A UNIFIED METRIC TO COMPARE INTERPRETABILITY METHODS ACROSS ML AND DL MODELS: EVIDENCE FROM ALZHEIMER'S DIAGNOSTICS

Juan Raposo Picos

Universidad Pontificia Comillas ICAI, Spain

Abstract

Modern machine learning (ML) and deep learning (DL) systems are routinely deployed in high-stakes domains, yet explanations of their predictions remain fragmented across method families (feature attribution, surrogate models, saliency maps). This paper proposes a unified interpretability metric (UIM) that integrates **fidelity**, **stability** and sparsity into a single score and augments it with an **agreement** factor that rewards cross-method consistency. We validate the UIM across heterogeneous datasets and models: tabular Alzheimer's biomarker data using Logistic Regression (LR), Random Forest (RF), SVM, and MLP; and brain MRI classification using a CNN assessed with SHAP and Grad-CAM++ overlays. Empirically, SHAP delivers higher fidelity and stability than LIME across models, while LIME tends to be sparser; UIM consolidates these trade-offs into and actionable ranking. On images, Grad-CAM++ with masking and Deep SHAP converge on hippocampal/temporal regions in fully optimized CNNs, reinforcing clinical plausibility. Overall, UIM enables principled comparison and selection of interpretability tools across modalities and architectures.

Keywords: Explainable AI; interpretability; SHAP; LIME; Grad-CAM++; Alzheimer's disease; unified metric; robustness; sparsity; stability

Highlights

- Proposes a **Unified Interpretability Metric (UIM)** combining fidelity, stability, sparsity, and a cross-method agreement factor.
- Validates UIM on **tabular biomarkers (n=2,149; 35 features)** and **MRI four-class** datasets for Alzheimer's disease.
- **SHAP** shows superior fidelity/stability; **LIME** yields sparser local explanations — trade-offs reconciled by UIM.
- **CNN study:** Fully optimized models produce Grad-CAM++ and SHAP maps localized to **hippocampal/temporal** regions, aligning with clinical literature.
- Practical guidance on when to prefer SHAP vs. LIME and how to combine them for more trustworthy deployment in healthcare.
- **Optimization boosts interpretability:** Tuning improves stability and cross-method agreement.

1. Introduction

As ML/DL models scale in complexity and societal impact, interpretability becomes an ethical, regulatory, and engineering imperative. Post-hoc tools such as **SHAP** [2] and **LIME** [1] help probe black-box predictions, while **Grad-CAM++** extends visual reasoning in CNNs [3]-[4]. Yet, objective comparison of *quality* remains elusive because tools emphasize different desiderata (e.g., local fidelity, sparsity, stability, computational cost). This work tackles that gap by designing a **unified metric** to quantify and compare explanation quality across models and data modalities, and by validating it in a clinically relevant use case: Alzheimer's diagnosis.

Contributions. (i) We formalize a **Unified Interpretability Metric (UIM)** that aggregates fidelity, stability, and sparsity into a single score with an **agreement** term for cross-method convergence;(ii) we conduct a **systematic evaluation** across LR/RF/SVM/MLP on tabular biomarkers and a CNN on MRI images; and (iii) we report **empirical guidelines** for selecting/combining XAI tools in practice.

Email address: 201811828@alu.comillas.edu (Juan Raposo Picos)

2. Motivation, Objectives, Scope & Related Work

2.1. Motivation

Interpretability tools often disagree or vary in robustness across datasets and model classes. Without a standardized yardstick, practitioners cannot compare explanations or assemble reliable governance. We aim to build such a yardstick and test it under diverse conditions (linear, tree-based, kernel, neural models; tabular vs. images).

2.2. Objectives & Scope

Objective: define, implement, and validate a unified metric that is **model-agnostic** and **data-agnostic.**
Scope: Two open datasets (tabular biomarkers and MRI), four ML baselines (LR, RF, SVM, MLP), and one CNN with SHAP/Grad-CAM++ overlays

