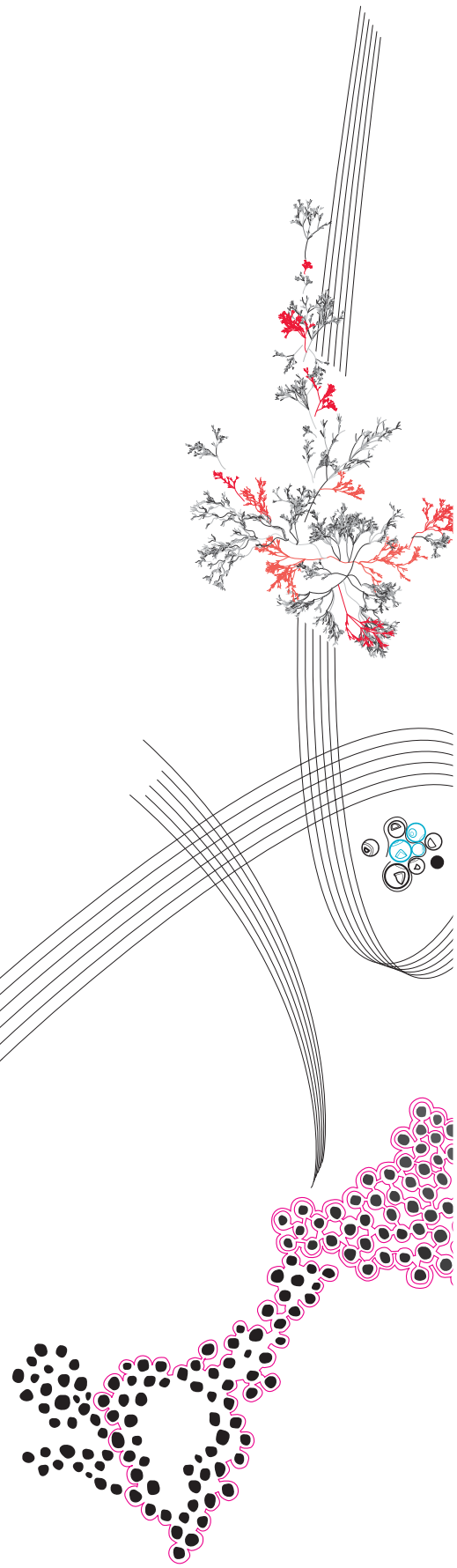


MSc Computer Science
Specialization: Data Science and Technology.
Master Thesis

From Unknown to Understood: Evolution of XAI Over Active Learning Cycles

Srivatsa Chidara

Supervisor: Dr. Faizan Ahmed
Examiner: Dr. Alex Stergiou

August.2025

Department of Computer Science
Faculty of Electrical Engineering,
Mathematics and Computer Science,
University of Twente

UNIVERSITY OF TWENTE.

Contents

Abstract

Deep Learning models require a huge pool of labeled data that often lack transparency and interpretability which can hinder trust in the AI systems. Explainable Active Learning (XAL) aims to address these challenges by selecting informative data samples for labeling and providing explanations for model decisions. However, little is known about how and when the quality of explanations evolve across successive active learning cycles. This research investigates this evolution by implementing a framework based on functionally grounded metrics like Continuity, Compactness, Correctness, Consistency, and an aggregated metric called XAI Quality score (XQS). These metrics are computed on two widely used XAI methods, SHAP and GradCAM, which are applied to ResNet-34 model that is trained via entropy-based uncertainty sampling on both PCAM (medical) and CIFAR-10 (natural) imaging datasets. Results show that while active learning enhances both predictive accuracy and explanation quality with few samples, evolution of explanation depends greatly on dataset domain, choice of XAI method and evaluation metrics employed. This thesis also addresses the challenges of using pre-trained Deep CNN and only a single active learning strategy as well as the feasibility of aggregated metric fine-tuned for medical imaging. It provides an initial framework for tracking and evaluating the evolution of explanations in XAL workflow that can be further enhanced using un-trained lightweight models, domain-adaptive metrics, computational optimization, and diverse active learning strategies.

Chapter 1

Introduction

1.1 Introduction

The use of Machine Learning and Deep Learning models has increased rapidly in recent years, achieving state-of-the-art performance across various fields. These models face a growing challenge regarding the amount of trust placed in their predictions due to lack of transparency. This lack of transparency is described as the "black-box" problem and has led to rising demand for interpretability especially in domains like healthcare. This leads to the emergence of Explainable Artificial Intelligence (XAI)[1, 4, 12]. These XAI methods provide explanations for model decisions that humans can interpret which enhance trust in model's performance.

Active Learning (AL) has gained popularity as an efficient strategy for reducing annotation costs without sacrificing model performance[13, 49]. It achieves this by selecting the data points with the most information for labeling which optimize data acquisition. However, traditional AL strategies such as uncertainty sampling, random sampling do not provide any insights into why specific samples are chosen for labeling. This lack of interpretability can undermine trust and usability in human-in-the-loop settings where practitioners seek explanations for both the predictions and sampling decisions[29, 40].

To address this, Explainable Active Learning (XAL) is introduced by incorporating XAI techniques into AL workflow to enhance interpretability of the training process. This interpretability can be invaluable in high stake domains such as medical imaging where understanding both model predictions and annotation choices is critical. Existing XAI research employs metrics such as Faithfulness, Stability, and Human evaluation[16, 34, 50, 24] and Fidelity, Consistency, Comprehensibility[14, 39]. However, there is no standardized framework to capture the evolution of explanations with the evaluation metrics over active learning cycles.

This thesis presents a systematic investigation into evolution of XAI explanations during active learning. We hypothesize that explanation quality evolves in a meaningful and potentially predictable way as model progressively learns from more informative samples. To test this, functionally grounded proxy metrics and an aggregated metric are computed for two widely used XAI methods over active learning cycles in both medical and natural image domains. The results are compared across datasets to find if the evolution patterns are consistent or domain-specific. Detailed

analyses are conducted to answer both the main research question and sub-research questions outlined below. While this work establishes an initial framework for measuring explanation evolution, it is limited by the focus on only two XAI methods, a single AL strategy, significant computational overhead, use of proxy metrics rather than human assessment, and dependence on pre-trained CNN model for training.

1.2 Research Questions

This section outlines the main and the sub research questions guiding this thesis. The questions focus on understanding the evolution of XAI explanations, and how these changes relate to the model performance. This thesis aims to provide insights into the dynamics of explanation quality and its implications on different domains with varied XAI methods.

MAIN RESEARCH QUESTION:

RQ: How does the quality of XAI explanations evolve across active learning cycles?

SUB RESEARCH QUESTIONS:

SRQ1: How do quantitative XAI metrics (correctness, compactness, consistency, continuity) evolve through Active learning cycles?

SRQ2: How closely do improvements in explanation quality align with gains in model performance?

SRQ3: How do explanation trends differ between SHAP and GradCAM during active learning?

SRQ4: How does dataset domain (medical vs. natural images) influence the progression of explanation metrics?

1.3 Thesis Structure

The introduction is followed by the background section. This provides an overview of key areas such as XAI, Active learning, Evaluation metrics. This is followed by the research methodologies and implementation framework. Next section presents the experiments conducted and their results. It is followed by a discussion that addresses the findings, limitations and future work. Finally, the thesis is concluded by a summary of how research questions are answered.

Chapter 2

Background

This chapter provides the foundational concepts, context, and reasoning to understand the research in thesis. It begins with the core principles of XAI, AL, and CNN. The chapter further explores the evaluation strategies for both model performance and explanation quality. It highlights the unique challenges associated with assessing the XAI methods. By systematically reviewing these topics, the section puts the research question within the broader landscape of ML, and XAI.

2.1 Explainable AI

Explainable Artificial Intelligence (XAI) refers to a set of methods designed to make the decision making of ML models transparent and interpretable to humans. As AI systems become more complex and opaque which are often described as black boxes [12], it is difficult to understand how certain predictions are made [1, 4]. This lack of transparency poses a significant challenge in high-stakes domains such as healthcare, finance, and other systems. Here understanding the reason behind prediction is essential for trust and regulatory compliance [4]. XAI addresses these challenges with the use of advanced tools and techniques that can decipher the model's inner workings. In this study, XAI methods act as the backbone of whole research as every method is built around them for qualitative and quantitative insights.

2.1.1 SHAP

Shapley Additive exPlanations (SHAP) is a model-agnostic XAI technique that attributes the output of a ML model to its input features using principles from cooperative game theory [12]. SHAP computes the scores for every feature by representing its average contribution to the model's prediction across all feature combinations(Equation 2.1). It provides both local and global insights. SHAP can highlight the pixels that have a strong influence on the model's decision. It is a valuable tool for debugging and validating model behavior [33]. Gradient SHAP [33] which is specifically designed for deep learning models is used in this research. It combines SHAP and integrated gradients by attributing the output to its input features by averaging integrated gradients with randomly perturbed baselines. It provides pixel level explanations that are quantitatively evaluated through AL pro-

cess. The names SHAP and Gradient SHAP are used interchangeably throughout this document.

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[v(S \cup \{i\}) - v(S) \right] \tag{2.1}$$

where,

ϕ_i : Shapley value for feature i
F : Set of all input features
S : Subset of features not containing i
$|S|$: Cardinality (number of elements) of set S
$v(S)$: Model output when only features in S are present

2.1.2 GradCAM

Gradient-weighted Class Activation Mapping (GradCAM) is a model-specific XAI method tailored for convolutional neural networks (CNN). It provides visual explanations for classification decisions [47]. GradCAM works by computing the gradients of a target class score with respect to the feature maps in the last convolutional layer (Equation: 2.2). It produces a heatmap that highlights the most influential regions of the image for model's prediction [17]. GradCAM++ [10] an improved variant which is tailored to CNN is used in the thesis. It is computed by a weighted combination of positive partial derivatives of last convolutional layer feature map for better localization, and reduce noise in explanations. It provides region level explanations that are quantitatively evaluated through AL process. The names GradCAM and GradCAM++ are interchangeably used across the document.

$$L^c_{\text{GradCAM++}} = \text{ReLU}\left(\sum_k \alpha_k^c A^k \right), \alpha_k^c = \frac{\sum_{i,j} \frac{\partial^2 y^c}{(\partial A_{ij}^k)^2}}{2 \sum_{i,j} \frac{\partial^2 y^c}{(\partial A_{ij}^k)^2} + \sum_{i,j} A_{ij}^k \frac{\partial^3 y^c}{(\partial A_{ij}^k)^3}} \tag{2.2}$$

where,

ReLU : Rectified Linear Unit function
α_k^c : Weight for feature map k and class c
A^k : Activation map from the k-th channel of the last convolutional layer
y^c : Score for class c
A_{ij}^k : Activation at spatial location (i, j) in feature map k
$\frac{\partial^2 y^c}{(\partial A_{ij}^k)^2}$: Second-order partial derivative of y^c w.r.t. A_{ij}^k
$\frac{\partial^3 y^c}{(\partial A_{ij}^k)^3}$: Third-order partial derivative of y^c w.r.t. A_{ij}^k

2.2 Active Learning

2.2.1 Entropy based Uncertainty Sampling

Active learning is a ML paradigm that aims to reduce annotation costs by iteratively selecting the most informative samples for labeling [49, 29]. The approach is particularly valuable in domains where the labeling costs are very high as it focuses annotation efforts on ambiguous or uncertain data. Widely used strategy is entropy-based uncertainty sampling in which the model computes the Shannon entropy of its predicted class probabilities for all unlabeled data points and selects those with highest entropy for labeling by human. The entropy for a sample x with predicted class probabilities $p_1, p_2, ..., p_K$ is given by the equation 2.3. It targets data points where the model is least confident thus accelerating learning and improving efficiency [2, 34].

$$H(x) = -\sum_{k=1}^{K} p_k \log p_k \tag{2.3}$$

where,
$H(x)$: Entropy of the model's prediction for sample x
K: Number of classes
p_k: Predicted probability for class k

2.3 Convolutional Neural Network

Convolutional Neural Networks (CNN) are a class of deep learning models that are the standard for image recognition due to their ability to learn feature representations directly from raw pixel [2]. CNN utilize convolutional layers to extract spatial features. Among all the architectures, the ResNet family introduced the concept of residual learning. It allows for training of much deeper networks by incorporating shortcut connections to mitigate vanishing gradient problem. ResNet-34 is a widely used architecture that contains 34 layers with residual blocks. This strikes a balance between complexity and efficiency while delivering strong performance [18]. In this thesis, ResNet-34 is employed as the backbone model, leveraging its proven effectiveness.

2.4 Evaluation

This section outlines the evaluation strategies used in the research for both the model performance and the quality of XAI explanations. Standard classification metrics such as accuracy, precision, recall, F1-score, and AUC-ROC are used to evaluate model performance. Metrics such as consistency, compactness, correctness, and continuity are used for XAI explanations.

2.4.1 Model Evaluation

The ML model evaluation metrics are quantitative measures used to assess the predictive performance. In this study, they are applied to image classifications tasks on PCAM and CIFAR-10 datasets. These metrics include accuracy, which measures the proportion of correctly classified images. Precision evaluates the fraction of true positive among all the positive predictions. It is crucial for minimizing false alarms in domains like medical imaging. Recall quantifies the proportion of actual positives correctly identified by the model. It is especially significant for PCAM because missing a cancerous patch could have serious implications. F1-score balances precision and recall. AUC-ROC measures the model's ability to distinguish between classes across all possible threshold. It is particularly important for binary classification like PCAM. These metrics are derived from confusion matrix which summarizes true positive, false positive, true negative, and false negative. Their relationships and formulas are illustrated in Figure2.1. Together, these metrics ensure a comprehensive evaluation of model performance as it evolves after each AL cycle.

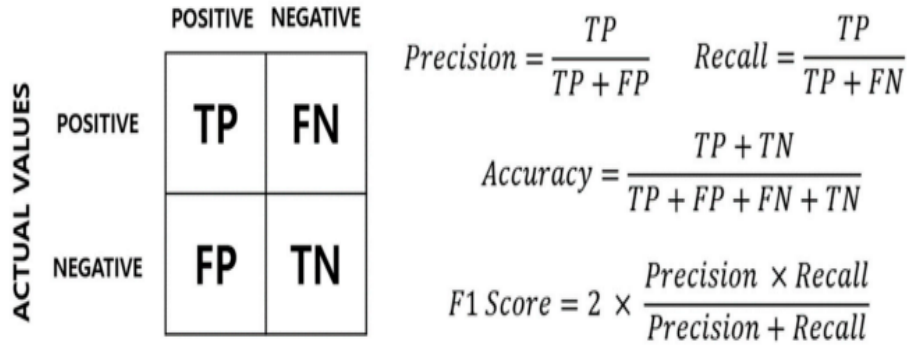


FIGURE 2.1: Evaluation Metrics in terms if True/False, Positive/Negative [48]

2.4.2 XAI Evaluation Metrics

To quantitatively assess the quality of explanations produced by XAI methods, a set of metrics are employed by following recommendations from CO-12 framework Nauta et al. for rigorous XAI evaluation. Specifically suited for image datasets and active learning. These metrics are continuity, consistency, compactness, and correctness.

- **Continuity:** It evaluates whether the explanations remain stable when input image is subjected to minor perturbations like changes in pixel values or slight noise addition. High continuity indicates robust and reliable explanations.
- **Compactness:** It is also referred as sparseness. Quantifies how concentrated the attribution values are within the explanations. A compact explanation that focuses on small, meaningful pixels or regions makes the interpretation easier. This is particularly valuable for image analysis, as overly diffused explanations can obscure important insights.

- **Correctness:** Correctness quantifies how well an explanation captures the model's actual decision making process during image classification. It is measured by altering image regions deemed as important by XAI methods, and observing the resulting change in the model's prediction. A huge change in the value indicates the explanation accuracy. This in turn reflects the features the model relies on. It provides an objective measure of alignment between explanation, and model reasoning as it evolves over AL cycles.

- **Consistency:** Consistency measure the stability of the explanations across different AL cycles. As the model is updated and retrained with new data points at each cycle, explanation for a given image should remain similar if the model's understanding is stable. High consistency indicates interpretable and reliable explanations.

By employing these metrics, this thesis ensures a comprehensive and objective evaluation of XAI methods. Moving beyond visual assessments and supporting reproducible, quantitative comparison of explanation quality over AL cycles.

2.5 Challenge of Evaluating Explanations

2.5.1 Ground Truth Problem

A persistent challenge in the evaluation of XAI methods for image classification tasks is the absence of ground truth for explanations [15, 38, 41]. Unlike standard supervised learning, where model predictions can be compared to known class labels. There are no universally accepted references for what constitutes a correct explanations [32]. It makes robust and reproducible evaluation difficult. This complicates the assessment if an XAI method accurately reflects the model's internal reasoning or merely producing outputs. It often leads to reliance on subjective and anecdotal evaluation methods that are not scalable or consistent [41].

To address the problem, we adopt functionally grounded proxy metrics that quantitatively assess explanation quality without requiring human-annotated ground truths [41]. These metrics enable systematic, objective, and reproducible evaluation of XAI methods as model evolves through AL cycles. It directly tackles the ground truth problem and supports scientific comparison of explanation quality. However, it is important to acknowledge that while these metrics offer a practical solution they also have limitations. They may not fully capture semantic or domain-specific relevance of explanations. They can overlook subtle aspects that only human experts would recognize. Despite these limitations, these metrics provide a significant step towards more robust and transparent evaluation of XAI methods.

Chapter 3

Methodology

This chapter presents the methodological approach adopted to achieve the research goals of the thesis. It begins by outlining the adoption of CRISP-DM and DSR methodologies providing the foundation for a reproducible, rigorous, and systematic research process. The chapter describes and justifies the overall research strategy, framework, and processes that are guiding the study. It mentions specific methods employed for data collection, model development, and evaluation.

3.1 Research Methodologies

We combine Design Science Research and CRISP-DM methodologies to systematically investigate the evolution of XAI over AL cycles. The following subsections detail how these methodologies structure the research process.

3.1.1 Design Science Research (DSR)

Design Science Research (DSR) is a research paradigm which focuses on the systematic creation and rigorous evaluation of the artifacts, such as models, methods, or frameworks to address complex and real-world problems in computer science [20]. Hevner et al. formalize this paradigm by proposing seven guidelines for conducting research. These guidelines ensure that DSR projects are scientific and practical. The research process is typically iterative involving cycles of problem identification, artifact development, and empirical evaluation refined at each iteration by theoretical insights and empirical feedback [20].

DSR provides the methodological framework for investigating the RQ on how the explanations evolve over AL cycles. The research is structured around the iterative development and assessment of the experimental pipeline that quantifies XAI explanation quality using functionally grounded metrics. The artifact i.e., the XAI evaluation pipeline is applied across two distinct datasets. The evaluation strategy is explicitly guided by comprehensive framework of Venable et al. which emphasizes context sensitive and multi-faceted artifact assessment. The framework advocates for careful alignment of evaluation strategies with the RQ and the artifact characteristics. It distinguishes between *ex ante and ex post* evaluations along with artificial and naturalistic settings. Ex post and artificial evaluations are conducted

systematically by applying pipeline to benchmark the datasets in controlled environments. It enables robust comparison of XAI metric evaluation over AL cycles is both empirically rigorous and practically relevant [20, 54].

3.1.2 CRISP-DM

The Cross-Industry Standard Process for Data Mining (CRISP-DM) is widely adopted methodology that structures the data science workflow into six steps [9, 37]. It provides a hierarchical and iterative process that ensures each stage is systematically planned, executed and documented. Notably, there is significant overlap between DSR and CRISP-DM as they both emphasize the iterative process, evaluation, and objective alignment. Employing both methodologies is justified as DSR offers a scientific foundation for artifact creation and evaluation. While CRISP-DM structures the workflow ensuring methodological rigor and reproducibility [9, 37].

