Explainable AI: GradCAM, LIME, and SHAP Applied to Image and Tabular Data

Alexandre Cotorobai
107849
Complementos de Aprendizagem Automática 24/25
MEI, DETI
University of Aveiro
Aveiro, Portugal
alexandrecotorobai@ua.pt

André Oliveira
107637
Complementos de Aprendizagem Automática 24/25
MEI, DETI
University of Aveiro
Aveiro, Portugal
andreaoliveira@ua.pt

***Abstract*—This study presents a comprehensive comparative analysis of three prominent Explainable Artificial Intelligence (XAI) methods: Gradient-weighted Class Activation Mapping (Grad-CAM), Local Interpretable Model-agnostic Explanations (LIME), and SHapley Additive exPlanations (SHAP). We evaluate these methods across two distinct machine learning paradigms: computer vision tasks using a multi-class waste classification dataset with deep convolutional neural networks (ResNet, VGG, DenseNet), and structured data analysis using the AIDS Clinical Trials Group Study 175 dataset with Random Forest ensemble methods. Through systematic experimentation, we assess each method's ability to provide meaningful explanations for model predictions, examining their computational efficiency, consistency across different model architectures, and interpretability quality. Our findings reveal distinct strengths and limitations: Grad-CAM excels in providing spatially coherent visual explanations for image classification, LIME offers intuitive local explanations through perturbation-based analysis, while SHAP provides theoretically grounded feature importance with additive properties. The comparative evaluation demonstrates that explanation quality varies significantly across data modalities and model complexities, highlighting the importance of selecting appropriate XAI methods based on specific application requirements and stakeholder needs.**

***Index Terms*—Explainable AI, GradCAM, LIME, SHAP, image classification, tabular data, deep learning, interpretability**

I. Introduction

The increasing complexity of modern machine learning models has created a critical trade-off between performance and interpretability. While deep neural networks achieve state-of-the-art results across numerous domains, their black-box nature poses significant challenges for trust, accountability, and regulatory compliance [1], [2].

Explainable Artificial Intelligence (XAI) has emerged as a crucial field addressing these concerns by developing methods that provide insights into model decision-making processes [3]. Among the most prominent XAI techniques are Gradient-weighted Class Activation Mapping (Grad-CAM) [4], Local Interpretable Model-agnostic Explanations (LIME) [5], and SHapley Additive exPlanations (SHAP) [6].

Each method employs distinct approaches: Grad-CAM leverages gradient information to highlight important image regions, LIME generates local explanations through systematic input perturbation, and SHAP provides theoretically grounded feature attributions based on cooperative game theory. However, systematic comparisons of these methods across different data modalities remain limited in the literature.

This study conducts a comprehensive comparative analysis of these three XAI methods across two domains: image classification using a waste classification dataset with multiple CNN architectures, and tabular data analysis using clinical trial data with ensemble methods. Our objectives are to evaluate explanation quality, computational efficiency, and consistency across different model architectures, ultimately providing evidence-based recommendations for XAI method selection in practical applications.

II. Related Work

The field of Explainable Artificial Intelligence has witnessed significant growth in recent years, driven by the increasing deployment of complex machine learning models in critical applications. This section reviews the foundational work and recent advances in XAI methods, with particular focus on the three techniques analyzed in this study.

A. Foundations of Explainable AI

Early work in model interpretability focused primarily on inherently interpretable models such as linear regression and decision trees [7]. However, the superior performance of complex models like deep neural networks created a fundamental tension between accuracy and interpretability, leading to the development of post-hoc explanation methods [8].

Ribeiro et al. [9] established key principles for explanation methods, emphasizing the importance of fidelity, interpretability, and model-agnosticism. Their framework distinguished between global explanations that describe overall model behavior and local explanations that clarify individual predictions, a distinction that remains central to XAI research.

B. Gradient-Based Explanation Methods

Gradient-based methods exploit the differentiable nature of neural networks to identify important input features. Early

approaches included vanilla gradients [10] and guided backpropagation [11], which highlighted input regions with high gradient magnitudes.

Grad-CAM [4] emerged as a significant advancement by focusing on convolutional layers rather than input pixels, producing more interpretable spatial heatmaps. Subsequent work extended Grad-CAM to various architectures and tasks [12], while Grad-CAM++ improved localization accuracy through weighted gradient computations.

Recent studies have evaluated Grad-CAM's performance across different CNN architectures. Wang et al. [13] demonstrated varying explanation quality depending on network depth and architecture design, while Adebayo et al. [14] raised important questions about the reliability of gradient-based methods through sanity checks.

C. Perturbation-Based Explanation Methods

LIME [5] introduced the influential concept of learning local surrogate models through systematic input perturbation. The method's model-agnostic nature and intuitive approach to generating explanations made it widely adopted across diverse domains [15].

Extensions of LIME have addressed various limitations of the original approach. SHAP [6] provided a unifying theoretical framework based on Shapley values from cooperative game theory, ensuring explanation consistency and additivity. Kernel SHAP and TreeSHAP variants optimized computational efficiency for specific model types [16].

Recent work has focused on improving perturbation strategies and addressing stability issues. Slack et al. [17] demonstrated potential vulnerabilities in LIME explanations, while Garreau and Luxburg [18] provided theoretical analysis of LIME's consistency properties.

D. Comparative Studies and Evaluation Frameworks

Several studies have attempted to compare XAI methods across different criteria. Samek et al. [19] proposed evaluation metrics for explanation methods in computer vision, emphasizing the importance of localization accuracy and sensitivity analysis.

Doshi-Velez and Kim [20] established a framework for evaluating interpretability through application-grounded, human-grounded, and functionally-grounded approaches. This taxonomy has influenced subsequent evaluation methodologies in XAI research.

More recent comparative studies have revealed significant variations in explanation quality across methods and domains. Bhatt et al. [21] evaluated multiple XAI techniques on medical imaging tasks, finding that method effectiveness varies substantially with dataset characteristics and model architectures.

E. Domain-Specific Applications

XAI methods have been extensively applied in computer vision tasks. For image classification, comparative studies have shown that different explanation methods can produce conflicting interpretations of the same prediction [22]. In medical imaging, the reliability and consistency of explanations have become critical concerns [23].

For tabular data analysis, SHAP has gained particular prominence due to its theoretical foundations and practical effectiveness [24]. Studies comparing feature importance methods have shown that SHAP often provides more stable and consistent explanations compared to permutation-based approaches [2].