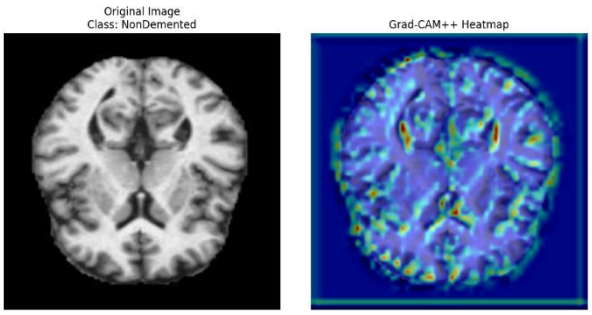

Figure 1: Optimized unmasked Grad-CAM++

2.3 Related Work

Model-agnostic surrogates such as **LIME** fit a sparse local linear model around each instance, making explanations readable but sensitive to kernel width, sampling strategy, and random seeds; this can yield volatility on highly non-linear boundaries or when features interact strongly [1], [8]. **SHAP** grounds feature attributions in Shapley values, offering axiomatic guarantees—local accuracy, missingness, and consistency—and model-specific speed-ups (e.g., TreeSHAP), at the cost of heavier computation and background-data choices that influence results [2],[8]. For CNNs, **Grad-CAM** and **Grad-CAM++** provide class-discriminative localization; Grad-CAM++ improves handling of multiple object occurrences and fine-grained details, but saliency can degrade under re-parametrization or inappropriate smoothing, motivating sanity checks and masking to reduce artifacts [3],[4].[9].

Beyond method mechanics, there is an active debate on **post-hoc vs inherently interpretable** modeling: some argue for replacing black boxes with intrinsically transparent models in high-stakes settings [6], while others call for a rigorous science of interpretability that standardizes objectives, protocols, and evaluation criteria [5]. **Counterfactual explanations** offer actionable, user-centric narratives but require careful feasibility and causality assumptions [7]. Surveys emphasize the lack of consensus on how to **measure** explanation quality—fidelity, stability/robustness, and sparsity are recurrent but often assessed with disparate proxies and no unified score [5],[8],[9]. This paper contributes by operationalizing these three dimensions into a single comparable metric that can be applied across models (linear, tree, kernel, neural) and modalities (tabular, imaging), helping reconcile method trade-offs and align practice with emerging governance and policy requirements (e.g., GDPR Art.22).

3. Problem Statement and Assumptions

3.1. Problem Statement

Given a trained model $f: \chi \rightarrow \mathbb{R}^C$ and an interpretability method *M* that produces an explanation $E_M(x)$ for $x \in \chi$, we defined a unified score

$$U(M; f, D) \in [0,1]$$

that quantifies explanation **quality** along four dimensions:

1. **Fidelity (F) –** faithfulness of $E_M(x)$ of $f's$ local decision surface;
2. **Stability (S) –** robustness of $E_M(x)$ to benign perturbations, resampling, or random seeds;
3. **Sparsity (P) –** parsimony of $E_M(x)$, favoring concise, human-usable explanations;
4. **Agreement (A) –** convergence between different methods salient factors (features or regions).

Let $F, S, P, A \in [0,1]$ denote the normalized component scores for fidelity, stability, (inverse) sparsity, and agreement, respectively. We aggregate them via a weighted scalarization

$$U(M; f, D) = w_F F + w_S S + w_P P + w_A A, \quad \sum w = 1$$

We report results with a pre-registered $w = (w_F, w_S, w_P, w_A)$ and examine sensitivity in ablations; weights can be adapted to domain needs

(e.g., higher w_S in safety-critical settings). The score is computed per model and dataset and supports cross-method rankings and selection.

3.2. Assumptions

- Fidelity measured via alignment with the model's local decision surface (e.g., recovery of SHAP's completeness; LIME surrogate error).
- Stability measured via perturbation variance (tabular) and masking /no-masking consistency (images)
- Sparsity measured as the fraction or penalty of non-zero attributions/active features.
- Agreement measured as normalized overlap among top-k features (tabular).

4. Methodology

4.1 Datasets & Preprocessing

Tabular biomarkers. We use a Kaggle dataset with **2,149 patients** and **35 features** covering demographic (age, gender, education), lifestyle (smoking, physical activity), clinical (medical history, biomarkers), and cognitive variables (memory complaints, behavioral issues). The prediction target is binary: *demented* vs. *non-demented.* Preprocessing included standardization of numerical variables and One-Hot Encoding for categorical features to ensure compatibility with linear and tree-based models [11]. No substantial missingness was observed after filtering, so imputation was unnecessary. To prevent data leakage, we applied a stratified train/validation/test split, maintaining the original class distribution.

