American Journal of Sciences and Engineering Research

E-ISSN– 2348–703X, Volume 9, Issue 5, 2026

SHAP versus LIME in Explaining a Dynamic Mortality Prediction Model for Hepatocellular Carcinoma: Cross-Cohort and Temporal Consistency

Ipek Balikci Cicek[1,*], Zeynep Kucukakcali[2]

[1,2] *Inonu University Faculty of Medicine, Department of Biostatistics and Medical Informatics, 44280, Malatya, Turkey*

** Corresponding author: Ipek Balikci Cicek, e-mail: ipek.balkci@inonu.edu.tr*

Abstract: *Background: SHAP and LIME are among the most widely used explainability methods in clinical machine learning, but they rely on fundamentally different approximation strategies and are rarely compared systematically. It also remains unclear whether feature-importance explanations from a single fitted model remain consistent across independent patient cohorts and over time, an important consideration when explainability findings are generalized to different clinical settings.*

Methods: Using an externally validated, landmark-based Random Forest model for 12-month mortality prediction in patients with intermediate-stage hepatocellular carcinoma treated with transarterial chemoembolization, we characterized global and local feature attribution using SHAP (TreeExplainer) and LIME. SHAP feature importance was compared across the Derivation, Internal Testing, and Multicentric Testing cohorts, between early and late landmarks, and between 6- and 12-month prediction horizons to assess cross-cohort and temporal stability. Agreement between SHAP and LIME was quantified in 50 randomly sampled observations using Pearson and Spearman correlations of attribution values and Spearman correlation of feature-importance rankings.

Results: Tumor diameter, alpha-fetoprotein, and lesion number were the dominant predictors, showing non-linear, threshold-like relationships with predicted risk. SHAP rankings were nearly identical across the three cohorts (Spearman rank correlation with the Derivation Cohort: r=0.998 for Internal Testing and r=0.995 for Multicentric Testing; both p<0.0001). The top seven features showed complete overlap between the Derivation and Internal Testing cohorts (7/7) and substantial overlap between the Derivation and Multicentric Testing cohorts (6/7). Rankings were similarly stable between early and late landmarks and between the 6- and 12-month prediction horizons. When binary and one-hot-encoded categorical variables were explicitly specified as categorical features in LIME rather than processed using default continuous-variable discretization, SHAP and LIME showed strong agreement at the attribution-value level (Pearson r=0.871; Spearman r=0.825) and moderate-to-strong agreement in feature-importance rankings (Spearman r=0.721). Nevertheless, differences persisted for a subset of predictors, notably ECOG performance status, a common, non-sparse predictor.

Conclusion: Feature attributions from this model were highly stable across independent cohorts, across the disease course, and across prediction horizons, supporting the consistency of the model's dominant prognostic drivers across the evaluated settings. SHAP-LIME agreement improved substantially when categorical variables were explicitly specified in LIME rather than relying on default settings, demonstrating that attribution comparisons are sensitive to explainer configuration. However, differences persisted for a subset of predictors

 Received-31-08-2026 Accepted- 12-09-2026

even after categorical-feature specification. Explainability findings in clinical machine learning should therefore be reported with explicit attention to both the attribution method and its configuration and, where feasible, corroborated using more than one approach before informing clinical interpretation.

Keywords: *explainable artificial intelligence; SHAP; LIME; feature attribution; hepatocellular carcinoma; machine learning; reproducibility*

I. Introduction

Explainable artificial intelligence (XAI) methods are now routinely used to render clinical machine learning models interpretable, with SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) among the most widely adopted approaches[1-3]. Both methods assign an importance value to each input feature for a given prediction and are frequently presented in the clinical literature as broadly interchangeable tools for "opening the black box." In practice, however, SHAP and LIME rely on fundamentally different approximation strategies: SHAP is grounded in cooperative game theory and, for tree-based models, can be computed exactly via TreeExplainer [2], whereas LIME fits a local linear surrogate model to a perturbed neighborhood around each prediction. These approaches are not mathematically guaranteed to agree, and a growing methodological literature has described a broader "disagreement problem" in explainable machine learning, in which different attribution methods applied to the same model and prediction can yield materially different explanations [4-6]. Despite this, direct quantitative comparisons of SHAP and LIME within a single clinical oncology model remain uncommon, and most clinical XAI studies report only one method without testing whether its conclusions would hold under an alternative approach.