F. Research Gaps and Motivation

Despite extensive research on individual XAI methods, systematic comparative evaluations across different data modalities and model architectures remain limited. Most existing studies focus on single domains or specific model types, limiting the generalizability of findings.

Furthermore, there is a lack of comprehensive analysis examining how explanation quality varies with model complexity and architecture design. This gap is particularly relevant given the diversity of modern deep learning architectures and their varying internal representations.

This study addresses these limitations by conducting a systematic comparison of three prominent XAI methods across both image and tabular data domains, using multiple model architectures to assess explanation consistency and reliability.

III. Dataset Description

This study employs two distinct datasets to evaluate XAI methods across different data modalities.

A. Waste Classification Dataset

The image classification component utilizes a comprehensive waste classification dataset [25] containing images categorized into six classes: cardboard, glass, metal, paper, plastic, and trash.

Dataset Characteristics: The dataset exhibits class imbalance typical of real-world scenarios, with fewer samples in the trash category. Images vary significantly in background complexity, lighting conditions, and object positioning, creating diverse visual contexts that challenge both model performance and explanation quality.

Practical Relevance: Waste classification represents a socially relevant application where model interpretability is crucial for building trust in automated sorting systems.

B. AIDS Clinical Trials Group Study 175 Dataset

For tabular data analysis, we employ the AIDS Clinical Trials Group Study 175 dataset from the UCI Machine Learning Repository [26], containing 2,139 HIV-infected patients with 23 clinical and demographic features.

Feature Composition: The dataset includes treatment indicators, baseline and follow-up CD4 counts, age, gender, and various clinical measurements. The target variable indicates disease progression, creating a binary classification problem with significant clinical implications.

Clinical Significance: Understanding feature importance in HIV treatment outcomes has direct implications for medical decision-making, making this dataset valuable for assessing XAI explanations in healthcare contexts.

IV. Machine Learning Methodology

This section outlines the comprehensive experimental framework designed to evaluate explainable AI methods across different data modalities. Our methodology employs distinct approaches for image classification and tabular data analysis, each tailored to the specific characteristics and requirements of the respective domains.

A. Image Classification Methodology

We analyze three convolutional neural network (CNN) architectures previously trained for a multi-class waste classification task. To explore how explanation methods behave under different accuracy levels and architectural designs, we selected one low-performing model and two high-performing models from prior studies. Specifically, we include ResNet with early stopping as the least accurate baseline, and VGG16 and DenseNet-121 as the two best-performing alternatives. These choices provide a balanced framework for evaluating explanation consistency and reliability across varying levels of model capacity and robustness.

Our explainability pipeline for image classification consists of three sequential steps. First, we apply Gradient-weighted Class Activation Mapping (GradCAM) to visualize the internal focus of each CNN, yielding heatmaps that reflect gradient-based spatial attention. This offers a model-specific, architecture-dependent baseline for interpretability.

Following GradCAM, we apply Local Interpretable Model-agnostic Explanations (LIME) to produce perturbation-based surrogate models centered on individual predictions. For each image, we evaluate LIME under various configurations to measure how its explanations change with segmentation granularity (using both 20 and 100 superpixel segments), number of top features highlighted (5 or 10), and different numbers of perturbation samples. These controlled variations allow us to analyze the sensitivity of LIME explanations with respect to key parameters.

Lastly, we generate SHapley Additive exPlanations (SHAP) for each prediction using superpixel-based input segmentation. SHAP provides additive feature attributions that are theoretically grounded and enables comparison of consistent contribution values across architectures. Together, this multi-method pipeline allows us to evaluate spatial coherence, attribution consistency, and parameter robustness of the different XAI techniques in a unified framework.

B. Tabular Data Methodology

For the structured data domain, we employ the AIDS Clinical Trials Group Study 175 dataset, which includes clinical and demographic variables from HIV-positive patients. The data undergoes preprocessing through standardization of numerical features and appropriate encoding of categorical attributes. A Random Forest classifier is trained with hyperparameter optimization via grid search and cross-validation to ensure robust baseline performance.

To interpret the Random Forest model, we evaluate both global and local feature importance using SHAP and LIME. We analyze global feature importance by aggregating SHAP values across the dataset, and inspect local explanations for individual instances using LIME. Additionally, we employ SHAP interaction values to assess how pairs of features jointly influence the model's predictions.

A key part of our analysis involves manipulating the dataset to probe the robustness of these explainability methods. We identify the most influential feature—*time*—and remove it to observe how the explanations shift in its absence. This allows us to examine whether LIME and SHAP preserve interpretability and maintain alignment with known clinical patterns when dominant features are suppressed. The goal is to understand how the structure of feature attributions changes under altered input contexts, and whether meaningful explanations can still be extracted when the model's most predictive feature is excluded.

V. Explainable AI Methods

To interpret the predictions made by the image and tabular models described in Section IV, we employ three widely used explainability techniques: Gradient-weighted Class Activation Mapping (Grad-CAM), Local Interpretable Model-agnostic Explanations (LIME), and SHapley Additive exPlanations (SHAP). Each method offers complementary perspectives on model behavior, enabling both global and local interpretability across modalities.

A. Gradient-weighted Class Activation Mapping (Grad-CAM)

Grad-CAM [4] is a saliency-based technique designed for convolutional neural networks. It generates class-discriminative heatmaps by computing the gradient of the output class score with respect to feature maps in the final convolutional layer. These gradients are used to produce a weighted combination of the feature maps, highlighting the most influential regions in the input image.

In our implementation, Grad-CAM is applied as the first step in the image explanation pipeline. For each architecture—ResNet, VGG16, and DenseNet-121—the last convolutional layer is automatically detected to ensure architectural consistency. The resulting activation maps are overlaid on the input images to produce intuitive spatial visualizations of model attention.

Grad-CAM offers fast, visually intuitive explanations aligned with human perception, but is limited to CNNs and may produce diffuse or inconsistent attention maps depending on network depth and design.

B. Local Interpretable Model-agnostic Explanations (LIME)

LIME [5] explains individual predictions by fitting locally faithful surrogate models around the input instance. The approach perturbs the input features and observes the corresponding changes in model output to estimate the influence of each feature.

For image data, LIME uses SLIC superpixel segmentation to define interpretable regions. Explanations are generated by systematically masking these regions and training a linear

model on the resulting perturbed samples. In our analysis, we explore the effects of varying the number of segments, top features, and perturbation samples to evaluate explanation stability and interpretability across configurations.

