

A Machine Learning Model Based on Health Records for Predicting Recurrence After Microwave Ablation of Hepatocellular Carcinoma

Chao An^{1,*}, Hongcai Yang^{1,2,*}, Xiaoling Yu¹, Zhi-Yu Han¹, Zhigang Cheng¹, Fangyi Liu¹, Jianping Dou¹, Bing Li³, Yansheng Li⁴, Yichao Li⁴, Jie Yu¹, Ping Liang¹

¹Department of Ultrasound, PLA Medical College & 5th Medical Center of Chinese PLA General Hospital, Beijing, 100853, People's Republic of China;

²School of Medicine, Nankai University, Tianjin, People's Republic of China; ³National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing, People's Republic of China; ⁴DHC Mediway Technology CO, Ltd, Beijing, People's Republic of China

*These authors contributed equally to this work

Correspondence: Ping Liang; Jie Yu, Department of Ultrasound, PLA Medical College & 5th Medical Center of Chinese PLA General Hospital, Beijing, 100853, People's Republic of China, Tel +86-10-66939530, Fax +86-10-68161218, Email liangping301@126.com; yu-jie301@hotmail.com

Background and Aim: Early recurrence (ER) presents a challenge for the survival prognosis of patients with hepatocellular carcinoma (HCC). The aim of this study was to investigate machine learning (ML) models using clinical data for predicting ER after microwave ablation (MWA).

Methods: Between August 2005 and December 2019, 1574 patients with early-stage HCC underwent MWA at four hospitals were reviewed. Then, 36 clinical data points per patient were collected, and the patients were assigned to the training, internal, and external validation set. Apart from traditional logistic regression (LR), three ML models—random forest, support vector machine, and eXtreme Gradient Boosting (XGBoost)—were built and validated for their predictive ability with the area under ROC curve (AUC). Algorithms such as SHapley Additive exPlanations (SHAP) and local interpretable model-agnostic explanations (LIME) were used to realize their interpretability.

Results: The three ML models all outperformed LR ($P < 0.001$ for all) in predictive ability. When nine variables (tumor number, platelet, α -fetoprotein, comorbidity score, white blood cell, cholinesterase, prothrombin time, neutrophils, and etiology) were extracted simultaneously using recursive feature elimination with cross-validation, the XGBoost model achieved the best discrimination among all models, with an AUC value 0.75 (95% CI [confidence interval]: 0.72–0.78) in the training set, 0.74 (95% CI: 0.69–0.80) in the internal validation set, and 0.76 (95% CI: 0.70–0.82) in the external validation set, and it was interpreted depending on the visualization of risk factors by the SHAP and LIME algorithms. The predictive system of post-ablation recurrence risk stratification was provided on online (<http://114.251.235.51:8001/>) based on XGboost analysis.

Conclusion: The XGBoost model based on clinical data can effectively predict ER risk after MWA, which can contribute to surveillance, prevention, and treatment strategies for HCC.

Keywords: microwave ablation, hepatocellular carcinoma, recurrence, machine learning, risk stratification

Lay Summary

Microwave ablation (MWA) is an effective treatment for patients with early-stage hepatocellular carcinoma (HCC). However, in fact, the early hepatic recurrence rate including local recurrence and hepatic distant recurrence can reach to approximately 30%. Moreover, 5- or 10-mm ablative margin is always related to local recurrence, which attract the attention of scholars. Because of the lack of highly reliable model that predicts post-ablation recurrence, we developed an artificial intelligence system based on machine learning (ML) models using large clinical datasets extracted directly from electronic health records to predict recurrence and provide guidance for MWA. In particular, we used the SHapley Additive exPlanations and local interpretable model-agnostic explanations algorithms to realize the interpretability of the

ML model. This ML system can improve the likely success of the MWA treatment and guide clinical management after thermal ablation of HCC.

Introduction

Hepatocellular carcinoma (HCC), the fifth most common malignancy, results from hepatitis viral infections and causes considerable mortality and morbidity in China.^{1–3} Locoregional ablation (LA) is recommended as a first-line treatment in patients with early-stage HCC, as well for surgical resection (SR) and liver transplantation (LT),^{4–6} according to multiple international guidelines. In addition, LA as a primary and alternative treatment is also widely applied in patients with HCC who are ineligible for surgery, and is used as bridge to LT,^{7,8} with the advantage of less trauma, better repeatability, fewer complications, and better cost-effectiveness.

Microwave ablation (MWA) is an effective LA technology owing to the higher intratumoral temperature and a more massive region of cell necrosis, with a comparable 5-year overall survival (OS) for early-stage HCC, similar to SR.^{9,10} MWA is a curative option for small HCC lesions, but the post-ablation recurrence occurred commonly by the progression of residual de novo carcinoma and dissemination of microscopic lesions. Unfortunately, the approximately 50% of 5-years intrahepatic recurrence (local recurrence and distant recurrence) rate after MWA remains a great challenge in clinical practice. Especially, early recurrence (ER) (considered generally less than 2 year) may cause poor prognosis. Among them, while local recurrence rates are low in experienced hands the rate is variable, being able to predict who is likely to have local recurrence could allow for improved treatment options. For example, post-ablation recurrence risk stratification can help interventional radiologists design the ablation strategy and post-ablation medication (eg, preventive transcatheter arterial chemoembolization [TACE] or multi-targeted tyrosine kinase inhibitors [TKIs]).¹¹

A large amount of clinical data can be obtained easily from an electronic medical record system. However, these data interfere with each other commonly and weaken the predictive power of traditional statistical methods (eg, Cox regression and logistic regression [LR]) because of their poor calculation ability to analyze multidimensional clinical features. It is difficult to accurately identify the optimal biological indicators for recurrence. Machine learning (ML) is a branch of artificial intelligence that employs statistical, probabilistic, and optimization techniques to train a machine how to learn.^{12,13} ML algorithms can learn from clinical data, identify patterns, and make decisions with minimal human intervention by automating analytical model building, which has been used to calculate the probability of the future outcome of a specific individual so as to obtain risk prediction models.

In this study, we developed and validated multiple ML models using clinical data extracted directly from electronic health records to predict ER risk after MWA for early-stage HCC, with the aim of building a favorable method to provide guidance for MWA indication selection and HCC preoperative planning.

Materials and Methods

Study Design and Patient Data

This retrospective, multi-center study protocol was approved by the Ethics Committee of all participating institutions (Chinese PLA General Hospital, The Third Affiliated Hospital of Harbin Medical University, The Second Hospital of Nanjing, and The Tianyou Hospital of Wuhan University of Science and Technology). Because this is a retrospective study, the requirement for written informed consent was waived after obtained the approval of each ethics committee. All participants in this study promised to maintain the confidentiality of all patient data and strictly abide by the Declaration of Helsinki.

Between August 2005 and December 2019, a total of 2423 patients with HCC who subsequently underwent ultrasound-guided percutaneous microwave ablation (US-PMWA) were taken from a medical database at four high-volume medical centers in China and reviewed. The follow-up duration was terminated in July 2020. HCC was diagnosed according to the guidelines of the European Association for the Study of the Liver (EASL) and the American Association for the Study of Liver Diseases (AASLD).^{14,15} Details of patient recruitment are outlined according to the following inclusion criteria: a) patients with HCC who achieved complete ablation; b) patients with an Eastern Cooperative Oncology Group (ECOG) performance status of 0 or 1; c) patients with Child–Turcotte–Pugh

(CTP) grades A or B; d) patients with a single tumor ≤ 5 cm or a maximum diameter of tumors ≤ 3 cm when the number of tumors ≤ 3 ; e) patients without major vascular infiltration or extrahepatic metastasis. Exclusion criteria: a) patients who had HCC combined with other malignancies; b) patients with incomplete ablation; c) patients with previous treatment before MWA, excluding SR and LT; d) patients lost to follow-up > 6 months. Finally, the data of 1574 patients with HCC (307 females and 1267 males; mean age, 58.4 ± 11.2 years) with 2100 HCC lesions (mean diameter, 2.4 ± 0.9 cm) in medical records were screened for eligibility. Figure 1 shows the patient enrollment pathway and the flowchart of building the ML model based on clinical text data. The US-PMWA procedure, post-ablation assessment, and follow-up protocol are described in detail in [Supplementary Methods 1.1](#) and [1.2](#).