A second, related assumption is rarely tested explicitly: whether a model's explanation, once characterized, remains consistent beyond the dataset in which it was computed. Clinical prediction models are increasingly validated in external cohorts, and reporting standards such as TRIPOD emphasize rigorous assessment and transparent reporting of model performance [7]; the same scrutiny is seldom applied to their explanations, despite calls for more rigorous evaluation of interpretability claims [8-11] and specific concerns about the clinical readiness of current XAI approaches in healthcare [12, 13]. If feature-importance rankings differ substantially between an external cohort and the cohort in which a model was developed, or between different stages of a patient's clinical course, conclusions drawn from explainability analyses, and the clinical narratives built upon them, may be less robust than they initially appear.

We address these questions using a landmark-based Random Forest model for dynamic 12-month mortality prediction in patients with intermediate-stage hepatocellular carcinoma (HCC), staged according to BCLC criteria [14] and treated with transarterial chemoembolization (TACE) in accordance with EASL guidelines [15]. The present study focuses specifically on the explainability of this model. We (1) characterize global and local feature attribution using SHAP; (2) assess whether SHAP-based feature importance is consistent across three independent cohorts, including a geographically distinct multicentric cohort; (3) evaluate whether SHAP-based feature importance remains stable across the disease course (early versus late clinical assessments) and across two prediction horizons; and (4) formally quantify agreement between SHAP and LIME at both the individual attribution-value and feature-ranking levels. Rather than proposing a new predictive algorithm, this study provides an empirical evaluation of the consistency of model explanations across cohorts and temporal settings and their agreement across explainability methods.

II. Methods

2.1 Study Design and Data Source

This is a secondary explainability analysis using a publicly available, multi-cohort dataset (Dryad repository, DOI: 10.5061/dryad.pd44k8r) [16] of patients with intermediate-stage HCC treated with TACE. The dataset comprises a Derivation Cohort (979 patients, 3,774 repeated clinical assessments, termed landmarks), an Internal Testing Cohort (627 patients, 2,569 assessments), and a geographically independent Multicentric Testing Cohort recruited from separate hospitals (414 patients, 1,742 assessments). The present analysis focuses specifically on the explainability and stability of feature attributions derived from a landmark-based machine learning model using these cohorts.

2.2 Landmarking Framework

Each clinical assessment (landmark) was treated as an independent prediction origin, following the landmarking framework for dynamic risk prediction [17], allowing risk estimates to be updated as new clinical information became available over the disease course. At each landmark, the primary outcome was death within 12 months from that assessment, with a 6-month horizon used as a sensitivity analysis. Landmarks with insufficient follow-up to resolve the outcome were excluded from the corresponding horizon analysis. Patient identity was used as the grouping variable in all analyses to prevent data leakage from repeated within-patient measurements.

2.3 Feature Engineering and Preprocessing

Twenty-four clinical, laboratory, and tumor-related variables common to all three cohorts were used as model inputs: demographic (age, sex), laboratory (alpha-fetoprotein [AFP], albumin, aspartate aminotransferase, bilirubin, prothrombin time), tumor burden (diameter, number of lesions, new lesions, categorical location), disease extent (portal vein invasion, hepatic vein invasion, vena cava invasion, distant metastasis, lymph node metastasis, ascites), Eastern Cooperative Oncology Group (ECOG) performance status, and landmark-specific timing variables (landmark time since first admission and time-slice index). A composite Child-Pugh score was evaluated but excluded from the final feature set due to collinearity with its constituent laboratory variables, which was found to distort SHAP-based attribution of the individual components without improving discrimination.

2.4 Model Underlying the Explainability Analysis

The explainability analyses were performed using a Random Forest model [18] developed for 12-month mortality prediction in the Derivation Cohort using patient-grouped 5-fold cross-validation. The fitted model was subsequently evaluated without modification in the Internal Testing and Multicentric Testing cohorts, achieving AUCs of 0.865 and 0.868, respectively. All SHAP and LIME analyses reported in the present study were performed using this single Random Forest model.

2.5 Explainability Analysis

Global and local feature attribution was assessed using SHAP (TreeExplainer, providing exact attribution for tree-based models) and LIME (local linear surrogate models fitted via perturbation sampling), computed on the Derivation Cohort. To assess cross-cohort consistency, SHAP values were additionally computed for the same fitted model in the Internal Testing and Multicentric Testing cohorts, and feature rankings were compared using Spearman rank correlation. To assess temporal stability, SHAP rankings were compared (i) between early (landmark time ≤6 months) and late (>6 months) assessments within the 12-month-horizon model and (ii) between the 12-month-horizon model and a separately trained 6-month-horizon model fitted using the same feature set and algorithm but a 6-month rather than 12-month outcome window.