For tabular data, LIME perturbs numerical and categorical features by sampling from the training distribution, creating a neighborhood around the instance. A local surrogate model then assigns weights to each feature, reflecting their contribution to the prediction.

LIME is highly versatile and model-agnostic, but its explanations can be unstable and sensitive to the choice of perturbation parameters and sampling strategy.

C. SHapley Additive exPlanations (SHAP)

SHAP [6] is a model-agnostic framework grounded in cooperative game theory that attributes a model's output to individual features in a way that satisfies formal properties such as local accuracy, consistency, and additivity.

We apply different SHAP explainers depending on model type. For image-based CNNs, SHAP explanations are computed on superpixel-segmented inputs using a KernelExplainer, allowing us to visualize the additive contribution of each region. For tabular data, TreeExplainer is used with the Random Forest model to compute both individual feature attributions and pairwise interaction values.

SHAP explanations are used not only to identify important features and feature interactions but also to assess robustness under feature perturbation. In particular, we investigate the impact of removing the most influential feature (time) and observe how the model explanations shift, revealing the sensitivity and adaptability of SHAP to changing input contexts.

SHAP provides theoretically sound and consistent explanations across model types, but can be computationally expensive and less practical for high-dimensional or complex models.

VI. Results and Discussion

A. Image Data Classification

1) ***Gradient-weighted Class Activation Mapping (GradCAM) Analysis****:* GradCAM provides spatially coherent visual explanations by highlighting the regions that most strongly influence CNN predictions. Our analysis across three different architectures—ResNet, VGG, and DenseNet-121—reveals distinct patterns in both explanation quality and localization accuracy, offering valuable understanding on the relationship between model complexity and interpretability.

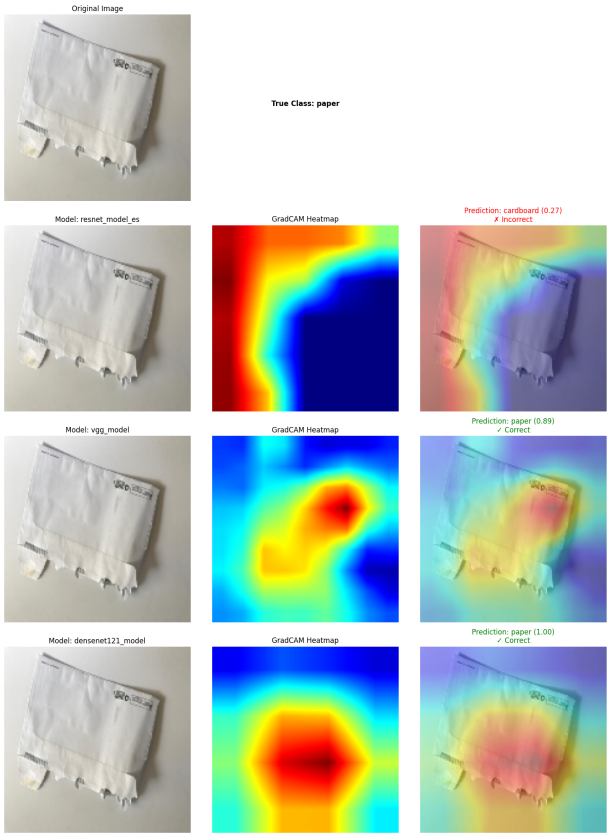

Fig. 1: GradCAM explanations across different CNN architectures for paper bag classification, showing varying prediction confidence and spatial attention patterns

a) Model Performance and Spatial Localization: The evaluation demonstrates significant variation in prediction accuracy and spatial attention patterns across different model architectures. For the paper bag sample (Figure 1), ResNet achieved the lowest confidence with an incorrect prediction of cardboard (0.27), while VGG and DenseNet-121 both correctly identified the material as paper with high confidence scores of 0.89 and 1.00, respectively. This performance difference suggests that deeper, more complex architectures like DenseNet-121, with their feature reuse through skip connections, provide more robust classification performance.

The spatial localization quality varies dramatically across architectures. ResNet shows broad, diffuse activation patterns with relatively uniform distribution across images, where the gradient-based attention appears less focused and potentially contributes to misclassification in challenging cases. In contrast, VGG demonstrates more concentrated activation regions, particularly focusing on object boundaries and texture-rich areas with clearer spatial coherence and stronger activations on material-specific features. DenseNet-121 exhibits the most precise spatial localization, with highly concentrated activation regions that closely align with object boundaries, suggesting that dense connectivity enables more discriminative feature representations.

b) Material-Specific Explanation Patterns: Different material types elicit distinct explanation patterns that provide insights into how models make classification decisions. For paper-based items (Figures 1 and 2), successful predictions consistently show concentrated activations on text regions and surface textures. The magazine cover examples demonstrate how models often rely on printed graphics and text as discriminative features for paper classification. Notably, while both VGG and DenseNet-121 correctly identified the material, VGG exhibited more precise spatial focus on the magazine's title and textual elements. In contrast, DenseNet-121's activation map appeared more diffuse and less aligned with the most informative content. This behavior may partially stem from the architectural design of DenseNet, which emphasizes feature reuse and global integration, potentially at the cost of fine-grained spatial focus. Additionally, the model was reused from previous work without dedicated hyperparameter tuning for this specific interpretability task. As such, the lack of targeted optimization may also contribute to the less localized attention observed in its GradCAM outputs.

