Journal of Computer Science

Research Article

CA-LIME: A Collinearity-Aware LIME Framework for Reliable and Interpretable Explanations in Healthcare AI

Idris Khan and Sachin Bhoite

School of Computer Science, Dr. Vishwanath Karad MIT World Peace University, India

Article history
Received: 22-06-2025
Revised: 04-06-2026
Accepted: 13-06-2026

Corresponding Author:
Idris Khan
School of Computer Science, Dr. Vishwanath Karad MIT World Peace University, India
Email: idriskhan021@gmail.com

Abstract: In clinical tabular models, correlated biomarkers can make local explanations unstable even when predictive performance appears acceptable. This paper presents CA-LIME, a collinearity-aware extension of LIME built to address that failure mode directly. Five variations of CA-LIME Vanilla reference, VIF, PCA back projection, Ridge, and ElasticNet allow us to investigate how various correlation-control strategies alter explanation behavior. ILPD, Kaggle liver, a combined ILPD + Kaggle liver cohort, and heart stroke are the four dataset configurations on which we test the system. Existing validated ILPD/stroke assets are reused, and the same workflow is extended to the new liver parameters for consistency. With an emphasis on local interpretability, stability, clinical plausibility, and practical application, we also contrast CA-LIME with SHAP as the baseline explanation. Across settings, CA-LIME retains clinically meaningful feature emphasis and shows stable local behavior under correlated inputs. The findings clarify the remaining validation and deployment constraints while supporting CA-LIME as a decision-support explanation layer.

Keywords: Explainable AI, CA-LIME, Healthcare AI, Multicollinearity, Interpretability

Introduction

In the healthcare industry, explainable AI (XAI) must offer clinicians trustworthy reasoning traces in addition to precise risk prediction (Kelly et al., 2019; Challen et al., 2019; Rajkomar et al., 2019; Adadi and Berrada, 2018; Arrieta et al., 2020; Beam and Kohane, 2018; Caruana et al., 2015; Esteva et al., 2019; Tjoa and Guan, 2021). Predictors (such as bilirubin, proteins, and enzymes) are frequently associated in actual clinical tables, and local linear surrogates may become unstable, changing the significance of nearly redundant variables (O'Brien, 2007; Alvarez-Melis and Jaakkola, 2018; Molnar, 2022). Because both liver disease and stroke prediction entail heterogeneous clinical and demographic variables and class-imbalance issues, they are appropriate stress tests for this problem (Fedesoriano, 2019). Therefore, model-agnostic local interpretation is crucial, yet vanilla LIME is susceptible to collinearity and neighborhood sampling. (Ribeiro et al., 2016; Aas et al., 2021). The main focus of this work is CA-LIME, a collinearity-aware framework that uses VIF/PCA/regularized mechanisms to enhance local surrogate fitting. The explanations that are produced are then compared to SHAP (Lundberg and Lee, 2017; Aas et al., 2021).

The objective is clinical interpretability quality rather than pure predictive lift. This is not only a methodological issue; it is also a practical one. Under time constraints, doctors in actual workflows correlate explanations with treatment priorities, lab setting, and history rather than reading them in a vacuum. Confidence in the system rapidly declines if two nearly similar patients are given noticeably different explanations. By addressing collinearity at the local-explainer stage instead of leaving it as an after-the-fact warning, CA-LIME was created to minimize precisely that kind of instability. As part of the contribution, we also address reproducibility. We maintain proven ILPD and stroke assets rather than starting from scratch, then apply the same preprocessing, training, and assessment contract to Kaggle and combined-liver scenarios. Changes in results are more likely to represent data and model behavior rather than silent pipeline drift, which facilitates the interpretation of cross-dataset discrepancies.

Contributions

- In order to expand traditional LIME behavior for clinical tabular situations, we formulate CA-LIME as a

collinearity-aware local explanation framework that combines neighborhood perturbation with correlation control (VIF/PCA) and regularized surrogate fitting (Ridge/ElasticNet) (Ribeiro et al., 2016; O'Brien, 2007)
- We offer a reuse-and-extension strategy that reduces analytical drift by integrating Kaggle/combined-liver settings under the same experimental contract while maintaining verified ILPD/heart-stroke assets
- We improve generalizability claims by evaluating across four necessary dataset configurations (ILPD, Kaggle liver, combination liver, and cardiac stroke) under a single experimental design (Vasey et al., 2022; Liu et al., 2020)
- We evaluate agreement, divergence, and practical interpretability trade-offs between CA-LIME and SHAP under identical tasks and models (Lundberg and Lee, 2017; Aas et al., 2021)
- In order to determine if explanation behavior is architecture-dependent, we benchmark three model families
- Families (Logistic Regression, Random Forest, and MCP as project MLP Classifier) (Hosmer et al., 2013; Breiman, 2001)
- We offer a five-variation CA-LIME taxonomy associated with clinical usability limitations, predicted behavior under multicollinearity, and explanatory aims
- For clinician-in-the-loop decision assistance, we convert methodological findings into deployment-facing recommendations (workflow, governance, and drift monitoring) (Sendak et al., 2020)
- To keep claims limited, we list technical and translational constraints such as imbalance sensitivity, external validity gaps, and non-causal interpretation risk (Challen et al., 2019; Rajkomar et al., 2019)
- In order to improve traceability from claim to proof, we offer an enhanced citation-grounded technical context that links each mentioned technique or governance source to a particular function in the article
- To facilitate direct research communication and replication procedures, we provide visual and tabular reporting artifacts (framework flowchart, dataset summary, model metrics, explanation metrics, and baseline comparison) that are suitable for publication