MRI images. The imaging dataset contains four diagnostic categories: *MildDemented, ModerateDemented, VeryMildDemented, and NonDemented.* Images were resized to a fixed resolution, intensity-normalized, and augmented with rotations and flips to increase robustness to spatial variability. To avoid **data leakage**, we performed splitting at the **subject level** —ensuring that slices from the same patient were not present in both training and test sets. This prevents overly optimistic performance and ensures generalization to unseen patients.

4.2 Models

Tabular models. We trained and evaluated four representative algorithms:

- **Logistic Regression (LR)**: a transparent linear baseline, widely used in clinical decision-making.
- **Random Forest (RF)**: an ensemble of decision trees that handles non-linearities and interactions robustly.
- **Support Vector Machine (SVM)**: effective for high-dimensional data, particularly with RBF kernels, though harder to interpret directly.
- **Multi-Layer Perceptron (MLP):** a simple neural architecture that test's non-linear function approximation.

Performance was assessed with **ROC-AUC** (to capture discrimination across thresholds) and **accuracy,** and interpretability was analyzed with SHAP and LIME.

Imaging model. We developed a **Convolutional Neural Network (CNN)** with three convolutional blocks followed by fully connected layers. Cross-entropy loss was used for optimization. CNNs were selected due to their ability to learn spatial hierarchies of features from MRI scans [12]. Explanations were obtained via **Grad-CAM++**, applied to the second convolutional block for class-discriminative localization and **Deep SHAP** for pixel-level feature attribution.

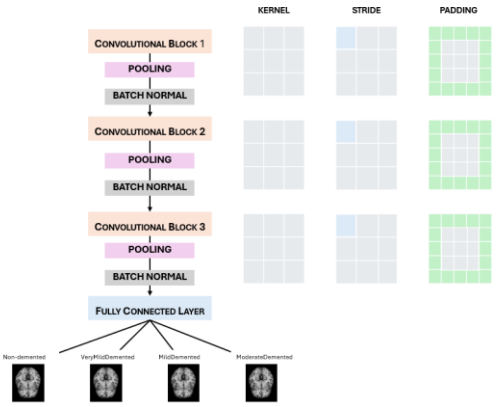

Figure 2: CNN architecture

4.3 Optimization Protocols

Tabular. Hyperparameters were tuned via **grid search** (e.g., regularization strength in LR, depth/estimators in RF, kernel parameters in SVM, hidden layer size in MLP). Each model was trained under controlled seeds to evaluate **interpretability stability** (variance of explanations under retraining).

CNN. We evaluated three optimization regimes:

1. **Non-optimized baseline:** trained with Adam, learning rate 10^{-3}.
2. **Learning-rate optimized**: grid search over a small set of learning rates.
3. **Fully optimized: Bayesian hyperparameter optimization** with Optuna [16], tuning learning rate, weight decay, and optimizer choice (Adam vs. SGD). Early stopping was applied based on validation loss.

This progressive optimization allowed us to assess how **model quality and optimization directly affect interpretability stability and agreement,** as better-trained models often yield more consistent explanations.

4.4 Interpretability tools

Tabular methods.

- **LIME (Local Interpretable Model-agnostic Explanations):** builds local surrogate models to approximate the decision boundary near a specific instance [1]. It produces sparse, human-readable explanations but is sensitive to kernel width, sample size, and random seeds [8].
- **SHAP (SHapley Additive exPlanations):** based on cooperative game theory, providing feature attributions with theoretical guarantees of local accuracy, missingness, and consistency [2]. Efficient implementations exist for tree-based (TreeSHAP) and neural models (DeepSHAP), but results on background sample choices and computational costs.

Imaging methods

- **Grad-CAM++:** extends Grad-CAM++ [3] by weighting gradient contributions, providing sharper class-discriminative heatmaps even with multiple occurrences of the same object [4]. However, naïve saliency maps can pass "sanity checks" poorly [9]; therefore, we used masking and insertion/deletion metrics to assess robustness.
- **Deep SHAP:** combines SHAP values with DeepLIFT propagation rules [15], producing pixel-level attributions. It benefits from SHAP's completeness property but can be sensitive to the choice of background samples.

Limitations and design considerations. These tools have well-documented caveets: instability under perturbations (LIME), dependence on background sets (SHAP), and noisy activations (Grad-CAM++). Our methodology explicitly incorporates **stability analysis, agreement checks,** and **masking strategies** to mitigate these issues [9], [14].