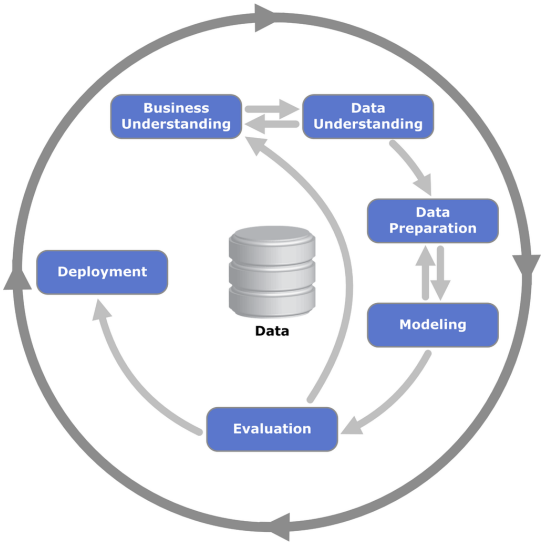


FIGURE 3.1: CRISP-DM Methodology [9]

The scope of this thesis aligns with the CRISP-DM framework. It is because of the organized research around the phases that mirror the structure of the methodology. It starts from defining the research questions to evaluation and deployment. Figure 3.1 illustrates the cyclical and interconnected nature of the methodology. Arrows between the phases indicate the frequent iteration and feedback. It allows the process to adapt dynamically to new insights. The approach ensures that insights learned at any stage can inform the earlier phases [9].

3.1.3 Applying the Methodologies

This subsection presents mapping of DSR and CRISP-DM methodologies to the steps of the study. The following table 3.1 shows the alignment between each methodological phase and the specific research tasks. Here DSR phases from Peffers et al. are used as Hevner et al. provides guidelines rather than the phases.

Table 3.1: Application of the Methodologies

DSR	**CRISP-DM**	**Thesis Implementation**
Problem Identification	Business Understanding	Defined Research Questions and selected tools.
Define Objectives	Data Understanding	Explored medical and natural image datasets, assessed data quality, and identified limitations and potential of the data.
Design& Development	Data Preparation	Preprocessed images, initialized subsets for experiments, and set up pipeline for reproducibility.
Demonstration	Modeling	Trained ML model with AL sampling, applied XAI methods, and tracked 100 samples per dataset.
Evaluation	Evaluation	Evaluated model performance and XAI explanations across cycles, performed correlation analysis, discussed whether RQs are answered.
Communication	Deployment	Synthesized findings, documented code, formulated recommendations for XAI evaluation in AL workflows.

This mapping provides a clear reference on how each stage of the thesis aligns with the methodological framework. It sets a stage for detailed implementation which is described in the next section.

3.2 Implementation Framework

Building on the methodological mapping of DSR and CRISP-DM, Figure 3.2 presents the experimental framework developed. The framework is specifically designed on central research question. The framework integrates both medical image and natural image datasets as reliance on single data domain can induce bias and limit generalization. The choice of ResNet-34 as the model architecture reflects the balance between complexity and interpretability. The use of both SHAP for pixel-wise explanations and GradCAM for region based explanations ensures a comprehensive comparison of XAI models.

Given the absence of ground-truth for explanations of XAI methods, framework uses a set of functionally grounded proxy metrics to evaluate explanation quality. It addresses a major challenge in current XAI landscape. The cyclical nature of the framework helps iterative data selection, model updating, and evaluation.

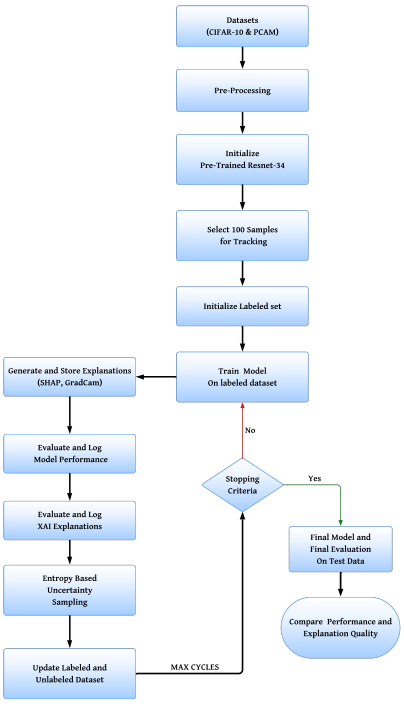


FIGURE 3.2: Framework Architecture

3.2.1 Implementation Details

All the experiments are conducted in Python by utilizing PyTorch and torchvision for model development and training with ResNet-34. Active learning was implemented using entropy-based uncertainty sampling for iterative data selection in both PCAM and CIFAR-10 datasets. SHAP (Captum) and GradCAM (PyTorch) provide XAI explanations. The framework computes and logs standard classification metrics (Sklearn) alongside XAI metrics (Quantus [30]) for each cycle. It tracks similar images across cycles for reproducible analysis of explanation quality.

The following measures are taken to ensure reproducibility and fair comparison across the experiments:

- **Fixed Random Seed:** A fixed random seed (42) was set for Python, NumPy, PyTorch to control stochasticity in data shuffling, weight initialization, and other random processes.
- **Stratified Data Splits:** Training, validation, and test subsets were created using stratified sampling to ensure class distributions. Indices for these splits were saved using pickle module and used across all the executions.
- **XAI Tracking Subset:** A set of 100 validation images per dataset are reserved using stratified sampling, and their indices are stored to ensure same images are used in every cycle to maintain consistency.

3.2.2 Dataset Description

PatchCamelyon (PCAM)

The first dataset used is PCAM [53] which is a binary image classification dataset from the Camelyon16 [6] challenge. It is aimed at detecting metastatic tissue in histopathological lymph node images [52]. It comprises 327,680 H&E-stained color image patches of 96×96 pixels. It is labeled based on the presence of tumor tissue in the central 32×32 region of the image. The design supports fully convolutional models without zero-padding, ensuring consistency on whole slide-images. The dataset is fully balanced with approximately 50% being negative (no tumor), see Figure 3.4 and 50% being positive (tumor), see Figure 3.3. It is split into training set of 262,144 images (6.8 GB), validation and test sets of 32,768 images each (865 MB). Stratified subsets of 40,000 training, 10,000 validation and test sets each were used to ensure minimal computational overload are used in this research. We sourced the dataset via PyTorch's torchvision datasets to ensure compatibility with the rest of the architecture that is built around PyTorch framework.

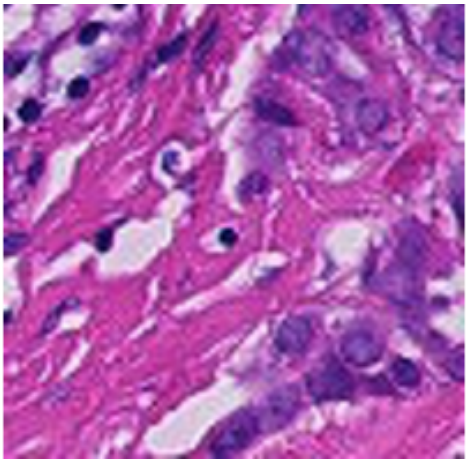


FIGURE 3.3: Positive Sample

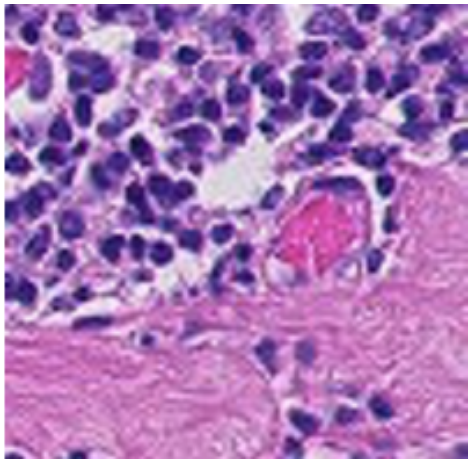


FIGURE 3.4: Negative Sample

CIFAR-10

The second dataset used is CIFAR-10. It is a well established benchmark dataset for image classification [26]. It consists of 60,000 tiny color images ($32 \times 32 pixels$) distributed across 10 classes namely airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. See Figures 3.5,3.6 for example images. Each class contains exactly 6,000 images resulting in a perfectly balanced dataset with 50,000 (150 MB) training and 10,000 (30 MB) testing images without predefined validation set. Original training set was split into 40,000 training, and 10,000 validation samples using stratified sampling. CIFAR-10 is also downloaded directly from PyTorch's torchvision datasets. This dataset acts as an ideal benchmark that provides a controlled environment with established baselines for evaluation of model performance and explanation quality.

FIGURE 3.5: Frog

FIGURE 3.6: Ship

Both datasets exhibit a measurable spread across samples and classes. It confirms that subsets used in this thesis are neither overly dispersed nor tightly clustered. The global variance and standard deviation for PCAM ($\sigma^2 = 1.22$, $\sigma = 1.09$) and CIFAR-10 ($\sigma^2 = 1.16$, $\sigma = 1.07$) indicates substantial diversity in data points. Class-level variance for PCAM(σ^2 : 1.71&1.30) and CIFAR-10 (σ^2 : $0.84 - 1.40$) across ten classes further reveal intra-class variability. These quantitative results along with dispersed points seen in the corresponding t-SNE plots (see Figure 3.7, 3.8) demonstrate that observed performance in AL pipeline is the result of framework rather than subsets having low-complexity, non-diverse samples.

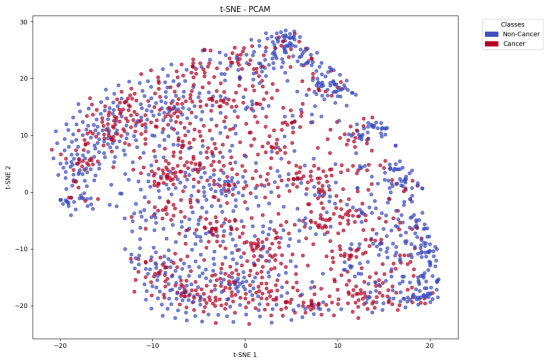


FIGURE 3.7: t-SNE PCAM

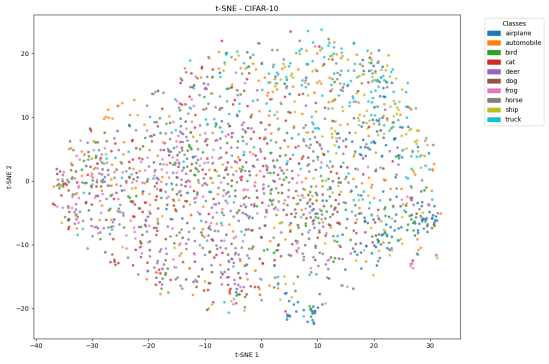


FIGURE 3.8: t-SNE CIFAR-10

3.2.3 Data Pre-Processing

This step ensures data compatibility with pre-trained ResNet-34 architecture and consistency across AL cycles. We normalized all images using ImageNet statistics ($\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$) as recommended by PyTorch for pretrained models [51]. CIFAR-10 images were up-sampled from 32×32 to 224×224 pixels to match ResNet's input requirements. We resized PCAM images from 96×96 to 224×224 pixels using bi-cubic interpolation to minimize artifacts. PCAM training transformations include random resized cropping (scale $0.8 - 1.0$),

horizontal/vertical flips ($p = 0.5$), rotation ($\pm 15°$), and conservative color jittering (brightness/contrast ± 0.1). This handles orientation, slide preparation, and staining variations. It preserves H&E diagnostic integrity by avoiding saturation and hue modifications. Only resizing and normalization are applied to test and validation sets for both the datasets to ensure deterministic evaluation conditions.

We adapted data augmentation strategies to address the specific requirements of each dataset. Minimal transformations are applied for CIFAR-10 to preserve its benchmarking ability. Medically appropriate augmentations are applied for PCAM which are chosen to make model robust without risking any loss of key diagnostic details. We kept training and testing transformations separate to help model generalize better during training and ensuring evaluation happens under standardized conditions. This approach enables reliable and consistent evaluation across the experiments.

3.2.4 Machine Learning Model

The training framework is customized for iterative nature of AL while maintaining consistency across cycles. ResNet-34 is initialized with pretrained weights with a fully connected layer adjusted for dataset specific classes. Dataset specific hyperparameter are determined after extensive experimentation. Adam optimizer was used with a learning rate of 5×10^{-5} for PCAM and 1×10^{-4} for CIFAR-10. Scheduler was configured with patience of 2 epochs, and factor of 0.1 for learning rate adaptation. For CIFAR-10 class weights are computed at each cycle based on distribution of labeled samples to compensate for class imbalance (see Figure 3.9) caused by uneven sample selection. PCAM employs fixed class weights to prioritize the positive class to increase recall for medical imaging as AL consistently favors negative samples for labeling as seen in Figure 3.10.

Active Learning parameters are selected based on established research and experimentation. An initial pool of 2000 stratified samples selected representing 5% of training pool [8] and queried 100 new samples per cycle to maintain class coverage. Training followed a cycle-dependent approach with 30 epochs for $1 - 5$ cycles and 20 epochs for beyond. It helps the model to learn on smaller initial dataset before adapting to more data points. Training metadata including hyper-parameters, validation metrics, AL labeled samples, class weights, and model checkpoints were systematically saved after each cycle for comprehensive analysis.

3.2.5 Explainability Methods

At the beginning of experiments a fixed stratified subset of 100 validation images is selected. Explainability maps are generated for these same 100 images at every AL cycle to ensure consistent comparison across cycles. Post-hoc methods such as Gradient-SHAP and GradCAM++ are applied to track the evolution of XAI across AL cycles. Detailed explanations of these methods are outlined in section 2.1.

- **Gradient-SHAP:** It creates pixel-level attribution maps by computing integrated gradients with baseline and input images. Baseline images are randomly

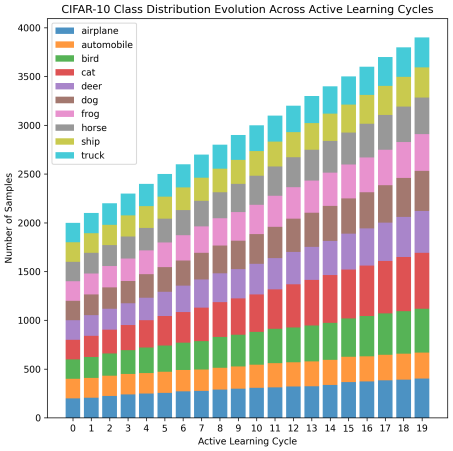


FIGURE 3.9: CIFAR-10 class distribution across AL cycles

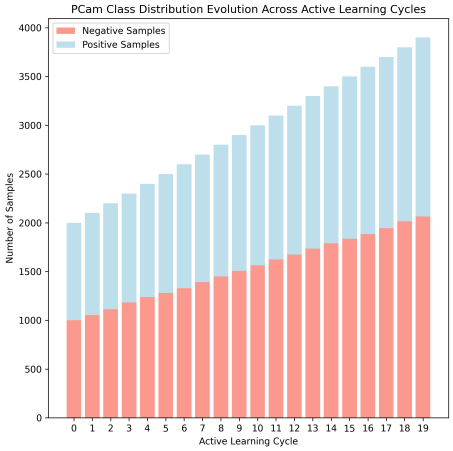


FIGURE 3.10: PCAM class distribution across AL cycles

stratified samples from validation subset. Attribution maps are normalized and superimposed on top of original images to highlight focus areas.

- **GradCAM++:** It creates class discriminative heatmaps by computing weighted gradients with respect to last convolutional layer. The heatmaps are superimposed on top of original images to visualize the focus area.

Pixel-level and region-level evaluation is conducted by combining these methods. All the visualizations are stored along with the original image for inspection and analysis.

3.2.6 Evaluation Approach

As outlined in Section 2.4, a diverse set of evaluation metrics are used for ML and XAI evaluation. The primary metrics used for PCAM are Recall, and AUC-ROC for clinical safety and model quality. The primary metric for CIFAR-10 is Accuracy because it reflects overall model performance. Both the datasets are supplemented by precision and F1-score. For XAI evaluation: continuity, consistency, compactness, correctness, and an aggregated mean of the metrics called XQS (XAI Quality Score) are tracked. These metrics are computed and logged at each AL cycle for comparison and analysis.

The full source code, including scripts for pre-processing, training, evaluation, and XAI analysis, is available at: Github [11]

Chapter 4

Results

This chapter presents the results of the methodology provided in previous section. It begins with an overview of the experiments conducted followed by detailed presentation of the results for both the datasets. Model performance, XAI evaluation, and series of analysis are reported for each dataset.

4.1 Overview of Experiments

This section provides an overview of the experiments conducted. As discussed in Table 4.1, there are three experiments including a baseline. Except for Continuity metric, all the experiments are duplicated for both the datasets. In all experiments mean and median aggregations of tracked image scores are calculated. Specific correlation functions are used for each experiment as mentioned in table. For continuity metric, gaussian noise (PCAM) and image rotation (CIFAR) are added as perturbation. For compactness, there are no parameters that are subjected to change. For correctness, perturbation baseline is a mask applied to original image to replace important pixels. Black refers to setting pixel value to zero. Mean refers setting to mean pixel value of the image. Uniform refers to setting pixel values randomly from uniform distribution. Steps refers to the number of pixels masked at each iteration. For consistency, different correlation functions are used for different experiments. XAI Quality Score(XQS), the aggregated score is calculated by finding mean of mean, median values and weighted sum of mean, and median values of all the metrics. A total of 10 variations are calculated for each dataset, 2 experiments for each metric. All the experiments are compared with the baseline and analyzed in the following sections.

4.2 PCAM

This section presents the key findings for PCAM dataset like model performance and XAI metric evaluation. Due to the large number of experiments and results, significant results are discussed in this section. Additional and secondary results can be viewed in Appendix A.

TABLE 4.1: Overview of Experiments

Name	Continuity (PCAM/CIFAR)	Compactness	Correctness	Consistency	XQS
Baseline	Perturbation Function : Noise ($std : 0.05$) /Rotation($10°$). Mean& Median. Spearman.	Mean& Median	Perturbation Baseline: Black, Steps:112, Mean & Median.	Spearman. Mean & Median.	Mean, Median & Weighted-XQS.
One	Perturbation Function : Noise ($std : 0.01, 0.2$) / Rotation($5°, 15°$). Mean&Median. Spearman.	Same as Baseline.	Perturbation Baseline: (mean,uniform), Steps:112, Mean & Median.	Pearson & Kendall Tau. Mean & Median.	—
Two	Perturbation Function : Noise ($std : 0.05$) /Rotation($10°$). Mean&Median. Pearson &Kendall Tau.	Same as Baseline.	Perturbation Baseline: Black, Steps: $(56, 224)$, Mean & Median.	Same as Baseline.	—

4.2.1 Model Performance

Model performance is calculated on both test and validation sets. The early stopping is based on val metrics. As seen in the Figure 4.2, there is a notable improvement in all metrics over the cycles. With a rapid increase during first $5 - 10$ cycles, and stabilized later. However as shown in Figure 3.10 the AL method increasingly selects negative sampling. This results in class imbalance that causes fluctuations in recall and precision as the model encounter fewer positive samples in later cycles. Despite this the model achieved final test AUC-ROC of 94% (Figure 4.1) and final val AUC-ROC of 95%. This indicates strong generalization and no overfitting. Especially this high AUC-ROC and reasonable recall values are critical for medical images. It reflects robust performance across decision threshold and accurate classification of True Positives (Cancer) with a labeled set of $4,000$ images.