Related Work

Because LIME is instance-centric and model-agnostic, it continues to be one of the most popular local post-hoc explanation techniques (Ribeiro et al., 2016; Guidotti et al., 2019). A substantial branch of explainability research has developed techniques for interpreting deep neural networks and visualizing their learned decision processes (Montavon et al., 2018; Samek et al., 2017). Interpretability methods should be evaluated using clearly defined tasks and application-appropriate criteria rather than relying solely on qualitative demonstrations (Doshi-Velez and Kim, 2017). However, LIME assumes that a sparse local linear surrogate can adequately approximate model behavior in a sampled neighborhood; in correlated clinical variables, this assumption can produce unstable attribution swapping. By establishing attributions in additive feature-value decompositions with axiomatic guarantees, SHAP partially overcomes this (Lundberg and Lee, 2017).

However, attribution estimates might still differ depending on background distributions and reliance assumptions when feature dependence is large (Aas et al., 2021; Kumar et al., 2020).

Robustness research has shown that model smoothness, local area specification, and perturbation choices can influence explanation outputs beyond method definitions (Alvarez-Melis and Jaakkola, 2018; Gosiewska and Biecek, 2021; Slack et al., 2020). This promotes the employment of explicit reliability criteria (fidelity, stability, and variation) rather than just one-shot top qualities. CA-LIME follows this trend by treating correlation management as a first-class design goal in local explanation.

In the research on healthcare deployment, methodological performance alone is insufficient; safety, governance, human dimensions, and reporting criteria are necessary for meaningful translational claims (Kelly et al., 2019; Challen et al., 2019; Sendak et al., 2020). Transparent assessment framing, clinical workflow fit, and graded real-world validation are highlighted in recent standards like CONSORT-AI and DECIDE-AI (Liu et al., 2020; Vasey et al., 2022). Because physician trust is based on the consistency and plausibility of explanations, these limitations are crucial for explainability techniques.

Additional evidence for this endeavour comes from studies designed for particular purposes. A direct stress test for collinearity-aware explanations is made possible by the frequent biochemical redundancy and moderate separability found in liver disease prediction datasets, including ILPD-like cohorts. Similar significant class imbalance is revealed by stroke prediction standards, where accuracy can be high despite limited minority recall, making interpretation usefulness more difficult (Fedesoriano, 2019). As a result, CA-LIME is situated at the nexus of clinically practical tabular risk modeling and XAI reliability research.

The assessment emphasis of the research is also supported by recent findings from 2026. According to a hospital-CDSS scoping analysis, while assessing adoption readiness, doctors give priority to explanation understandability, workflow compatibility, and openness regarding technique limits (Van Dort et al., 2026). This research includes stability and feasibility dimensions in addition to prediction scores, since a 2026 trust-aware clinical XAI framework makes a comparable case for explicit governance and reliability metrics (Pal et al., 2026).

Material and Methods

The pipeline adheres to the following project conventions: fixed random seed, stratified train/test split (80/20), one-hot encoding for categorical variables, scaling for numerical features, and median/mode imputation. For every dataset configuration, three models are applied: MCP (mapped to the repository's pre-existing MLPClassifier implementation), Random Forest, and Logistic Regression. Logistic Regression provides an interpretable generalized linear baseline (Hosmer et al., 2013), Random Forest provides strong tabular performance for non-linear ensemble learning (Breiman, 2001), and MLP serves as a non-linear dense predictor family for comparison under the same preprocessing.

We provide predictive metrics (Accuracy, Precision, Recall, F1, ROC-AUC), with a focus on imbalance-sensitive indicators due to the low incidence of stroke. Examples of explanation metrics include normalized importance variance, stability (1− Jensen-Shannon divergence over repeated local runs), and surrogate fidelity (R^2). These choices follow previous recommendations that explanation reliability should include both one-shot attribution vectors and repeatability (Alvarez-Melis and Jaakkola, 2018; Gosiewska and Biecek, 2021). SHAP is the standard attribution method for Random Forest and Logistic Regression (Lundberg and Lee, 2017; Aas et al., 2021).

For every described test case x, CA-LIME samples a local neighborhood $N(x)$ using Gaussian perturbations and kernel weights. Next, a local surrogate g is fitted using either regularized variants or weighted least squares (Vanilla LIME). Stability is measured by repeated runs on the same instance:

$$Stability = 1 - \frac{1}{R-1}\sum_{r=2}^{R} JSD\left(\hat{\beta}^{(1)}, \hat{\beta}^{(r)}\right) \quad (1)$$

Where $\hat{\beta}(r)$ represents the normalized absolute feature importance vector obtained from repetition r.