LIME explanations were generated using the LimeTabularExplainer implementation (lime package) [3]. Given that the comparative behavior of SHAP and LIME was a primary focus of the study, the LIME configuration was specified in detail. For each explained instance, 5,000 synthetic neighborhood samples were generated (num_samples=5000). The default kernel width, defined as $\sqrt{n\_features} \times 0.75$ (approximately 3.67 for the 24-variable feature set), was used to determine the exponential weighting of perturbed samples according to their proximity to the instance being explained. Continuous variables were discretized into quartile-based bins before fitting the local linear surrogate (discretizer='quartile'), consistent with the package default for

continuous features. The eight binary variables and five one-hot-encoded tumor-location indicators were explicitly specified as categorical features using the categorical_features argument rather than being processed through the default continuous-variable discretization pathway. In a preliminary analysis, omission of this specification resulted in quartile-based discretization of binary 0/1 variables and reduced SHAP-LIME agreement. A fixed random_state of 42 was used for both the explainer's internal sampling and the selection of the 50 explained instances to ensure reproducibility. Attributions were extracted for the predicted-death class (label=1, corresponding to death within the 12-month horizon) using LIME's index-based feature-to-coefficient mapping (as_map()), thereby ensuring unambiguous alignment with SHAP feature ordering. Both SHAP and LIME attributions were expressed on the scale of contribution to the predicted probability of 12-month mortality, and no additional rescaling was applied. Only SHAP satisfies the efficiency property, whereby the sum of feature attributions for a given prediction equals the deviation of that prediction from the model's average output; LIME's local linear surrogate does not guarantee this property. Accordingly, the two methods' attributions were considered comparable in scale but were not assumed to be numerically additive in the same manner.

Agreement between SHAP and LIME was quantified in a random sample of 50 observations (50 × 24 = 1,200 paired feature-level attribution values) using Pearson and Spearman correlations of raw attribution values and Spearman correlation of feature-level mean absolute importance rankings. The sample size was constrained by the computational cost of LIME, which requires an independent local surrogate model to be fitted for each explained observation.

2.6 Statistical Software

All analyses were performed in Python 3.12 using scikit-learn [19], SHAP [2] and LIME [3]. Statistical significance was set at two-sided α=0.05.

III. Results

The combined study population comprised 2,020 patients (979 Derivation, 627 Internal Testing, and 414 Multicentric Testing) followed across 8,085 landmark assessments. Briefly, the Random Forest model achieved a cross-validated AUC of 0.844 (95% CI 0.823–0.865) in the Derivation Cohort and AUCs of 0.865 (95% CI 0.834–0.894) and 0.868 (95% CI 0.835–0.898) in the Internal Testing and Multicentric Testing cohorts, respectively, with well-preserved discrimination across all three cohorts. The present analysis focuses specifically on the explainability of this model.

3.1 Global Feature Importance (SHAP)

Tumor diameter and AFP were the dominant predictors of 12-month mortality, followed by the number of intrahepatic lesions, tumor location, albumin, and ECOG performance status (Figure 1). Higher tumor diameter, AFP, and lesion number were consistently associated with increased predicted risk, whereas higher albumin levels were associated with lower predicted risk.