Comparison with Full Dataset Training

To compare AL with the complete dataset the same framework and parameters are used for both methods. PCAM full dataset contains 262,144 training , 32,768 validation , and 32,768 test images. While AL used only 40,000 training, 10,000 validation, and 10,000 test images and 4000 labeled samples at the end of AL cycle. As seen in the Figure 4.3, all metrics are nearly identical for both datasets. There is a

```
Classification Report:
              precision    recall  f1-score   support

    Negative       0.85      0.91      0.88      5002
    Positive       0.90      0.84      0.87      4998

    accuracy                           0.87     10000
   macro avg       0.87      0.87      0.87     10000
weighted avg       0.87      0.87      0.87     10000

AUC-ROC: 0.9400
```

FIGURE 4.1: Test Set Evaluation

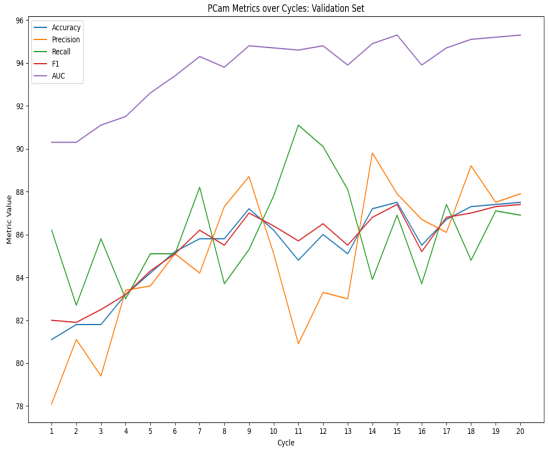


FIGURE 4.2: Val Set Evaluation

slight increase in AUC-ROC to 94.5% for complete dataset. The test set is balanced which shows the robustness of these results. This demonstrates the effectiveness of AL in achieving state-of-the-art performance with limited labeled samples that is valuable for medical imaging.

```
Classification Report:
              precision    recall  f1-score   support

    Negative       0.85      0.90      0.87     16391
    Positive       0.89      0.84      0.86     16377

    accuracy                           0.87     32768
   macro avg       0.87      0.87      0.87     32768
weighted avg       0.87      0.87      0.87     32768

AUC-ROC: 0.9449
```

FIGURE 4.3: Evaluation on Complete PCAM Dataset

Comparison with Baseline

Random sampling is used as a baseline to compare against entropy-based sampling. All parameters and framework were held constant with only AL sampling strategy varied. Random sampling represents basic form of AL, where the algorithm selects data points randomly from pool of unlabeled data. As seen in both the Figures 4.4, 4.5 it is clear that there is a minimal increase of approximately 3% in the evaluation of entropy from baseline as expected. However, this small margin states that pre-trained ResNet-34 contributes to the majority of performance by strong feature representation and classification.

```
Classification Report:
              precision    recall  f1-score   support

    Negative       0.80      0.88      0.84      5002
    Positive       0.87      0.80      0.83      4998

    accuracy                           0.84     10000
   macro avg       0.84      0.84      0.84     10000
weighted avg       0.84      0.84      0.84     10000

AUC-ROC: 0.9249
```

FIGURE 4.4: Test Set Evaluation

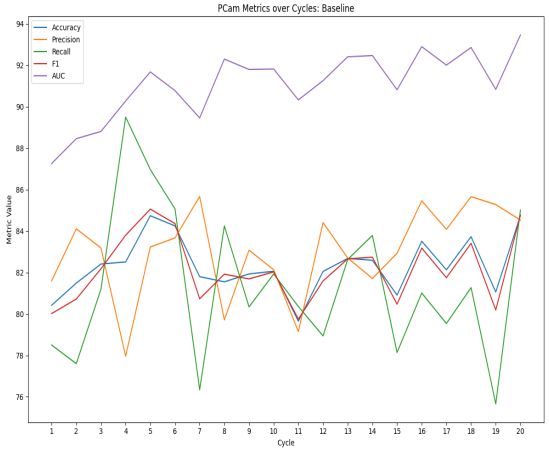


FIGURE 4.5: Val Set Evaluation

4.2.2 XAI Evaluation

This section evaluates the quality of XAI explanations using four metrics: continuity, compactness, correctness, and consistency. Baseline results for both SHAP and GradCAM are presented as detailed in Table 4.1. Continuity measures stability of explanations under minor perturbations, compactness quantifies the concentration of attribution in explanation, correctness measures the faithfulness when top features are masked, and consistency measures the similarity of explanations across AL cycles. These metrics were chosen because of their strong alignment with image data and AL framework.

The representative explanations for SHAP and GradCAM are illustrated in Figure 4.6, and 4.7. The green box at the center is the only place to have cancer cells as defined by dataset construction [52]. Both XAI methods concentrate their attribution within and around green box for cancerous image. While for normal image the attributions are located outside of the region as normal images do not have cancerous cells. These visualizations illustrate the regions that are considered important and align with dataset definition as cancer image has maximum attribution in center (green box) and non cancer image has maximum attribution in the background. In following subsections baseline metrics are presented and discussed.

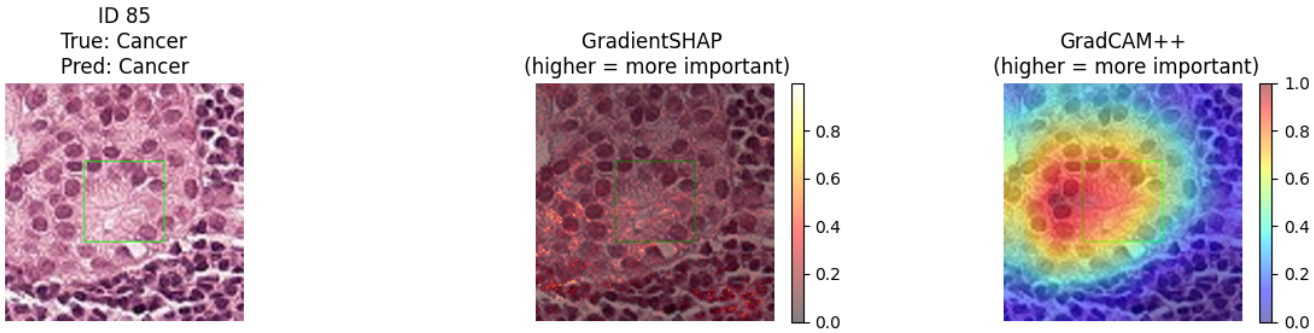


FIGURE 4.6: XAI Output: Cancer Image

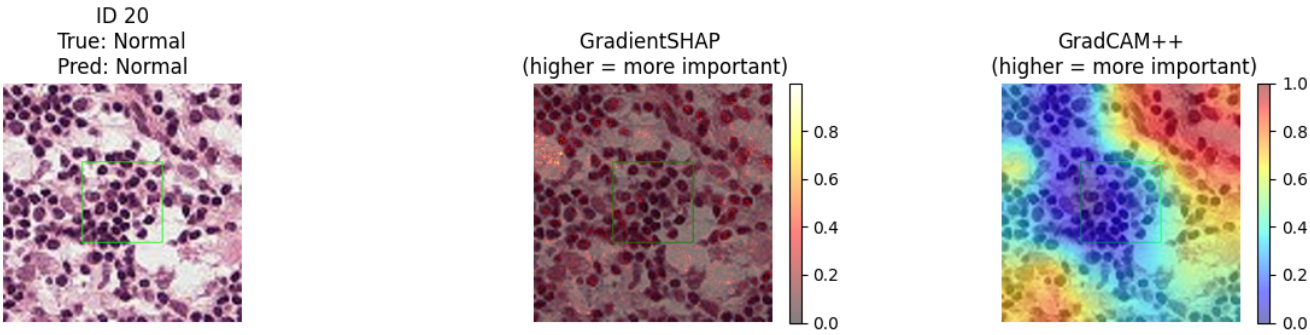


FIGURE 4.7: XAI Output: Normal Image

SHAP

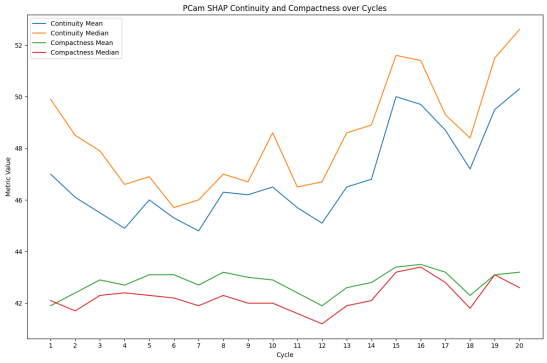


FIGURE 4.8: Continuity & Compactness

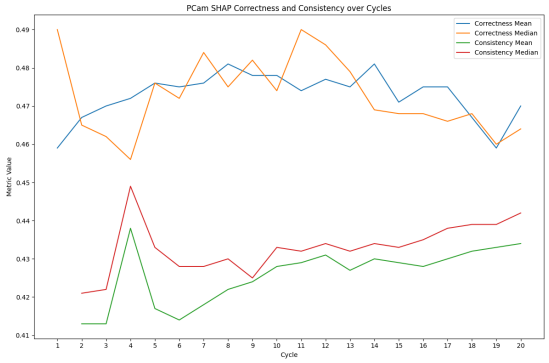


FIGURE 4.9: Correctness & Consistency

Continuity: LocalLipschitzEstimate from Quantus was used to calculate the similarity between original and gaussian noise perturbed images. Continuity scores are ranged from $0.45 - 0.50$ across cycles. This indicates moderate stability under input perturbations. This low values compared to GradCAM reflects the SHAP pixel-level sensitivity that any small perturbations can notably alter the pixel importance. This is an expected behavior for gradient based attribution like SHAP in medical imaging [22]. As per Figure 4.8 final cycle (20) shows peak continuity which suggests explanation is stabilized as model is exposed to more labeled data.

Compactness: Sparseness from Quantus was used to quantify the concentration of attribution in explanations. Figure 4.8 shows that compactness has a consistent high scores around $0.42 - 0.43$ across the cycles. This indicates that SHAP has a well localized attributions that focused on relevant regions of image. The stable range of values suggests that SHAP has a robust feature importance ranking. This property is crucial for medical imaging as it is important to highlight anatomical regions consistently.

Correctness: PixelFlipping from Quantus was used to quantify the faithfulness of explanations by masking important pixels. Correctness shows a gradual improvement of $0.45 - 0.48$ (Figure 4.9) across the cycles. This indicates the increasing

faithfulness as the model learn the features. This upward trend is expected in AL as model learns more from informative samples, SHAP attribution gets aligned with expected outputs. Improvement in correctness is valuable in medical imaging as explanations become more reliable to identify important regions.

Consistency: It is calculated by measuring similarity between explanations for every image across all cycles. As seen in Figure 4.9, consistency for cycle 1 has no value since there are no prior images to compare. Consistency values increase from $0.41 - 0.43$, with major changes occurring between cycle $5 - 10$. Gradual increase of consistency suggests that SHAP explanations are getting increasingly stable across AL cycles. This metric is useful as consistent explanations are important in medical imaging as stability of explanations is very important.

As seen in both Figures 4.8, 4.9, mean and median values for each metric exhibit minimal divergence. This indicated that SHAP attributions follow a near normal distribution without major outliers. The largest mean-median difference is in continuity with $\Delta = 0.02$ which is minimal. This indicates that SHAP's performance is robust and stable with respect to choice of aggregation method, and results are not influenced by complex or outlier cases.

GRADCAM

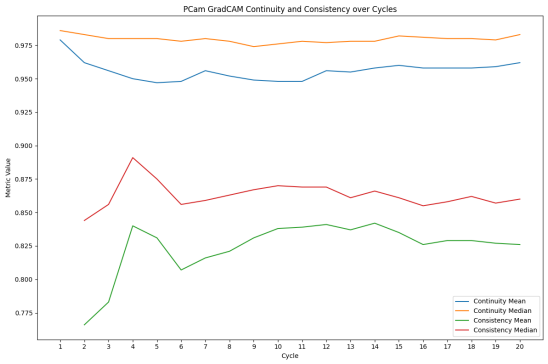


FIGURE 4.10: Continuity & Consistency

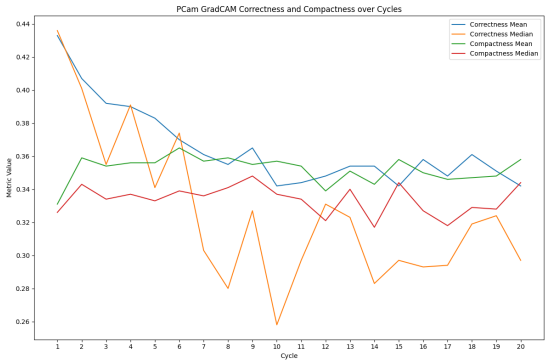


FIGURE 4.11: Correctness & Compactness

Continuity: As seen in Figure 4.10, GradCAM achieves a high continuity scores ranging from $0.94 - 0.98$. This higher stability reflects GradCAM region based attribution which produce consistent spatial localizations. The minimal fluctuation indicates reliable explanation patterns important for medical imaging.

Compactness: As seen in Figure 4.11, compactness scores remain constant and moderate around $0.33 - 0.37$. This indicates that explanations are focused enough for interpretability, but also not so sparse as to miss important information. This is an expected behavior of GradCAM as it produces coarse and low resolution heatmaps.

Correctness: As seen in Figure 4.11, correctness scores range from $0.28 - 0.43$ with notable fluctuations and a declining trend from early to later cycles. This suggests GradCAM identifies important regions initially although its faithfulness decreased with the addition of newer data points.

Consistency: As seen in Figure 4.10, value scores are pretty high and constant from $0.82 - 0.84$ for all the cycles. This suggests that explanations become highly stabilized once the model develops basic feature representations.

As seen in both Figures 4.10, 4.11, all the metrics have minimal divergence between mean and median except for correctness. The largest mean-median difference is $\Delta = 0.09$. This difference suggests that GradCAM is a little sensitive to outliers, reflecting its region based approach facing difficult spatial patterns in certain images.

4.2.3 XAI Quality Score (XQS)

The XAI Quality Score (XQS) is an aggregated metric designed to provide assessment of explanation quality by combining the XAI metrics used in this thesis. This enables a direct comparison of various explainability methods across AL cycles. The types of aggregation methods used are,

- **Mean & Median XQS:** The arithmetic mean of mean/median values of metrics.

$$\text{XQS} = \frac{\text{Continuity} + \text{Compactness} + \text{Correctness} + \text{Consistency}}{4} \tag{4.1}$$

- **Weighted Mean & Median XQS:** Each metric is multiplied by a specific weight reflecting its importance, and the results are summed. Table 4.2 provides the weights.

$$\text{Weighted XQS} = w_1 m_1 + w_2 m_2 + w_3 m_3 + w_4 m_3 \tag{4.2}$$

TABLE 4.2: Weights for weighted XQS calculation.

Method	m_1	m_2	m_3	m_4
SHAP	0.15 (w_1)	0.15 (w_2)	0.40 (w_3)	0.30 (w_4)
GradCAM	0.25 (w_1)	0.20 (w_2)	0.30 (w_3)	0.25 (w_4)

m_1 : *Continuty*, m_2 : *Compactness*, m_3 : *Correctness*, m_4 : *Consistency*

We carefully selected the weights for each metric in weighted XQS (Table 4.2) to reflect the strengths of each XAI methods and need for medical imaging. SHAP places high emphasis on correctness and consistency to prioritize faithful and reproducible explanations for clinical trust. GradCAM uses a more balanced scheme with higher weights on continuity and compactness to capture its region-based interpretability. All weights are normalized to sum to 1 for comparability. While there is no universal gold standard for such weights, their selection is guided by both established practices and empirical observations. This ensures that evaluation focuses on most critical aspects for generating explanations that can be trusted in medical contexts.

SHAP

As consistency does not have value for cycle 1, XQS for first cycle is calculated with three metrics and weighted XQS is set to none.

As seen in Figure 4.12, XQS methods exhibit a small initial decline from cycle 1 (0.449) to cycle 2(0.441). This represents an approximate decrease of 2% in the explanation quality as model does not calculate consistency. In later cycles, XQS stabilizes around $0.44 - 0.46$ with gradual improvements. It reaches final values of 0.460 (mean), 0.465 (median), 0.462 (weighted mean), and 0.461 (weighted median) by cycle 20. The minimal divergence between mean and median scores ($\Delta = 0.01$) indicates robust explanation quality and minimal outliers. Although the close alignment between weighted and unweighted aggregation metrics provides no significant changes in overall assessments. This pattern reflects the increase of XQS and explainability of SHAP with increasing AL cycles.

GRADCAM

As seen in Figure 4.13, metrics show a jump from cycle 1 (0.581) to cycle 2 (0.624). This is an increase of approximately 7.4% in explanation quality. From cycle 3 to 20 the values stabilize around $0.62 - 0.63$ with minimal fluctuation. The final value converge to 0.622 (mean), 0.621 (median), 0.621 (weighted mean), and 0.619 (weighted median). Similar to SHAP, GradCAM also exhibit a minimal diverge between mean and median. The close alignment between weighted and unweighted scores indicates domain and method specific weighting schemes does not substantially transform overall assessment for the dataset.

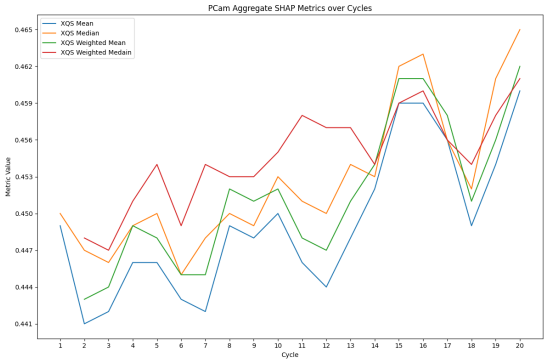


FIGURE 4.12: SHAP Aggregated Metrics

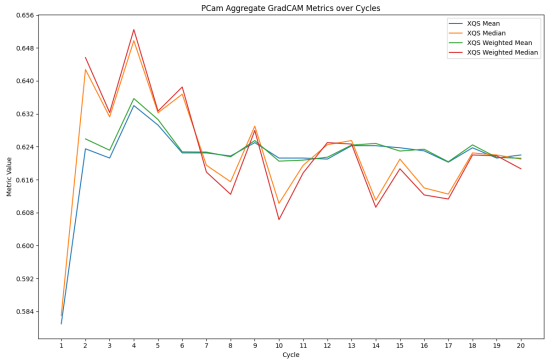


FIGURE 4.13: GradCAM Aggregated Metrics

4.2.4 Correlation Analysis

As seen in the Figure 4.14, SHAP correlation analysis reveals that XQS has high positive correlation with model performance metrics. Accuracy ($r = 0.780, p = 0.0001$), F1-score ($r = 0.730, p = 0.0003$), and AUC ($r = 0.736, p = 0.0002$). This indicates that as model becomes more accurate, SHAP explanations improves substantially. There is strong positive correlation between XQS and continuity ($r = 0.779, p =$

0.0001), and compactness ($r = 0.683, p = 0.0009$). This suggests spatial smooth and focused SHAP attribution maps directly contribute to XQS. Consistency metric also shows positive correlation with XQS ($r = 0.698, p = 0.0006$) which shows importance of stable explanations. Detail findings are provided in Appendix Table A.2.

As seen in the Figure 4.15, GradCAM correlation analysis reveals the strong negative correlation between model performance metrics and correctness. Such as accuracy, correctness ($r = \neg 0.71, p = 0.0004$) and F1, correctness ($r = \neg 0.76, p = 0.0001$). This reflects tradeoff where explanation faithfulness decrease with model improvement. Consistency is slightly negatively correlated with continuity ($r = \neg 0.57, p = 0.0092$). Continuity and compactness generally show weak correlation with XQS, and XQS does not depend strongly on a single XAI metric. This highlight the complexity of XAI evaluation with GradCAM. Detailed findings are provided in Appendix Table A.1.

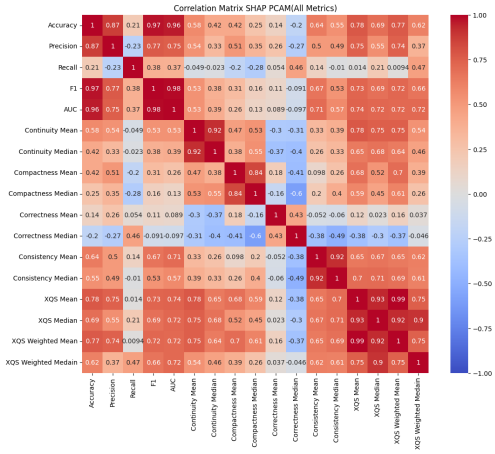


FIGURE 4.14: SHAP Correlation Matrix

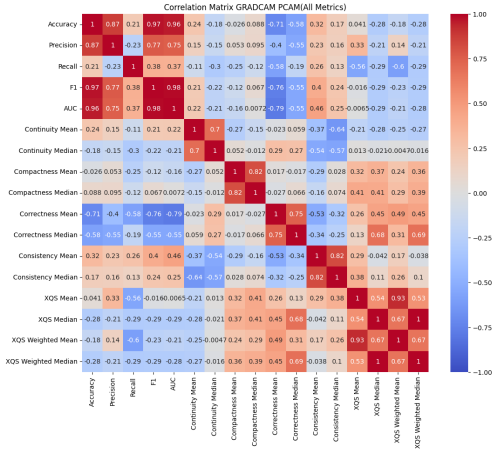


FIGURE 4.15: GradCAM Correlation Matrix

4.2.5 Per-Image Analysis

Until now the analysis is on aggregated metrics across 100 validation images at each cycle resulting in a single score for each metric and each XAI method. This subsection analyses the distribution of scores for each individual image to reveal the variability that is masked by aggregation. Per image distributions from Appendix Tables A.3,A.4 reveals that GradCAM produces highly smooth and generally stable explanations with moderate compactness and variability in correctness. This indicates that most images are well explained with a notable subset of poor attributions. In contrast, SHAP explanations are less smooth, less stable and have a overall low scores. Yet, they show a higher average correctness with good spread. This reflects that some images are explained very faithfully while many are not. These distributions highlight the diversity of XAI performance across images.