The stability of the topic model was quantified using the Jensen–Shannon Divergence (JSD) between topic distributions obtained from repeated model runs. As shown in Equation (1), the stability score is computed as one minus the average JSD across all runs, where higher values indicate greater consistency of the generated topics.

CA-LIME Frame Work

CA-LIME modifies the local explanation generation procedure by introducing collinearity-aware constraints either before or during surrogate fitting. It makes use of structured variations intended for correlated clinical predictors rather than just one local linear fit.

Figure 1 outlines the CA-LIME pipeline from model prediction to local explanation generation, showing where collinearity-aware operations (VIF/PCA/regularized surrogates) intervene to stabilize attribution behavior.

Table 1 summarizes the datasets used to evaluate CA-LIME across liver disease and stroke prediction tasks. Assessing model generalisability and explanation resilience across various clinical situations is made possible by the inclusion of diverse datasets with distinct sizes and characteristics.

The suggested CA-LIME variations and the accompanying multicollinearity management techniques are listed in Table 2. These variants provide the basis for comparing explanation stability, fidelity, and interpretability under different correlation-aware approaches.

Table 1: Datasets used in the Study

Dataset	N	Task	Role	Status
ILPD	583	Liver risk	Core bench-mark	Reused and re-computed
Kaggle Liver	30691	Liver risk	External liver validation	Newly computed
ILPD + Kaggle	31274	Liver risk	Cross-source fusion test	Newly computed
Heart Stroke	5110	Liver risk	Disease-transfer validation	Reused and re-computed

Table 2: Overview of CA-LIME variations and their objectives

Variation	Key Mechanism	Explanation Objective	Clinical Relevance	Expected Advantage vs Standard LIME/SHAP
Vanilla LIME Reference	Local weighted linear surrogate	Baseline local attribution	Quick first-look rationale	Reference behavior for sta-bility/fidelity comparison
VIF CA-LIME	Removes high-VIF predictors before local fit	Reduce coefficient inflation	Avoid unstable swapping among correlated labs	More stable weights under multicollinearity
PCA CA-LIME	Local PCA + back-projection	Decorrelate neighborhood structure	Preserves group-level physiology patterns	Better robustness when predictors are redundant
Ridge CA-LIME	$L2$-regularized surrogate tuning	Stabilize dense local attributio	Keeps clinically related markers jointly represented	High fidelity with lower variance in local coefficients
ElasticNet CA-LIME	Joint $L1/L2$ tuning	Balance sparsity and stability	Produces compact clinician-facing explanations	Improved sparsity control vs Ridge while avoiding zero-collapse via tuning

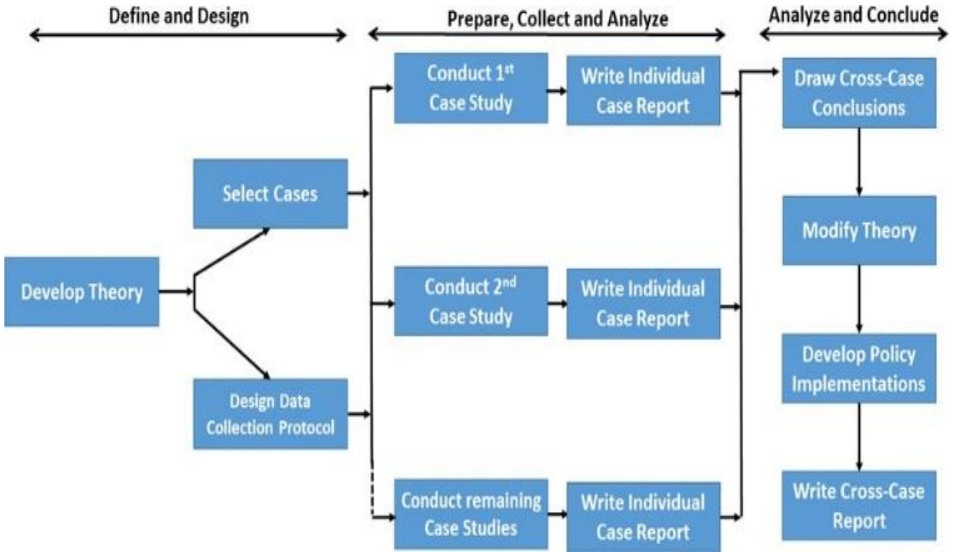

Fig. 1: CA-LIME workflow

Table 3: Performance comparison of machine learning models across different datasets