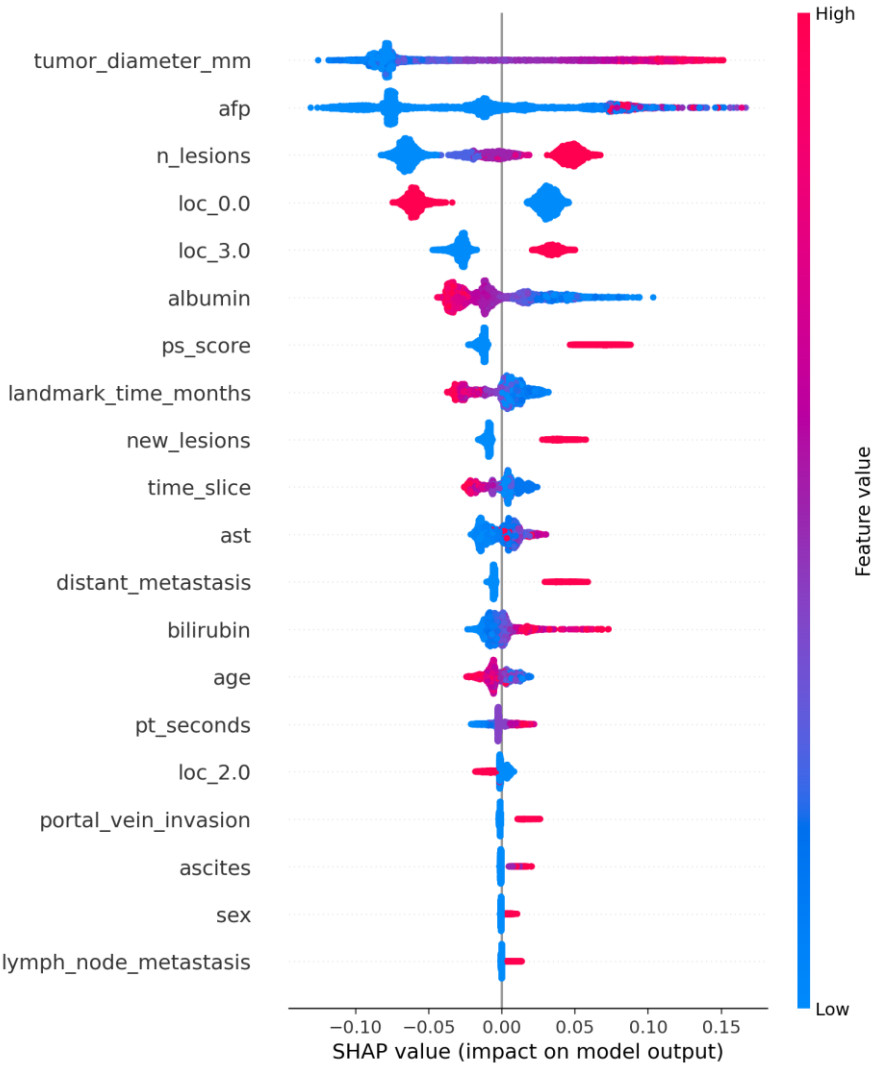

Figure 1. Global SHAP feature importance for 12-month mortality prediction.

SHAP dependence analysis (Figure 2) revealed non-linear, threshold-like relationships for the top predictors. Tumor diameter showed a steep increase in predicted risk up to approximately 100 mm, beyond which the SHAP contribution plateaued. AFP showed a sharp increase in predicted risk from low values followed by a plateau at higher concentrations. Albumin showed a strong, monotonic inverse relationship with predicted risk, whereas ECOG performance status ≥1 was associated with a large, discrete increase in predicted risk.

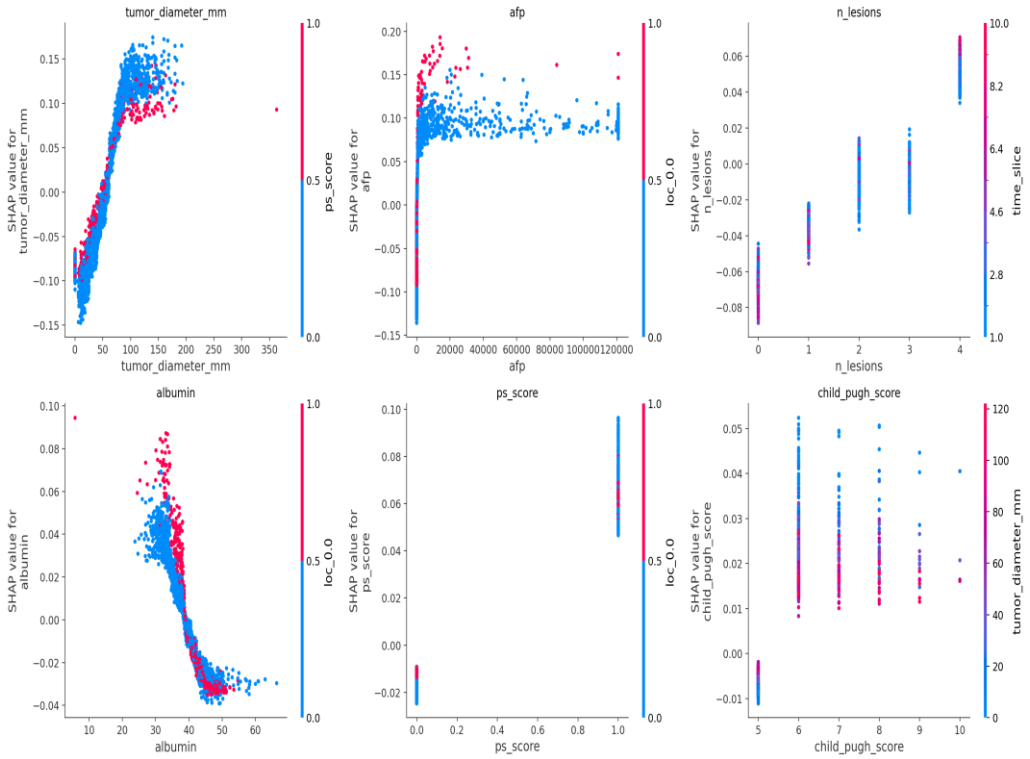

Figure 2. SHAP dependence plots for the six most important predictors.

