (i.e., candidate biomarkers) but also gain an insight into the pattern-driven by features underlying the prediction of a particular patient. 3) Data and validation: The present study will confirm the methodology and case study the publicly available HCC clinical data. The author will adhere to a strict machine learning workflow and assess the performance of the models using cross-validation and performance measurements. In the end, the SHAP analysis will produce a report on biomarker discovery that is decipherable.

2 Related Work

This part examines and compares available methods of machine learning that are applied in biomarker discovery with the aim of evaluating their strengths and weaknesses and offers a basis upon which to place the proposed method in this paper.

2.1 Comparison and Limitations of the Traditional Statistical Models and High-performance Black-Box Models

Early survival analysis and risk models Cox proportional hazards models, logistic regression and LASSO regression are commonly used to analyze early survival and risk based on genomic or clinical variables2. These are easy to understand and are much more interpretable with model coefficients directly proportional to feature hazard ratios or association strength, which aligns with the traditional medical research, but cannot be applied to high-dimensional data and the multi-omics interaction complexity is limited; the performance of these models can be easily outperformed by machine learning models when using multi-omics data10. Conversely, deep neural networks and sophisticated ensemble models (e.g., uninterpretable XGBoost and random forests) have been used to make HCC imaging diagnoses, predict outcome and response to therapy, in which high AUC scores have often been reported2, 10. They are more predictive as they are able to learn the nonlinear patterns and interactions in the data that might be complex, and they can automatically do this without special human intervention which is apparent when the data is large and high-dimensional. However, the apoliteness of interpreting is a significant obstacle in clinical practice since doctors cannot know the foundation of choices, which is why discovering faults or prejudice and eliciting novel biological understanding is challenging.

2.2 Trend in the Integration Pathway and Clinical Practicality of Explainable AI

In order to achieve a tradeoff between performance and interpretability, explainable AI integration techniques have now been made mainstream, with intrinsic interpretable techniques (e.g., decision trees) and post-hoc interpretation techniques (e.g., LIME, SHAP, etc.) being the most popular representatives of these approaches. In particular, model-agnostic algorithms such as SHAP can be used in any complex model, and provide a single, theory-backed and theory-inspired measure of contributions to features and a visualization that can be easy to understand. XGBoost with SHAP has also been successfully used in the discovery of biomarkers in HCC despite being costly to compute, may have misleading interpretations, and may not be able to present interpretations to clinicians smoothly, conveniently and unambiguously8, 5, 6. The state-of-the-art is now extended to multimodal integrated XAI, which comprises the multi-omics data fusion, algorithmic fairness-based XAI (e.g., FairShap), and frameworks of clinical effectiveness of XAI interpretations. These methods have the benefit to mine an in-depth perspective of ailing conditions in the form of multimodal information, yielding more complex biological understanding, and clinical usefulness. Nevertheless, heterogeneity of data and challenges in its fusion as well as the complexity of methods used demand co-operation on an interdisciplinary level. Recent multimodal XAI models combined with clinical, imaging, and immunological characteristics have shown the ability to enhance performance in detecting HCC9. Positioning: In short, the effort of this paper fits in the category of explainable AI integration approach mentioned above and is geared towards the goal of achieving clinical practicality. To demonstrate the end-to-end process of biomarker discovery, this research selects XGBoost-SHAP integration, which is a comparatively advanced technological couple that has successfully been applied to clinical data pools in the last few years, to discover the biomarkers.

3 Methods

3.1 Overall Study Design

This research has a retrospective case-control design. The main part is to build and explain a machine learning model that can predict or classify HCC. In general, the process of work is as follows: preliminary clinical data is first preprocessed. The feature engineering and dataset partitioning are followed next. After that, an XGBoost model is trained and optimized, with the performance of the model being reviewed. According to this, the SHAP technique is used in the interpretation of the model, both globally and locally. Lastly, the findings go through an in-depth analysis and validation.

3.2 Data Preprocessing and Sources

The Kaggle dataset "Predict Liver Cancer from Clinical Features, " was the major source of primary data used as the central focus of analysis. This data contains a number of clinical and demographic variables of HCC patients. In the dataset preprocessing the following measures were implemented to maintain the quality of the model: In the first place, there is Missing value handling where the median (in case of numerical variables) or mode (in case of categorical variables) was used to impute the missing values; In the second place, there is the Missing values handling where features with missing rates exceeding 40% were excluded. Second is Outlier handling: There was the use of statistical outlier handling (IQR rule) to identify outliers. Third

is Feature encoding: Categorical variables (e.g., gender, stages of liver cirrhosis) have been coded with one-hot encoding or label encoding. Fourth Data normalization: Continuous numerical variables were normalized (using Z-score) with the aim of removing the effect of different scales. Fifth is Dataset splitting: The dataset was split randomly to create training set and an independent test set in the ratio of about 8:2 to allow uniformity in allocation.