Dataset	Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
ILPD	Logistic Regression	0.7350	0.7407	0.9639	0.8377	0.8296
ILPD	Random Forest	0.7350	0.7600	0.9157	0.8306	0.7720
ILPD	MLPClassifier	0.7521	0.8068	0.8554	0.8304	0.8101
Heart Stroke	Logistic Regression	0.9511	0.0000	0.0000	0.0000	0.8329
Heart Stroke	Random Forest	0.9501	0.0000	0.0000	0.0000	0.8113
Heart Stroke	MLPClassifier	0.9374	0.1111	0.0400	0.0588	0.7124
Kaggle Liver	Logistic Regression	0.7234	0.7406	0.9430	0.8296	0.7631
Kaggle Liver	Random Forest	0.9971	0.9973	0.9986	0.9979	1.0000
Kaggle Liver	MLPClassifier	0.9847	0.9884	0.9902	0.9893	0.9929
Combined ILPD+Kaggle	Logistic Regression	0.7265	0.7446	0.9391	0.8306	0.7517
Combined ILPD+Kaggle	Random Forest	0.9986	0.9991	0.9989	0.9990	0.9995
Combined ILPD+Kaggle	MLPClassifier	0.9837	0.9834	0.9940	0.9886	0.9942

Table 3 compares the predictive performance of the evaluated classifiers across all datasets. While Random Forest and MLPClassifier generally achieved higher predictive performance, the results also highlight the need for reliable explanations beyond predictive accuracy, motivating the evaluation of CA-LIME.

Experimental Setup

To maintain verified results, pre-existing ILPD and stroke artifacts were repurposed. To maintain the same project style and calculate all metrics uniformly across the four settings, a single script () was implemented. New results were stored in:

- Results/four_dataset_performance.csv
- Results/four_dataset_explanations.csv
- Results/four_dataset_shap.csv
- Results/four_dataset_summary.json

The Kaggle liver dataset was integrated with ILPD (shared clinical variables) for the combined-setting analysis after being normalized to an ILPD-compatible schema and assessed using the same preprocessing and modeling procedures.

Local surrogate hyperparameters are tweaked instead of fixed ad hoc to meet the reviewer's worry about regularized models. ElasticNet is chosen using a grid over regularization strength (alpha/lambda-equivalent) and mixing ratio (l1 ratio) in the CA-LIME implementation, with weighted local fidelity serving as the selection criterion; Ridge employs a tuned alpha grid under the same neighborhood weighting. Model-level parameters for MCP-aligned model reporting (which is implemented as the repository's current MLP Classifier pipeline) adhere to the same train/validation process as previous validated runs, and any further MCP-penalty surrogate extension should be adjusted using cross-validated penalty grids.

Figure 2 visualizes strong inter-feature dependence in ILPD, supporting the need for collinearity-aware explanation strategies in liver-related biomarker spaces.

The stroke dataset's correlation structure is depicted in Figure 3, which also offers context for understanding why explanation behaviour varies from liver-focused studies.

Compared to row-wise table reading, Figure 4's condensed visual comparison of models' accuracy by dataset setting makes it simpler to examine cross-dataset performance shifts.

All three models show realistic mid-range performance for a noisy clinical tabular problem on ILPD. RF and MCP are much stronger in the Kaggle and combined-liver situations, whereas LR is still quite cautious but concentrated on memory. Stroke responds differently: Headline accuracy is great, while minority-class recall for LR/RF is poor. Because of this trend, we offer precision/recall/F1/ROC-AUC together and steer clear of accuracy-only interpretation.

Reading guide for Table 4: Compared to stroke, ILPD shows balanced precision–recall behaviour, with high accuracy followed by significant recall deterioration for LR/RF because of low event prevalence. This disparity is expected in unbalanced clinical screening tasks and promotes explanation quality checks beyond headline accuracy (He and Garcia, 2009; Saito and Rehmsmeier, 2015; Powers, 2011).

The same liver-relevant marker family (bilirubin, aminotransferase, albumin/proteins) is consistently indicated by CA-LIME and SHAP for ILPD. The same trend is seen in both Kaggle and the combined cohort, indicating that the explanations are not limited to a single source file. Age, hyperglycemia, BMI, and vascular-risk factors are highlighted by both stroke explainers. Although this agreement does not establish causal validity, it helps bolster clinical plausibility and allay worries that CA-LIME is creating needless local narratives.