3.2 Cross-Cohort Stability of SHAP Rankings

To assess whether these feature importance patterns were specific to the Derivation Cohort or remained consistent across settings, SHAP values were additionally computed for the same fitted model in the Internal Testing and Multicentric Testing cohorts (Table 1). Feature rankings were nearly identical across all three cohorts (Spearman rank correlation with Derivation: r=0.998 for Internal Testing, r=0.995 for Multicentric Testing; both p<0.0001). The seven most important variables were identical across the Derivation and Internal Testing cohorts (7/7) and largely concordant across the Derivation and Multicentric Testing cohorts (6/7), with new lesion development ranking modestly higher in the Multicentric Testing Cohort (Figure 3).

Table 1. Cross-cohort comparison of SHAP feature importance.

Feature	Derivation |SHAP|	Internal Testing |SHAP|	Multicentric Testing |SHAP|
Tumor diameter (mm)	0.0761	0.0765	0.0765
AFP (ng/mL)	0.0567	0.0553	0.0590
No. of lesions	0.0462	0.0446	0.0496
Tumor location (category 0)	0.0431	0.0424	0.0447
Tumor location (category 3)	0.0303	0.0305	0.0302
Albumin (g/L)	0.0248	0.0238	0.0255
ECOG performance status	0.0197	0.0196	0.0203
Landmark time (months)	0.0136	0.0129	0.0152
New lesions (yes/no)	0.0133	0.0147	0.0265
Time-slice index	0.0102	0.0097	0.0106

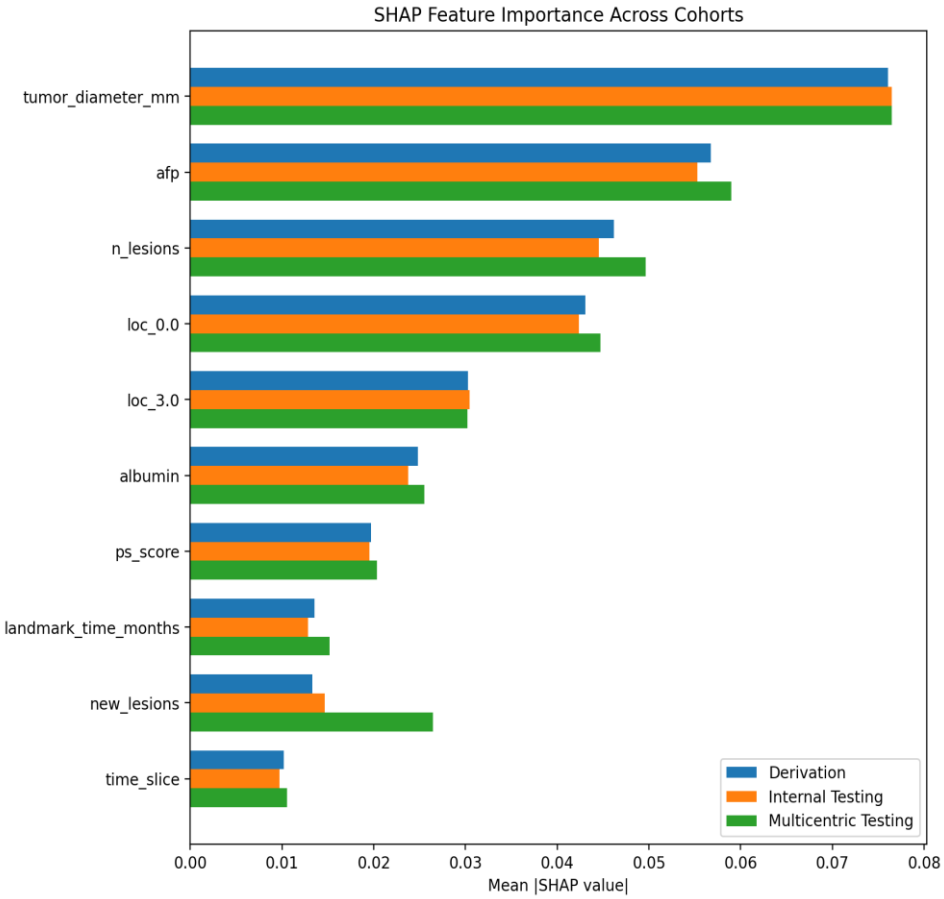

Figure 3. Cross-cohort stability of SHAP feature importance.