Outlier Analysis

In this subsection, outliers are identified and explained for both SHAP and GradCAM. Outliers are determined by ranking all images according to each XAI metric and XQS. Then selecting top and bottom five images with scores and occurrence in number of cycles. One representative image with consistent entry in top and bottom groups is selected for detailed explanation. Full distribution of metric values are visualized in Appendix Figures A.1, A.2 offering additional context.

SHAP and GradCAM showcase a distinct pattern in their top and bottom outlier images which can be seen in Figures 4.16, 4.17. The highest attribution region for both top outliers is not in the cancerous green box as annotated but is in the background tissue. This suggest despite the accurate model prediction, XAI explanation may highlight contextual feature rather than relevant feature. In contrast bottom outliers for both methods show diffused and low intensity attributions that are scattered across the image. This diffused pattern in normal samples reflects absence of distinctive features. These findings suggests that high explanation scores does not guarantee medically relevant attributions. While, low scores in normal samples may simply reflect the lack of distinctive medical features rather than failure of method.

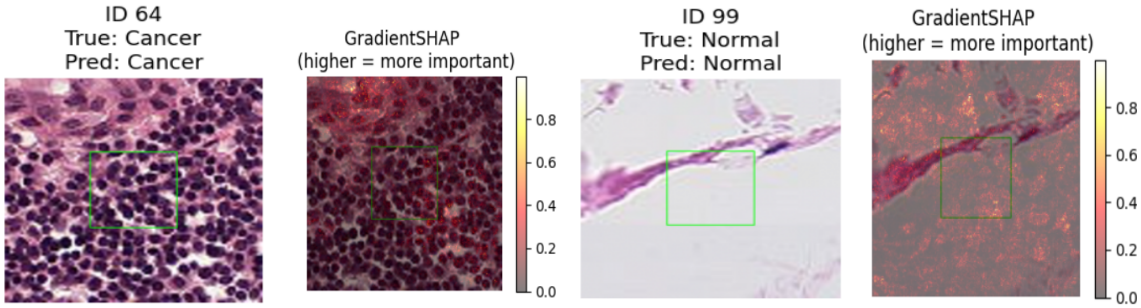


FIGURE 4.16: Top and bottom outliers from SHAP

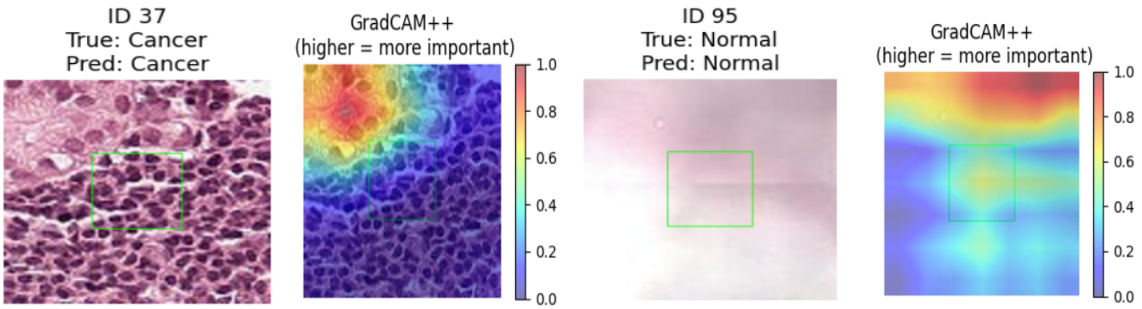


FIGURE 4.17: Top and bottom outliers from GradCAM

4.2.6 Sensitivity Analysis

Sensitivity analysis is conducted by systematically changing the key parameters for XAI metrics across 10 experiments which are mentioned in Table 4.1 over AL cycles. The main results are summarized in Table 4.3. The affect of these changes are quantified by Mean Absolute Difference (MAD) and Root Mean Square Error (RMSE). Detailed results and visualizations for each experiments are provided in Appendix A.3, A.5, A.6, A.7.

As seen in Table 4.3, both SHAP and GradCAM are highly robust to typical noise and step size changes by showing minimal impact on continuity and correctness. However both methods, especially SHAP is sensitive to disruptive shifts like correlation and baseline perturbations. Here explanation stability and faithfulness can degrade significantly. These findings underscore that while both XAI methods are reliable under standard conditions, their reliability can be diminished by more disruptive shifts.

Table 4.3: Sensitivity Analysis Results

Experiment	**Change**	**SHAP**	**GradCAM**
Continuity:1	gaussian noise:0.01	MAD: 0.0012, RMSE: 0.0017, (Highly Robust)	MAD: 0.0015, RMSE: 0.0018, (Highly Robust)
Continuity:2	gaussian noise:0.2	MAD: 0.0015, RMSE: 0.0017, (Highly Robust)	MAD: 0.0015, RMSE: 0.0019, (Highly Robust)
Continuity:3	corr:Pearson	MAD: 0.0741, RMSE: 0.0742, (Sensitive)	MAD: 0.0069, RMSE: 0.0074, (Robust)
Continuity:4	corr:kendal tau	MAD: 0.1410, RMSE: 0.1411, (Highly Sensitive)	MAD: 0.1011, RMSE: 0.1012, (Highly Sensitive)
Correctness:1	steps:56	MAD: 0.0009, RMSE: 0.0013, (Highly Robust)	MAD: 0.0002, RMSE: 0.0004, (Highly Robust)
Correctness:2	steps:224	MAD: 0.0008, RMSE: 0.0010, (Highly Robust)	MAD: 0.0006, RMSE: 0.0007 (Highly Robust)
Correctness:3	baseline:mean	MAD: 0.1156, RMSE: 0.1176, (Highly Sensitive)	MAD: 0.0075, RMSE: 0.0087, (Robust)
Correctness:4	baseline:uniform	MAD: 0.0652, RMSE: 0.0695, (Sensitive)	MAD: 0.0098, RMSE: 0.0119, (Robust)
Consistency:1	corr:Pearson	MAD: 0.0833, RMSE: 0.0857, (Sensitive)	MAD: 0.0227, RMSE: 0.0234, (Sensitive)
Consistency:2	corr:kendal tau	MAD: 0.1236, RMSE: 0.1268, (Highly Sensitive)	MAD: 0.1491, RMSE: 0.1531, (Highly Sensitive)

4.3 CIFAR-10

This section presents the key findings for CIFAR-10 dataset including model performance with AL and XAI metric evaluation for both XAI methods. All the experiments, methods, algorithms, and parameters used for PCAM are applied identically to CIFAR-10 dataset to ensure a direct and fair comparison. To avoid redundancy, detailed descriptions are not repeated in this section. Please refer to PCAM results section for comprehensive experimental detail.

4.3.1 Model Performance

Model performance on CIFAR-10 improved rapidly during the first few AL cycles. Accuracy, and F1-score overlap and settled above 92% by cycle 10 as seen in the Figure 4.19. The final test results (see Figure 4.18) support this finding as overall accuracy reaches 92% and most classes show above 90% F1-score. The exceptions are "cat" and "dog" which are challenging to differentiate in this dataset. These results highlight the efficiency and reliability of the AL setup with strong generalization.

Classification Report:

	precision	recall	f1-score	support
airplane	0.94	0.91	0.92	1000
automobile	0.95	0.96	0.95	1000
bird	0.92	0.89	0.90	1000
cat	0.86	0.85	0.85	1000
deer	0.90	0.95	0.92	1000
dog	0.89	0.86	0.88	1000
frog	0.89	0.99	0.94	1000
horse	0.98	0.89	0.94	1000
ship	0.94	0.97	0.96	1000
truck	0.95	0.94	0.95	1000
accuracy			0.92	10000
macro avg	0.92	0.92	0.92	10000
weighted avg	0.92	0.92	0.92	10000

FIGURE 4.18: Test Set Evaluation

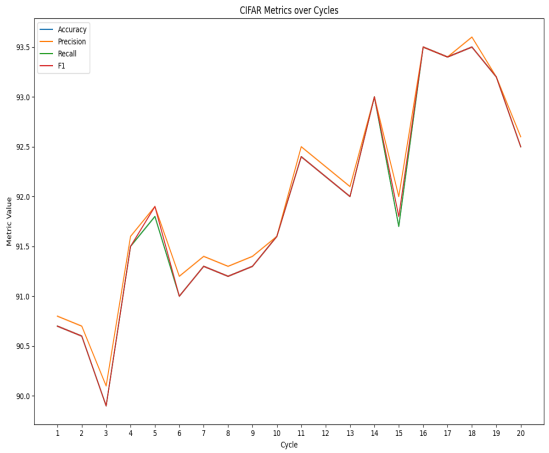


FIGURE 4.19: Val Set Evaluation

Comparison with Full Dataset Training

CIFAR-10 complete dataset includes 40,000 training, 10,000 validation, and 10,000 test images. AL was conducted using the same amount of test and validation sets but 2000 labeled samples with 100 more at each cycle. As seen in the Figure 4.20, an overall accuracy and F1-score of 97% is achieved along with all metrics for every class(bird, cat, etc.) ranging from $93\% - 99\%$. Especially the results from AL closely match with full dataset with minor differences in certain classes. This highlights the efficiency of AL with far fewer labeled samples that proves its practical value for large-scale image classification tasks.

Classification Report:

	precision	recall	f1-score	support
airplane	0.97	0.97	0.97	1000
automobile	0.98	0.97	0.98	1000
bird	0.97	0.96	0.97	1000
cat	0.93	0.93	0.93	1000
deer	0.97	0.97	0.97	1000
dog	0.95	0.94	0.94	1000
frog	0.99	0.99	0.99	1000
horse	0.98	0.98	0.98	1000
ship	0.97	0.98	0.98	1000
truck	0.97	0.97	0.97	1000
accuracy			0.97	10000
macro avg	0.97	0.97	0.97	10000
weighted avg	0.97	0.97	0.97	10000

FIGURE 4.20: Evaluation on Complete CIFAR-10 Dataset

Comparison with Baseline

Random sampling is used as a baseline to compare against entropy-based sampling. All parameters and framework are held constant with only AL sampling strategy varied. As seen in both the Figures 4.21, 4.22 it is clear that there is a minimal increase of approximately 2% in the evaluation of entropy from baseline as expected. However, this small margin states that pre-trained ResNet-34 contributes to the majority of performance by strong feature representation and classification.

Classification Report:

	precision	recall	f1-score	support
airplane	0.92	0.90	0.91	1000
automobile	0.95	0.94	0.95	1000
bird	0.94	0.82	0.88	1000
cat	0.77	0.87	0.81	1000
deer	0.88	0.94	0.91	1000
dog	0.87	0.84	0.86	1000
frog	0.95	0.92	0.93	1000
horse	0.91	0.92	0.91	1000
ship	0.94	0.95	0.94	1000
truck	0.93	0.93	0.93	1000
accuracy			0.90	10000
macro avg	0.91	0.90	0.90	10000
weighted avg	0.91	0.90	0.90	10000

FIGURE 4.21: Test Set Evaluation

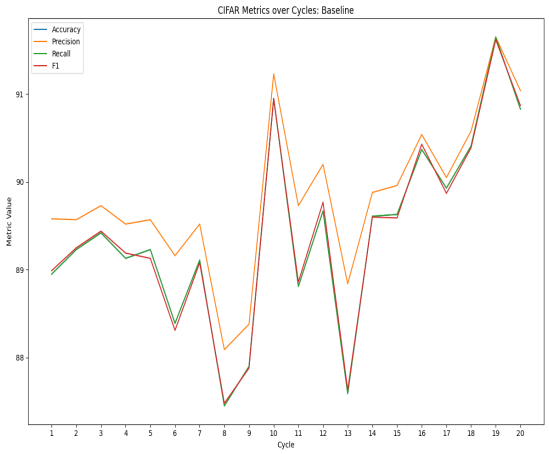


FIGURE 4.22: Val Set Evaluation

4.3.2 XAI Evaluation

This section examines the quality of XAI explanations for CIFAR-10 focusing on continuity, compactness, correctness, and consistency. Figures 4.23, 4.24 illustrate a typical explanation of image for both "cat" and "dog" images. In both the images SHAP and GradCAM tend to highlight facial features such as eyes and ears. Yet, attributions are concentrated on snout for the "cat", and head for the "dog". These visualizations help to identify what model deems important for classification. The following subsections present and discuss the evolution of XAI evaluation metrics over AL cycles.

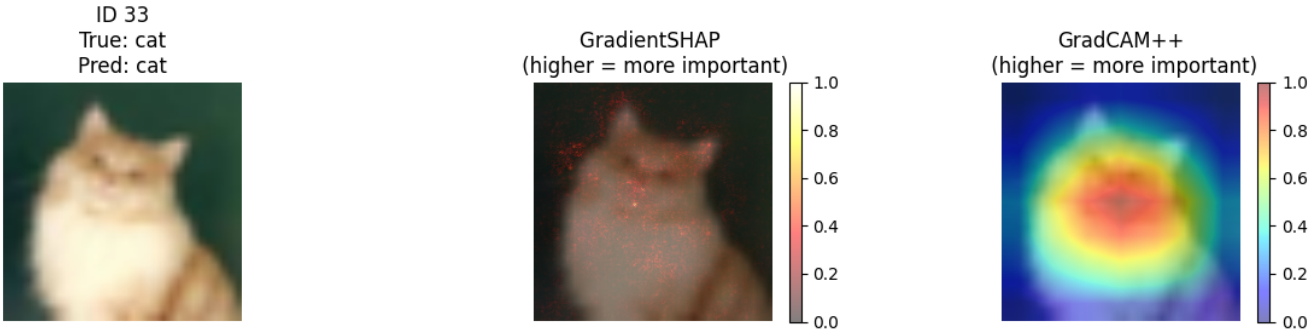


FIGURE 4.23: XAI Output: Cat Image

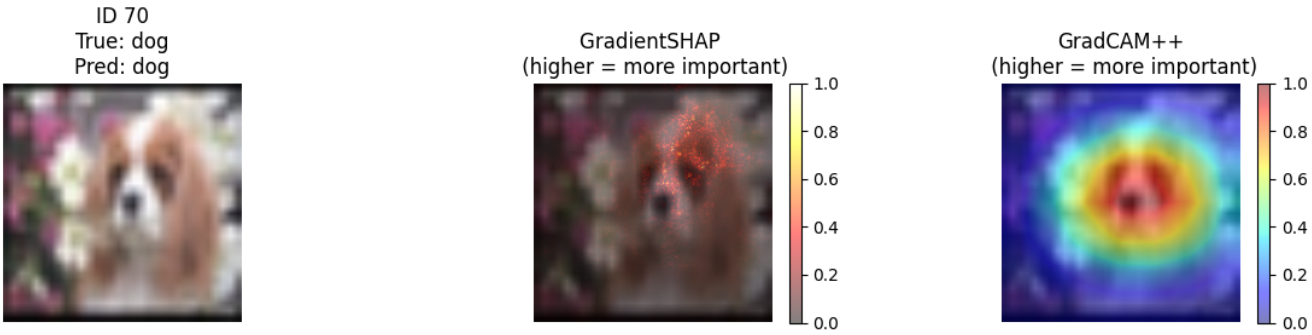


FIGURE 4.24: XAI Output: Dog Image

SHAP

Continuity: As seen in Figure 4.25, continuity scores remain consistently low across all cycle around $0.315 - 0.378$. They peak at cycle 10 but declines at the end. This points to a moderate sensitivity to small input perturbations with little gains as more data is labeled.

Compactness: As seen in Figure 4.26, compactness scores are very consistent with little fluctuations with the values ranging from $0.477 - 0.497$. The peak can be

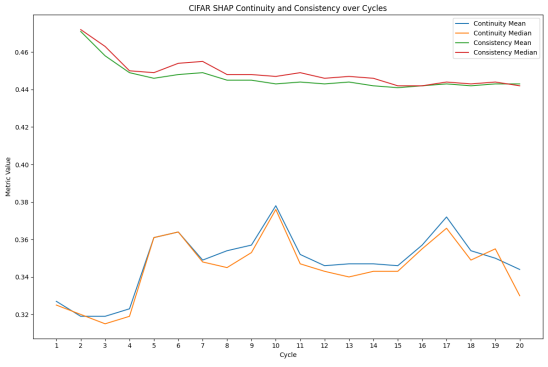


FIGURE 4.25: Continuity & Consistency

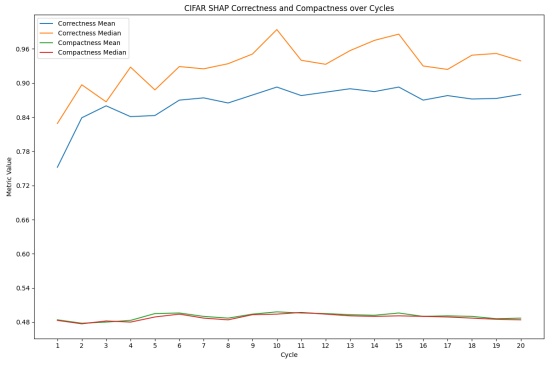


FIGURE 4.26: Correctness & Compactness

observed at cycle 10. This suggests that SHAP attributions maintain a consistent concentration and spread without much shift as training progresses.

Correctness: As seen in Figure 4.26, correctness demonstrates the highest improvement arc starting from $0.752 - 0.898$. It peaks at cycle 10 and remain consistent after that. This suggests that explanations become more faithful to model prediction as AL process unfolds.

Consistency: As seen in Figure 4.25, consistency scores are almost similar across cycles with little fluctuation ranging from $0.443 - 0.477$. This suggests that SHAP explanations become repeatable from cycle to cycle with minimal amount of labeled data.

As seen in both figures, mean and median values for each metric exhibit minimal divergence. This indicates that SHAP attributions follow a near normal distribution without major outliers. The largest mean-median difference is in correctness with $\Delta = 0.09$ which suggest the presence of minimal outliers. This stability between both aggregation methods indicates robustness of SHAP's performance.

GRADCAM

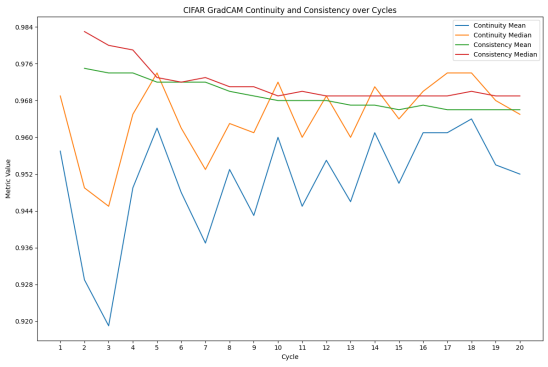


FIGURE 4.27: Continuity & consistency

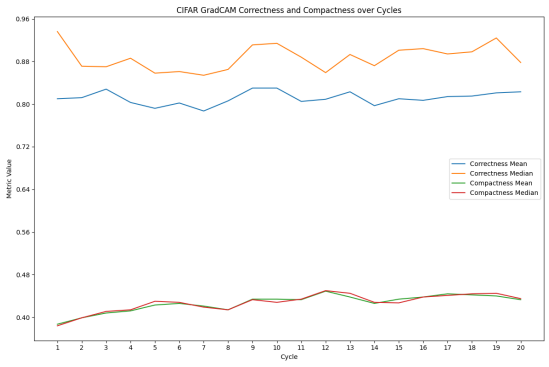


FIGURE 4.28: Correctness & Compactness

Continuity: As seen in Figure 4.27, continuity scores are consistently high ranging from $0.929 - 0.974$ across cycles. This suggests that GradCAM explanations remains stable under minor perturbations reflecting its robust explanation quality outperforming SHAP.