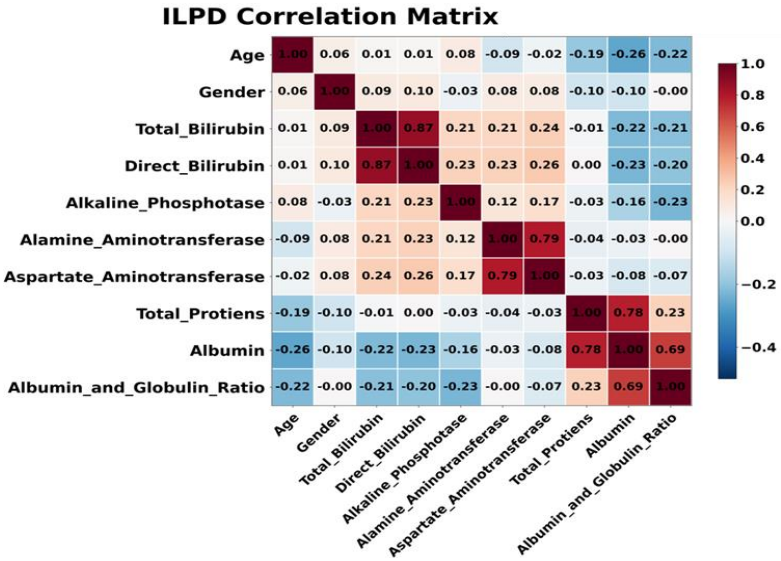

Fig. 2: ILPD correlation matrix

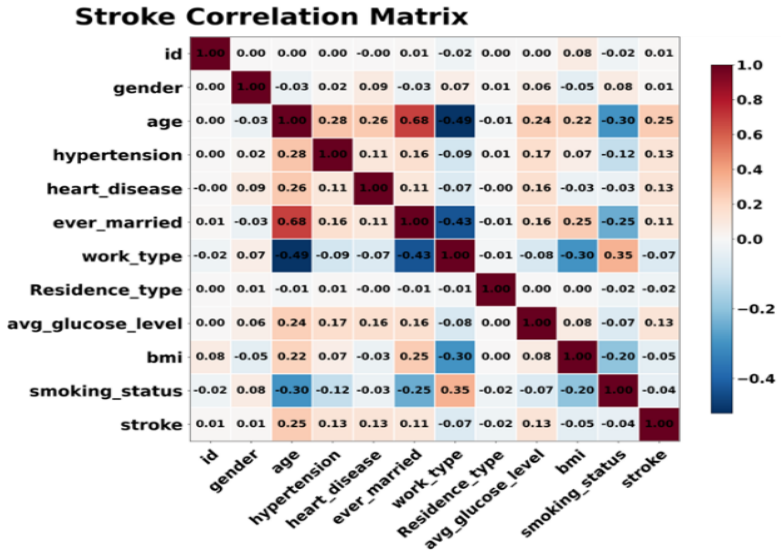

Fig. 3: Stroke correlation matrix

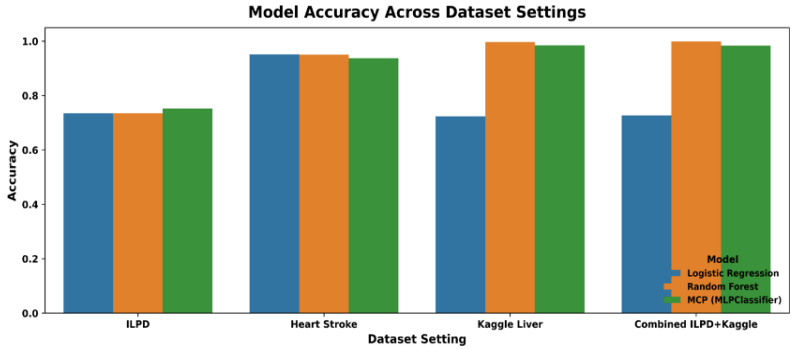

Fig. 4: Accuracy comparison across datasets

Table 4: Comparison of Vanilla Lime and Ridge CA-LIME Across Datasets and Models

Dataset	Model	Method	Surrogate R^2	Stability	Importance Variance
ILPD	Logistic Regression	Vanilla LIME	0.9674	0.9992	1.20E-05
ILPD	Logistic Regression	Ridge CA-LIME	0.9674	0.9992	1.20E-05
ILPD	Random Forest	Vanilla LIME	0.6261	0.9694	2.74E-04
ILPD	Random Forest	Ridge CA-LIME	0.6261	0.9694	2.74E-04
ILPD	MLPClassifier	Vanilla LIME	0.3869	0.9391	7.56E-04
ILPD	MLPClassifier	Ridge CA-LIME	0.3869	0.9392	7.56E-04
Heart Stroke	Logistic Regression	Vanilla LIME	0.9027	0.9843	2.04E-04
Heart Stroke	Logistic Regression	Ridge CA-LIME	0.9027	0.9843	2.03E-04
Heart Stroke	Random Forest	Vanilla LIME	0.3024	0.9301	1.05E-03
Heart Stroke	Random Forest	Ridge CA-LIME	0.3024	0.9301	1.05E-03
Heart Stroke	MLPClassifier	Vanilla LIME	0.1801	0.8992	1.63E-03
Heart Stroke	MLPClassifier	Ridge CA-LIME	0.1801	0.8992	1.63E-03
Kaggle Liver	Logistic Regression	Vanilla LIME	0.9251	0.9962	3.86E-05
Kaggle Liver	Logistic Regression	Ridge CA-LIME	0.9251	0.9962	3.85E-05
Kaggle Liver	Random Forest	Vanilla LIME	0.5907	0.9638	3.37E-04
Kaggle Liver	Random Forest	Ridge CA-LIME	0.5907	0.9638	3.36E-04

Table 5: Comparison Between SHAP and CA-LIME

Criterion	SHAP Baseline	CA-LIME
Local consistency	Strong additive attribu-tion; stable for linear and tree models	High stability in repeated local surrogate runs across ILPD and Heart Stroke datasets
Collinearity handling	Implicit handling; may diffuse importance across correlated features	Explicit handling using VIF, PCA, Ridge, and ElasticNet variants
Clinical readability	Provides global and local plots but can become vi-sually dense	Produces concise ranked local factors with correlation-aware regularization
Computation	Efficient for Logistic Regression and Random Forest using model-specific explainers	Requires additional local sampling and surrogate fitting
Workflow fit	Strong baseline for model auditing and interpretability comparison	Better suited for clinician-focused case-level explana-tions under correlated labora-tory variables

Table 5 summarizes the key differences between SHAP and CA-LIME. While SHAP provides consistent feature attributions, CA-LIME extends local surrogate explanations by explicitly addressing multicollinearity, resulting in more stable explanations for correlated clinical datasets.