We developed and validated the predictive models for recurrence to confirm the stability and repeatability of the ML models applied in the study. Then, 1069 patients were assigned to the training set, and 268 were assigned to the internal

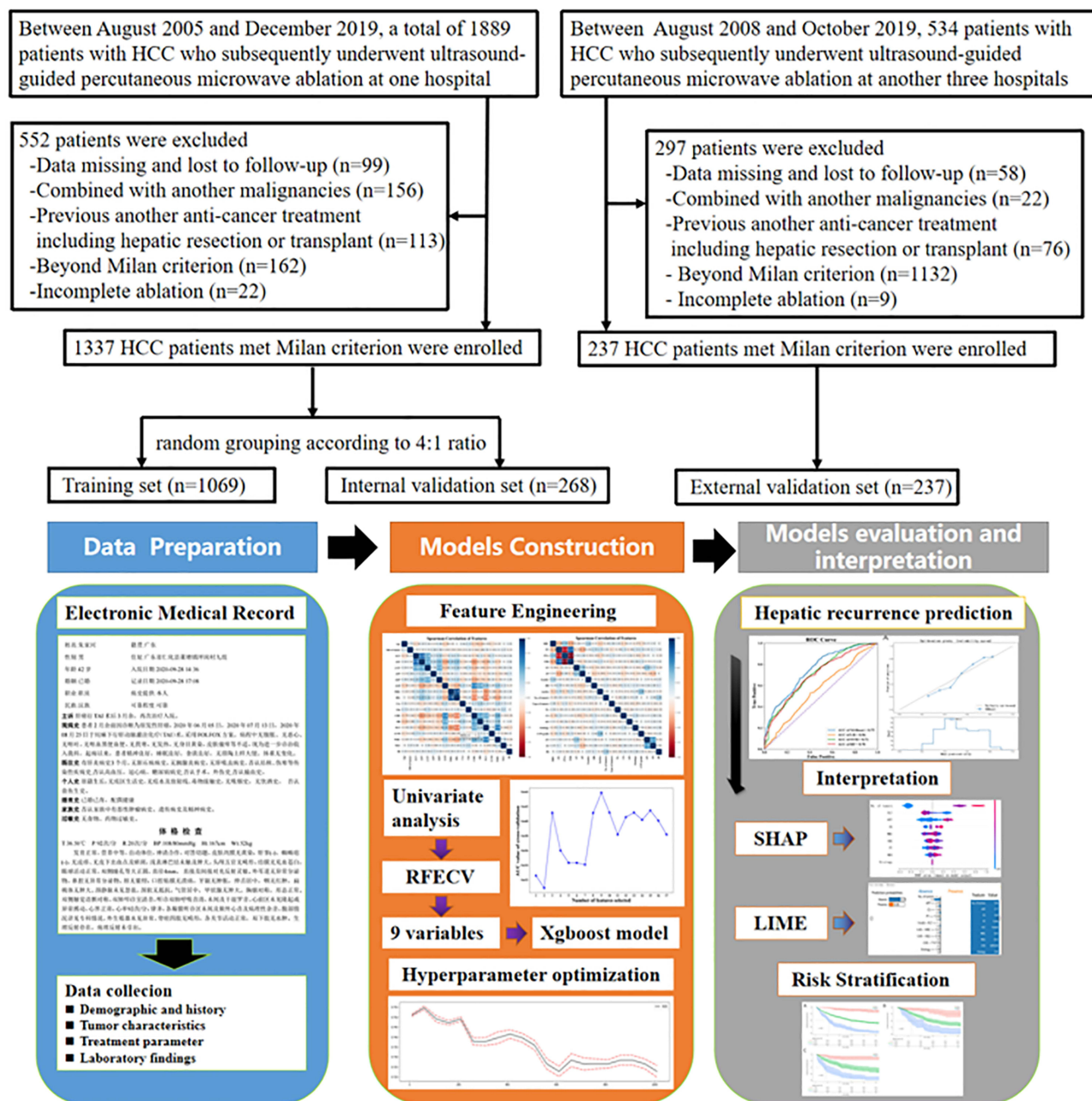


Figure 1 The patient enrollment pathway and the flowchart of building the ML model.

validation set based on a 4:1 ratio at hospital 1 (Chinese PLA General Hospital). Moreover, 237 patients with HCC from the other three hospitals (The Third Affiliated Hospital of Harbin Medical University, The Second Hospital of Nanjing, and The Tianyou Hospital of Wuhan University of Science and Technology) were assigned to the external validation set during the same period.

Variable Collection and Feature Selection

The variable data with a missing ratio greater than 20% were excluded in our study. A total of 36 variables were collected to develop the predictive models. First, 36 variables ([Table S1](#)) were analyzed by a univariate analysis using LR. Subsequently, the variables with significant differences were used to automatically adjust the number of variables through recursive feature elimination with cross-validation (RFECV) to obtain the best feature combination.¹⁶ RFECV includes two parts: elimination of recursive features and cross-validation. Given an external estimator, recursive feature elimination (RFE) selects features by recursively considering smaller and smaller sets of features. First, the estimator is trained on the initial set of features, and the initial importance of each feature is obtained either. Then, the least important features are pruned from the current set of features. The procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached. Cross-validation guarantees the reliability of each step of pruning.

Development and Validation of ML Predictive Models

In this study, one conventional statistical method, LR, and three conventional ML classification algorithms—random forest (RF), support vector machine (SVM), and eXtreme Gradient Boosting (XGBoost)—were used to develop and validate the predictive models.^{17,18} These models underwent continuous parameter optimization to compare the fitting effects and predictive ability of each model. For each base learner, the optimized XGBoost algorithm randomly samples 50% of the data and 80% of the features for training. Each iteration selects the feature with the largest current information gain as the node. When the predefined threshold is reached, this base learner is accomplished. Given a learning rate of 0.01, the predictive model completes the construction of 901 base learners, each of which is fitted based on the residual of the previous one. That is, we calculate the loss function of the current prediction and the actual value and minimize the Taylor expansion of the second order of the loss function. To raise the ability of model's generalization, XGBoost limits the maximum tree depth to 2 and optimizes the model's weight parameters in L1 and L2 regularization terms.

$$L^{(t)} = \sum_{i=1}^n l(y_i, y_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \quad (1)$$

The other three predictive models are introduced in detail in [Supplementary Methods 1.3](#) and [Figures S1–S3](#).

The validation queue data were used to evaluate the prediction performance of the model and calculate the area under receiver operator characteristic (AUC). A calibration curve was drawn, the calibration degree of the model was measured, and the degree of consistency between the predicted risk and the actual risk of the model was evaluated.