3.3 Temporal Stability of Feature Importance

Feature importance rankings remained highly stable across the disease course. Comparing SHAP rankings between early (landmark time ≤6 months) and late (>6 months) assessments, the seven most important variables retained an identical rank order; only lower-ranked variables showed minor re-ordering, without substantial changes in their overall contribution to model output.

This temporal stability was also observed when comparing the 6-month and 12-month horizon models (Table 2). The ten most important variables showed near-identical rankings across prediction horizons, with tumor diameter, AFP, and lesion number consistently ranked first through third.

Table 2. SHAP feature importance across 6- and 12-month prediction horizons.

Feature	Mean |SHAP|, 12-month model	Mean |SHAP|, 6-month model
Tumor diameter (mm)	0.0761	0.0936
AFP (ng/mL)	0.0567	0.0567
No. of lesions	0.0462	0.0521
Tumor location (category 0)	0.0431	0.0416
Tumor location (category 3)	0.0303	0.0264
Albumin (g/L)	0.0248	0.0323
ECOG performance status	0.0197	0.0244
Landmark time (months)	0.0136	0.0121
New lesions (yes/no)	0.0133	0.0173
Time-slice index	0.0102	0.0081

3.4 Agreement Between SHAP and LIME

Across a random sample of 50 observations, with binary and one-hot-encoded categorical variables explicitly declared to LIME as categorical features (Section 2.5), SHAP and LIME attributions showed strong agreement at the level of individual attribution values (Pearson r=0.871, p<0.0001; Spearman r=0.825, p<0.0001; Figure 4, left panel) and moderate-to-strong agreement at the level of feature importance rankings (Spearman r=0.721, p<0.0001). This represented a material improvement over the preliminary analysis in which categorical variables were processed using LIME's default continuous-variable discretization (value-level Pearson r=0.775; ranking Spearman r=0.585), indicating that SHAP-LIME agreement was sensitive to the handling of categorical and binary variables.

The two most important variables, tumor diameter and AFP, were consistently identified as dominant by both methods, although LIME assigned them larger mean absolute attribution values (Figure 4, right panel). Differences persisted for a subset of variables after categorical-feature specification. ECOG performance status ranked 6th by SHAP but 3rd by LIME, whereas distant metastasis ranked 11th by SHAP but 4th by LIME. Notably, ECOG performance status was a common, non-sparse predictor, indicating that the observed differences between SHAP and LIME were not limited to sparse binary variables.

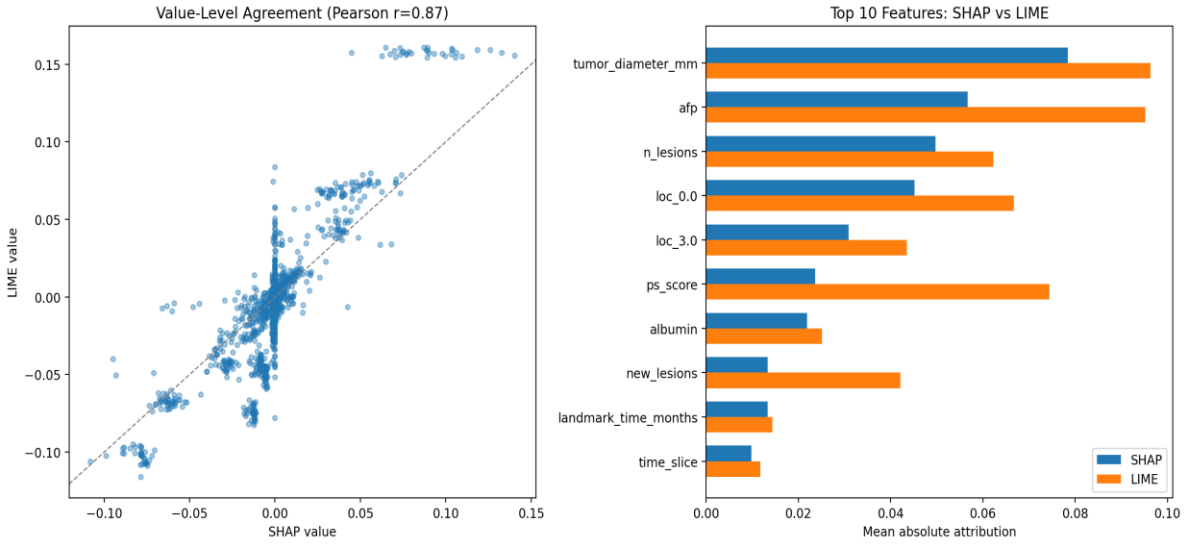

Figure 4. Agreement between SHAP and LIME feature attributions.