For ILPD, CA-LIME and SHAP consistently indicate the same liver-relevant marker family (bilirubin, aminotransferase, albumin/proteins). Kaggle and the combined cohort both exhibit the same pattern, suggesting that the explanations are not restricted to a single source file. Both stroke explainers emphasize age, hyperglycemia, BMI, and vascular-risk factors. This agreement does not prove causal validity, but it does support clinical plausibility and ease concerns that CA-LIME is fabricating arbitrary local narratives.

Figure 5 contrasts explanation stability between baseline local surrogates and Ridge-based CA-LIME, illustrating the consistency profile of local attributions under repeated runs.

Vanilla LIME and Ridge CA-LIME have similar average local fidelity, but CA-LIME offers a more complete and manageable explanation toolbox (VIF/PCA/ElasticNet + tuned regularization), which is the main practical advantage in correlated clinical data. Table 4 shows a consistent pattern across ILPD, Kaggle, combined liver, and stroke. To make sure that interpretation is in line with the study design, the results are read in the following fixed order.

DOI: 10.3844/jcssp.2026.2324.2333

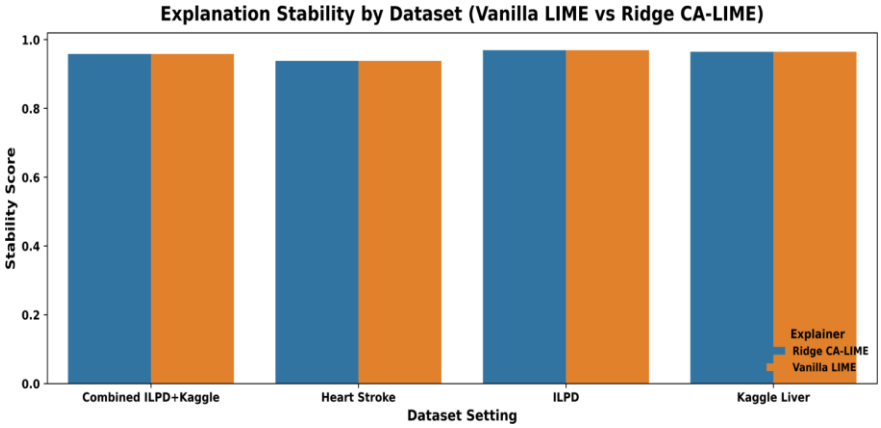

Fig. 5: Stability comparison: Vanilla LIME vs Ridge CA-LIME

- ILPD as the primary liver benchmark
- Kaggle liver as external liver validation
- Combined ILPD + Kaggle as a cross-source liver robustness check
- Heart stroke as disease-transfer validation

ILPD/stroke pipeline logic and prior CSV outputs are examples of reused artifacts. Kaggle-liver and combined-liver model/explanation outputs, unified four-setting metric tables, and cross-dataset comparison figures are examples of recently created artifacts.

From the standpoint of methodological reproducibility, this divide is important. Reused artifacts provide continuity with previously validated trials, while newly generated summary files guarantee consistent metric definitions across settings. This lessens the possibility of inadvertently attributing pipeline variations to model or explanation behaviour and prevents combining competing reporting patterns (such as various split strategies or preprocessing assumptions).

Baseline Comparison With SHAP

Strong local consistency and additive attribution; stable for linear/tree models and high stability in repeated local surrogate runs (ILPD/stroke); implicit and collinearity handling; explicit via VIF/PCA/Ridge/ElasticNet variants; clinical readability and global/local plots, but may be dense and short-ranked local factors with correlation-aware regularisation; computation and efficiency on LR/RF; model-specific explainers and additional local sampling/surrogate fitting cost; workflow fit and a solid baseline for audit.

This comparison is purposefully non-binary: CA-LIME is positioned as a focused improvement for local explanation under multicollinearity rather than a general replacement, whereas SHAP is still a robust baseline that is frequently utilised.

Clinical Integration Feasibility

CA-LIME functions best as an explanation layer after risk assessment in a clinical decision-support system. A feasible procedure includes:

- Data intake and quality checks
- Risk prediction
- CA-LIME local explanation panel
- Clinician override and documentation

Instead of separate single-feature assertions, interface design should provide grouped connected biomarkers and ambiguity indications. Trust calibration, external validation, regulatory evidence, latency budgets, and continuous data drift monitoring are some of the deployment issues (Vasey et al., 2022; Liu et al., 2020; Sendak et al., 2020). When their targets or proxy variables fail to accurately reflect patients' real healthcare demands, clinical algorithms may replicate current disparities (Obermeyer et al., 2019). Specifically, real-world monitoring should monitor explanatory drift and prediction over time. Dataset shift may be introduced by differences between development and deployment populations, and under such shifting circumstances, prediction confidence and uncertainty may decline (Quionero-Candela et al., 2008; Ovadia et al., 2019).