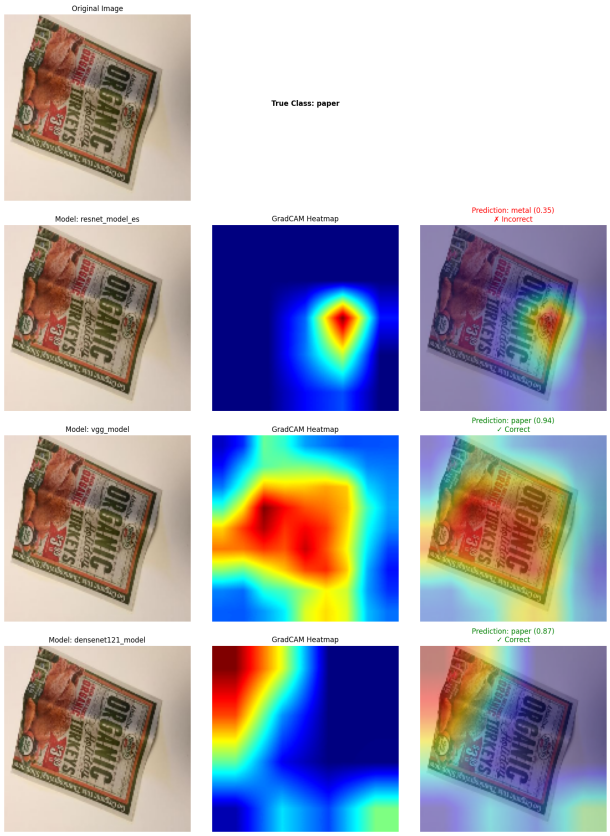

Fig. 2: GradCAM analysis of magazine cover classification showing consistent focus on textual and graphical elements across architectures

Cardboard materials (Figure 3) show consistent performance across all three models, with correct predictions and confidence scores ranging from 0.40 to 1.00. The heatmaps effectively highlight structural edges and surface textures characteristic of corrugated cardboard, demonstrating how different material properties are captured by the various architectures.

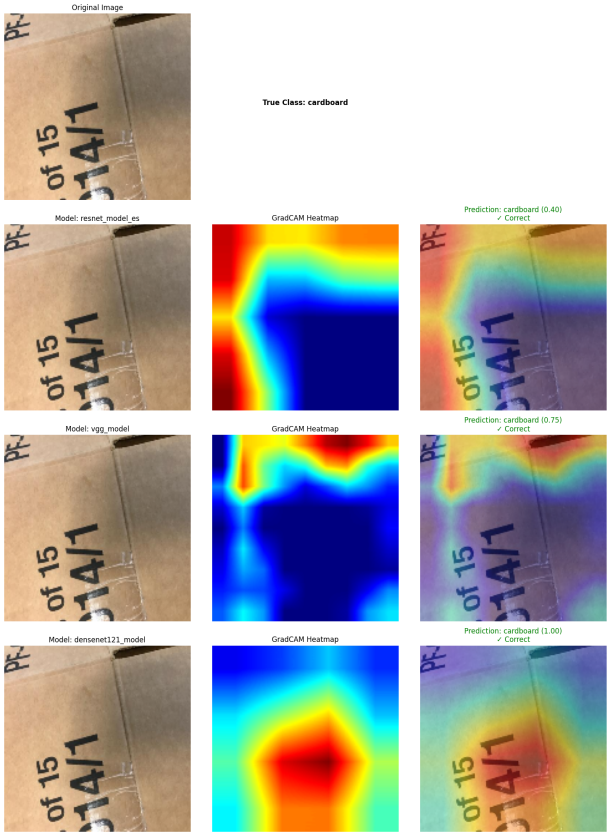

Fig. 3: Consistent cardboard classification performance across all three CNN architectures with focused attention on structural features

c) Architecture-Dependent Consistency and Limitations: Our analysis reveals that explanation quality is strongly dependent on model architecture across several dimensions. DenseNet-121 consistently achieved the highest confidence scores, suggesting more decisive feature representations, while more complex architectures like DenseNet-121 and VGG produced more spatially coherent explanations compared to the standard ResNet implementation. Deeper networks demonstrated superior ability to focus on material-specific characteristics rather than background elements.

The ResNet model's misclassification of the paper bag as cardboard (Figure 1) illustrates important limitations that extend beyond simple accuracy metrics. The diffuse activation pattern suggests difficulty in distinguishing between materials with similar visual characteristics, while lower confidence scores indicate uncertainty in the decision-making process. Most critically, the explanation quality appears correlated with prediction accuracy, raising important questions about the reliability of explanations for incorrect predictions.

d) Implications for Practical Deployment: These findings highlight several critical considerations for deploying GradCAM in practical applications. The choice of CNN ar-

chitecture significantly impacts both prediction accuracy and explanation quality, suggesting that applications requiring high interpretability should prioritize architectures that demonstrate consistent spatial coherence. The correlation between explanation quality and prediction accuracy indicates that Grad-CAM explanations should be interpreted cautiously for low-confidence predictions, particularly in high-stakes applications where explanation reliability is paramount.

Furthermore, the varying performance across material types suggests that GradCAM effectiveness may require domain-specific optimization and evaluation. This finding emphasizes the importance of comprehensive testing across diverse scenarios within a target domain before deployment. The trade-off between model complexity and explanation interpretability also suggests that practitioners must carefully balance performance requirements with interpretability needs, potentially accepting lower accuracy for more reliable explanations in critical applications.

*2) **Local Interpretable Model-agnostic Explanations (LIME) Analysis**:* LIME generates explanations through systematic perturbation of superpixel segments, creating local surrogate models that approximate the decision boundary around individual predictions. Our comprehensive analysis examines how segmentation parameters influence explanation quality and interpretability across different CNN architectures, revealing critical insights for practical deployment.

a) Experimental Configuration and Parameter Settings: Our LIME analysis employs a systematic experimental design with carefully controlled parameters to ensure robust and comparable results. For each image sample, we evaluate two distinct segmentation granularities: coarse segmentation with 20 superpixel segments and fine segmentation with 100 superpixel segments. Within each segmentation configuration, we generate explanations using both 5 and 10 top features, allowing us to assess how the number of highlighted regions affects explanation quality and interpretability. This creates a comprehensive evaluation matrix of four parameter combinations per sample: (20 segments, 5 features), (20 segments, 10 features), (100 segments, 5 features), and (100 segments, 10 features). Perturbations are created by systematically masking different combinations of superpixel segments, enabling the construction of local surrogate models. This experimental design facilitates direct comparison of explanation consistency across different architectural complexities while maintaining controlled conditions for parameter sensitivity analysis.

b) Impact of Segmentation Strategy on Explanation Quality: The superpixel segmentation approach fundamentally determines both the granularity and quality of LIME explanations. Our evaluation across different segmentation configurations reveals a clear trade-off between interpretability and precision that has significant implications for practical applications.

Coarse segmentation using 20 segments (Figure 4) produces more interpretable explanations with clearer regional boundaries that align well with human visual perception. This configuration offers reduced computational complexity due to fewer perturbation combinations, higher stability in explanation generation across multiple runs, and better robustness against local noise and artifacts. The resulting explanations provide intuitive visualizations that are particularly valuable when communicating with non-technical stakeholders.