IV. Discussion

In this explainability analysis of an externally validated, landmark-based mortality prediction model for HCC, tumor diameter, AFP, and lesion number were consistently identified as the dominant predictors, with non-linear, threshold-like relationships with predicted risk. These rankings were highly stable across three independent cohorts, including a geographically distinct multicentric cohort, as well as across the disease course (early versus late landmarks) and prediction horizons (6 versus 12 months). SHAP and LIME showed strong agreement when categorical variables were explicitly specified in LIME, whereas agreement was lower when categorical variables were processed using LIME's default continuous-variable discretization. These findings indicate that the apparent degree of agreement between explainability methods can be influenced by implementation choices. This issue is particularly relevant in healthcare, where recent reviews have identified substantial heterogeneity in the application and evaluation of XAI methods and have emphasized the need for greater attention to robustness, validity, and standardized reporting of explanations [20, 21].

A body of methodological work has described a general “disagreement problem” in explainable machine learning, in which different attribution methods, or even the same method under different implementation choices, can yield materially different explanations for the same model and prediction [4–6].

Similar concerns regarding the stability and reliability of post-hoc explanations have increasingly been highlighted in healthcare applications [20-22].

Our findings provide a quantitative clinical illustration of two aspects of this problem. First, SHAP-LIME agreement was sensitive to explainer configuration: explicitly specifying binary and one-hot-encoded variables as categorical features in LIME increased value-level Pearson correlation from r=0.775 to r=0.871 and feature-ranking Spearman correlation from r=0.585 to r=0.721. This suggests that a substantial component of the observed disagreement was related to explainer configuration rather than differences in the underlying model or data. Second, residual disagreement persisted even after categorical variables were explicitly specified. ECOG performance status, a common and clinically important predictor, remained more highly ranked by LIME than by SHAP, suggesting that differences between the two methods cannot be explained by feature sparsity or categorical-feature handling alone. These observations are consistent with the distinct theoretical foundations of the two methods, with Shapley-value attribution underlying SHAP and local linear approximation underlying LIME, and emphasize the importance of both method selection and implementation when interpreting feature-attribution results.

Three practical implications follow. First, the strong cross-cohort and temporal stability of SHAP rankings in this model supports the use of feature attribution for communicating the model's dominant prognostic drivers to clinicians, particularly for the small number of variables that consistently dominate model output. Such stability is relevant to broader calls for clinically meaningful and rigorously evaluated explainability in healthcare [10-13]. Second, the sensitivity of SHAP-LIME agreement to LIME's categorical-feature configuration indicates that clinical XAI studies should explicitly report explainer settings, consistent with broader calls for transparent and standardized reporting of clinical prediction models and their explanations [20, 21]. Comparisons conducted using default settings without careful consideration of variable type may overstate disagreement between attribution methods. Third, the persistence of differences for a common, non-sparse predictor after categorical-feature specification supports caution when clinical interpretations are based on secondary predictors identified by a single attribution method. Where feasible, such findings should be corroborated using more than one attribution approach, or their uncertainty should be explicitly acknowledged, before being used to inform clinical interpretation or prospective validation.

V. Limitations

Several limitations merit consideration. First, LIME's computational cost restricted the SHAP-LIME comparison to a random sample of 50 observations. Although this yielded 1,200 paired feature-level attribution values, these values were derived from the same 50 observations, and a larger sample would provide more robust estimates of agreement, particularly for lower-ranked variables. Relatedly, only one explainer configuration choice, categorical-feature specification, was evaluated with respect to its effect on SHAP-LIME agreement. Other implementation choices, such as the number of perturbation samples or the discretization strategy for continuous variables, were not systematically varied. Accordingly, the reported agreement estimates should be interpreted as conditional on the specific LIME configuration described in Section 2.5 rather than as a general property of LIME.

Second, both SHAP and LIME are post-hoc attribution methods applied to an already-fitted model, and neither provides a ground-truth explanation against which the other can be validated [6]. Therefore, the present findings characterize agreement between two attribution approaches rather than the accuracy of either method in representing the model's underlying decision process. Moreover, adversarial analyses have demonstrated that post-hoc explanations can, in principle, be manipulated independently of a model's predictive behavior [5].