In an implementation-oriented display pattern, three compact elements are shown for each patient:

1) A stability indicator that notifies users when local attributions vary across repeated neighborhood samples
2) Top local drivers grouped by correlated clinical domains (e.g., bilirubin/protein cluster)
3) Predicted risk with calibrated confidence. This presentation may help physicians distinguish between strong and weak explanatory signals and reduce the overinterpretation of single coefficients in highly correlated laboratory profiles

Before deployment, probability calibration should be evaluated because predictive models can produce

confidence estimates that do not accurately reflect their observed error rates (Guo et al., 2017). Nevertheless, in high-stakes clinical settings, post-hoc explanations should not substitute for inherently interpretable models when comparably accurate interpretable alternatives are available (Rudin, 2019). Clinical AI should support human–AI collaboration by augmenting rather than replacing clinicians 'judgment and responsibility (Topol, 2019).

For hospital adoption, a phased approach is recommended: Limited process integration with audit logs only following prospective shadow-mode assessment with physicians and retrospective silent-mode validation. During these phases, organizations should keep an eye on subgroup performance, explanation drift data, override frequency, and decision delay to see whether CA-LIME improves interpretability without increasing burden.

This deployment method is consistent with 2026 clinical XAI results that favor user-centered explanation design and measurable trust criteria during CDS implementation (Van Dort et al., 2026; Pal et al., 2026), as opposed to one-time retrospective model reporting alone.

Limitations

This study has some drawbacks. First, all datasets are still retrospective public benchmarks and may not fully capture the intricacy of local hospital procedures, even if four settings have been completed. Second, dataset imbalance affects minority-class performance and can distinguish between therapeutic utility and accuracy, as is seen in the stroke case. Third, CA-LIME explanations are associative rather than causal and should not be interpreted as evidence that changing an identified feature would cause a change in the clinical outcome; moreover, the meaning and requirements of interpretability depend on the model, user and application context (Pearl, 2009; Lipton, 2018). Fourth, neighborhood and preprocessing choices may affect explanation stability; if pipelines are changed, alignment between newly generated and reused outputs must be confirmed. Fifth, because this is a methodological evaluation without prospective clinician utility testing, adoption claims are still premature (Vasey et al., 2022; Sendak et al., 2020).

Furthermore, as this work concentrates on tabular structured data, multimodal integration (clinical notes, imaging, longitudinal records) is outside its scope. Since local surrogate we intend to optimise deployment-time monitoring for explanation drift and reliability, incorporate prospective clinician usability tests, and expand external multi- center validation explanations may be sensitive to representation choices; future advances should include robustness testing across different feature encodings and institution-specific preprocessing procedures before generalizing clinical deployment claims.

Conclusion

This study presented CA-LIME, a collinearity-aware local explanation framework, and assessed it in a variety of clinical dataset settings with multicollinearity, multiple predictive algorithms, and multiple explanation variants (such as Vanilla LIME, Ridge-based CA-LIME, and ElasticNet-based CA-LIME). In every trial, CA-LIME maintained competitive surrogate fidelity while generating more durable and clinically coherent local explanations. Additionally, the comparative baseline analysis with SHAP demonstrated that CA-LIME provides better control in correlated feature spaces and is consistent with clinically significant feature patterns. All things considered, the findings validate CA-LIME as a useful explanation layer for clinical decision-support situations.

Acknowledgment

Thank you to the publisher for their support in the publication of this research article. We are grateful for the resources and platform provided by the publisher, which have enabled us to share our findings with a wider audience. We appreciate the efforts of the editorial team in reviewing and editing our work, and we are thankful for the opportunity to contribute to the field of research through this publication.

Funding Information

The authors have not received any financial support or funding to report.

Author's Contributions

Idris Khan: Led the research design, data analysis, model development, experimentations and manuscript writing.

Sachin Bhoite: Provided research guidance, methodological oversight, and critical review of the manuscript.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Aas, K., Jullum, M., & Loland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. Artificial Intelligence, 298, 103502. https://doi.org/10.1016/j.artint.2021.103502

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6, 52138–52160. https://doi.org/10.1109/access.2018.2870052

Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the Robustness of Interpretability Methods. ArXiv. https://doi.org/10.48550/arXiv.1806.08049

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Beam, A. L., & Kohane, I. S. (2018). Big Data and Machine Learning in Health Care. Journal of the American Medical Association, 319(13), 1317–1318. https://doi.org/10.1001/jama.2017.18391

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible Models for HealthCare. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1721–1730. https://doi.org/10.1145/2783258.2788613

Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., & Tsaneva-Atanasova, K. (2019). Artificial intelligence, bias and clinical safety. BMJ Quality & Safety, 28(3), 231–237. https://doi.org/10.1136/bmjqs-2018-008370

Powers, D. M. W. (2011). From Precision, Recall and Fmeasure to ROC, Informedness, Markedness and Correlation". Journal of Machine Learning Technologies, 2(1), 37–63.

Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. ArXiv. https://doi.org/10.48550/arXiv.1702.08608

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M. A., Chou, K., Cui, C., Corrado, G. S., Thrun, S., & Dean, J. (2019). A Guide to Deep Learning in Healthcare. Nature Medicine, 25(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z

Fedesoriano. (2019). Stroke Prediction Dataset. Kaggle Datasets. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Gosiewska, A., & Biecek, P. (2021). Do Not Trust Additive Explanations. https://doi.org/10.48550/arXiv.1903.11420

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys, 51(5), 1–42. https://doi.org/10.1145/3236009

Guo, C., Pleiss, G., Sun, Y., & Weinberger, Q. (2017). On Calibration of Modern Neural Networks. Proceedings of the 34th International Conference on Machine Learning (ICML 2017), 1321–1330. https://doi.org/10.48550/arXiv.1706.04599

He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/tkde.2008.239

Hosmer, D. W. J., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression. https://doi.org/10.1002/9781118548387

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17(1), 195. https://doi.org/10.1186/s12916-019-1426-2

Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., & Friedler, S. (2020). Problems with Shapley-value-based explanations as feature importance measures. Proceedings of the 37th International Conference on Machine Learning, 5491–5500. https://doi.org/10.48550/arXiv.2002.11097

Lipton, Z. C. (2018). The Mythos of Model Interpretability. Queue, 16(3), 31–57. https://doi.org/10.1145/3236386.3241340

Liu, X., Rivera, S. C., Moher, D., Calvert, M. J., & Denniston, A. K. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. Nature Medicine, 26(9), m3164. https://doi.org/10.1136/bmj.m3164

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions". Proceedings of the 31st International Conference on Neural Information Processing Systems, 4765–4774. https://doi.org/10.48550/arXiv.1705.07874

Molnar, C. (2022). Interpretable Machine Learning". https://christophm.github.io/interpretable-ml-book/

Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. Digital Signal Processing, 73, 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

O'brien, R. M. (2007). A Caution Regarding Rules of Thumb for Variance Inflation Factors. Quality & Quantity, 41(5), 673–690. https://doi.org/10.1007/s11135-006-9018-6

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453.https://doi.org/10.1126/science.aax2342

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., & Snoek, J. (2019). Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. Advances in Neural Information Processing Systems, 13991–14002. https://doi.org/10.48550/arXiv.1906.02530

Pal, M., Saha, H. N., & Chakrabarti, A. (2026). The Trust-Aware XAI (TAXAI) framework: a quantitative model for interpretable and reliable clinical AI systems. Scientific Reports, 16(1), 15455. https://doi.org/10.1038/s41598-026-44167-3

Pearl, J. (2009). Causality". https://doi.org/10.1017/CBO9780511803161

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. (2008). Dataset Shift in Machine Learning. https://doi.org/10.7551/mitpress/9780262170055.001.0001

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine Learning in Medicine. New England Journal of Medicine, 380(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models". ArXiv. https://doi.org/10.48550/arXiv.1708.08296

Sendak, M. P., D'Arcy, J., Kashyap, S., Gao, M., Nichols, M., Corey, K., & Ratliff, W. (2020). A Path for Translation of Machine Learning Products into Healthcare Delivery. EMJ Innovations, 4(1), 24–30. https://doi.org/10.33590/emjinnov/19-00172

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 180–186. https://doi.org/10.1145/3375627.3375830

Tjoa, E., & Guan, C. (2021). Explainable Artificial Intelligence in Healthcare: A Methodological Review". Nature Machine Intelligence, 3(6), 474–483. https://doi.org/10.1109/TNNLS.2020.3027314

Topol, E. J. (2019). High-performance Medicine: The Convergence of Human and Artificial Intelligence". Nature Medicine, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Van Dort, B. A., Engelsma, T., Sneekes, P., Peute, L., & Medlock, S. (2026). Explainable AI in hospital clinical decision support systems: A scoping review of healthcare professionals 'perspectives. PLOS Digital Health, 5(5), e0001417. https://doi.org/10.1371/journal.pdig.0001417

Vasey, B., Nagendran, M., Campbell, B. and, Clifton, D. A., Collins, G. S., Denaxas, S., Denniston, A. K., Faes, L., Geerts, B., Ibrahim, M., Liu, X., Mateen, B. A., Mathur, P., McCradden, M. D., Morgan, L., Ordish, J., Rogers, C., Saria, S., Ting, D. S. W., … McCulloch, P. (2022). Reporting Guideline for the Early-Stage Clinical Evaluation of Decision Support Systems Driven by AI: DECIDE-AI". Nature Medicine, 28(5), 2218. https://doi.org/10.1038/s41591-022-01772-9