Compactness: As seen in Figure 4.28, compactness values remain moderate and stable ranging from $0.387 - 0.449$. Gradual increase from early to mid cycles suggests that GradCAM's heatmaps become more focused as models evolve over cycles. But, the overall spatial concentration remains fairly consistent.

Correctness: As seen in Figure 4.28, correctness scores are high and variable ranging from $0.787 - 0.830$. This fluctuation suggests that even with the high scores GradCAM's faithfulness to model's prediction does not consistently improve unlike SHAP's clear upward trend.

Consistency: As seen in Figure 4.27, consistency scores are very high and stable ranging from $0.966 - 0.975$. This suggests that GradCAM explanations are highly repeatable from one cycle to the next.

As shown in both figures, consistency and compactness have minimal divergence between mean and median when compared to correctness, and continuity. The largest mean-median difference for correctness is $\Delta = 0.13$ and continuity is $\Delta = 0.03$. This pattern suggests that correctness distributions are skewed than other metrics indicating mild outliers.

4.3.3 XAI Quality Score (XQS)

The XAI Quality Score provide a comprehensive assessment of explanation quality by aggregating the four XAI metrics into a single score. Both XQS, and Weighted XQS are employed as detailed in the Section 4.2.3 for PCAM dataset. The mathematical formulas (see Formulas: 4.1, 4.2) and XQS weights (see Table 4.2) remains identical for CIFAR-10 enabling direct comparison.

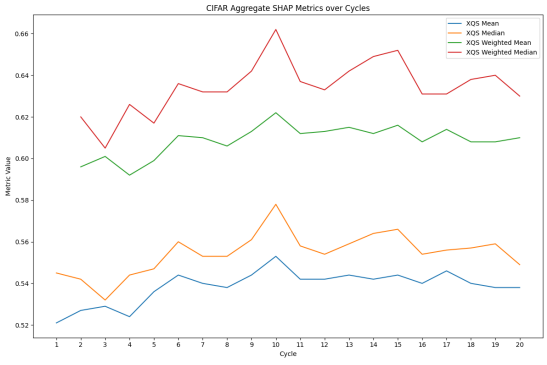


FIGURE 4.29: SHAP Aggregated Metrics

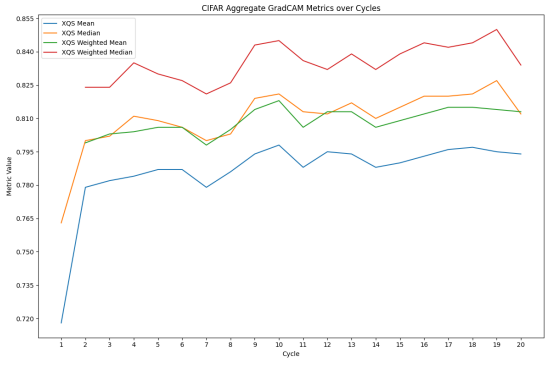


FIGURE 4.30: GradCAM Aggregated Metrics

SHAP

As seen in the Figure 4.29, XQS values demonstrate a clear two-phase pattern. Rapid movement from cycle 1 to 10 where mean score raises from $0.521 - 0.553$ followed by stabilizing at 0.54. Weighted XQS scores consistently outperforms unweighted scores by approximately 0.07 on an average, reflecting the emphasis of correctness in weight scheme. The minimal divergence of mean-median (unweighted $\Delta = 0.016$, weighted $\Delta = 0.026$) indicates robust explanation quality and minimal outliers. This suggests that SHAP explanations increase during initial AL cycles, and then reaches a stable plateau. Additional labeled data yields diminishing returns for explainability. This highlights the efficiency of AL for both predictive performance and explanation consistency.

GRADCAM

As seen in the Figure 4.30, XQS values exhibit a dramatic jump from cycle 1 (0.718) to cycle 2 (0.779). It maintains a consistent high performance around $0.79 - 0.80$ for remaining cycles. Weighted XQS scores consistently outperforms unweighted scores by approximately 0.019 on an average, reflecting the emphasis of continuity in weight scheme. The minimal divergence of mean-median (unweighted $\Delta = 0.023$, weighted $\Delta = 0.027$) indicates robust explanation quality and minimal outliers. This suggests that GradCAM explanations quickly reach good explanations with minimal labeled data and maintain it consistently.

4.3.4 Correlation Analysis

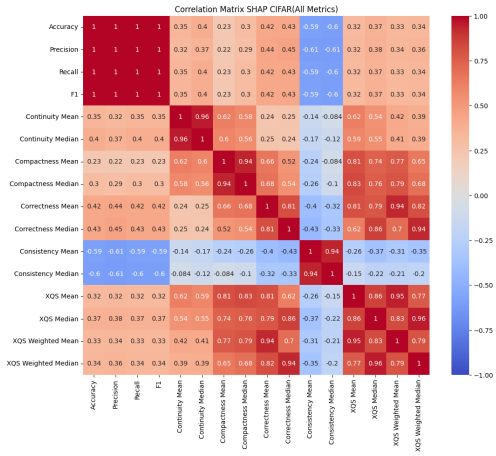


FIGURE 4.31: SHAP Correlation Matrix

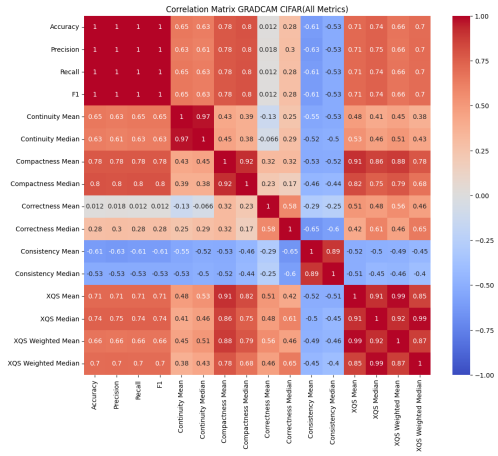


FIGURE 4.32: GradCAM Correlation Matrix

As seen in the Figure 4.31, SHAP correlation analysis reveals weak to moderate correlations between model performance and XAI metrics. Accuracy shows moderate correlation with continuity ($r = 0.351, p = 0.129$) and correctness ($r = 0.423, p = 0.063$). The strongest relationships is between XAI metrics themselves

mainly compactness and correctness ($r = 0.660, p = 0.002$). Interestingly consistency stands out with a clear negative correlation to model performance metrics ($r = \neg 0.593\, to\, \neg 0.613, p < 0.01$). This suggests that when accuracy increases, explanation stability tend to drop. This pattern indicates that better predictions do not always mean more reliable explanations for SHAP. These findings are provided in detail in Appendix A.14.

As seen in the Figure 4.32, GradCAM correlation analysis reveals strong positive correlations between model performance and certain XAI metrics. In particular model performance with continuity ($r = 0.776 - 0.804, p < 0.001$), and XQS Scores ($r = 0.665 - 0.743, p < 0.01$). Correctness shows minimal correlation with accuracy ($r = 0.012, p = 0.961$), while consistency shows notable negative correlation ($r = \neg 0.614\, to\, \neg 0.526, p < 0.05$). The strong correlation between compactness and accuracy highlights that focused attributions are tend to yield better classifications. Unlike SHAP, GradCAM's XQS scores maintain robust positive correlation with model performance. This indicates that explanation quality and model performance are more aligned. These findings are provided in detail in Appendix A.13.

4.3.5 Per-Image Analysis

Per image distributions from Appendix Tables A.8, A.9 reveals that GradCAM shows consistently strong performance with tightly grouped continuity scores and high consistency scores. This indicates robust stability and reliable explanations. Compactness is less variable and correctness shows a wider spread. This suggests faithful attributions for most images but degraded quality in complex cases. In contrast, SHAP shows a variable behavior with lower continuity and more spread. It achieves slightly higher correctness with more variability. The consistency differences are clearly visible. GradCAM maintains stable explanations while SHAP shows higher variability across the cycles. These distributions highlight the diversity of XAI performance across images. Full distribution of metric values are visualized in Appendix Figures A.4, A.5 offering additional context.

Outlier Analysis

The outlier analysis for CIFAR-10 reveals a distinctive pattern between top and bottom images as seen in Figures 4.33, 4.34. Top outlier in SHAP produces focused, high intensity attribution on specific object features. GradCAM generates a more concentrated heatmap covering the main image areas. Bottom outliers expose the failures of each methods. SHAP exhibits a scattered, low intensity attributions. GradCAM display weak, unfocused heatmap providing minimal interpretability. These contrasts demonstrate that methods can achieve superior explanation quality under optimal conditions, and they exhibit different failures. In failures SHAP showcases fragmented and noisy attributions, GRADCAM produces general and uninformative heatmaps. This pattern suggests that explanation quality improves overall with AL, yet some images remain challenging for consistent interpretability.

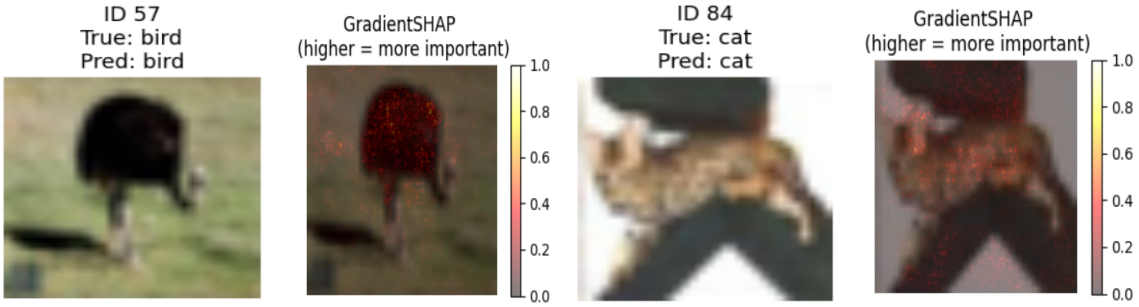


FIGURE 4.33: Top and bottom outliers from SHAP

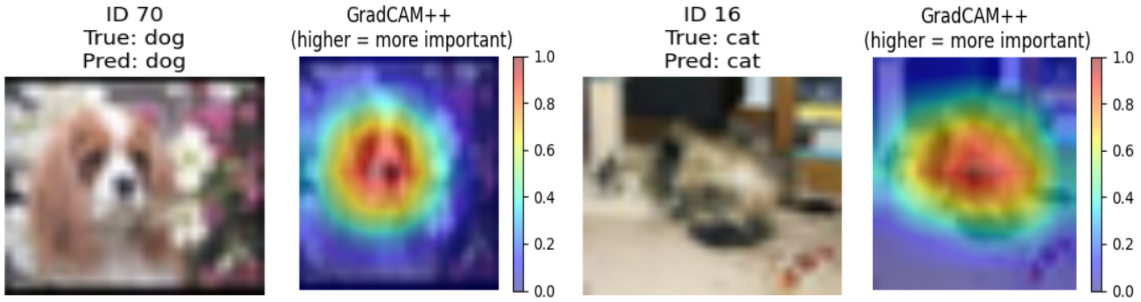


FIGURE 4.34: Top and bottom outliers from GradCAM

4.3.6 Sensitivity Analysis

We conducted sensitivity analysis by systematically changing the key parameters for XAI metrics across 10 experiments which are mentioned in Table 4.1. The main results are summarized in Table 4.4. Detailed results and visualizations for each experiments are provided in Appendix A.6, A.11, A.12, A.10.

As seen in Table 4.4, both SHAP and GradCAM are highly robust to step size variations of correctness. However both methods, especially SHAP are sensitive to correlation and baseline perturbations. Here, explanation stability can degrade significantly. These findings underscore that both XAI methods are reliable under standard conditions. Their robustness is challenged by more disruptive correlation changes and baseline shifts.

Table 4.4: Sensitivity Analysis Results

Experiment	**Change**	**SHAP**	**GradCAM**
Continuity:1	Image rotation:5°	MAD:0.0449, RMSE: 0.0452 (Sensitive)	MAD: 0.0106, RMSE: 0.0110 (Robust)
Continuity:2	Image rotation:15°	MAD: 0.0292, RMSE: 0.0295 (Sensitive)	MAD: 0.0177, RMSE: 0.0180 (Sensitive)
Continuity:3	corr:Pearson	MAD: 0.0784, RMSE: 0.0792 (Highly Sensitive)	MAD: 0.0019, RMSE: 0.0024 (Robust)
Continuity:4	corr:kendal tau	MAD: 0.1115, RMSE: 0.1116 (Highly Sensitive)	MAD: 0.1256, RMSE: 0.1259 (Highly Sensitive)
Correctness:1	steps:56	MAD: 0.0011, RMSE: 0.0016 (Highly Robust)	MAD: 0.0008, RMSE: 0.0009 (Highly Robust)
Correctness:2	steps:224	MAD: 0.0015, RMSE: 0.0017 (Highly Robust)	MAD: 0.0012, RMSE: 0.0013 (Highly Robust)
Correctness:3	baseline:mean	MAD: 0.0780, RMSE: 0.0820 (Highly Sensitive)	MAD: 0.0498, RMSE: 0.0525 (Sensitive)
Correctness:4	baseline:uniform	MAD: 0.0428, RMSE: 0.0456 (Sensitive)	MAD: 0.0371, RMSE: 0.0448 (Sensitive)
Consistency:1	corr:Pearson	MAD: 0.0203, RMSE: 0.0229 (Sensitive)	MAD: 0.0056, RMSE: 0.0062 (Robust)
Consistency:2	corr:kendal tau	MAD: 0.1328, RMSE: 0.1362 (Highly Sensitive)	MAD: 0.0951, RMSE: 0.0976 (Highly Sensitive)

Chapter 5

Discussion

This chapter discuss the results presented in previous chapter. It begins with the interpretation of results and its implications, then limitations of the methodology used and finally, the future directions of the methodology is discussed.

5.1 Interpretation of Results

5.1.1 Active Learning

The results from previous section confirm that active learning demonstrates the expected increase in performance on both the datasets as number of labeled points increases. The first ten cycles of AL on both the datasets show a rapid increase in the performance and plateaus in the later part of learning. While PCAM model achieves performance similar to training on the complete dataset, CIFAR-10 slightly under-performs compared to its full dataset. The results underscore AL efficiency which requires far fewer samples. This is crucial for the medical domain where labeling is expensive.

Entropy-based uncertainty sampling is used in the methodology, as it is a commonly used effective strategy for selecting informative samples. In both datasets, it repeatedly selects samples from certain classes in every cycle. This causes high class imbalance even if a balanced labeled dataset is used initially. Class imbalance can impact safety and reliability in medical imaging. It can also affect the stability and faithfulness of the explanations analyzed in later sections. This should be taken into consideration during training, and necessary changes should be made for better performance.

Active learning becomes computationally expensive with an increase in the samples to be labeled and the resolution of images. Although both the datasets start and end with same number of labeled data points, cycles for PCAM are significantly slower than CIFAR-10 as PCAM images are of higher resolution to begin with. In summary, it is clear that model performance is directly proportional to the amount of labeled data and class imbalance can have implications for XAI evaluations. These results set a foundation for understanding how explanation quality evolves along with performance and data distribution changes.

5.1.2 XAI Evaluation Metrics

The selection of three methods from Quantus toolkit to implement the XAI evaluation metrics is an approach to implement proxy metrics. While there are various other toolkits and methods available to calculate the same metrics. The calculation methods mentioned are used as they align with metric definition and dataset type (image), specific to this thesis. The four evaluation metrics are selected from CO-12 framework because of their orientation with AL, dataset type, and goals of the thesis. GradientSHAP and GradCAM++ are the variants used as they are effective with images, AL, and pre-trained ResNet-34.

Evaluation pattern reveals fundamental differences between dataset domain. While PCAM demonstrates gradual improvement across cycles suggesting medical imaging requires sustained learning for good explanation quality. While CIFAR-10 exhibits a clear two-phase pattern with rapid early improvement followed by stabilization around cycle 10. The findings suggests that in medical imaging additional labeled data provides better explanation quality. While in natural imaging explanation quality hit the maximum in early cycles.

SHAP & GradCAM

The most fluctuated evolution for SHAP on both datasets can be observed in continuity and least in compactness. The high fluctuation implies that the explanations created in certain cycles are more sensitive to minor training changes. Whereas less fluctuation in compactness implies that across the cycles explanations remain focused and less likely to shift the attention towards irrelevant regions. Correctness has high scores which means explanations are faithful to model performance. There is gradual increase in consistency indicating explanations are getting more similar and consistent. Unlike SHAP, fluctuations in GradCAM are dataset specific. In PCAM, Correctness has high fluctuated evolution and continuity has least fluctuated evolutions. In CIFAR-10, continuity has higher and compactness has lower fluctuations. Even the scores are highly dataset dependent, where PCAM has very less correctness scores and CIFAR-10 has high scores. This implies that explanation faithfulness cannot be transferred to other datasets. This understanding of pattern is crucial for development of application where there is need of reliable explanations along with pixel-level understanding and selection of explainability method based domain specific characteristics rather than assuming universal applicability across all domains.

SHAP Vs GradCAM

This thesis highlights a fundamental trade-off between SHAP's pixel-level precision and GradCAM's region based stability. SHAP's attribution deliver a highly detailed but volatile explanations. Its low continuity, noticeable sensitivity to correlation and baseline perturbations reveal that even a minor change can affect the marking of important pixels. In contrast, GradCAM's heatmaps achieve very high continuity and consistency reflects its spatially logical and repeatable explanations remain robust under minor changes but at the cost of losing fine-grained detail

needed for medical imaging. This trade-off observed across PCAM, and CIFAR-10 in this thesis states that practitioners must choose between granular faithfulness (SHAP) or explanation stability (GradCAM) depending on the domain needs and AL objectives.

XAI Quality Score (XQS)

Weighted XAI scores higher than unweighted consistently on both the datasets and XAI methods. Here, unweighted is average of all the metrics and weighted gives XAI method appropriate weighting for each metric. This captures the evaluation in a single metric that can be compared across the datasets and AL cycles. Ground truth problem is addressed by using these functionally grounded metrics and aggregation of these metrics into single score. As discussed before, XQS also has similar fluctuations and evolution path reflecting increased performance with increased AL cycles. This represents a standardized approach for comparing explanation methods and making data-driven decisions about AL stopping criteria based on explanation quality rather than subjective assessment.

In general, Minimal divergence between mean and median values, alongside weighted XQS high performance showcase the robustness of evaluation and confidence in the framework. This is crucial as it signifies that evaluation can reliably explained without being influenced by outliers or edge cases. The overall findings make it clear that explanation get more faithful to model prediction with increasing AL cycles. But, the evolution pattern is highly dependent on XAI methods, dataset domain and evaluation methods and SHAP has lower values in general compared to GradCAM.

5.1.3 Experimentation & Analysis

SHAP evaluation scores on PCAM are highly, and CIFAR-10 are moderately positively correlated with model performance. Even though individual XAI metric have variable correlation with model performance, but the aggregated metric (XQS) is highly correlated. The correlation between individual metric and XQS are highly skewed like correctness in PCAM has zero but highly positive correlation. Whereas GradCAM XAI evaluation is negatively correlated in PCAM and positively correlated in CIFAR-10 with model performance. Similar to SHAP individual metric results are skewed in GradCAM. Finally, the evolution pattern is much more complicated in GradCAM compared to SHAP and is highly variable based on dataset domain.

As expected in outlier analysis, top samples have the best highly focused attribution and bottom samples has the worst scattered attribution maps on both methods and datasets. But, in PCAM highest attribution region for top outlier is not in the cancerous spot as annotated but is in the background tissue. This suggests that despite the accurate model prediction, XAI explanation may highlight contextual feature rather than relevant feature. To conclude, explanation quality improves with AL. Yet some images remain challenging for consistent interpretability.

The experiments conducted are identical for both the datasets, with one exception in continuity metric with PCAM being gaussian noise and CIFAR-10 with

image rotation as perturbation for dataset appropriation. In both datasets, the explanation evolution is very robust to standard parameter changes but are very sensitive to change of correlation function, and correctness metric baseline function. In summary, the metrics are stable under normal variations, sensitive to meaningful shifts, consistent across domains and methods. These proxy metrics can provide reliable and reproducible XAI evaluation across AL cycles without human annotated ground truth.