Interpretation Methods of the Predictive Models

Advanced ML models are usually black boxes. Although these models retain good accuracy, such metrics can be misleading. In this study, we used the SHAP and LIME algorithms as interpretation algorithms of the ML black box model.^{19–21}

The SHAP algorithm is a game theoretical approach that explains the output of any ML model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory. The algorithm calculates the marginal contribution of a feature when it is added to the model and then considers whether the factor is different in all factor sequences. The marginal contribution fully explains the influence of all factors included in the model for prediction and distinguishes the attributes of the factors (risk factors/protective factors).

The LIME algorithm is used to locally approximate the black box model by giving weights to the disturbance input such that the observation model gives the basis for interpretation of the sample prediction results. In the present study, we

randomly sampled the validation set and used the LIME algorithm to fit the model's predictive behavior in order to verify the rationality of the model's basis for predicting results. This can greatly improve the credibility of human-computer interaction.

Statistical Analysis

Continuous variables were given as a median with interquartile range (IQR) and compared by the Kruskal–Wallis test. Categorical variables were presented as frequencies with percentages and compared by the chi-squared test. Survival curves were calculated using the Kaplan–Meier method and compared with the Log rank test. Univariate Cox regression analyses were applied to calculate the hazard ratios (HRs) and corresponding 95% confidence interval (CI) of variables and identify independent prognostic factors. The discrimination, predictive accuracy, and calibration of the model were assessed by AUC and calibration curve. Given that a quick evaluation of recurrence risk levels with cutoff values might be useful in routine clinical practice, based on the optimal cutoff value thresholds for the scores determined by the maximally selected log-rank statistics, we stratified patients into low-risk, middle-risk, and high-risk subgroups. All statistical analyses were conducted using R version 4.0.4. A two-tailed p-value of less than 0.05 was considered as statistically significant.

Results

Patient Characteristics

Over a 10-year follow-up period, the median follow-up durations were 25.7 months (IQR, 6.3, 160.7 months), 26.9 months (IQR, 6.5, 159.2 months), and 23.7 months (IQR, 7.2, 124.8 months) in the training set, internal validation set, and external validation set, respectively, with no significant differences ($P = 0.139$). The baseline characteristics of patients in the three datasets are outlined in [Table 1](#). The 512/1574 (32.5%) recurrence occurred 2 years after MWA. All HCC patients underwent complete ablation, and 35.1% (375/1069), 30.2% (81/268) and 28.7% (68/237) patients experienced recurrence at 2 years after MWA in the training set, internal validation set, and external validation set, respectively. Among them, local recurrence rate was 9.2% (98/1069), 7.1% (19/268) and 8.0% (19/237), respectively.

Variables Selected

By 36 risk factors for correlation analysis, the correlation coefficient matrix heat map of the features ([Figure 2](#)) showed that the top five features correlated negatively with the outcomes were platelet (PLT), ablation time, prothrombin activity (PTA), cholinesterase (CHE), and α -fetoprotein (AFP). The top five characteristics correlated positively with outcomes were number of tumors, Child grade, gender, etiology, and total bilirubin (Tbil). Given 36 variables might interfere with each other, thereby reducing the final predictive ability of the ML models, we used the univariate Cox regression to screen out 17 risk factors with statistical differences. The results from univariate Cox analysis for recurrence outcome are summarized in [Table 2](#). We input these 17 factors into the ML algorithms to develop and validate the predictive model for recurrence after MWA.

Building and Evaluation of Four Predictive Models

The four ML models and the codes and construction processes are detailed in [Table S2](#). Moreover, we list the 1-, and 2-year AUC values of the three datasets of the different applied predictive models in [Table 3](#). Censoring until 2 years after MWA, the AUC values of the XGBoost model were 0.74 (95% confidence interval [CI], 0.71–0.77), 0.67 (95% CI, 0.59–0.71), and 0.73 (95% CI, 0.66–0.79) for the training set, internal validation set, and external validation set, respectively, which had the best predictive ability among three ML models and significantly outperformed the LR model ($P < 0.001$) ([Figure 3](#)). The robustness of the ML predictive model was further explored in a subgroup analysis. The screening results of RFECV showed ([Figure S4](#)) that when the 9 variables were retained, XGBoost exhibited the best predictive accuracy and discrimination. XGBoost sorted the nine variables according to feature values in advance and stored them as a block structure, which is convenient for quick query every time a node is split. The calibration plots for the recurrence outcome had good predictive value and were validated well in the internal validation set ([Figure S5A](#)) and

Table I The Baseline Characteristics of Patients in Three Datasets

Variables	Training Set (n= 1069)	Internal Validation Set (n=268)	External Validation Set (n=237)	P value
Demographic and history				
Age (years) ^a	58.4 ± 11.2	58.1 ± 11.5	57.5 ± 10.4	0.653
Gender ^b				0.168
Female	219 (20.49%)	50 (18.66%)	38 (16.03%)	
Male	850 (79.51%)	217 (80.97%)	199 (83.97%)	
Etiology ^b				0.856
None	171 (16.00%)	51 (19.03%)	36 (15.19%)	
HBV	802 (75.02%)	191 (71.27%)	180 (75.95%)	
HCV	91 (8.51%)	23 (8.58%)	21 (8.86%)	
Alcohol-induced	5 (0.47%)	3 (1.12%)	0 (0%)	
Child-Pugh grade ^b				0.362
A	1039 (97.19%)	262 (97.76%)	227 (95.78%)	
B	30 (2.71%)	6 (2.24%)	10 (4.22%)	
Diagnosis basis ^b				0.631
Biopsy	492 (46.02%)	114 (42.54%)	112 (47.26%)	
Image	577 (53.98%)	83 (59.71%)	125 (52.74%)	
Tumor data				
Number of tumors ^b				0.933
1	667 (62.39%)	166 (61.94%)	150 (63.29%)	
2	269 (25.16%)	62 (23.13%)	56 (23.63%)	
3	133 (12.44%)	40 (14.93%)	31 (13.08%)	
Maximum diameter of tumors (cm) ^a	2.4 ± 0.9	2.3 ± 1.0	2.5 ± 0.9	0.783
Abutting organs ^b				0.572
None	586 (54.56%)	143 (52.83%)	133 (56.6%)	
Gastrointestinal tract	81 (7.62%)	18 (6.79%)	23 (9.79%)	
Gallbladder	54 (5.08%)	17 (6.42%)	9 (3.83%)	
Hilum hepatis	11 (1.03%)	3 (1.13%)	3 (1.28%)	
Large vessels (more than 3 mm)	94 (8.84%)	21 (7.92%)	22 (9.36%)	
Bile duct (above secondary branch)	11 (1.03%)	2 (0.75%)	4 (1.7%)	
Diaphragm	116 (10.91%)	31 (11.7%)	20 (8.51%)	
Liver surface	110 (10.34%)	27 (10.19%)	14 (5.96%)	
Pericardium	6 (0.56%)	6 (2.26%)	7 (2.98%)	
Treatment parameter				
Ablation time (s) ^c	480 (360.0, 780.0)	480 (360.0, 817.5)	480 (360.0, 750.0)	0.317
Ablation power (W) ^c	50 (50.0, 60.0)	50 (50.0, 60.0)	50 (50.0, 60.0)	0.471
Laboratory findings				
ALT (U/L) ^c	26.8 (18.2–40.3)	28.4 (19.5–41.5)	26.15 (20.75–39.35)	0.260
AST (U/L) ^c	26.3 (20.1–36.8)	25.6 (20.2–39.6)	27.0 (20.9–38.9)	0.271
AFP (ug/mL) ^c	9.88 (3.38–110.90)	12.29 (3.68–135.50)	10.54 (3.15–56.98)	0.216
γ-GT (U/L) ^c	43.4 (27.0–80.3)	41.5 (26.7–86.5)	41.1 (27.6–76.1)	0.227
AKP(U/L) ^c	14.0 (13.2–14.9)	14.0 (13.4–14.9)	13.8 (13.3–15.0)	0.203
ALB (g/L) ^c	39.8 (36.7–42.8)	39.5 (36.3–42.2)	40.2 (36.8–42.9)	0.367
TBIL (μmol/L) ^c	14.6 (11.0–19.8)	14.3 (10.4–19.9)	14.3 (10.9–20.0)	0.496
BIL (μmol/L) ^c	5.0 (3.5–7.2)	5.0 (3.2–7.2)	5.1 (3.4–7.5)	0.433
Cr (μmol/L) ^c	71.7 (62.1–82.7)	73.8 (64.8–83.1)	72.2 (63.8–83.3)	0.309
Glu (mmol/L) ^c	5.0 (4.57–5.81)	4.89 (4.57–5.80)	5.02 (4.49–5.66)	0.337

(Continued)

Table 1 (Continued).