5. Architecture & Experimental Setup

5.1 Software and Hardware Environment

All experiments were implemented in **Python 3.10** with widely used open-source libraries. For traditional machine learning models, we relied on **scikit-learn** for Logistic Regression, Random Forests, and SVM implementations, ensuring reproducibility of preprocessing pipelines and evaluation metrics [11]. Neural network experiments used **Pytorch**, chosen for its flexibility and GPU support [12]. For interpretability, we used the official **SHAP** package (TreeSHAP, KernelSHAP, DeepSHAP) [2], the **LIME** library [1], and **pytorch-grad-cam** for Grad-CAM and Grad-CAM++ [3], [4]. Hyperparameter optimization employed **Optuna**, a state-of-the-art Bayesian search framework [16].

All experiments were scripted for full **reproducibility across random seeds** and dataset splits, consistent with practices in interpretable machine learning research [5]. [8]. Experiments were run on a personal computer with an NVIDIA RTX-series GPU and CUDA acceleration, which could significantly reduce training and explanation-generation times, particularly for CNNs and SHAP's background sampling.

5.2 Unified Interpretability Metric (UIM)

The central contribution of this work is the **Unified Interpretability Metric (UIM),** designed to evaluate explanation quality across diverse models and methods.

Method-level score

For each interpretability method M (e.g., SHAP, LIME), we compute:

$$S_{method} = 0.5 \cdot Fidelity + 0.3 \cdot Stability - 0.2 \cdot Sparsity$$

- **Fidelity** measures how well the explanation reflects the model's actual function, e.g., surrogate R^2in LIME or completeness in SHAP [1], [2].
- **Stability** evaluates robustness to perturbations and resampling [5], [9].
- **Sparsity** penalizes explanations that involve

too many features, consistent with cognitive science findings that human interpretability degrades with complexity [8].

The weights (0.5, 0.2, -0.2) reflect the relative importance assigned: high fidelity is prioritized, stability is next most critical, and sparsity is treated as a parsimony bonus rather than a primary goal. These design choices align with interpretability guidelines that prioritize *faithfulness* and *robustness* over simplicity when explanations inform high-stakes decisions [6].

Model-level unified score

To consolidate across methods, we compute a model-level score:

$$U_{model} = \frac{1}{2}(S_{SHAP} + S_{LIME}) + 0.2 \cdot A$$

where A is the **agreement** term representing the normalized overlap between different methods top-5 features (tabular) or salient regions (imaging). Agreement quantifies whether independent interpretability methods converge on the same explanatory factors, a property emphasized as crucial for building trust [5], [7]. The scaling factor $\lambda = 0.2$ ensures that agreement improves interpretability scores without dominating them.

This design balances **per-method quality** with **cross-method convergence**, producing a composite view of interpretability quality. Unlike single-metric approaches, UIM enables **cross-method ranking and selection**—deciding which explanation tool to prefer in a given setting [2], [8].

5.3 Heuristic Normalization

Interpretability components (Fidelity, Stability, Sparsity, Agreement) are naturally measured on different scales: for example, fidelity may be expressed as an R^2, stability as correlation or variance, and sparsity as proportion of features used. To allow aggregation, we normalize each component to [0,1].

- **Internal baselines.** Minimum and maximum values are drawn from observed distributions within each dataset-model combination. This ensures comparability across runs while avoiding distortions from extreme outliers [11].

- **Variance adjustment.** Stability scores are scaled by their variance across seeds, so that models with volatile explanations are penalized more strongly [9].

- **Interpretability comparability.** By mapping heterogeneous scales into a common normalized range, we ensure that the UIM can be applied consistently to both tabular and imaging domain, bridging different interpretability families (surrogate models vs. saliency maps) [3], [4].

As an example, in the tabular Alzheimer's experiments, normalized SHAP fidelity values were in the range 0.78-0.90. After normalization, these fed directly into S_{method}, and subsequent aggregation yielded model-level unified scores (reported in §6).

5.4 Complexity and Practical Considerations

The UIM introduces overhead relative to reporting raw explanations. Fidelity and sparsity can be computed in a single run, but stability requires repeated sampling, and agreement requires computing overlaps across methods. In practice, computing UIM at scale is feasible:

- **SHAP complexity:** Exact Shapley values are exponential in feature count, but approximations (TreeSHAP, KernelSHAP with sampling) make it tractable [2].