5.1.4 Methodology Evaluation

The methodology in this thesis combines DSR and CRISP-DM. It is well-suited for developing, implementing, and evaluating the XAI explanation quality framework. DSR guided iterative artifact creation and evaluation pipeline. While CRISP-DM structured the data centric process from sample selection to deployment. The pipeline was executed consistently across both the datasets. It leverages fixed subsets, and reproducible configurations to ensure fairness and cross domain validity. The use of functionally grounded proxy metrics enabled objective explanation quality assessment in the absence of ground truth. The strength lies in content sensitive execution, strong foundation in empirical testing, and robustness to noise across the cycles. However it is limited in reflecting the semantic and visual clarity of the explanations which are important in human-centric designs. This highlights the future need to explore hybrid approaches combining quantitative evaluation with user studies. Overall, this methodology enabled detailed insights into evolution and variability of explanations across AL.

5.2 Limitations

Although this thesis offers several strengths and contributions, it also has limitations and trade-offs that originate from time and resource constraints. These limitations arise from the methodology, experimental design, and explainability evaluation. Recognizing these limitations is important for understanding the results and to guide future work to build on this thesis.

The first limitation is the reliance on a single sampling strategy for AL. As discussed, entropy-based sampling can cause class imbalance in sampling which is problematic for medical imaging. Furthermore, the datasets are upscaled to 224x224 pixels which may introduce artifacts. Even though 20 cycles show sufficient performance with current dataset size, number of cycles should be adjusted for dataset size. Similarly, only two post-hoc XAI methods are applied which may not capture the true decision-making process during model training.

Although ground truth problem is addressed, the functionally grounded metrics do not fully capture semantic and domain specific relevance that humans recognize. These metrics cannot assess completely if explanations support human decisions in real-world scenarios. The use of predefined weights optimized for medical imaging in weighted XQS may not reflect the relative importance of metrics in other domains. While the datasets cover distinct domains, the results cannot be generalized to other forms of medical imaging and do not reflect the real-world tasks. Finally, the

correlation between model performance and explanation quality is discussed while the causal relationships still remain unclear.

The next limitation of this thesis is the use of a single architecture (ResNet-34) with pre-trained weights. This inflates the AL performance compared to smaller alternative architectures or training a model from scratch. Although pre-trained models accounts for much of the efficiency, it aligns with real-world construction of AI solutions with limited data where transfer learning is not just beneficial but also essential. The fixed subsets for training, validation, testing and tracking will not fully capture the variability of real-world dynamics. Although many perturbation functions were used, other potential perturbations are not explored. Results are dependent on using specific toolkits like Captum and Quantus, replication under other tools should be verified. Training, testing, experimentation have a huge computational cost especially with computationally heavy methods like SHAP, XAI metrics and large datasets like PCAM. This combination would not cause a major issue if performed only once. However, the cyclical nature of methodology requires recalculating all the metrics in every cycle which rapidly increases computational burden.

The absence of domain-expert validation and user studies to determine whether the explanations enhance human understanding that is a key real-world constraint. In addition, understanding the framework performance on larger datasets and high resolution images remains insufficiently explored. The underlying reasons for explanation evolution are not discussed adequately.

5.3 Future Work

This thesis provides the first systematic analysis of XAI evolution, laying a foundation for further research. Limitations discussed in previous section highlight the opportunities for better methodological innovation, broader applicability, and deeper human-centric validation.

Implementation of the framework using more balanced AL sampling like certainty-based sampling, diversity sampling or a hybrid approach. Using uncertainty-based sampling for initial gains and gradually introducing explanation-driven sampling once the model stabilizes. These methods may mitigate class imbalance and can be compared with this thesis to find which sampling is better for this framework. Training a lightweight machine learning model from scratch instead of using pre-trained weights would isolate architecture effects and interpolation artifacts. Using multiple intrinsic explainable architectures to compare against post-hoc methods provide model's decision inherently. Integration of additional post-hoc methods like LIME, DLIME, LRP, to test if observed SHAP-GradCAM trade-off generalizes.

Developing adaptive XQS weighting scheme by meta-optimization and expert elicitation customized for each metric and domain. Incorporating semantic similarity measure or causal explanations score to capture domain relevance which is overlooked by functionally grounded metrics. Conducting user studies and human-in-the-loop experiments to measure annotation speed, trust, and decision quality of the framework. Experimentation with other medical domains and complex natural datasets would reveal whether the framework is transferable.

Extension of perturbation tests to adversarial attacks, illumination changes and domain shifts to map worst case explanation failures. Investigating the sensitivity of XAI attribution maps for different feature representations, especially for SHAP to quantify how pre-processing steps affect explanations. Developing a system to efficiently update attribution maps after each cycle by only recalculating those affected by new data or model changes instead of recomputing explanations from scratch. Explore GPU-efficient algorithms and batch metric computation to keep large-scale computations achievable. Finally, validating the results under alternative XAI libraries and XAI metric tools will ensure tool-agnostic robustness.

Chapter 6

Conclusion

This final chapter synthesizes the main research findings and revisits the research questions in light of presented results. It also reflects on the significance and limitations of the presented framework.

6.1 Sub-Research Question 1

How do quantitative XAI metrics evolve through active learning cycles: We tracked four key metrics such as continuity, compactness, correctness, and consistency, each display a unique evolution pattern throughout both the experimentation. The pattern is dependent on both XAI methods and datasets. For PCAM, SHAP shows steady but fluctuating gains across all the metrics. GradCAM achieves high continuity, consistency early but with declining correctness with added samples. For CIFAR-10, explanations stabilize faster with SHAP displaying a clear two-phase improvement with rapid early gains followed by a plateau and GradCAM showing an early plateau at higher explanation quality. Overall our experiments show that explanation quality measured by XQS has improved by an average of 4% for SHAP and by roughly 9% for GradCAM from the first to final cycle.

6.2 Sub-Research Question 2

How closely do improvements in explanation quality align with gains in model performance: Examining the relationship between explanation quality and model performance reveals partial alignment. In PCAM, SHAP explanations show a strong positive correlation ($r = 0.78$) with performance metrics like accuracy, F1-score, AUC indicating better explanations with increased metrics. On the other hand GradCAM shows a complex relationship with some negative scores ($\neg 0.71$) indicating high model accuracy corresponding to lower faithful explanations. In CIFAR-10, both XAI methods show a closer alignment but few inconsistencies remain while examining individual metrics. These results highlight the relationship between model performance and explanation quality is not universal. But it depends on specific combinations of XAI method, evaluation metrics and dataset

domain. Therefore, the explanation quality should be monitored independently from standard performance metrics in workflow.

6.3 Sub-Research Question 3

How do explanation trends differ between SHAP and GradCAM during active learning: SHAP and GradCAM exhibit fundamentally different evolution pattern reflecting their underlying methodological approaches. SHAP provides precise explanations and demonstrates high sensitivity with low stability and high correctness scores. SHAP shows a gradual improvement across most of the metrics but remains sensitive to perturbations and correlation changes. In contrast, GradCAM provides region-based heatmaps with high stability, continuity and lower correctness in later cycles. Our comparative analysis uncovers a major difference, with SHAP achieving pixel-level precision (correctness: $0.45 - 0.48$) and GradCAM maintaining regional stability (continuity: $0.94 - 0.98$). The choice between methods should be guided by application specific requirements between detailed pixel-level attribution and region-level focus maps.

6.4 Sub-Research Question 4

How does dataset domain influence the progression of explanation metrics: Dataset domain has noticeable influence on the explanation evolution on both medical and natural images showing distinct learning patterns. Medical imaging exhibits a gradual and sustained improvement, with greater sensitivity to class imbalance, and benefits from more labeled data. Natural images show rapid early gains followed by stabilization, indicating faster convergence and limited benefit from further labeling beyond this point. These patterns suggest that AL strategies and stopping criteria should be domain-specific. Where, medical tasks require longer training and natural images can stop earlier once explanations stabilize.

6.5 Main Research Question

How does the quality of XAI explanations evolve across active learning cycles: Our experiments confirm that XAI explanations for both SHAP and GradCAM exhibit trackable evolution patterns during AL cycles. These patterns are strongly dependent on XAI methods, evaluation metrics and dataset domain. The study establishes that explanations generally improve with more labeled data, but character and pace of improvement varies significantly. SHAP shows gradual gains in faithfulness and stability, while GradCAM achieves high stability and declining faithfulness over time. The proposed XQS captures these evolutionary trends with weighted aggregations outperforming unweighted. Overall, this research establishes a systematic framework for tracking explanation evolution in AL workflows. This enables practitioners to make informed decisions about AL stopping criteria and model deployment based on explanation quality rather than relying solely on model performance metrics.

Bibliography

[1] Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160.

[2] Adhane, G., Dehshibi, M. M., and Masip, D. (2022). On the use of uncertainty in classifying aedes albopictus mosquitoes. *IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING*, 16:224–233.

[3] Andresini, G., Pendlebury, F., Pierazzi, F., Loglisci, C., Appice, A., and Cavallaro, L. (2021). Insomnia: Towards concept-drift robustness in network intrusion detection. In *AISec 2021 - Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security, co-located with CCS 2021*, pages 111–122. Export Date: 11 December 2024; Cited By: 65.

[4] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115.

[5] Baur, T., Heimerl, A., Lingenfelser, F., Wagner, J., Valstar, M. F., Schuller, B., and André, E. (2020). explainable cooperative machine learning with nova. *KI - Kunstliche Intelligenz*, 34:143–164. Export Date: 11 December 2024; Cited By: 28.

[6] Bejnordi, B. E., Veta, M., Van Diest, P. J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J. A., Hermsen, M., Manson, Q. F., Balkenhol, M., et al. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210.

[7] Cantürk, F. and Aydoğan, R. (2023). Explainable active learning for preference elicitation.

[8] Chandra, A. L., Desai, S. V., Devaguptapu, C., and Balasubramanian, V. N. (2021). On initial pools for deep active learning. In *NeurIPS 2020 Workshop on Pre-registration in Machine Learning*, pages 14–32. PMLR.

[9] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., et al. (2000). Crisp-dm 1.0: Step-by-step data mining guide. *SPSS inc*, 9(13):1–73.

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

[11] Chidara, S. (2025). Explainable active learning. https://github.com/srivatsa07/ExplainableActiveLearning.

[12] Christoph, M. (2020). *Interpretable machine learning: A guide for making black box models explainable*. Leanpub.

[13] Cohn, D. A., Ghahramani, Z., and Jordan, M. I. (1996). Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145.

[14] Coroama, L. and Groza, A. (2022). Evaluation metrics in explainable artificial intelligence (xai). In *International conference on advanced research in technologies, information, innovation and sustainability*, pages 401–413. Springer.

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

[16] Ghai, B., Liao, Q. V., Zhang, Y., Bellamy, R., and Mueller, K. (2021). Explainable active learning (xal): Progressive disclosure: Empirically motivated approaches to designing effective transparency. *Proceedings of the ACM on Human-Computer Interaction*, 4. Export Date: 11 December 2024; Cited By: 38.

[17] Gildenblat, J. and contributors (2021). Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam.

[18] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[19] Hemelings, R., Elen, B., Barbosa-Breda, J., Lemmens, S., Meire, M., Pourjavan, S., Vandewalle, E., de Veire, S. V., Blaschko, M. B., Boever, P. D., and Stalmans, I. (2020). Accurate prediction of glaucoma from colour fundus images with a convolutional neural network that relies on active and transfer learning. *Acta Ophthalmologica*, 98:e94–e100. Export Date: 11 December 2024; Cited By: 87.

[20] Hevner, A. R., March, S. T., Park, J., and Ram, S. (2004). Design science in information systems research. *MIS quarterly*, pages 75–105.

[21] Holm, S. and Macedo, L. (2023). The accuracy and faithfullness of al-dlime - active learning-based deterministic local interpretable model-agnostic explanations: A comparison with lime and dlime in medicine. In Longo, L., editor, *EXPLAINABLE ARTIFICIAL INTELLIGENCE, XAI 2023, PT I*, volume 1901, pages 582–605. 1st World Conference on Explainable Artificial Intelligence (XAI), Lisbon, PORTUGAL, JUL 26-28, 2023.

[22] Hu, Y. and Chaddad, A. (2025). Shap-integrated convolutional diagnostic networks for feature-selective medical analysis. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

[23] Huang, Y., Zhao, Y., Wang, Z., Liu, X., Liu, H., and Fu, Y. (2023). Explainable district heat load forecasting with active deep learning. *Applied Energy*, 350. Export Date: 11 December 2024; Cited By: 15.

[24] Kalakoti, R., Nomm, S., and Bahsi, H. (2024). Enhancing iot botnet attack detection in socs with an explainable active learning framework. In Paul, R., Kundu, A., and Bhattacharyya, R., editors, *2024 IEEE 5TH ANNUAL WORLD AI IOT CONGRESS, AIIOT 2024*, pages 265–272. IEEE 5th Annual World AI IoT Congress (AIIoT), WA, MAY 29-31, 2024.

[25] Keele, S. et al. (2007). Guidelines for performing systematic literature reviews in software engineering. Technical report, Technical report, ver. 2.3 ebse technical report. ebse.

[26] Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.

[27] Kriznar, K., Rozanec, J. M., Fortuna, B., and Mladenic, D. (2023). Explainable artificial intelligence meets active learning: A novel gradcam-based active learning strategy. In Vrcek, N., Ortega, L. D., and Grd, P., editors, *CENTRAL EUROPEAN CONFERENCE ON INFORMATION AND INTELLIGENT SYSTEMS, CECIIS*, pages 399–406. 34th International Scientific Central European Conference on Information and Intelligent Systems (CECIIS), Univ Dubrovnik, Dubrovnik, CROATIA, SEP 20-22, 2023.

[28] Kuelzer, D. F., Debbichi, F., Stanczak, S., and Botsov, M. (2021). On latency prediction with deep learning and passive probing at high mobility. In *IEEE INTERNATIONAL CONFERENCE ON COMMUNICATIONS (ICC 2021)*. IEEE International Conference on Communications (ICC), ELECTR NETWORK, JUN 14-23, 2021.

[29] Kumar, S., Datta, S., Singh, V., Datta, D., Singh, S. K., and Sharma, R. (2024). Applications, challenges, and future directions of human-in-the-loop learning. *IEEE Access*, 12:75735–75760.

[30] Le, P. Q., Nauta, M., Van Bach Nguyen, S. P., Pathak, S., Schlötterer, J., and Seifert, C. (2023). Benchmarking explainable ai-a survey on available toolkits and open challenges. In *IJCAI*, pages 6665–6673.

[31] Lee, H. and Li, W. (2024). Improving interpretability of deep active learning for flood inundation mapping through class ambiguity indices using multi-spectral satellite imagery. *REMOTE SENSING OF ENVIRONMENT*, 309.

[32] Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

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

[34] Luo, Y., Yang, Z., Meng, F., Li, Y., Guo, F., Qi, Q., Zhou, J., and Zhang, Y. (2023). Xal: Explainable active learning makes classifiers better low-resource learners.

[35] Mandalika, S. and Nambiar, A. (2024). Segxal: Explainable active learning for semantic segmentation in driving scene scenarios.

[36] Mark Chignell, L. W. M.-H. C. (2020). *Interactive Machine Learning for Data Exfiltration Detection: Active Learning with Human Expertise*. IEEE.

[37] Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., Ramírez-Quintana, M. J., and Flach, P. (2019). Crisp-dm twenty years later: From data mining processes to data science trajectories. *IEEE transactions on knowledge and data engineering*, 33(8):3048–3061.

[38] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.

[39] Mirzaei, S., Mao, H., Al-Nima, R. R. O., and Woo, W. L. (2023). Explainable ai evaluation: a top-down approach for selecting optimal explanations for black box models. *Information*, 15(1):4.

[40] Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., and Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, 56(4):3005–3054.

[41] Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., Van Keulen, M., and Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42.

[42] Nishimura, Y., Takeda, N., Legaspi, R., Ikeda, K., Plötz, T., and Chernova, S. (2023). Allfa: Active learning through label and feature augmentation. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 158–165.

[43] Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, 24(3):45–77.

[44] Phan, T. D. and Giudici, P. (2025). Sustainability, accuracy, fairness, and explainability (safe) machine learning in quantitative trading. *Mathematics*, 13(3):442.

[45] Phillips, R. L., Chang, K. H., and Beckman, M. (2018). Interpretable active learning * †.

[46] Rožanec, J. M., Zajec, P., Trajkova, E., Šircelj, B., Brecelj, B., Novalija, I., Dam, P., Fortuna, B., and Mladenic, D. (2022). Towards a comprehensive visual quality inspection for industry 4.0. In *IFAC-PapersOnLine*, volume 55, pages 690–695. Export Date: 11 December 2024; Cited By: 12.

[47] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2019). Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359.

[48] Seol, D. H., Choi, J. E., Kim, C. Y., and Hong, S. J. (2023). Alleviating class-imbalance data of semiconductor equipment anomaly detection study. *Electronics*, 12(3):585.

[49] Settles, B. (2009). Active learning literature survey.

[50] Teso, S. and Kersting, K. (2019). Explanatory interactive machine learning. In *AIES '19: PROCEEDINGS OF THE 2019 AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY*, pages 239–245. 2nd AAAI/ACM Conference on AI, Ethics, and Society (AIES), Honolulu, HI, JAN 27-28, 2019.

[51] TorchVision maintainers and contributors (2021). torchvision.models: Deep learning models for image classification. https://docs.pytorch.org/vision/0.9/models.html. Version 0.9.

[52] Veeling, B. S., Linmans, J., Winkens, J., Cohen, T., and Welling, M. (2018a). Patchcamelyon (pcam) deep learning classification benchmark. https://github.com/basveeling/pcam. Accessed: 23-06-2025.

[53] Veeling, B. S., Linmans, J., Winkens, J., Cohen, T., and Welling, M. (2018b). Rotation equivariant cnns for digital pathology. In *Medical image computing and computer assisted intervention–mICCAI 2018: 21st international conference, granada, Spain, September 16-20, 2018, proceedings, part II 11*, pages 210–218. Springer.

[54] Venable, J., Pries-Heje, J., and Baskerville, R. (2012). A comprehensive framework for evaluation in design science research. In *Design Science Research in Information Systems. Advances in Theory and Practice: 7th International Conference, DESRIST 2012, Las Vegas, NV, USA, May 14-15, 2012. Proceedings 7*, pages 423–438. Springer.

[55] Wang, J., Chen, Y., and Giudici, P. (2025). Group shapley with robust significance testing and its application to bond recovery rate prediction. *arXiv preprint arXiv:2501.03041*.

[56] Zhang, J., Bo, X., Wang, C., Dai, Q., Dong, Z., Tang, R., and Chen, X. (2024). Active explainable recommendation with limited labeling budgets. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 5375–5379. Institute of Electrical and Electronics Engineers Inc.