Variables	Training Set (n= 1069)	Internal Validation Set (n=268)	External Validation Set (n=237)	P value
CHE (u/L) ^c	6053.6 (4726.1–7174.2)	5982.9 (4613.7–7177.7)	5841.7 (4608.2–7256.6)	0.275
HB (g/L) ^c	138.0 (126.0–149.0)	136.0 (125.0–146.0)	137.5 (124.7–148.0)	0.454
RBC ($\times 10^{12}$) ^c	4.37 (3.98–4.78)	4.30 (3.84–4.74)	4.33 (3.92–4.76)	0.287
WBC ($\times 10^9$) ^c	4.61 (3.56–5.72)	4.49 (3.42–5.68)	4.40 (3.47–5.81)	0.354
PLT ($\times 10^9$) ^c	114.0 (79.0–156.0)	115.0 (72.0–154.0)	109.0 (73.0–155.0)	0.290
LYM ^c	0.355 (0.281–0.453)	0.353 (0.292–0.462)	0.346 (0.285–0.417)	0.248
NEU ^c	0.596 (0.516–0.688)	0.582 (0.514–0.686)	0.593 (0.522–0.666)	0.485
PT (s) ^c	14.0 (13.2–14.9)	14.0 (13.4–14.9)	13.8 (13.3–15.0)	0.335
PTA (%) ^c	86.0 (76.0–94.0)	85.0 (76.0–92.1)	87.7 (76.0–95.0)	0.079
INR ^c	1.1 (1.04–1.19)	1.1 (1.05–1.1)	1.1 (1.04–1.20)	0.124
Follow-up (months)				0.139
Median	25.7	26.9	23.7	
Range	6.3–160.4	6.5–159.2	7.2–124.8	

Notes: ^aData are mean \pm standard deviation; ^bData are frequencies with percentages in parentheses; ^cData are median with interquartile range (IQR) in square brackets; P value<0.05 indicated a significant difference.

Abbreviations: HCC, hepatocellular carcinoma; HBV, hepatitis B virus; HCV, hepatitis C virus; ALT, alanine aminotransferase; AST, aspartate aminotransferase; AFP, α -fetoprotein; γ -GT, gamma glutamyltransferase; AKP, alkaline phosphatase; ALB, albumin; TBIL, total bilirubin; BIL, bilirubin; Cr, creatinine; Glu, glucose; CHE, cholinesterase; HB, hemoglobin; LYM, lymphocyte; NEU, neutrophils; PT, prothrombin time; PTA, prothrombin activity; INR, international normalized ratio.

the external validation set (Figure S5B). The XGBoost model continued to show good predictive accuracy and discrimination in different subgroups of patients stratified according to the nine variables (Table S3).

Risk Stratification

We identified two optimal cutoff values (34.12 and 73.35) of the XGBoost model to stratify patients into low-risk, middle-risk, and high-risk groups in the training cohorts. The cumulative 1-, 2-, 3-, 4-, and 5-year recurrence-free survival (RFS) rates were

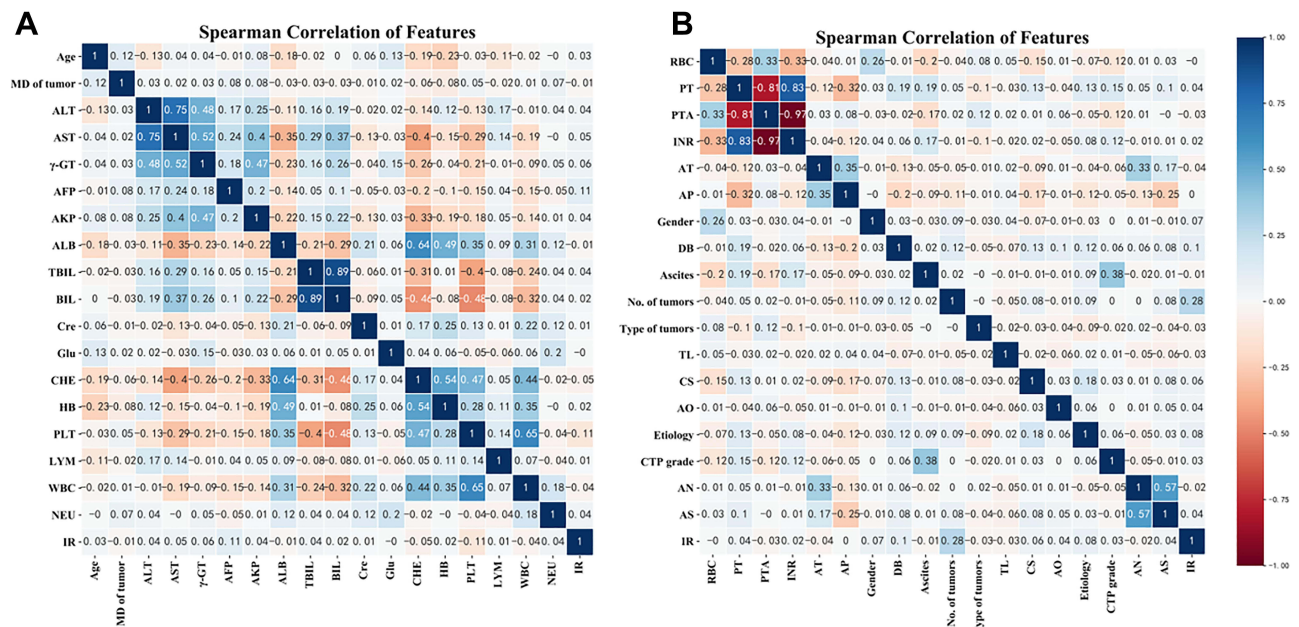


Figure 2 The correlation coefficient matrix heat map of all 36 variables. (A) The Correlation analysis of 18 variables including age; (B) the correlation analysis of the other 18 variables including RBC.