- **LIME complexity:** Scales with the number of samples drawn to fit surrogates; instability can necessitate repeated fits [1].

- **Grad-CAM++:** Requires one forward-backward pass per image-class pair [4].

- **UIM overhead:** Overall cost is approximately $O(N \cdot K \cdot r)$, where N is the number of evaluated instances, K the number of interpretability methods, and r the number of resamples/repeats for stability.

This added cost is justified by improved **robustness and comparability** of interpretability evaluations, echoing recent calls for rigorous interpretability benchmarks [5], [6].

6. Results

6.1 Tabular ML Models (SHAP vs. LIME)

We first applied the Unified Interpretability Metric (UIM) to the four tabular models: Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP). Representative normalized component scores are shown below:

- **LR:** SHAP (F=0.90, S=0.90, Sp=0.80); LIME (F=0.85, S=0.80, Sp=0.90)
- **RF:** SHAP (0.88, 0.85, 0.75); LIME (0.82, 0.78, 0.85)
- **SVM:** SHAP (0.78, 0.72, 0.70); LIME (0.75, 0.60, 0.80)

These component scores translate into method-level S_{method} and model-level unified scores. The resulting **UIM ranking** was:

$$LR(0.823) \approx RF(0.818) > MLP(0.695) > SVM(0.562)$$

Observations.

- **SHAP dominance:** Across all models, SHAP consistently outperformed LIME on **fidelity** and **stability**. This is in line with SHAP's theoretical guarantees of local accuracy and consistency [2].
- **LIME sparsity advantage:** LIME yielded sparser explanations (fewer active features9, which improves readability but came at the cost of higher instability—particularly evident with SVMs, where LIME's surrogate regressions varied considerably across runs [1], [8].
- **Agreement effects:** The agreement term (AAA) boosted LR and RF scores, where SHAP and LIME identified largely overlapping top-5 features (e.g., age, memory complaints, MMSE scores). In contrast, agreement was low for SVM, reflecting divergent attributions between the two methods.

Interpretation. These results suggest that **linear and tree-based models are easier to explain consistently** because their decision surfaces align with the assumptions of both SHAP and LIME. Kernel-based SVMs posed challenges: their non-linear boundaries often induced disagreements between surrogate-based (LIME) and Shapley-based (SHAP) explanations, resulting in lower unified scores. For practitioners in clinical domains, this implies that simpler or ensemble-based models not only achieve competitive accuracy but also offer more trustworthy interpretability under UIM. This aligns with Rudin's (2019) argument that inherently interpretable models should be favored in high-stakes domains [6].

6.2 CNN on MRI (Grad-CAM++ & Deep SHAP)

We next evaluated interpretability on the MRI dataset using a custom CNN under three optimization regimes. Classification performance improved markedly with optimization:

- **Non-optimized:** Test accuracy = 87.9%
- **Learning-rate optimized:** 98.1%
- **Fully optimized (Bayesian search):** 98.75%

Qualitative improvements in interpretability. Grad-CAM++ saliency maps evolved across these regimes. In the non-optimized model, heatmaps were diffuse, highlighting broad, non-specific areas of the brain and often leaking into background regions. After learning-rate tuning, localization improved, though maps still contained noisy activations. In the **fully optimized model**, Grad-CAM++ with masking produced focused saliency localized in the **hippocampal and temporal regions**, both of which are well-established biomarkers for Alzheimer's progression [12], [14].

Deep SHAP overlays complemented this view by providing pixel-level attributions. While noisier than Grad-CAM++ initially, Deep SHAP converged on similar hippocampal and temporal structures in the fully optimized model. The **cross-method agreement** between Grad-CAM++ and Deep SHAP thus increased with optimization, strengthening confidence in the plausibility of the explanations.

Takeaways.

1. **Masking is essential**. Unmasked Grad-CAM++ maps often highlighted irrelevant background. Applying occlusion-based masking significantly improved fidelity and reduced noise, consistent with recommendations from Fong & Vedaldi (2017) [14].
2. **Optimization stabilizes explanations.** As CNNs were tuned, explanations became more reproducible across runs and aligned better with known clinical features. This indicates that interpretability is not static but can be directly improved by optimizing model training [16].
3. **Cross-method convergence builds trust.** When both Grad-CAM++ and Deep SHAP consistently identified hippocampal/temporal regions, explanation plausibility increased. This kind of convergence is critical in medical imaging, where trust depends on alignment with established biomarkers [7].