3.3 XGBoost Model Construction

XGBoost is an effective tree ensemble algorithm, which continuously combines decision trees in order to correct the residue of the last tree. Not only does it have inbuilt regularization against overfitting, but it also inherently offers an initial indication of the importance of the features based on so-called gain, " cover, or frequency, which in turn forms the basis of further SHAP interpretation. In this research, 5-fold or 10-fold cross-validation, grid search or random search was employed to tune important hyperparameters. Parameters to be optimised are: learning rate (learning rate, learning rate), maximum tree depth (maximum depth, max depth), number of trees (n estimators, n estimators), subsample ratio and column sampling by tree (subsample, colsample by tree), and L1 L2 regularisation terms (reg alpha, reg Lambda). The author trained the final model on the training set after tuning the training parameters, and measured it on an unseen test set. The evaluation measures were made to be comprehensive (they comprised accuracy, precision, recall (sensitivity), F1 score, and area under the receiver operating characteristic curve (AUC-ROC). The major indicator of the discriminative power of binary classification models is AUC-ROC.

3.4 SHAP Interpretability Analysis and Discovery of biomarkers

SHAP values of every sample and every feature were calculated by using the SHAP library in Python. In the case of tree-based models, an efficient TreeExplainer algorithm is used to make computations faster. Each patient's SHAP force waterfall plot was created to understand how the predictions were made at a particular patient, indicating the contribution made by each feature to the overall movement of the base prediction (the contribution made by all features). This assists the clinicians to know the reason why a patient is expected to be high- or low-risk. Lastly the combination of the global SHAP summary plot rankings and directionality of features in the prediction outcome (positive/negative relationship) allows the identification of top-ranked features that are intensive to clinical knowledge or can contribute novel information as candidate biomarkers.

4 Results

4.1 Comparison of Experimental Results

The results of the experiment will be compared against the results of other similar studies conducted in the past.

In order to confirm the quality of the XGBoost-SHAP framework applied in the framework of the present research paper, this paper contrasted the predictive capabilities of this structure with the predictive capabilities of various baseline models. The identical dataset splits, preprocessing steps, and evaluation metrics were common in all the models. Table 1 and Table 2 demonstrate the model description and the results of the experiment.

Table 1. Comparative Models Description

Model	Description	Expected Advantages	Expected Disadvantages	Aim
Logistic Regression (LR)	Generalized linear model providing feature coefficients.	Highly interpretable, computationally efficient, serves as a strong baseline.	Unable to capture complex nonlinear relationships.	Performance comparison baseline to highlight the necessity of complex models.
Random Forest (RF)	Bagging ensemble method constructing multiple decision trees in parallel.	Strong resistance to overfitting, can provide feature importance.	Model size can be large, interpretability weaker than a single decision tree.	Comparison with XGBoost (Boosting) as another ensemble learning method.
Support Vector Machine (SVM)	Seeks the optimal separating hyperplane in high-dimensional space.	May perform excellently on high-dimensional, small-sample data.	Sensitive to parameter and kernel function selection, slow training on large-scale data.	Comparison as a classic machine learning representative.
XGBoost (Proposed)	Tree ensemble algorithm under the Gradient Boosting framework.	Predictive performance is typically leading, with built-in regularization to prevent overfitting.	Numerous hyperparameters requiring careful tuning.	Serves as the core model to evaluate its performance upper limit.

Table 2. Comparison of Results of Performance in the Models

Model Name	Accuracy	AUC-ROC	Precision	Recall	F1 Score
Logistic Regression (LR)	0.7500	0.8368	0.4831	0.6694	0.5614
Random Forest (RF)	0.8370	0.8747	0.6038	0.7198	0.6568
Support Vector Machine (SVM)	0.7620	0.8526	0.4708	0.7788	0.5868
XGBoost (Proposed)	0.8510	0.8730	0.6828	0.5853	0.6303

4.2 Results of the Ablation Experiment

In order to validate the efficiency of the SHAP explanation module and the significance of various feature sets in the model performance, this study has formulated ablation experiments.

Experiment 1: SHAP Consistency of Explainability.