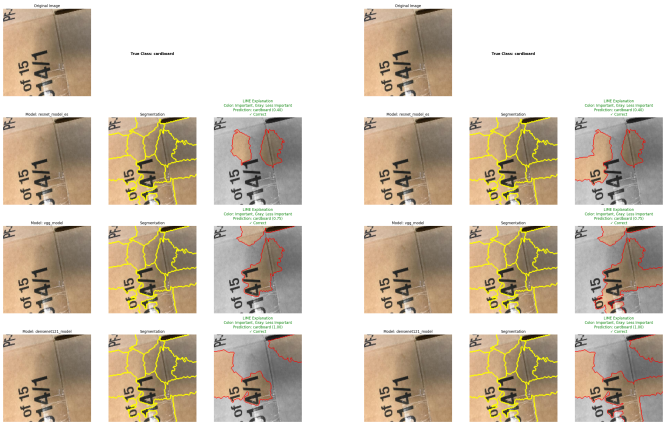

Fig. 4: LIME explanations with coarse segmentation (20 segments) showing clear regional boundaries: 5 features (left) vs. 10 features (right)

In contrast, fine segmentation using 100 segments (Figure 5) enables more precise localization of discriminative features and increased sensitivity to local texture variations. However, this precision comes at the cost of higher computational requirements and potential instability in explanations for ambiguous regions. The fine segmentation approach is particularly valuable when precise feature attribution is critical for understanding model decisions.

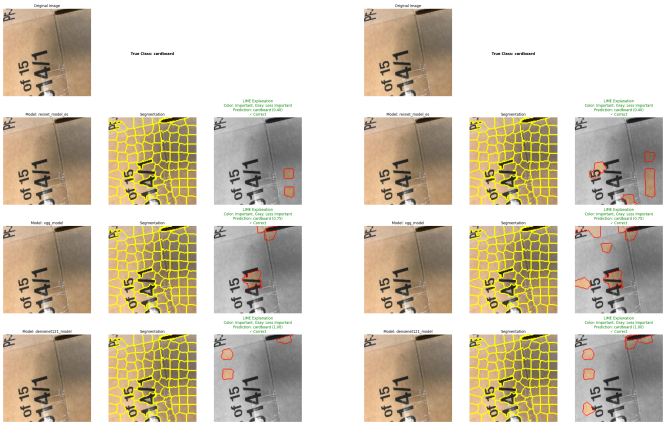

Fig. 5: LIME explanations with fine segmentation (100 segments) enabling detailed feature localization: 5 features (left) vs. 10 features (right)

c) Architecture-Dependent Response Patterns and Model Sensitivity: Different CNN architectures exhibit varying sensitivity to LIME's perturbation strategy, with clear patterns emerging across our evaluation. ResNet demonstrates consistent but relatively low-confidence predictions ranging from

0.40 to 0.75 across all segmentation configurations. The explanations show a preference for larger contiguous regions in coarse segmentation but become increasingly fragmented in fine segmentation, suggesting moderate stability across different perturbation schemes.

VGG shows improved confidence scores between 0.75 and 1.00 with more focused explanations that clearly identify material-specific features in coarse segmentation while better preserving explanation coherence in fine segmentation. The model maintains consistent focus on texture-rich regions regardless of segmentation granularity, indicating robust feature representation.

DenseNet-121 achieves the highest confidence of 1.00 across all configurations, producing the most stable explanations across different segmentation parameters. This architecture demonstrates precise feature localization in both coarse and fine segmentation with minimal sensitivity to perturbation variations, suggesting that the dense connectivity enables more robust decision boundaries.

d) Cross-Class Validation and Generalization: To assess generalization across different material types, we analyzed LIME explanations for paper classification under identical segmentation configurations. The paper examples (Figures 6 and 7) demonstrate consistent behavior patterns across material classes, with coarse segmentation highlighting larger text and surface regions while fine segmentation identifies specific textural elements and edge features.

Fig. 6: LIME explanations for paper classification with coarse segmentation: 5 features (left) vs. 10 features (right)

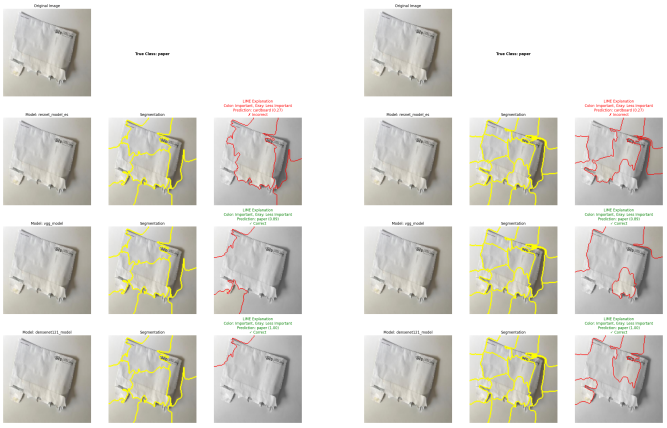

Fig. 7: LIME explanations for paper classification with fine segmentation (100 segments): 5 features (left) vs. 10 features (right)

The paper bag classification case study provides particularly valuable insights into LIME's behavior under challenging scenarios. Across all segmentation configurations, we observe distinct prediction patterns: ResNet consistently misclassifies the paper as cardboard with a confidence of 0.27, VGG correctly classifies it as paper with 0.89 confidence, and DenseNet-121 achieves perfect confidence in correct classification. As segmentation granularity increases, ResNet explanations become increasingly fragmented, highlighting scattered regions that contribute to incorrect predictions. In contrast, VGG explanations maintain focus on the main surface while progressively integrating edge detail, and DenseNet-121 explanations display consistent central focus with refined attention to material-specific features. Notably, LIME excels at capturing the texture boundaries within the image; the segmentation maps clearly delineate the paper's torn edges and the separation between the paper and the uniform background. This demonstrates LIME's ability to isolate semantically meaningful regions, which in turn improves explanation interpretability, especially for visual materials with clear textural transitions.