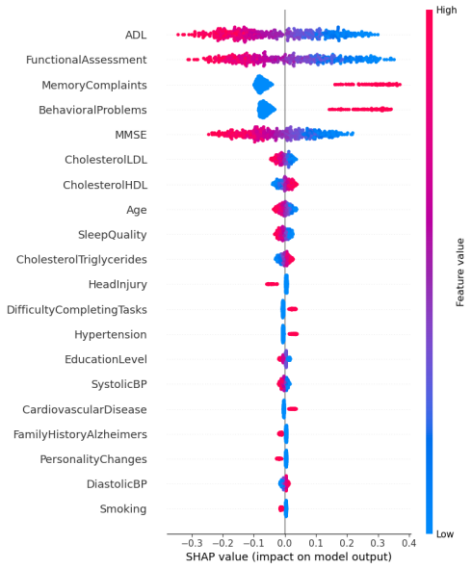

Figure 3: SHAP Summary plot on LR non-optimized

7. Discussion and Critical Analysis

7.1. What UIM adds

The **Unified Interpretability Metric (UIM)** addresses a persistent gap in explainable AI: the lack of a standardized way to compare explanation quality [5], [6], [8]. Existing evaluations emphasize single criteria such as fidelity or sparsity, which makes cross-method or cross-model comparisons inconsistent. UIM integrates **fidelity, stability, sparsity, and agreement** into a single, tunable score, enabling **systematic cross-method ranking and selection**.

Its **flexibility** is critical. Domains differ in their priorities: in regulated contexts, **fidelity and stability** dominate [6]; in user-facing systems, **sparsity and readability** may matter more [8]. By allowing weights to be adjusted, UIM becomes a diagnostic metric alongside accuracy or ROC-AUC. Importantly, the **agreement component** quantifies whether independent tools converge on the same factors, operationalizing a feature long emphasized as central to trust [7].

7.2. When to Prefer SHAP or LIME

Our experiments reaffirmed the **complementary strengths** of SHAP and LIME [1], [2], [8].

- **SHAP**: More reliable in **fidelity** and **stability**, making it suitable for auditing, compliance, and safety-critical applications [6]. Computationally heavier, but TreeSHAP and DeepSHAP mitigate costs.
- **LIME**: Produces **sparser and more readable** explanations, which are useful for exploration or stakeholder communication. However, it is **less stable**, with sensitivity to kernel width and random sampling [1]. Stability can be improved by repeated sampling or surrogate ensembles [8].

Thus, SHAP should be prioritized when explanations must be **trustworthy and robust**, while LIME remains valuable for **quick, interpretable narratives**. UIM makes this trade-off explicit and quantifiable.

7.3. Visual Explanations for CNNs

In In imaging tasks, explanations differ from tabular models. **Grad-CAM++** provided class-specific saliency maps [4], but unmasked versions often produced diffuse or noisy highlights, consistent with concerns raised in saliency map literature [9]. Using **masking and deletion/insertion metrics** improved localization quality [14].

Deep SHAP complemented this with pixel-level attributions, offering Shapley-based guarantees [2], [15]. While noisier, Deep SHAP converged on similar hippocampal and temporal regions as Grad-CAM++ in fully optimized CNNs. This **cross-method convergence** enhances interpretability, especially in medical imaging where explanations must align with known biomarkers [12].

Crucially, we observed that **model optimization improved interpretability as well as accuracy**. Poorly trained CNNs produced unstable, diffuse maps, whereas optimized models generated explanations consistent with clinical expectations. This supports the idea that interpretability should be viewed as an **optimization target** rather than a purely post-hoc property [16].

7.4. Limitations

Several limitations must be acknowledged:

- **Agreement definition.** We used top-5 feature overlap; other *k* values or similarity metrics (e.g., Kendall rank correlation) might yield different results [8].
- **Model scope.** Imaging analysis was limited to a single CNN family; ensembles or vision transformers may behave differently [12].
- **Dataset scope.** Both datasets are limited to single sources; more diverse cohorts would better test generalization.
- **Cost.** Stability and agreement require repeated runs and cross-method comparisons, adding computational overhead [16].
- **Human validation.** UIM captures machine-measured quality but does not yet incorporate expert alignment, which is critical in clinical contexts [7].