Table 2 The Results from Univariate Analysis for Recurrence Outcome

Variables	ER Group (n= 694)	Non-ER Group (n= 643)	P value
Gender^a			0.031
Female	123 (17.72%)	146 (22.71%)	
Male	571 (82.28%)	497 (77.29%)	
Etiology^a			0.003
None	92 (13.26%)	130 (20.22%)	
HBV	537 (77.38%)	456 (70.92%)	
HCV	62 (8.93%)	52 (8.09%)	
Alcohol-induced	3 (0.43%)	5 (0.78%)	
Number of tumors^a			<0.001
1	345 (49.71%)	488 (75.89%)	
2	213 (30.69%)	118 (18.35%)	
3	136 (19.60%)	37 (5.75%)	
Abutting organs^a			0.030
None	370 (53.31%)	359 (55.83%)	
Gastrointestinal tract	63 (9.08%)	36 (5.60%)	
Gallbladder	48 (6.92%)	23 (3.58%)	
Hilum hepatitis	4 (0.58%)	10 (1.56%)	
Large vessels	51 (7.35%)	64 (9.95%)	
Bile duct	8 (1.15%)	5 (0.78%)	
Diaphragm	75 (10.81%)	72 (11.20%)	
Liver surface	68 (9.80%)	69 (10.73%)	
Pericardium	7 (1.01%)	5 (0.78%)	
Tumor location^a			0.013
Left	157 (22.62%)	146 (22.71%)	
Right	537 (77.38%)	497 (77.29%)	
Comorbidity score^a			0.034
Low	432 (62.25%)	426 (66.25%)	
High	262 (37.75%)	217 (33.75%)	
AST (U/L)			0.013
>40	454 (65.40%)	321 (49.90%)	<0.001
≤40	240 (34.60%)	322 (51.10%)	
AFP (ng/mL)			<0.001
>400	305 (62.25%)	156 (66.25%)	
≤400	389 (37.75%)	487 (33.75%)	
γ-GGT (U/L)			0.003
>60	415 (59.80%)	332 (51.60%)	
≤60	279 (40.20%)	311 (48.40%)	
AKP (U/L)			<0.001
>150	521 (75.10%)	211 (32.80%)	
≤150	173 (24.90%)	432 (67.20%)	
TBIL (umol/mL)			<0.001
>17.1	483 (69.60%)	335 (52.10%)	
≤17.1	211 (30.40%)	308 (47.90%)	
BIL (umol/mL)			0.049
>7.0	384 (55.00%)	388 (60.30%)	
≤7.0	314 (45.00%)	255 (39.70%)	
CHE (U/L)			<0.001
>4000	512 (73.80%)	387 (60.20%)	
≤4000	182 (26.20%)	256 (39.80%)	

(Continued)

Table 2 (Continued).

Variables	ER Group (n= 694)	Non-ER Group (n= 643)	P value
WBC ($\times 10^9$)			0.004
>4.0	488 (70.30%)	404 (62.80%)	
≤ 4.0	206 (29.70%)	239 (37.20%)	
PLT ($\times 10^9$)			<0.001
>100	492 (62.25%)	418 (66.25%)	
≤ 100	202 (37.75%)	225 (33.75%)	
NEU ($\times 10^9$)			0.021
>400	432 (70.90%)	426 (65.00%)	
≤ 400	262 (29.10%)	217 (35.00%)	
PT (s)			0.004
>15	468 (67.40%)	385 (59.90%)	
≤ 15	226 (32.60%)	258 (40.10%)	

Notes: ^aData are frequencies with percentages in parentheses; P value<0.05 indicated a significant difference.

Abbreviations: HCC, hepatocellular carcinoma; IR, intrahepatic recurrence; HBV, hepatitis B virus; HCV, hepatitis C virus; AST, aspartate aminotransferase; AFP, α -fetoprotein; γ -GT, gamma glutamyl-transferase; AKP, alkaline phosphatase; TBIL, total bilirubin; BIL, bilirubin; CHE, cholinesterase; NEU, neutrophils; PT, prothrombin time.

Table 3 Summary of Predictive Performance of Multiple Models

Prediction Models	1-Year AUC (95% CI)	2-Year AUC (95% CI)	P value
Training set			
XGBoost	0.75 (0.73, 0.79)	0.74 (0.71, 0.77)	-
SVM	0.67 (0.62, 0.72)	0.66 (0.59, 0.71)	0.028
RF	0.70 (0.66, 0.74)	0.70 (0.67, 0.73)	<0.001
LR	0.57 (0.53, 0.61)	0.56 (0.52, 0.59)	<0.001
Internal validation set			
XGBoost	0.68 (0.64, 0.72)	0.67 (0.60, 0.73)	-
SVM	0.53 (0.48, 0.62)	0.50 (0.44, 0.57)	0.022
RF	0.74 (0.67, 0.81)	0.72 (0.65, 0.78)	0.216
LR	0.60 (0.52, 0.69)	0.58 (0.51, 0.65)	<0.001
External validation set			
XGBoost	0.74 (0.70, 0.83)	0.73 (0.66, 0.79)	-
SVM	0.53 (0.49, 0.61)	0.52 (0.44, 0.60)	<0.001
RF	0.66 (0.57, 0.75)	0.70 (0.63, 0.78)	0.130
LR	0.60 (0.52, 0.69)	0.58 (0.51, 0.65)	0.010

55%, 42%, 13%, 7%, and 3% in the low-risk group, 74%, 62%, 34%, 25%, and 19% in the middle-risk group, and 97%, 85%, 59%, 48%, and 30% in the high-risk group, respectively, which had a significant statistical difference among the three groups ($P < 0.001$) (Figure 4A). Both in the internal validation (cutoff value, 62.31 and 112.21) and external validation cohort (cutoff value 60.98 and 117.30), the cumulative 1-, 2-, 3-, 4-, and 5-year RFS rates among the high-risk, middle-risk, and low-risk groups also had a significant statistical difference (both, $P < 0.001$) (Figure 4B and C). An AI system for recurrence prediction after MWA accessed online is available at <http://114.251.235.51:8001/> (Figure S6).

Interpretation of the XGBoost Model

We used the SHAP and LIME algorithms to visualize the influence of each variable in the XGBoost model on the outcome variable in Figure 5. Based on the SHAP algorithm, the influence of nine variables on the outcome in the

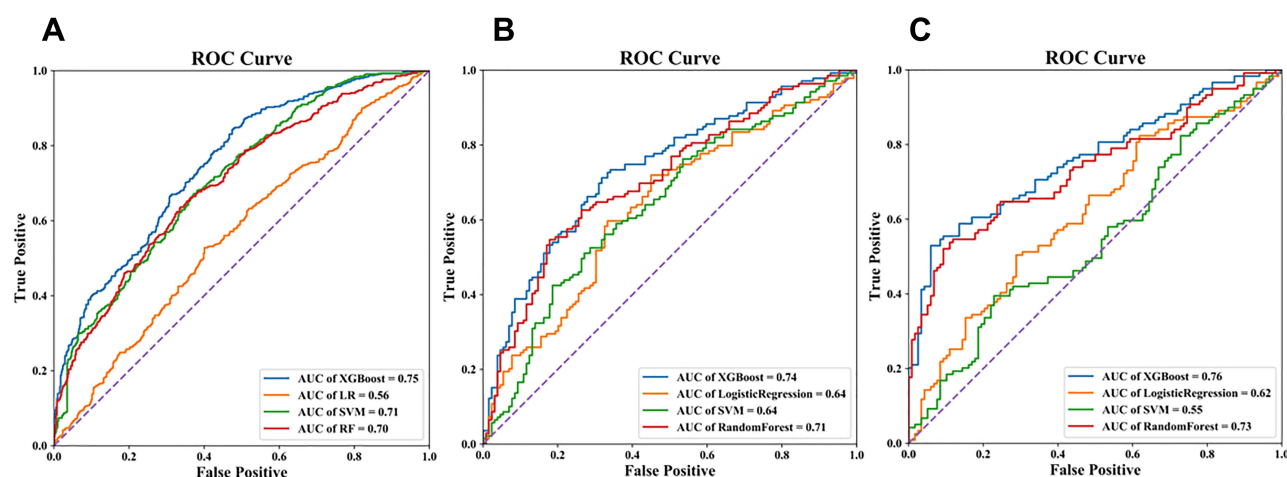


Figure 3 The ROC curves of four models in three datasets, (A) the training set; (B) the internal validation set; (C) the external validation set.