Appendix A

Tables and Figures

A.1 PCAM

A.1.1 Correlation Analysis

Table A.1: PCAM GradCAM Correlation Results

Metric Pair	**r value**	**p value**
Accuracy & Correctness Mean	-0.712	0.0004
Accuracy & Correctness Median	-0.582	0.0071
Precision & Correctness Mean	-0.404	0.0770
Precision & Correctness Median	-0.552	0.0117
Recall & Correctness Mean	-0.583	0.0070
Recall & XQS Mean	-0.561	0.0100
Recall & XQS Weighted Mean	-0.596	0.0056
F1 & Correctness Mean	-0.759	0.0001
F1 & Correctness Median	-0.548	0.0125
AUC & Correctness Mean	-0.787	0.0000
AUC & Correctness Median	-0.555	0.0112
Continuity Mean & Consistency Median	-0.644	0.0022
Continuity Median & Consistency Mean	-0.543	0.0134
Continuity Median & Consistency Median	-0.566	0.0092
Correctness Mean & Consistency Mean	-0.528	0.0166
XQS Mean & Recall	-0.561	0.0100
XQS Weighted Mean & Recall	-0.596	0.0056

Table A.2: PCAM SHAP Correlation Results

Metric Pair	**r value**	**p value**
Accuracy & Precision	0.869	0.0000
Accuracy & F1	0.971	0.0000
Accuracy & AUC	0.958	0.0000
Accuracy & XQS Mean	0.780	0.0001
Accuracy & XQS Median	0.691	0.0007
Accuracy & XQS Weighted Mean	0.770	0.0001
Precision & F1	0.771	0.0001
Precision & AUC	0.753	0.0001
Precision & XQS Mean	0.748	0.0001
Precision & XQS Weighted Mean	0.742	0.0002
F1 & AUC	0.985	0.0000
F1 & XQS Mean	0.730	0.0003
F1 & XQS Median	0.685	0.0009
F1 & XQS Weighted Mean	0.720	0.0003
AUC & Consistency Mean	0.707	0.0005
AUC & XQS Mean	0.736	0.0002
AUC & XQS Median	0.722	0.0003
AUC & XQS Weighted Mean	0.718	0.0004
AUC & XQS Weighted Median	0.717	0.0004
Continuity Mean & Continuity Median	0.921	0.0000
Continuity Mean & XQS Mean	0.779	0.0001
Continuity Mean & XQS Median	0.752	0.0001
Continuity Mean & XQS Weighted Mean	0.755	0.0001
Compactness Mean & Compactness Median	0.839	0.0000
Compactness Mean & XQS Mean	0.683	0.0009
Compactness Mean & XQS Weighted Mean	0.698	0.0006
Consistency Mean & Consistency Median	0.921	0.0000
Consistency Median & XQS Mean	0.698	0.0006
Consistency Median & XQS Median	0.709	0.0005
Consistency Median & XQS Weighted Mean	0.687	0.0008
XQS Mean & XQS Median	0.932	0.0000
XQS Mean & XQS Weighted Mean	0.995	0.0000
XQS Mean & XQS Weighted Median	0.754	0.0001
XQS Median & XQS Weighted Mean	0.923	0.0000
XQS Median & XQS Weighted Median	0.902	0.0000
XQS Weighted Mean & XQS Weighted Median	0.747	0.0002

A.1.2 Per-Image Analysis

Table A.3: PCAM GradCAM Per-Image Statistic Results

Metric	Count	Mean	Std	Min	25%	50%	75%	Max
XQS	2000	0.614	0.108	0.325	0.543	0.601	0.688	0.879
XQS Weighted	2000	0.615	0.115	0.311	0.536	0.597	0.695	0.890
Continuity	2000	0.956	0.067	0.380	0.950	0.980	0.992	1.000
Compactness	2000	0.352	0.097	0.161	0.285	0.334	0.399	0.717
Correctness	2000	0.365	0.256	0.010	0.135	0.333	0.563	0.958
Consistency	2000	0.784	0.226	0.010	0.736	0.857	0.926	0.995

Table A.4: PCAM SHAP Per-Image Statistic Results

Metric	Count	Mean	Std	Min	25%	50%	75%	Max
XQS	2000	0.444	0.146	0.168	0.321	0.425	0.576	0.748
XQS Weighted	2000	0.445	0.198	0.131	0.265	0.387	0.642	0.795
Continuity	2000	0.469	0.114	0.158	0.401	0.485	0.546	0.760
Compactness	2000	0.428	0.043	0.347	0.401	0.422	0.448	0.621
Correctness	2000	0.473	0.413	0.007	0.040	0.482	0.898	0.992
Consistency	2000	0.405	0.136	0.010	0.350	0.427	0.487	0.674

A.1.3 Sensitivity Analysis

Table A.5: PCAM Consistency Experiments Sensitivity Results

Perturbation	**Metric**	**S-MAD**	**S-RMSE**	**G-MAD**	**G-RMSE**
Pearson	Mean	0.0833	0.0857	0.0227	0.0234
Pearson	Median	0.0847	0.0871	0.0278	0.0303
Kendall	Mean	0.1236	0.1268	0.1491	0.1531
Kendall	Median	0.1242	0.1275	0.1649	0.1693

Table A.6: PCAM Continuity Experiments Sensitivity Results

Perturbation	Metric	S-MAD	S-RMSE	G-MAD	G-RMSE
0.01	Mean	0.0012	0.0017	0.0015	0.0018
0.01	Median	0.0041	0.0052	0.0017	0.0021
0.2	Mean	0.0015	0.0017	0.0015	0.0019
0.2	Median	0.0052	0.0062	0.0014	0.0017
Pearson	Mean	0.0741	0.0742	0.0069	0.0074
Pearson	Median	0.0830	0.0833	0.0047	0.0053
Kendall	Mean	0.1410	0.1411	0.1011	0.1012
Kendall	Median	0.1459	0.1461	0.0956	0.0959

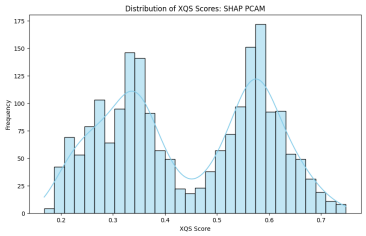
S: SHAP, G: GradCAM, MAD: Mean Absolute Difference, RMSE: Root Mean Square Error

Table A.7: PCAM Correctness Experiments Sensitivity Results

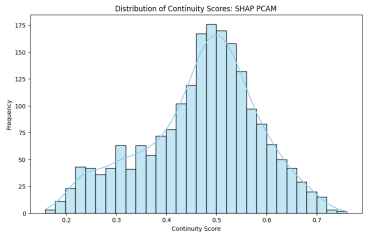
Perturbation	Metric	S-MAD	S-RMSE	G-MAD	G-RMSE
56.0	Mean	0.0009	0.0013	0.0002	0.0004
56.0	Median	0.0068	0.0089	0.0003	0.0005
224.0	Mean	0.0008	0.0010	0.0006	0.0007
224.0	Median	0.0070	0.0095	0.0009	0.0011
mean	Mean	0.1156	0.1176	0.0075	0.0087
mean	Median	0.1405	0.1518	0.0383	0.0438
uniform	Mean	0.0652	0.0695	0.0098	0.0119
uniform	Median	0.0606	0.0647	0.0340	0.0395

S: SHAP, G: GradCAM, MAD: Mean Absolute Difference, RMSE: Root Mean Square Error

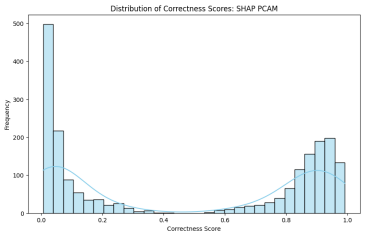
A.1.4 Distribution of Scores



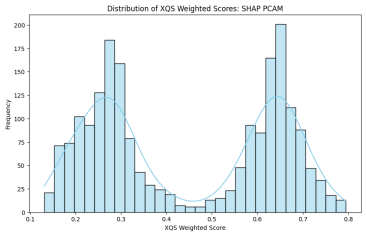
(A) SHAP XQS Distribution



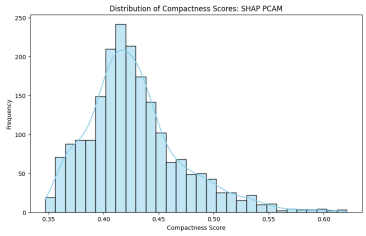
(C) SHAP Continuity Distribution



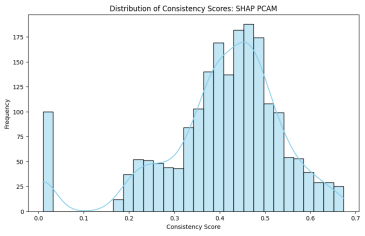
(E) SHAP Correctness Distribution



(B) SHAP Weighted XQS Distribution

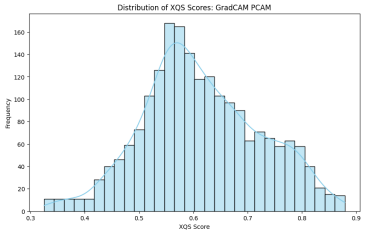


(D) SHAP Compactness Distribution

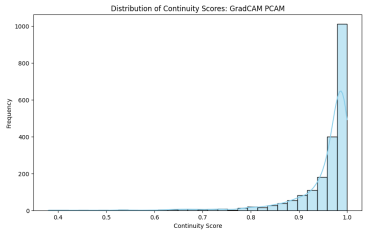


(F) SHAP Consistency Distribution

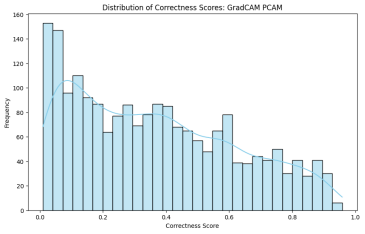
FIGURE A.1: SHAP Per-Image Metrics Distribution



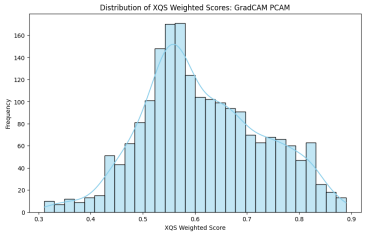
(A) GradCAM XQS Distribution



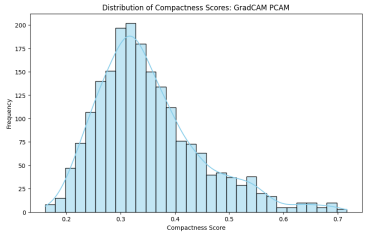
(C) GradCAM Continuity Distribution



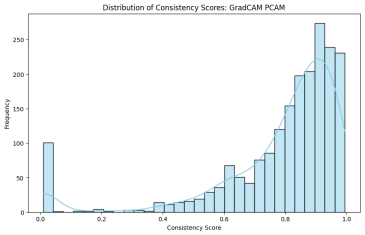
(E) GradCAM Correctness Distribution



(B) GradCAM Weighted XQS Distribution

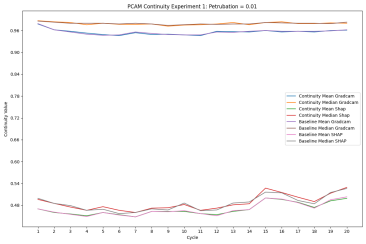


(D) GradCAM Compactness Distribution

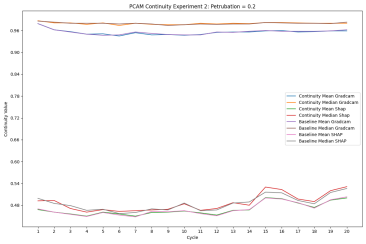


(F) GradCAM Consistency Distribution

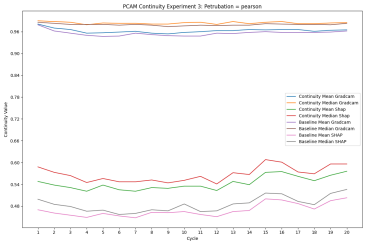
FIGURE A.2: GradCAM Per-Image Metrics Distribution



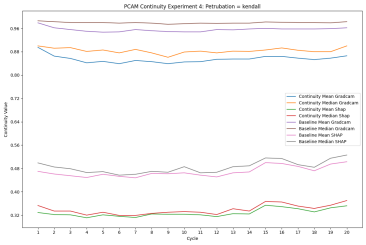
(A) Continuity:1



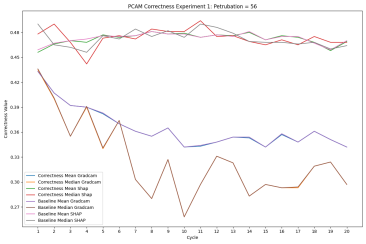
(B) Continuity:2



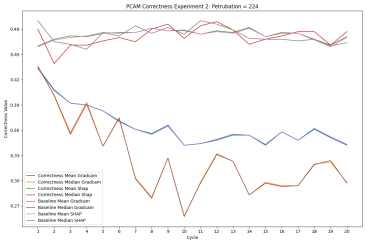
(C) Continuity:3



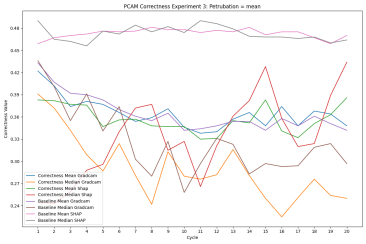
(D) Continuity:4



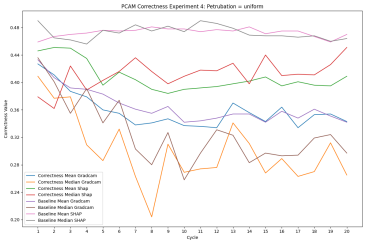
(E) Correctness:1



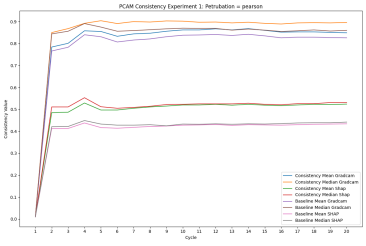
(F) Correctness:2



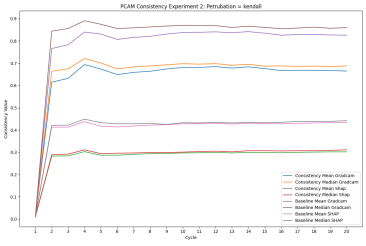
(G) Correctness:3



(H) Correctness:4



(I) Consistency:1



(J) Consistency:2

FIGURE A.3: PCAM Sensitivity Analysis

A.2 CIFAR-10

A.2.1 Per-Image Analysis

Table A.8: CIFAR-10 GradCAM Per-Image Statistic Results

Metric	Count	Mean	Std	Min	25%	50%	75%	Max
XQS	2000	0.777	0.079	0.351	0.753	0.804	0.828	0.884
XQS Weighted	2000	0.797	0.086	0.342	0.770	0.827	0.853	0.903
Continuity	2000	0.950	0.050	0.484	0.940	0.965	0.980	0.997
Compactness	2000	0.427	0.051	0.265	0.390	0.427	0.461	0.595
Correctness	2000	0.811	0.206	0.005	0.768	0.886	0.950	0.999
Consistency	2000	0.921	0.210	0.010	0.959	0.970	0.981	0.998

Table A.9: CIFAR-10 SHAP Per-Image Statistic Results

Metric	Count	Mean	Std	Min	25%	50%	75%	Max
XQS	2000	0.532	0.066	0.203	0.505	0.544	0.574	0.678
XQS Weighted	2000	0.600	0.087	0.148	0.578	0.624	0.653	0.738
Continuity	2000	0.348	0.086	0.080	0.291	0.346	0.403	0.581
Compactness	2000	0.490	0.033	0.405	0.468	0.487	0.509	0.613
Correctness	2000	0.866	0.190	0.023	0.855	0.941	0.976	1.000
Consistency	2000	0.425	0.113	0.010	0.395	0.444	0.486	0.611

A.2.2 Sensitivity Analysis

Table A.10: CIFAR-10 Consistency Experiments Sensitivity Results

Perturbation	**Metric**	**S-MAD**	**S-RMSE**	**G-MAD**	**G-RMSE**
Pearson	Mean	0.0203	0.0229	0.0056	0.0062
Pearson	Median	0.0172	0.0207	0.0056	0.0061
Kendall	Mean	0.1328	0.1362	0.0951	0.0976
Kendall	Median	0.1343	0.1378	0.0965	0.0990

S: SHAP, G: GradCAM, MAD: Mean Absolute Difference, RMSE: Root Mean Square Error

Table A.11: CIFAR-10 Continuity Experiments Sensitivity Results

Perturbation	Metric	S-MAD	S-RMSE	G-MAD	G-RMSE
5°	Mean	0.0449	0.0452	0.0106	0.0110
5°	Median	0.0475	0.0481	0.0082	0.0085
15°	Mean	0.0292	0.0295	0.0177	0.0180
15°	Median	0.0299	0.0311	0.0155	0.0162
Pearson	Mean	0.0784	0.0792	0.0019	0.0024
Pearson	Median	0.0800	0.0813	0.0042	0.0047
Kendall	Mean	0.1115	0.1116	0.1256	0.1259
Kendall	Median	0.1110	0.1112	0.1242	0.1246

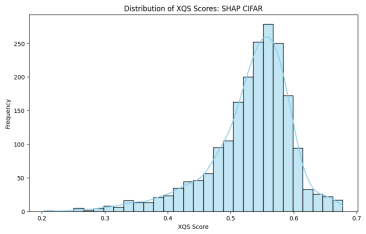
S: SHAP, G: GradCAM, MAD: Mean Absolute Difference, RMSE: Root Mean Square Error

Table A.12: CIFAR-10 Correctness Experiments Sensitivity Results

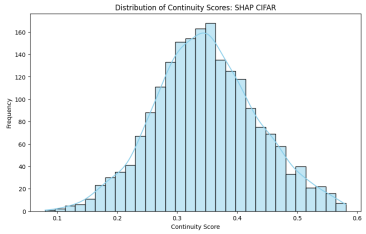
Perturbation	Metric	S-MAD	S-RMSE	G-MAD	G-RMSE
56	Mean	0.0011	0.0016	0.0008	0.0009
56	Median	0.0026	0.0039	0.0009	0.0010
224	Mean	0.0015	0.0017	0.0012	0.0013
224	Median	0.0038	0.0071	0.0019	0.0020
mean	Mean	0.0780	0.0820	0.0498	0.0525
mean	Median	0.0771	0.0851	0.0435	0.0541
uniform	Mean	0.0428	0.0456	0.0371	0.0448
uniform	Median	0.0379	0.0449	0.0474	0.0614

S: SHAP, G: GradCAM, MAD: Mean Absolute Difference, RMSE: Root Mean Square Error

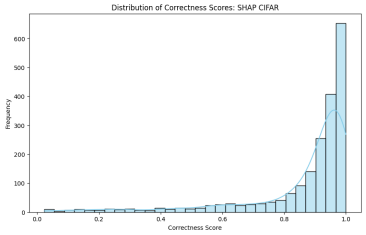
A.2.3 Distribution of Scores



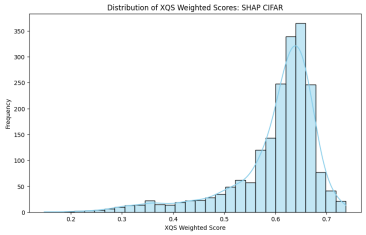
(A) SHAP XQS Distribution



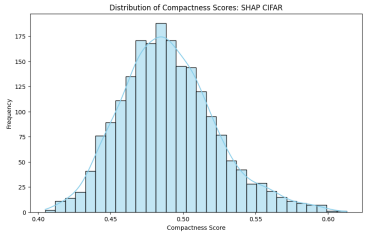
(C) SHAP Continuity Distribution



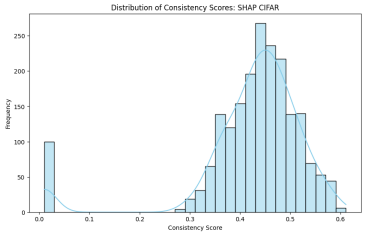
(E) SHAP Correctness Distribution



(B) SHAP Weighted XQS Distribution

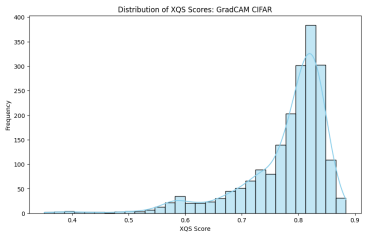


(D) SHAP Compactness Distribution

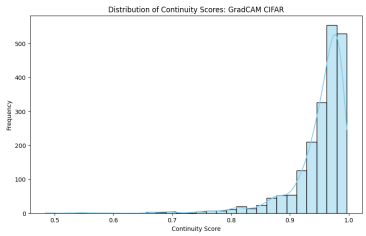


(F) SHAP Consistency Distribution

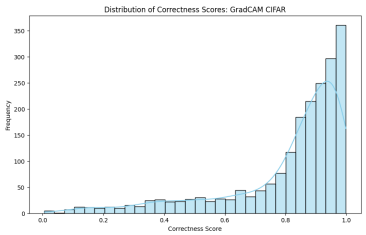
FIGURE A.4: SHAP Per-Image Metrics Distribution