*3) **SHapley Additive exPlanations (SHAP) Analysis**:* SHAP provides theoretically grounded explanations for model predictions by computing feature attributions based on cooperative game theory principles. Our analysis examines SHAP explanations across three CNN architectures using superpixel-based segmentation to understand how different models focus on distinct image regions for classification decisions.

a) Experimental Configuration and Methodology: Our SHAP analysis employs a systematic approach using superpixel segmentation to create interpretable image regions for explanation generation. Each image is segmented into 16 superpixels using the SLIC algorithm, which groups spatially coherent pixels based on color similarity and spatial proximity. This coarser segmentation creates larger, more interpretable regions that align well with human visual perception while maintaining computational efficiency for SHAP value calculation. The resulting segments form clear geometric boundaries that facilitate straightforward interpretation of feature impor-

tance across distinct image areas.

The explanation process involves systematically masking different combinations of superpixel segments and observing the resulting changes in model predictions. SHAP values are computed for each segment, indicating whether that region increases (positive values, shown in pink/red) or decreases (negative values, shown in blue) the model's confidence for the predicted class. This approach enables direct comparison of how different architectures prioritize various image regions for their classification decisions.

b) Architecture-Dependent Explanation Patterns: Our analysis reveals distinct explanation patterns across the three CNN architectures, demonstrating how architectural design influences feature attribution and decision-making processes. The comparison across ResNet, VGG, and DenseNet-121 architectures provides insights into how model complexity and connectivity patterns affect interpretability.

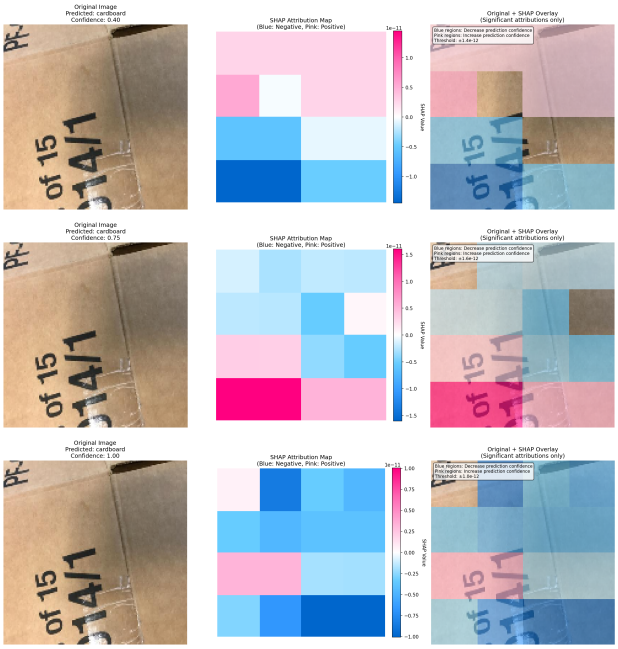

Fig. 8: SHAP explanations for cardboard classification across three architectures: ResNet (top), VGG (middle), and DenseNet-121 (bottom). Each row shows the original image, SHAP attribution map, and overlay visualization.

For cardboard classification (Figure 8), all three models correctly classify the image, but exhibit notable differences in confidence and attribution focus. ResNet achieves the lowest confidence score (0.40), with a more diffuse attribution pattern and scattered positive and negative SHAP values, indicating uncertainty and less discriminative feature usage. VGG improves upon this with a moderate confidence score (0.75), showing more concentrated positive attributions on the printed text and structural edges, while minimizing irrelevant background influence. DenseNet-121 achieves perfect confidence (1.00), producing sharply focused attributions on the most informative visual features, such as the central text and fold lines, while assigning strong negative contributions to the background. These differences reveal how architectural capacity and connectivity influence both confidence calibration and the clarity of learned visual cues.

c) Cross-Class Validation and Prediction Consistency: The analysis extends to paper classification scenarios, revealing how the same architectures handle different material types and the consistency of their explanation patterns across classes. This cross-class examination provides insights into model generalization and the stability of explanation quality across different prediction scenarios.

Fig. 9: SHAP explanations for paper classification.

The paper bag example (Figure 9) reveals significant differences in model behavior and explanation quality. ResNet misclassifies the paper bag as cardboard with low confidence (0.27), showing predominantly negative attributions across most image regions, which reflects the model's uncertainty and poor feature discrimination. VGG correctly identifies the material as paper with high confidence (0.89), demonstrating clear positive attributions on the bag surface and torn regions while maintaining negative attributions in background areas. DenseNet-121 achieves perfect confidence (1.00) in correct paper classification, with highly focused positive attributions on the bag's central features and strong negative attributions elsewhere, indicating superior discriminative capability.

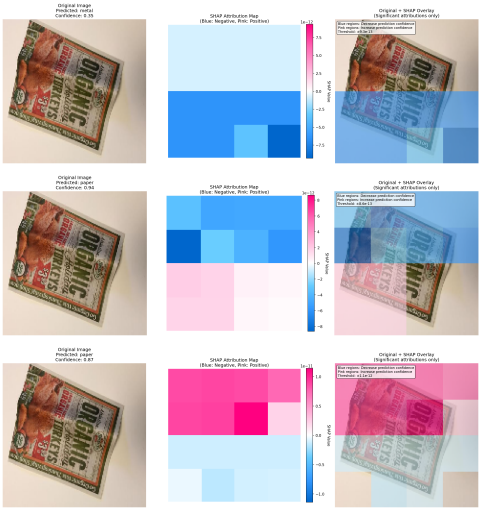

Fig. 10: SHAP explanations for magazine classification demonstrating consistent paper prediction across all architectures with varying confidence levels and attribution patterns.

The magazine classification case (Figure 10) shows all models correctly predicting the label as paper, but with differing confidence and attribution clarity. ResNet yields the lowest confidence (0.35) and scattered attributions, indicating less certainty. VGG achieves the highest confidence (0.94). DenseNet-121 (0.87) highlights informative regions, with consistent attribution to the printed cover and suppression of background influence.

d) Explanation Quality and Model Performance Correlation: The relationship between prediction accuracy and explanation quality emerges as a critical finding across our SHAP analysis. Models that achieve higher confidence scores consistently produce more interpretable and focused explanations, while lower confidence predictions correlate with diffuse or contradictory attribution patterns.

*4) **GradCam vs LIME vs SHAP on Image-Based XAI Methods**:* The collective analysis of GradCAM, LIME, and SHAP across CNN architectures reveals consistent trends that highlight each method's strengths and practical limitations. GradCAM excels in producing fast, spatially smooth heatmaps that align with feature activations in convolutional layers. It performs particularly well with deeper networks like DenseNet-121, offering high confidence and interpretable spatial focus. However, its reliance on internal gradients makes it less flexible and less precise in outlining discrete object boundaries, especially for misclassified inputs or shallow architectures like ResNet.