This experiment will be directed to check the reasonableness of the SHAP explanations (Figure 1). This paper serially took out the most significant feature by contribution of age and the least

significant feature by contribution of hepatitis B (Bottom 1) based on the global SHAP beaswarm plot and monitored the alterations in model performance. The results of the consistency validation of SHAP in a table are provided in Table 3.

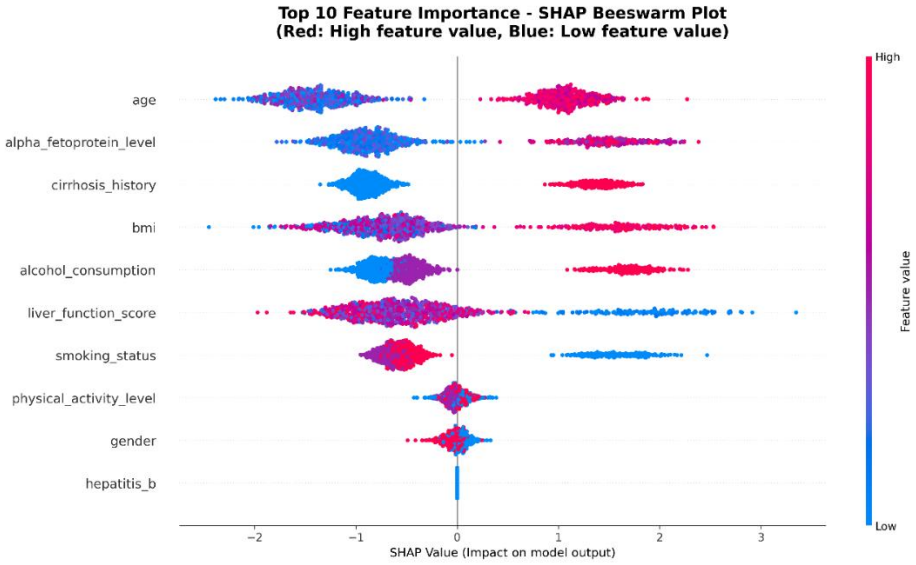

Fig. 1. Shap Beeswarm Plot (Picture credit: Original)

Table 3. SHAP Explanation Consistency Validation Results

Feature Set Configuration	Accuracy	AUC-ROC	Number of Retained Features	Performance Change Analysis
All Features (Baseline)	0.8510	0.8730	13	Baseline performance
Remove SHAP Top 1 Feature	0.8360	0.8353	12	Accuracy decreased by 0.015, AUC change: -0.0377
Remove SHAP Bottom 1 Feature	0.8540	0.8721	12	Accuracy increased by 0.003, AUC change: -0.0009

Experiment 2: SHAP-based Feature Selection Effectiveness Criterion validation.

The given experiment is going to discover the possibility of SHAP as a feature selection tool. This study discarded the lowest K features and retained the remaining K features to retrain the XGBoost model according to the SHAP importance. The effectiveness of feature selection as shown in Table 4 is validated on SHAP basis.

Table 4. Validation Results of Feature Selection Effectiveness Based on SHAP

Feature Selection Method	Number of Selected Features (K)	Accuracy	AUC-ROC	Feature Dimensionality Reduction Rate
Baseline: All Features	13	0.8510	0.8730	0%
SHAP Importance Top K	5	0.8340	0.8290	61.5%
SHAP Importance Top K	10	0.8450	0.8727	76.9%

5 Discussion

To build a strong baseline model the algorithm of XGBoost was employed as the fundamental classifier that demonstrated high predictive power in the detection of liver cancer. The XGBoost model continues to present a better performance than the traditional machine learning models by offering the XGBoost greater accuracy of 0.8510 and AUC-ROC of 0.8730, as indicated in Table 2. More specifically, XGBoost was also found to be more precise (0.6828) and F1 (0.6303), which means it is more specific and overall discriminating, with high sensitivity, which will support further interpretation analysis.

To learn more about the biological workings of model decision-making, SHAP (Shapley Additive exPlanations) values are presented as an importance evaluation method to operationalize the local and global interpretability analysis of model predictions. The SHAP value uses the Shapley value of game theory, which can reasonably apportion the value of every feature to the model result, therefore has a theoretical rigor and quantifiable foundation of biomarker screening.

This was done in the experiment of feature selection (see Table 4), where the impact of this of the Top-K feature subset as a result of SHAP importance ranking on a model performance was systematically evaluated. As can be seen in the obtained experimental outcomes, the model can be trained to reach a score of 0.8450 and AUC-ROC of 0.8727 (nearly identical to the baseline model using all 13 features) having retained only the 10 features with the largest SHAP score (reduction rate of features used 76.9). The outcome of the result is strong evidence that there are not many distinguishing attributes that possess the primary discriminant data within the diagnostic task of liver cancer, and the others are possibly either redundant or even noisy. Moreover, a reduction in the model performance is observed when only top-5 features are used (accuracy 0.8340, AUC 0.8290) but this figure is within the clinically reasonable range that indicates that it is possible to construct efficient and cost-effective liver cancer screening models using very few features.