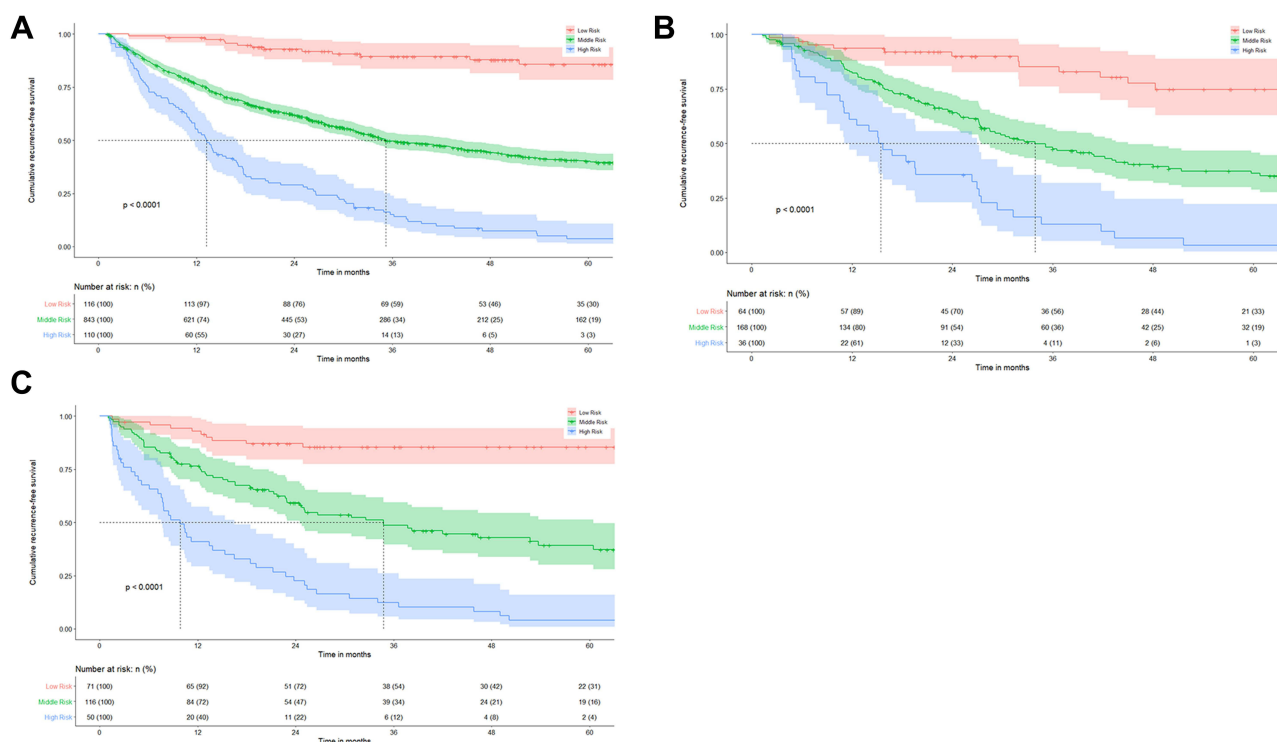


Figure 4 The risk stratification of the XGBoost Model. (A) The training set; (B) the internal validation set; (C) the external validation set.

XGBoost model is shown in [Figure S7](#). The feature-importance ranking was as follows ([Figure S8](#)): number of tumors, PLT, AFP, comorbidity score (CS), WBC, CHE, PT, NEU, and etiology. The number of tumors, PLT, and CS had approximately linear effects on the outcome; the number of tumors and CS were positively correlated with the outcome. It is worth noting that the CS stabilized after five points; PLT was negatively correlated with the outcome; when the value was greater than 150, the impact became stable. The effects of AFP, WBC, and CHE on the outcome all had peaks and troughs, and beyond the SHAP value, gradually stabilized. The influence of PT and NEU on the outcome was slightly more complicated. The SHAP value of etiology was near 0, which had little effect on the outcome. The LIME algorithm explained the predictions of the XGBoost model on each sample and summarized the predictions of the model in the training set, internal validation set, and external test set, showing the distribution of four types of results: true positive,

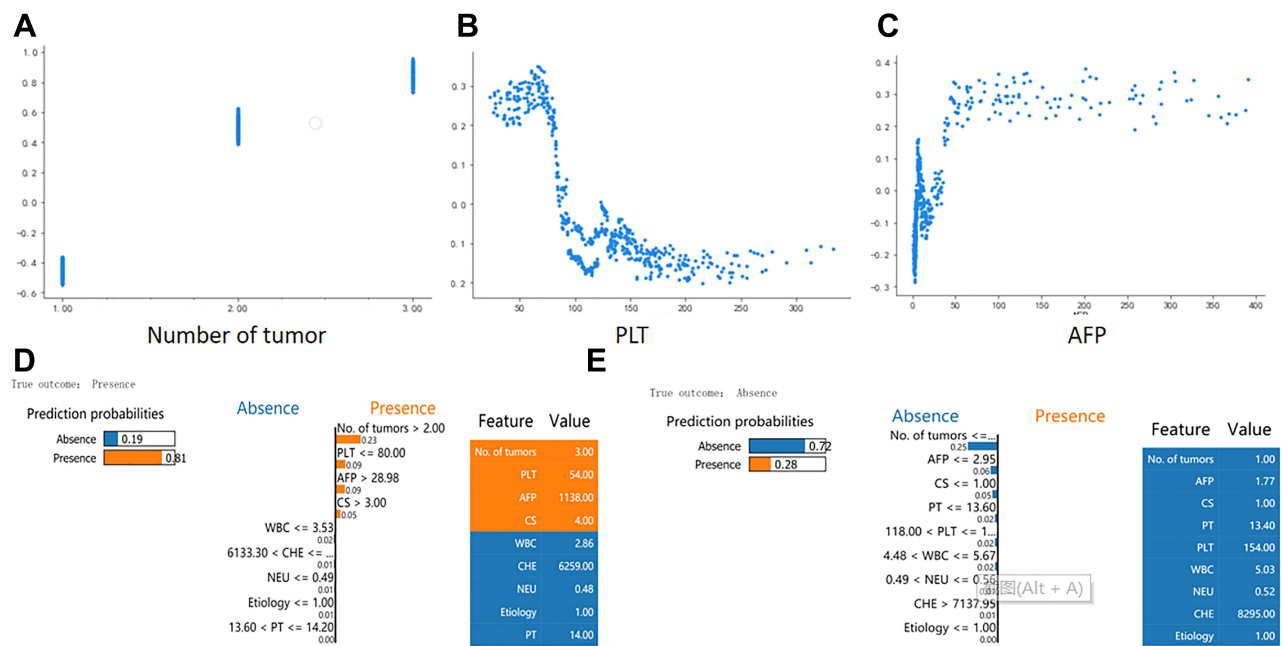


Figure 5 Interpretation of the XGBoost model. (A–C) SHAP algorithms to visualize the influence of each variable in the XGBoost model; (D and E) LIME algorithms to visualize the influence of each variable in the XGBoost model.

true negative, false positive, and false negative. Overall, the number of samples predicted to be true positive and negative both accounted for 70%, indicating that the performance of the XGBoost prediction model is relatively stable. In addition, we randomly selected four types of samples (true positive, true negative, false positive, false negative) and visualized them separately. When the prediction probability of the model is far from 0.5, the physician only needs to judge whether the prediction result is reasonable based on the model prediction basis given by LIME. If the model prediction basis is unreasonable, the result will not be adopted; when the predicted probability of the model is close to 0.5, such samples are usually clinically difficult to judge.