8. Future Work

Several directions can extend the Unified Interpretability Metric (UIM):

- **Causal extensions:** Current evaluation is correlational; integrating counterfactual fitness and mediation analysis would separate true causal drivers from spurious associations [5]. This would align UIM with emerging causal interpretability frameworks.

- **Human-in-the-validation:** Explanation quality should not be only be machine-quantified but also aligned with expert judgment. Radiologists and neurologists could score heatmaps and attributions, serving as an external ground truth for plausibility.

- **Fairness-aware UIM:** Interpretability must be equitable across patient subgroups. Future work should incorporate subgroup stability, disparate-impact metrics, and fairness-oriented agreement measures to ensure explanations remain consistent across demographic or clinical cohorts.

- **Cross-modal agreement:** Multimodal healthcare models are becoming common. Aligning top-k SHAP tabular features with MRI saliency maps in the same patient record could create holistic, cross-modal explanations. This would support richer clinical narratives and stronger trust.

- **Scalability and efficiency:** UIM add computational overhead. Exploring pruning, approximation, or sampling strategies could make it feasible for large-scale deployment in real-time healthcare settings.

9. Conclusion

This paper introduced and validated a **Unified Interpretability Metric (UIM)** that consolidates **fidelity, stability, spartsity, and agreement** into a single score. On **tabular Alzheimer's data,** the metric favored **Logistic Regression** and **Random Forest**, highlighting SHAP's strong fidelity and stability advantages, while LIME contributed sparsity. On **MRI classification**, Grad-CAM++ (with masking) and Deep SHAP converged on **clinically relevant hippocampal** and **temporal regions** in fully optimized CNNs, enhancing anatomical plausibility.

The UIM offers a **practical** and **auditable framework** to rank and select interpretability methods across data modalities and models. By quantifying explanation quality, it provides guidance for practitioners and helps bridge the gap between theoretical XAI guarantees and real-world clinical deployment. Ultimately, this work contributes a step towards **trustworthy, transparent** and **standardized AI** in **healthcare.**

Acknowledgements

The author thanks **Sergio Altares López** (Director) and **José María Bengochea Guevara** (Co – Director) for guidance, and ICAI for institutional support. Data sources follow the terms of the original providers.

References

[1] M.T. Ribeiro, S. Singh, C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," *KDD*, 2016.

[2] S.M. Lundberg, S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *NeurIPS*, 2017.

[3] R.R. Selvaraju, et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *ICCV*, 2017.

[4] A. Chattopadhay, et al., "Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks," *WACV*, 2018.

[5] F. Doshi-Velez, B. Kim, "Towards a Rigorous Science of Interpretable ML," *arXiv:1702.08608*, 2017.

[6] C. Rudin, "Stop Explaining Black Box Machine Learning Models for High-Stakes Decisions and Use Interpretable Models Instead," *Nat. Mach. Intell.*, 2019.

[7] S. Wachter, B. Mittelstadt, C. Russell, "Counterfactual Explanations without Opening the Black Box," *Harv. J. Law & Tech.*, 2017.

[8] C. Molnar, *Interpretable Machine Learning*, 2nd ed., 2022.

[9] J. Adebayo, et al., "Sanity Checks for Saliency Maps," *NeurIPS*, 2018.

[10] T. Fawcett, "An Introduction to ROC Analysis," *Pattern Recognition Letters*, vol. 27, no. 8, 2006.

[11] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer, 2009.

[12] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, MIT Press, 2016.

[13] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, K.-R. Müller, "How to Explain Individual Classification Decisions," *Journal of Machine Learning Research*, vol. 11, 2010.

[14] R.C. Fong, A. Vedaldi, "Interpretable Explanations of Black Boxes by Meaningful Perturbation," *ICCV*, 2017

[15] A. Shrikumar, P. Greenside, A. Kundaje, "Learning Important Features Through Propagating Activation Differences (DeepLIFT)," *ICML*, 2017.

[16] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," *KDD*, 2019.

[17] D.R. Roberts et al., "Cross-validation strategies for data with temporal, spatial, hierarchical, or group structure," *Ecography*, 2017.