In this study, an ablation experiment was developed to confirm the biological rationality of SHAP interpretation (see Table 3). The findings of the experiment prove that by removing the top-1 feature with the highest SHAP score the model score becomes smaller by 0.015 and AUC-ROC is reduced by far greater numbers 0.0377, which indicates that this feature has a decisive impact on the prediction of the model. Nevertheless, any one of the Bottom-1 features with the smallest SHAP score would raise the accuracy of the model by 0.003, and reduce the AUC by 0.0009 (nearly negligible). The outcomes of this sympathetic verification are, in a significant

way, in agreement with the SHAP theory which completely substantiates the credibility, as well as the sensitivity of SHAP method in revealing the primary biomarkers.

6 Conclusion

Overall, this paper has managed to develop a framework of interpretable artificial intelligence for predicting biomarkers in hepatocellular carcinoma that combines an XGBoost algorithm with the SHAP interpretability analysis. Suggested XGBoost model has better predictive accuracy of 0.8510, and AUC-ROC of 0.8730 compared to old machine learning models like logistic regression, random forest and support vector machine and these models. More to the point, using SHAP values, this paper will give both the global and local explanations of how the model makes its decision and convert the predictions of the black-box model into a clinical interpretation.

The SHAP-based interpretability analysis not only confirms the biological relevance of the specified key features, but also allows obtaining a ranked list of possible biomarkers. This study shows that a small set of features, including those that are the top 10 features in this case study, can be used to gather most of the discriminatory power, which implies that SHAP is very potent in simplifying clinically actionable information on high-dimensional data. Cutting top-contributing features had a big influence on the performance of the model, but cutting low-contributing features did not, which again confirms the validity of SHAP in identifying biologically significant variables.

The study fills the gap existing between high-performance prediction and clinically relevant mechanistic interpretation by providing a clear and clinically implemented tool in the risk stratification of HCC. The list of candidate biomarkers that have been identified allows one to get a clear-cut roadmap of further experimental validation through experimental methods like qPCR, immunohistochemistry, or mass spectrometry. Finally, this framework not only increases physician confidence in AI-based clinical support systems, but also opens the path to individual diagnosis and specific therapeutic interventions in hepatocellular carcinoma.

References

[1] Sung, H., et al.: Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 71(3). pp. 209–249 (2021)

[2] Feng, S., et al.: Current Status and Analysis of Machine Learning in Hepatocellular Carcinoma. J Clin Transl Hepatol. 11(5). pp. 1184–1191 (2023)

[3] Pan, Q., et al.: Explainable machine learning model identified potential biomarkers in liver cancer survival prediction. Biomed Signal Process Control. 96. 106504 (2024)

[4] Yagin, F.H., et al.: Metabolomics Biomarker Discovery to Optimize Hepatocellular Carcinoma Diagnosis: Methodology Integrating AutoML and Explainable Artificial Intelligence. Diagnostics. 14(18). 2049 (2024)

[5] Wang, R., et al.: A transformer-based deep learning survival prediction model and an explainable XGBoost anti-PD-1/PD-L1 outcome prediction model based on the cGAS-STING-centered pathways in hepatocellular carcinoma. Brief Bioinform. 26(1). bbae686 (2025)

[6] Lundberg, S.M., Lee, S.I.: A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems. pp. 4765–4774 (2017)
[7] Sun, C., et al.: Successful construction of a "Liver Cancer Immune Early Warning System" for predicting HCC recurrence. Nature. (2025)
[8] Gutierrez-Chakraborty, E., et al.: Discovering novel prognostic biomarkers of hepatocellular carcinoma using eXplainable Artificial Intelligence. Expert Syst Appl. 255. 124456 (2024)
[9] Liang, T., et al.: Artificially Intelligent Model for Accurate Detection of HCC by Integrating Clinical, Radiological, and Peripheral Immunological Features. ClinicalTrials.gov. NCT06637059 (2024)
[10] Kandoriya, J., Degadwala, S.: A Comprehensive Review on Machine Learning Methods for Categorizing Liver Tumors. Int J Sci Res Comput Sci Eng Inf Technol. 9(10). pp. 332–338 (2023)