Finally, these explainability findings were derived from a dataset encompassing a limited number of centers; their generalizability to other populations, clinical settings, cancer types, and treatment contexts remains to be established.

VI. Conclusion

Feature attributions from an externally validated, landmark-based mortality prediction model for hepatocellular carcinoma were highly stable across independent cohorts, across the disease course, and across prediction horizons, supporting the consistency of SHAP-derived rankings for identifying the model's dominant prognostic drivers. SHAP-LIME agreement was substantially improved by explicitly declaring categorical variables to LIME rather than relying on default settings, although differences persisted for a subset of predictors, including at least one common, non-sparse variable. Explainability findings in clinical machine learning should therefore be reported with explicit attention to both the attribution method and its configuration and, where feasible, corroborated using more than one approach before informing clinical interpretation.

VII. References

[1] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems,* vol. 30, 2017.

[2] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair*, et al.*, "From Local Explanations to Global Understanding with Explainable AI for Trees," *Nat Mach Intell,* vol. 2, pp. 56-67, Jan 2020.

[3] M. T. Ribeiro, S. Singh, and C. Guestrin, "" 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.

[4] S. Krishna, T. Han, A. Gu, S. Wu, S. Jabbari, and H. Lakkaraju, "The disagreement problem in explainable machine learning: A practitioner's perspective," *arXiv preprint arXiv:2202.01602,* 2022.

[5] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling lime and shap: Adversarial attacks on post hoc explanation methods," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 180-186.

[6] I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler, "Problems with Shapley-value-based explanations as feature importance measures," in *International conference on machine learning*, 2020, pp. 5491-5500.

[7] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. Moons, "Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement," *Ann Intern Med,* vol. 162, pp. 55-63, Jan 6 2015.

[8] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608,* 2017.

[9] C. Molnar, *Interpretable machine learning*: Lulu. com, 2020.

[10] C. Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *Nat Mach Intell,* vol. 1, pp. 206-215, May 2019.

[11] Z. C. Lipton, "The mythos of model interpretability," *Communications of the ACM,* vol. 61, pp. 36-43, 2018.

[12] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *Lancet Digit Health,* vol. 3, pp. e745-e750, Nov 2021.

[13] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, "What clinicians want: contextualizing explainable machine learning for clinical end use," in *Machine learning for healthcare conference*, 2019, pp. 359-380.

[14] M. Reig, A. Forner, J. Rimola, J. Ferrer-Fàbrega, M. Burrel, Á. Garcia-Criado*, et al.*, "BCLC strategy for prognosis prediction and treatment recommendation: The 2022 update," *J Hepatol,* vol. 76, pp. 681-693, Mar 2022.

[15] "EASL Clinical Practice Guidelines: Management of hepatocellular carcinoma," *J Hepatol,* vol. 69, pp. 182-236, Jul 2018.

[16] L. Shen, Q. Zeng, P. Guo, J. Huang, C. Li, T. Pan*, et al.*, "Dynamically prognosticating patients with hepatocellular carcinoma through survival paths mapping based on time-series data," *Nat Commun,* vol. 9, p. 2230, Jun 8 2018.

[17] H. C. Van Houwelingen, "Dynamic prediction by landmarking in event history analysis," *Scandinavian Journal of Statistics,* vol. 34, pp. 70-85, 2007.

[18] L. Breiman, "Random forests," *Machine learning,* vol. 45, pp. 5-32, 2001.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel*, et al.*, "Scikit-learn: Machine learning in Python," *the Journal of machine Learning research,* vol. 12, pp. 2825-2830, 2011.

[20] J. Caterson, A. Lewin, and E. Williamson, "The application of explainable artificial intelligence (XAI) in electronic health record research: A scoping review," *Digit Health,* vol. 10, p. 20552076241272657, Jan-Dec 2024.

[21] Y. Abas Mohamed, B. Ee Khoo, M. Shahrimie Mohd Asaari, M. Ezane Aziz, and F. Rahiman Ghazali, "Decoding the black box: Explainable AI (XAI) for cancer diagnosis, prognosis, and treatment planning-A state-of-the art systematic review," *Int J Med Inform,* vol. 193, p. 105689, Jan 2025.

[22] R. Shankar, Z. Goh, F. Devi, and Q. Xu, "A systematic review of explainable artificial intelligence methods for speech-based cognitive decline detection," *NPJ Digit Med,* vol. 8, p. 724, Nov 26 2025.