LIME complements this by delivering highly interpretable segment-based visualizations through superpixel masking. Its sensitivity to segmentation granularity enables both broad overviews and fine-grained focus, with notable ability to delineate texture boundaries—critical for distinguishing materials like paper and cardboard. While LIME provides more human-aligned and stable explanations in high-performing models, its dependency on perturbation strategies introduces variability in low-confidence predictions.

SHAP further enriches interpretability by attributing directional importance to input regions, distinguishing helpful and harmful features for the prediction. This method is particularly effective in highlighting inconsistencies in model confidence and decision-making. For instance, ResNet's low-confidence outputs frequently show diffuse or contradictory SHAP values, whereas DenseNet-121's high-confidence predictions correspond to compact, semantically meaningful attributions.

Together, these findings suggest that no single method is universally superior. GradCAM offers speed and general spatial reasoning, LIME excels in segment interpretability and localization, and SHAP provides principled attribution with polarity. Their comparative use reveals not just how decisions are made, but how confidently and coherently different architectures justify those decisions. For robust deployment, particularly in high-stakes scenarios, combining these methods enables a comprehensive and layered interpretability framework that supports both debugging and trust-building in model behavior.

B. Tabular Data Classification

This section evaluates the model performance and interpretability, using SHAP and LIME explainability methods.

*1) **SHAP Analysis (Global Feature Importance)**:* As shown in Figure 11, the time variable dominates with a negative SHAP value of -0.12, while all other features have contributions below 0.019. Although the relative ranking among clinical variables remains stable, their explanatory power is markedly reduced.

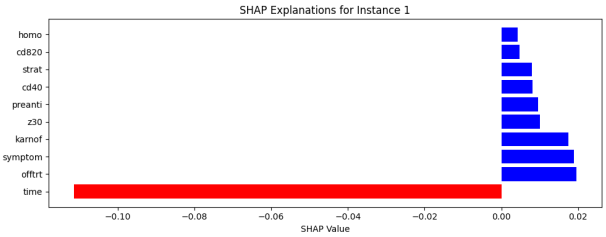

Fig. 11: SHAP explanations: temporal dominance over clinical features

To recover meaningful clinical insights, we re-trained the model without the temporal variable. Figure 12 shows a balanced distribution of clinical feature contributions. Symptom status has the highest positive SHAP value $(\text{SHAP} = 0.038)$, followed by Karnofsky performance score $(\text{SHAP} = 0.018)$, which reflects the functional deterioration associated with disease progression. CD4 counts $(\text{cd40}, \text{cd420})$ contribute protective effects, consistent with clinical literature. Treatment and demographic variables show negative SHAP values, suggesting a protective role in this context.

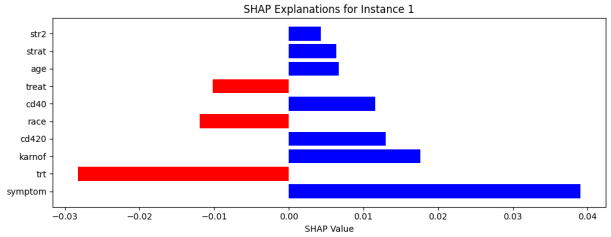

Fig. 12: SHAP explanations without the time feature: balanced clinical contributions

*2) **LIME Analysis (Local Interpretability)**:* Figure 13 shows that temporal thresholds dominate. The range $696 < \text{time} \leq 998$ has the strongest contribution (≈ 0.09). Treatment status ($\text{offtrt} \leq 0.00$) remains somewhat relevant (≈ 0.07), but other clinical features become negligible (≤ 0.03).

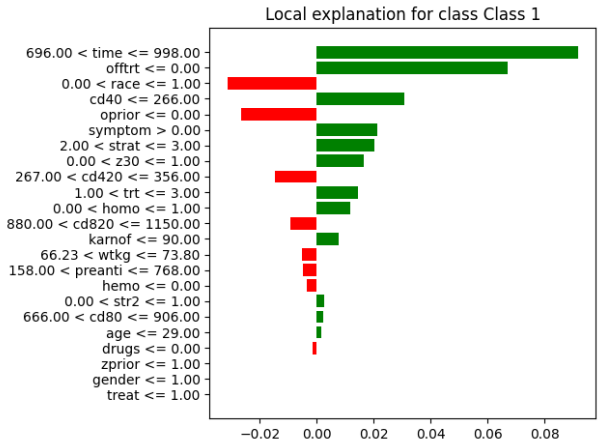

Fig. 13: LIME explanations: temporal threshold dominance

When time is excluded, LIME reveals clearer clinical thresholds. In Figure 14, the strongest risk contributor is $\text{cd40} \leq 266$ (contribution ≈ 0.07). A Karnofsky score ≤ 90 suggests functional decline (contribution ≈ 0.04), while the presence of symptoms adds moderate risk (≈ 0.03). An intermediate CD4 range $267 < \text{cd420} \leq 356$ appears protective.

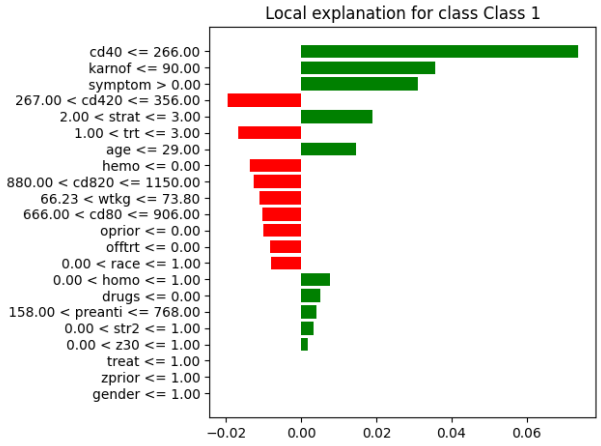

Fig. 14: LIME explanation without the time feature: clear clinical thresholds

*3) **SHAP vs. LIME and Clinical Insight Recovery**:* SHAP and LIME offer complementary interpretability strengths: SHAP provides consistent, global additive attributions, while LIME offers actionable, instance-specific rules with explicit thresholds.