Discussion

In this study, we developed and validated ML models using clinical data extracted directly from electronic health records to predict individual ER outcomes after MWA based on a long-term follow-up cohort comprising 1574 patients with early-stage HCC who were recommended as MWA candidates. These analysis results suggest that the XGBoost model has more favorable performance and discrimination than the other three predictive ML models, which identify three ER risk strata and help further clinicians analyze HCC patient survival benefit.

A real world study based on a consecutive multi-center cohort was performed, with a large sample size from four high-volume medical centers by setting three cohorts, which ensured the stability of the ML algorithm. The clinical data available on biomarkers for prediction of ER were obtained. Previously identified risk factors for hepatic ER after LA mainly included tumor size, number, and AFP level.^{21–23} However, these clinical variables that predict post-ablation recurrence remain controversial because of individual differences in patients and tumor heterogeneity. Moreover, traditional staging systems, such as tumor-node-metastasis (TNM), Barcelona Clinic Liver Cancer (BCLC) grade, and Okuda score, were used to try to predict HCC outcomes,²⁴ but they had insufficient predictive ability on postoperative recurrence. Aggressive pathological factors, including microvascular invasion (MVI), microsatellite lesions, and pathological grade, could provide valuable information on tumor invasion ability, which is associated with early recurrence after SR.^{25,26} However, preoperative diagnosis mainly relies on biopsy during MWA procedure, and complete pathological tissue cannot be obtained, so it was difficult to collect effective pathological information for MWA.

The imaging feature commonly reflects the aggressive characteristics of a tumor; therefore, several radiomics models were built in prediction of HCC recurrence in the last 5 years. Among them, two radiomic predictive models of post-

ablation recurrence were developed. Yuan et al reported that a radiomics nomogram based on a contrast-enhanced CT was developed to predict the early recurrence of HCC cases eligible for curative ablation, and the C-index of the radiomics model and radiomics, combined with a clinical model, were 0.72 and 0.79, respectively.²⁷ Shen et al reported that a radiomics model based on a post-treatment CT image was built to predict the early recurrence after ablation, and the AUC value was 0.89.²⁸ As the abovementioned radiomics model takes 200–300 cases, it is not enough to provide big-data information and verification, causing inadequate stability and robustness. In addition, feature extraction is complicated and dependent on the model, which is difficult to apply and lacks image visualization. As a cost-effective and non-invasive manner, clinical text data are a better choice. Therefore, we decided to establish a ML model for ER risk prediction based on clinical text data.

The predictive models based on the XGBoost algorithm, with advantages including efficient computing power and the ability to handle missing values, were built and tested via two validation sets, which further ensured the stability. In our study, the XGBoost model could reduce eigenvalues from a great number of electronic health records compared with the other models. In terms of missing value processing, the XGBoost algorithm could handle missing data automatically by adding a default direction based on learning during the training procedure in each tree node. When a value was missing in the validation data, the instance was classified into the default direction. In this study, 36 variables were input into the XGBoost model, and 9 variables were output to develop the optimal predictive performance in this study, with AUC values of 0.74 and 0.73 in the training and validation datasets, respectively. ML has remarkable computing power and will not be affected by large samples or multi-dimensional data. Development of the predictive model using ML to help physicians formulate follow-up treatment plans in clinical practice has been increasing.^{29,30} For example, a precise prognostic assessment of COVID-19 in an ICU was built by ML methods that can accurately screen out high-risk groups and guide physicians in effectively prevention and treatment.³¹ In addition, an ML model to predict kidney outcomes in IgA nephropathy was developed in multiple centers, which can help physicians in rational use of drug according to these predictive results.³² Similarly, our XGBoost algorithm model showed the ability to generalize across populations and screening protocols. This suggests that in future clinical deployments, the XGBoost model offers strong baseline performance but may benefit from fine-tuning with local data. Interventional radiologists can use preventive TACE or neoadjuvant chemotherapy treatment (NCT) for high-risk recurrence HCC cases based on ML model prediction and can reduce the recurrence risk after MWA further, thereby “nipping it in the bud”, as the saying goes.

In this study, we used the SHAP and ITME algorithms to explain the XGBoost model because the black boxes used to understand the principles behind ML model could be accessed online conveniently. After full explanation by the SHAP and LIME algorithms, the XGBoost model showed accurate and stable prediction ability in recurrence. However, parts of clinical variable data that should not be used may have accidentally leaked into the training data. The SHAP algorithm calculates the marginal contribution of a feature when it is added to the model and then considers whether the variables are different in all variable sequences. The marginal contribution fully explains the influence of all variables included in the model prediction and distinguishes the attributes of the factors (risk/protective factors). Moreover, we randomly sampled the test set and used the LIME algorithm to fit the model’s predictive behavior to verify the rationality of the model’s basis for predicting results. This approach can greatly improve the trust value of human–computer interaction.

There are some limitations to our study. First, the risk of selection bias is unavoidable in observational studies. However, this risk was minimized by the inclusion of all consecutive patients and a large cohort of early HCC cases for MWA to date. Second, the cohort was not from a prospective therapeutic trial because of the nature of retrospective research; therefore, part of the data was lacking. Fortunately, the XGBoost model had the ability to deal with missing values in this study. Third, the prediction model based on clinical text prediction is effective and convenient, but if it is combined with imaging information, more accurate results may be obtained; a prospective randomized trial is underway to confirm this hypothesis.

In conclusion, with an easy-to-use XGBoost model consisting of nine clinical variables, the model exhibited adequate performance and individualized predictive ability that could stratify patients into three ER risk groups with a referable cumulative recurrence rate. Therefore, the visual XGBoost model may help physicians with decision-making before MWA for HCC in clinical practice and trials.

Ethics

This retrospective, multi-center study protocol was approved by the Ethics Committee of all participating institutions (Chinese PLA General Hospital, The Third Affiliated Hospital of Harbin Medical University, The Second Hospital of Nanjing, and The Tianyou Hospital of Wuhan University of Science and Technology). Because this is a retrospective study, the requirement for written informed consent was waived after obtained the approval of each ethics committee. All participants in this study promised to maintain the confidentiality of all patient data and strictly abide by the Declaration of Helsinki.