When the time feature is removed, both methods identify clinically validated risk factors: CD4 counts are confirmed as central predictors, with LIME highlighting $\text{cd40} \leq 266$ as a critical threshold. Karnofsky score ≤ 90 serves as a strong indicator of declining functional status. Symptom presence consistently aligns with disease progression. Treatment variables reveal protective patterns, underscoring therapeutic effectiveness.

*4) **Model Performance and the Role of Temporal Features**:* Including a temporal variable significantly improved the model's predictive accuracy: 87% with the time feature, compared to 77% without it, representing a 10% absolute gain. This highlights the predictive power of temporal information but raises questions about interpretability.

*5) **Impact on Model Interpretability**:* The inclusion of time leads to a clear feature masking effect. Time becomes the most dominant feature, five to six times more influential than any clinical variable, reducing the visibility and actionability of other important predictors. As a result, meaningful clinical patterns become compressed or obscured, limiting their value for medical interpretation.

*6) **Clinical and Methodological Implications**:* There is a clear trade-off between interpretability and performance. Removing the time feature reduces accuracy by 10%, but enables much richer clinical insight. The agreement between SHAP and LIME further reinforces the model's reliability and alignment with established medical knowledge.

From a practical standpoint, this analysis emphasizes that high performance is not always synonymous with clinical utility. Removing temporal bias restores transparency, making the model more trustworthy and useful for decision support in healthcare.

Conclusion: Temporal variables enhance predictive accuracy but reduce interpretability by overshadowing clinical features. Their removal reveals actionable, evidence-based medical insights, highlighting the importance of balancing accuracy with transparency in healthcare machine learning applications.

VII. CONCLUSION

This study presents a comprehensive comparative analysis of three prominent Explainable Artificial Intelligence methods—GradCAM, LIME, and SHAP—across both image classification and tabular data domains. Through systematic experimentation on waste classification using CNN architectures and HIV clinical trial data with Random Forest models, we have identified distinct strengths, limitations, and optimal use cases for each XAI technique.

Our findings reveal that explanation quality is strongly dependent on both model architecture and data modality. For image classification tasks, GradCAM excels in providing spatially coherent visual explanations, particularly with deeper networks like DenseNet-121, but shows reduced effectiveness with lower-performing models. LIME demonstrates superior interpretability through segment-based visualizations with notable sensitivity to segmentation parameters, offering intuitive explanations that align well with human visual perception. SHAP provides theoretically grounded feature attributions with clear directional importance, effectively highlighting inconsistencies in model confidence and decision-making processes.

A critical finding across both domains is the strong correlation between prediction accuracy and explanation quality. High-confidence predictions consistently yield more interpretable and focused explanations, while low-confidence outputs often result in diffuse or contradictory attribution patterns. This relationship has significant implications for deploying XAI methods in high-stakes applications where explanation reliability is paramount.

For tabular data analysis, our investigation into temporal feature effects reveals a fundamental trade-off between predictive performance and interpretability. The inclusion of temporal variables improved model accuracy from 77% to 87% but created a feature masking effect that obscured clinically meaningful patterns. Removing temporal bias restored interpretability, enabling both SHAP and LIME to identify clinically validated risk factors such as CD4 counts, Karnofsky performance scores, and symptom status, demonstrating strong alignment with established medical knowledge.

The comparative evaluation demonstrates that no single XAI method is universally superior. Instead, method selection should be guided by specific application requirements: GradCAM for rapid spatial analysis in computer vision, LIME for intuitive local explanations requiring human interpretation, and SHAP for principled global feature importance analysis. For robust deployment, particularly in critical applications, our findings suggest that combining multiple XAI methods provides a comprehensive interpretability framework that supports both model debugging and trust-building.

A. Key contributions

This work includes: a systematic comparison of XAI methods across multiple data modalities and model architectures, empirical evidence of the relationship between model performance and explanation quality, and practical guidelines for XAI method selection based on application requirements.

B. Future research directions

Future research should explore the application of these techniques to time-series and multimodal datasets, investigate newer methods such as Integrated Gradients and attention-based explanations, and develop standardized evaluation frameworks for XAI methods across different domains. Additionally, research into hybrid explanation approaches that combine the strengths of multiple XAI techniques could further enhance interpretability in complex machine learning systems.

VIII. WORK LOAD

Each student worked 50% of the project.

ACKNOWLEDGMENT

The authors would like to thank professor Petia Georgieva, regent of CAA course, for her support and assistance throughout the project.

REFERENCES

[1] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[2] C. Molnar, *Interpretable Machine Learning*, 3rd ed., 2025. [Online]. Available: https://christophm.github.io/interpretable-ml-book

[3] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (xai)," *IEEE access*, vol. 6, pp. 52 138–52 160, 2018.

[4] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[5] 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.

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

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

[8] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," *arXiv preprint arXiv:1606.05386*, 2016.

[10] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[11] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[12] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 839–847.

[13] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-cam: Score-weighted visual explanations for convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 24–25.

[14] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," *Advances in neural information processing systems*, vol. 31, 2018.

[15] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

[16] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature machine intelligence*, vol. 2, no. 1, pp. 56–67, 2020.

[17] 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.

[18] D. Garreau and U. Luxburg, "Explaining the explainer: A first theoretical analysis of lime," in *International conference on artificial intelligence and statistics*. PMLR, 2020, pp. 1287–1296.

[19] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 11, pp. 2660–2673, 2016.

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

[21] U. Bhatt, A. Weller, and J. M. Moura, "Evaluating and aggregating feature-based model explanations," *arXiv preprint arXiv:2005.00631*, 2020.

[22] J. Adebayo, M. Muelly, I. Liccardi, and B. Kim, "Debugging tests for model explanations," *arXiv preprint arXiv:2011.05429*, 2020.

[23] N. Arun, N. Gaw, P. Singh, K. Chang, M. Aggarwal, B. Chen, K. Hoebel, S. Gupta, J. Patel, M. Gidwani *et al.*, "Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging," *Radiology: Artificial Intelligence*, vol. 3, no. 6, p. e200267, 2021.

[24] R. Marcinkevičs and J. E. Vogt, "Interpretability and explainability: A machine learning zoo mini-tour," *arXiv preprint arXiv:2012.01805*, 2020.

[25] CCHANG, "Garbage classification," 2018. [Online]. Available: https://www.kaggle.com/ds/81794

[26] "AIDS Clinical Trials Group Study 175," UCI Machine Learning Repository, 2024.