Author Contributions

All authors made substantial contributions to conception and design, acquisition of data, or analysis and interpretation of data; took part in drafting the article or revising it critically for important intellectual content; agreed to submit to the current journal; gave final approval of the version to be published; and agree to be accountable for all aspects of the work.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. *CA Cancer J Clin*. 2020;70:7–30. doi:10.3322/caac.21590
2. Forner A, Reig M, Bruix J. Hepatocellular carcinoma. *Lancet*. 2018;391:1301–1314. doi:10.1016/S0140-6736(18)30010-2
3. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68:394–424. doi:10.3322/caac.21492
4. Dimitroulis D, Damaskos C, Valsami S, et al. From diagnosis to treatment of hepatocellular carcinoma: an epidemic problem for both developed and developing world. *World J Gastroenterol*. 2017;23:5282–5294. doi:10.3748/wjg.v23.i29.5282
5. Vitale A, Trevisani F, Farinati F, Cillo U. Treatment of hepatocellular carcinoma in the Precision Medicine era: from treatment stage migration to therapeutic hierarchy. *Hepatology*. 2020;72(6):2206–2218. doi:10.1002/hep.31187
6. Marrero JA, Kulik LM, Sirlin CB, et al. Diagnosis, staging, and management of hepatocellular carcinoma: 2018 practice guidance by the American Association for the Study of Liver Diseases. *Hepatology*. 2018;68:723–750. doi:10.1002/hep.29913
7. Eilard MS, Naredi P, Helmersson M, et al. Survival and prognostic factors after transplantation, resection and ablation in a national cohort of early hepatocellular carcinoma. *HPB*. 2021;23:394–403. doi:10.1016/j.hpb.2020.07.010
8. Sangiovanni A, Colombo M. A therapeutic conundrum: delaying ablation of small nonresectable early hepatocellular carcinoma to facilitate liver transplantation. *Liver Transpl*. 2016;22:161–162. doi:10.1002/lt.24382
9. Liang P, Yu J, Yu XL, et al. Percutaneous cooled-tip microwave ablation under ultrasound guidance for primary liver cancer: a multicentre analysis of 1363 treatment-naïve lesions in 1007 patients in China. *Gut*. 2012;61:1100–1101. doi:10.1136/gutjnl-2011-300975
10. Liu W, Zou R, Wang C, et al. Microwave ablation versus resection for hepatocellular carcinoma within the Milan criteria: a propensity-score analysis. *Ther Adv Med Oncol*. 2019;11:1758835919874652. doi:10.1177/1758835919874652
11. Zaitoun M, Elsayed SB, Zaitoun NA, et al. Combined therapy with conventional trans-arterial chemoembolization (cTACE) and microwave ablation (MWA) for hepatocellular carcinoma >3–<5 cm. *Int J Hyperthermia*. 2021;38:248–256. doi:10.1080/02656736.2021.1887941
12. Liu X, Lu J, Zhang G, et al. A machine learning approach yields a multiparameter prognostic marker in liver cancer. *Cancer Immunol Res*. 2021;9:337–347. doi:10.1158/2326-6066.CIR-20-0616
13. Zou ZM, Chang DH, Liu H, Xiao YD. Current updates in machine learning in the prediction of therapeutic outcome of hepatocellular carcinoma: what should we know. *Insights Imaging*. 2021;12:31. doi:10.1186/s13244-021-00977-9
14. European Association for the Study of the Liver. Electronic address: easloffice@easloffice.eu, European Association for the Study of the Liver. EASL clinical practice guidelines: management of hepatocellular carcinoma. *J Hepatol*. 2018;69:182–236. doi:10.1016/j.jhep.2018.03.019
15. Heimbach JK, Kulik LM, Finn RS, et al. AASLD guidelines for the treatment of hepatocellular carcinoma. *Hepatology*. 2018;67:358–380. doi:10.1002/hep.29086
16. Lu P, Zhuo Z, Zhang W, Tang J, Tang H, Lu J. Accuracy improvement of quantitative LIBS analysis of coal properties using a hybrid model based on a wavelet threshold de-noising and feature selection method. *Appl Opt*. 2020;59:6443–6451. doi:10.1364/AO.394746
17. Tahmassebi A, Wengert GJ, Helbich TH, et al. Impact of machine learning with multiparametric magnetic resonance imaging of the breast for early prediction of response to neoadjuvant chemotherapy and survival outcomes in breast cancer patients. *Invest Radiol*. 2019;54:110–117. doi:10.1097/RLI.0000000000000518
18. Chang W, Liu Y, Xiao Y, et al. A machine-learning-based prediction method for hypertension outcomes based on medical data. *Diagnostics*. 2019;9(4):178. doi:10.3390/diagnostics9040178
19. Cha Y, Shin J, Go B, et al. An interpretable machine learning method for supporting ecosystem management: application to species distribution models of freshwater macroinvertebrates. *J Environ Manage*. 2021;291:112719. doi:10.1016/j.jenvman.2021.112719
20. Oh S, Park Y, Cho KJ, Kim SJ. Explainable machine learning model for glaucoma diagnosis and its interpretation. *Diagnostics*. 2021;12(1):11. doi:10.3390/diagnostics12010011
21. Lin MY, Li CC, Lin PH, et al. Explainable machine learning to predict successful weaning among patients requiring prolonged mechanical ventilation: a retrospective cohort study in central Taiwan. *Front Med*. 2021;8:663739. doi:10.3389/fmed.2021.663739

22. Yang Y, Chen Y, Ye F, et al. Late recurrence of hepatocellular carcinoma after radiofrequency ablation: a multicenter study of risk factors, patterns, and survival. *Eur Radiol.* 2021;31:3053–3064. doi:10.1007/s00330-020-07460-x
23. Kim CG, Lee HW, Choi HJ, et al. Development and validation of a prognostic model for patients with hepatocellular carcinoma undergoing radiofrequency ablation. *Cancer Med.* 2019;8:5023–5032. doi:10.1002/cam4.2417
24. Kao WY, Su CW, Chiou YY, et al. Hepatocellular carcinoma: nomograms based on the albumin-bilirubin grade to assess the outcomes of radiofrequency ablation. *Radiology.* 2017;285:670–680. doi:10.1148/radiol.2017162382
25. Lee S, Kang TW, Song KD, et al. Effect of microvascular invasion risk on early recurrence of hepatocellular carcinoma after surgery and radiofrequency ablation. *Ann Surg.* 2021;273:564–571. doi:10.1097/SLA.0000000000003268
26. Yamashita YI, Imai K, Yusa T, et al. Microvascular invasion of single small hepatocellular carcinoma ≤ 3 cm: predictors and optimal treatments. *Ann Gastroenterol Surg.* 2018;2:197–203. doi:10.1002/ags3.12057
27. Yuan C, Wang Z, Gu D, et al. Prediction early recurrence of hepatocellular carcinoma eligible for curative ablation using a Radiomics nomogram. *Cancer Imaging.* 2019;19:21. doi:10.1186/s40644-019-0207-7
28. Shen JX, Zhou Q, Chen ZH, et al. Longitudinal radiomics algorithm of posttreatment computed tomography images for early detecting recurrence of hepatocellular carcinoma after resection or ablation. *Transl Oncol.* 2021;14:100866. doi:10.1016/j.tranon.2020.100866
29. Heo J, Yoon JG, Park H, Kim YD, Nam HS, Heo JH. Machine learning-based model for prediction of outcomes in acute stroke. *Stroke.* 2019;50:1263–1265. doi:10.1161/STROKEAHA.118.024293
30. Wang S, Pathak J, Zhang Y. Using electronic health records and machine learning to predict postpartum depression. *Stud Health Technol Inform.* 2019;264:888–892. doi:10.3233/SHTI190351
31. Pan P, Li Y, Xiao Y, et al. Prognostic assessment of COVID-19 in the intensive care unit by machine learning methods: model development and validation. *J Med Internet Res.* 2020;22:e23128. doi:10.2196/23128
32. Chen T, Li X, Li Y, et al. Prediction and risk stratification of kidney outcomes in IgA nephropathy. *Am J Kidney Dis.* 2019;74:300–309. doi:10.1053/j.ajkd.2019.02.016

Journal of Hepatocellular Carcinoma

Dovepress

Publish your work in this journal

The Journal of Hepatocellular Carcinoma is an international, peer-reviewed, open access journal that offers a platform for the dissemination and study of clinical, translational and basic research findings in this rapidly developing field. Development in areas including, but not limited to, epidemiology, vaccination, hepatitis therapy, pathology and molecular tumor classification and prognostication are all considered for publication. The manuscript management system is completely online and includes a very quick and fair peer-review system, which is all easy to use. Visit <http://www.dovepress.com/testimonials.php> to read real quotes from published authors.

Submit your manuscript here: <https://www.dovepress.com/journal-of-hepatocellular-carcinoma-journal>