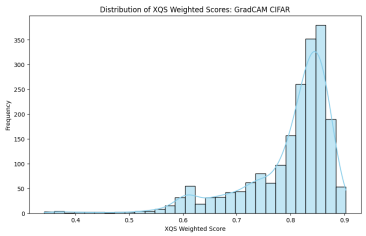
(A) GradCAM XQS Distribution



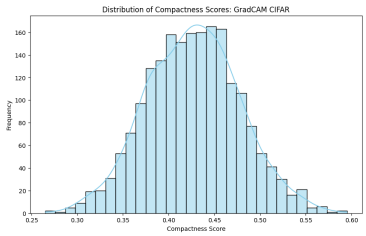
(C) GradCAM Continuity Distribution



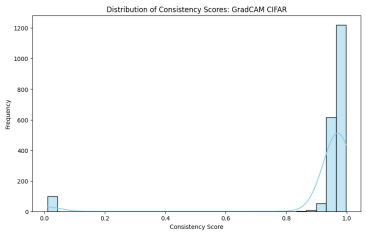
(E) GradCAM Correctness Distribution



(B) GradCAM Weighted XQS Distribution



(D) GradCAM Compactness Distribution



(F) GradCAM Consistency Distribution

FIGURE A.5: GradCAM Per-Image Metrics Distribution

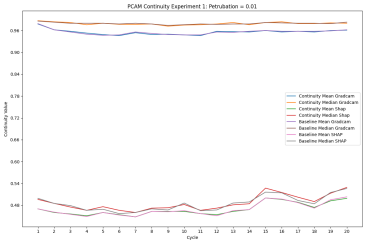
A.2.4 Correlation Analysis

Table A.13: CIFAR-10 GradCAM Correlation Results

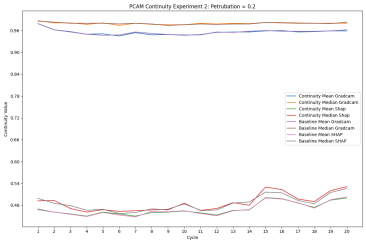
Metric Pair	**r value**	**p value**
Accuracy & Compactness Mean	0.776	0.0001
Accuracy & Consistency Mean	-0.614	0.0040
Accuracy & Continuity Mean	0.653	0.0018
Accuracy & Continuity Median	0.631	0.0029
Accuracy & XQS Mean	0.714	0.0004
Accuracy & XQS Median	0.743	0.0002
Compactness Median & XQS Median	0.754	0.0001
Correctness Median & Consistency Mean	-0.649	0.0020
Correctness Median & Consistency Median	-0.598	0.0053
Continuity Mean & Precision	0.635	0.0027
Continuity Mean & Recall	0.653	0.0018
Continuity Mean & F1	0.653	0.0018
Continuity Median & Precision	0.614	0.0040
Continuity Median & Recall	0.631	0.0029
F1 & Compactness Mean	0.776	0.0001
F1 & Consistency Mean	-0.614	0.0040
F1 & Continuity Mean	0.653	0.0018
F1 & Continuity Median	0.631	0.0029
F1 & XQS Mean	0.714	0.0004
F1 & XQS Median	0.743	0.0002
Precision & Consistency Mean	-0.630	0.0029
Precision & Continuity Mean	0.635	0.0027
Precision & Continuity Median	0.614	0.0040
Precision & XQS Mean	0.714	0.0004
Precision & XQS Median	0.746	0.0002
XQS Median & Compactness Median	0.754	0.0001
XQS Weighted Mean & Precision	0.662	0.0015
XQS Weighted Mean & Recall	0.665	0.0014
XQS Weighted Mean & F1	0.665	0.0014

Table A.14: CIFAR-10 SHAP Correlation Results

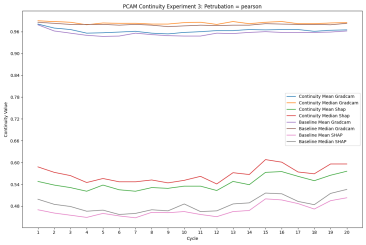
Metric Pair	**r value**	**p value**
Accuracy & Consistency Mean	-0.593	0.0059
Accuracy & Consistency Median	-0.601	0.0050
Compactness Mean & Compactness Median	0.584	0.0069
Compactness Mean & Continuity Mean	0.621	0.0035
Compactness Mean & Continuity Median	0.596	0.0056
Compactness Mean & Correctness Mean	0.660	0.0016
Compactness Mean & XQS Median	0.743	0.0002
Compactness Mean & XQS Weighted Mean	0.767	0.0001
Compactness Mean & XQS Weighted Median	0.649	0.0019
Compactness Median & Correctness Mean	0.678	0.0010
Compactness Median & XQS Median	0.764	0.0001
Compactness Median & XQS Weighted Median	0.681	0.0010
Consistency Mean & Precision	-0.606	0.0047
Consistency Mean & Recall	-0.593	0.0059
Consistency Mean & F1	-0.593	0.0059
Consistency Median & Precision	-0.613	0.0040
Consistency Median & Recall	-0.601	0.0050
Consistency Median & F1	-0.601	0.0050
Continuity Mean & Compactness Median	0.584	0.0069
Continuity Mean & XQS Mean	0.622	0.0034
Continuity Median & Compactness Mean	0.596	0.0056
Continuity Median & XQS Mean	0.593	0.0058
Correctness Mean & Compactness Median	0.678	0.0010
Correctness Median & XQS Mean	0.621	0.0035
Correctness Median & XQS Weighted Mean	0.701	0.0006
Precision & Consistency Mean	-0.606	0.0047
Precision & Consistency Median	-0.613	0.0040
Recall & Consistency Mean	-0.593	0.0059
Recall & Consistency Median	-0.601	0.0050
XQS Mean & Continuity Median	0.593	0.0058
XQS Mean & Correctness Median	0.621	0.0035
XQS Mean & XQS Weighted Median	0.773	0.0001
XQS Median & Compactness Mean	0.743	0.0002
XQS Median & Compactness Median	0.764	0.0001
XQS Weighted Mean & Correctness Median	0.701	0.0006
XQS Weighted Mean & Compactness Mean	0.767	0.0001
XQS Weighted Median & Compactness Mean	0.649	0.0019
XQS Weighted Median & Compactness Median	0.681	0.0010
XQS Weighted Median & XQS Mean	0.773	0.0001



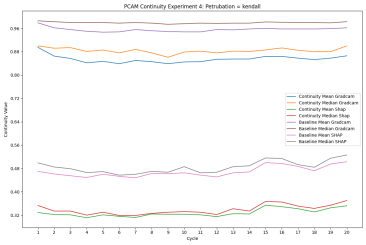
(A) Continuity:1



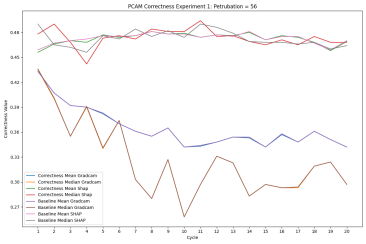
(B) Continuity:2



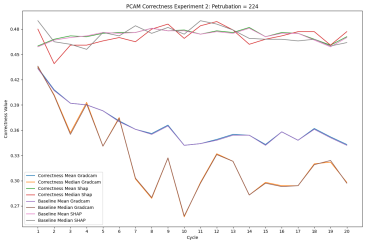
(C) Continuity:3



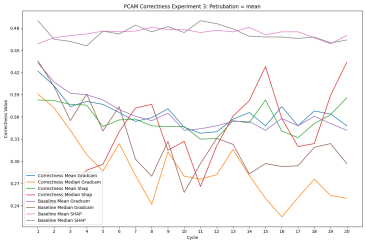
(D) Continuity:4



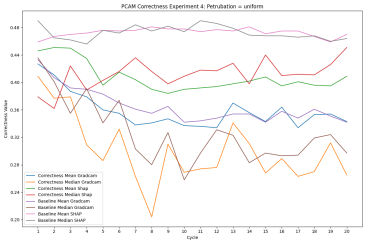
(E) Correctness:1



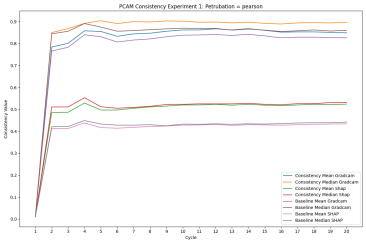
(F) Correctness:2



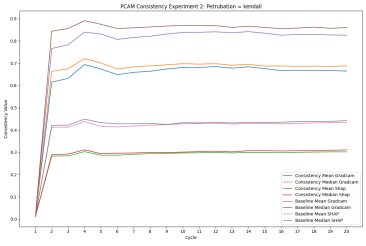
(G) Correctness:3



(H) Correctness:4



(I) Consistency:1



(J) Consistency:2

FIGURE A.6: CIFAR Sensitivity Analysis

Appendix B

Literature Review

This literature review examines the current landscape of XAL by identifying existing strategies, challenges and evaluation methods.This review is performed using the Kitchenham[25] methodology. It requires specifying research questions, search strategy, inclusion & exclusion criteria, and the information extracted from selected papers.

B.1 Research Questions and Search Strategy

The following research questions are formulated for the literature review, consists of the primary research question and sub-research questions.

RQ: What are the existing approaches for integrating explainability into active learning workflows (XAL) for ML and DL models, and and how is explainability evaluated?

SQ1: What are the existing sampling techniques and query strategies commonly used in Explainable active learning models (XAL)?

SQ2: What are the existing XAI strategies used to interpret the output of XAL models?

SQ3: How do XAL improve interpretability and model performance compared to traditional active learning strategies as reported in the literature?

SQ4: Which metrics and evaluation frameworks are used to assess the effectiveness of explainability techniques in active learning workflows?

Multiple academic databases were used to source the literature To address these questions. The databases includes, Scopus, Web of Science, and IEEE Xplore. The selected scientific papers are accessed using the network of University of Twente's institutional access. Advanced search features of the database are utilized to construct custom query that includes pre-defined keywords and their synonyms related to the research context.

The keywords are as follows:

- Explainable Active Learning, Active learning, XAL

- Explainable AI, XAI, Explainable Artificial Intelligence
- Explainability, Explainable models
- Machine learning, Deep Learning, Interpretable machine learning
- Semi supervised learning, Semi-supervised learning

These key-words are combined using the logical operators such as "AND", and "OR" to generate the search query. It corresponds to the research context of Explainable Active Learning. The query results were exported in the BibTeX format (.bib) for further processing and reference management. The search query used is:

- *("Explainable Active Learning" OR "Active learning" OR "XAL") AND ("Explainable AI" OR "XAI" OR "Explainable Artificial Intelligence" OR "Explainability" OR "Explainable Models") AND ("Machine Learning" OR "Deep Learning" OR "Interpretable Machine Learning" OR "Semi Supervised Learning" OR "Semi-supervised Learning")*

B.2 Study Selection and Data Extraction

The collected data(BibTeX file) from all the databases are combined and managed in Mendeley. It is a specialized reference management tool that streamlines importing references from multiple sources. It is used to detect and remove duplicates while allowing user to manually review and retain relevant references. This ensures that only unique and relevant studies are included in final dataset. Notion is another versatile tool used for this review. It allow users to create and organize notes, tasks, databases, and projects in a single workspace.

B.2.1 Inclusion Criteria

TABLE B.1: Inclusion criteria for literature selection.

1. The paper is written in English.
2. The paper aligns with the research of this review.
3. Papers identified through manual searches, recommendations, and citation tracking are included.
4. The paper contains pre-defined keywords relevant to the title, abstract, or keyword list.
5. The paper addresses one or more research questions.

B.2.2 Exclusion Criteria

TABLE B.2: Exclusion criteria for literature selection.

1. Literature published before 2016, except for foundational works directly relevant to research.
2. Papers that cannot be accessed through UT services.
3. Duplicates of the original paper.
4. Title or abstract of the paper that are irrelevant and do not address the research context.
5. Papers which are in a language other than English.
6. The paper addressing active learning without focusing on explainable AI and vice versa.

B.2.3 Study Selection

During the search process, some inclusion and exclusion criteria are applied using the advanced filter in the databases. These filters include language of the paper, selection of the paper which contain pre-defined keywords in title or abstract or keyword list, publishing year. After these initial filters the count of the papers was reduced from 2500 to 160. That is reduction of 93% papers from search query that are irrelevant to the research scope. The next step involved importing all references into Mendeley. It removed duplicates and left with 110 papers i.e., 30% are duplicates. The next step was abstract and conclusion screening reducing count to 28 papers which is around 25% of filtered papers. The next step was to read the whole paper from abstract to appendices leaving 20 papers which are highly relevant to the research questions accounting for 0.8% of the original search results. In addition to the database filtered papers, few papers are obtained manually which comprises of AL foundation, XAI book and foundation, Evaluation of XAI.

B.2.4 Data Extraction

In this stage, the following information is extracted from every paper.

- Research Domain: Natural Language Processing (NLP), Image Processing and Computer Vision (IPCV), Time Series Analysis (TS), Explainable Active Learning (XAL).
- References: Relevant works are referenced in each domain.
- Remarks: Observations regarding the application focus and its alignment with XAL.

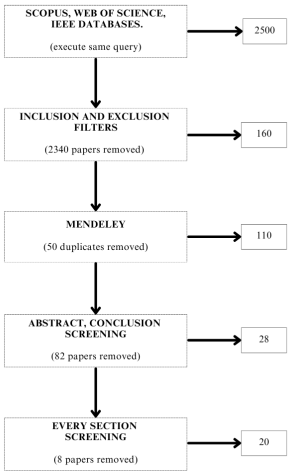


FIGURE B.1: Systematic Literature Review flow diagram

B.3 Results

In this section the information extracted from the papers are, Main contributions of the paper, Active Learning strategies/ queries that are used. Explainability strategies used to explain AL model output along with the evaluation of the quality of the explanations if applicable to the paper. Machine learning or Deep learning algorithms that are used to train the model. Evaluation metrics that are used to evaluate the performance of the classification/ regression task of AL model. It is further discussed in appendices **B.6**.

B.3.1 Distribution of Literature

As we can see from the Figure:B.2, the largest number of papers are published in 2023 in XAL domain indicating growing interest in the field. The trend line showcase us the interest in XAL has been increasing from past few years. The oldest paper is from 2018 which is around 7 years old. It states that XAL research is very recent and growing very fast. From the Figure:B.3, we can see that Image Processing and Computer Vision (IPCV) has the highest share in the research. This is because most of papers deal with the Medical diagnosis, and Real-time events. They need a huge pool of labeled data and it is expensive for human labeling. Then it is followed by Natural Language Processing (NLP) , Cyber Security (CS), XAL (papers suggesting new XAL framework without being related to any domain), and Time Series (TS).

B.3.2 Active Learning Strategies in XAL

This literature review identifies six main type of AL strategies used in XAL. They are, uncertainty sampling, random sampling, query by committee, density sampling, novel sampling strategies, certainty sampling. These strategies are compared in table B.3.

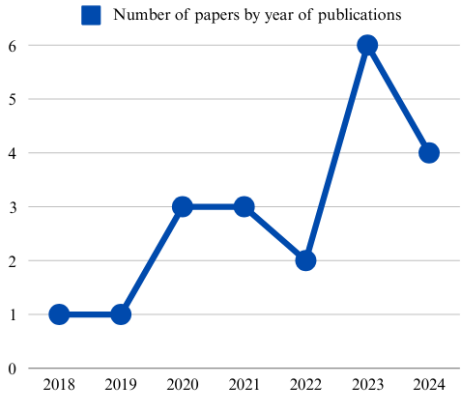


FIGURE B.2: Temporal Distribution

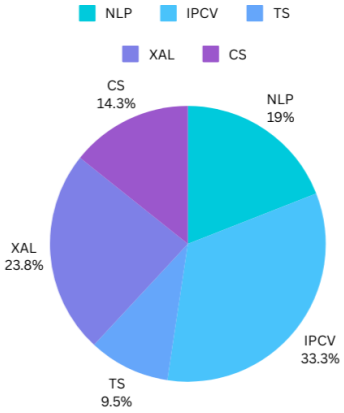


FIGURE B.3: Research Domain Distribution

Uncertainty based sampling is most commonly adopted strategy in this research. It selects the data points for labeling where the model is most unsure about its prediction that is based on the metrics (entropy, confidence, margin, etc.). Entropy-based sampling is the popular technique used that calculates entropy for sample selection[16][34][35][56][36][31]. Luo.et.al [34] explores different criteria that uses clustering algorithm for sample selection using uncertainty like Monte carlo dropout [2], Margin sampling[42][31], Confidence-based [5] [23], and Logistic regression [21]. There are several studies employing uncertainty sampling that omit explicit criteria mentioning in the studies[45][3][46][50][24].

Other AL strategies used are scenario specific. Certainty sampling [28] selects the data points of highest confidence of the model for labeling which is the opposite of uncertainty sampling. It is used to reinforce known region of data space. Density sampling [31] selects the data from high density regions that ensure representation of clusters and underlying distributions. Query by committee [24] uses a committee of models for selection of most important data points for labeling by maximizing the disagreement of the model predictions. Random sampling[19][31][27] selects data points for labeling stochastically without considering any criteria. It is mostly used as a baseline for comparison. Some papers propose novel sampling methods, DICE sampling [35] selects samples by using DICE similarity coefficient as criteria for labeling in semantic segmentation. GradCam sampling [27] selects images for labeling uses GradCAMavg metric that averages activation map explanations.

B.3.3 Explainable AI Strategies in XAL

This literature review categorizes XAI strategies used in XAL into mainly three types. They are post-hoc explanations, visualization based explanations, and intrinsic explanations. These strategies are compared in Table B.4.

Post-hoc explanations are provided after the model prediction that offers insights into the decision making process. LIME (Local Interpretable Model-Agnostic Explanations)[45][50][24] is used to provide local explanations by approximating

Table B.3: Active Learning Strategies

AL Strategy	Selection Method	Strengths	Limitations
Uncertainty Sampling	Selects samples with highest uncertainty.	Reduce annotation cost by targeting ambiguous regions.	Ignores diverse data points and struggles with imbalanced dataset.
Certainty Sampling	Selects samples with highest confidence.	Reinforces confidence in known patterns.	Fails to explore new data regions and risks overfitting.
Density Sampling	Selects samples from high density regions.	Diverse sample representation.	Selection of redundant data is possible and overlook sparse clusters.
Query-by-committee	Uses committee of models for maximizing disagreement.	Reduces bias and robust to outliers.	Computationally expensive and issues of scalability with model complexity.
Random Sampling	Random selection of data points.	Unbiased baseline and easy to implement.	Inefficient as it miss informative samples.
Novel Sampling	Custom strategies for specific tasks.	Optimized for domain specific needs.	Limited generalization and often task dependent.

complex models with interpretable models. It can also be integrated with additional explanation method such as iNNvestigate [5] for DL explanations. Active Learning Deterministic LIME (AL-DLIME) [21] offers local and deterministic explanations. This works on tweaked version of LIME with ridge linear regression as surrogate model for decision boundary approximation. SHAP (Shapely Additive explanations)[42][28][24] is also a prominent explanation technique. This works on the principles of game theory which compute feature importance and provide robust explanations. Bayesian Layer-wise Relevance Propagation (B-LRP) [2] provide local explanations by generating saliency maps to highlight most influential features in the model‘s decision making. DALEX (moDel Agnostic Language for Exploration and eXplanation) [3] provide global explanations and track features over time.

Visualization based explanations operate by providing graphical representation that illustrate and interpret the features influencing prediction. Local feature importance visualization through Bar charts [16] provide a simple method to represent the impact of the feature on the model's output. GradCAM[46][27] is used in convolutional neural networks that provides local explainability by visualizing the image region that influence model predictions with Heat-maps [35]. It can visualize local feature importance in spatial terms of the image. This can reveal areas which contribute for model prediction. Saliency maps[2][19] is a well known visualization

tool which provides local explanations by highlighting specific regions in the images that influence decision making of the model.

Intrinsic Explanations are interpretability methods that are built into structure of ML models eliminating the use of any post-hoc methods. GPT-based explanations [34] that employs a pre trained uni-directional decoder which can generate local NLP explanations for model understanding. Logistic Regression [7] is an interpretable model that is transparent and provide explanations based on the feature weights to determine the decision boundary. PETER (Personalized Transformer for Explainable Recommendation) [56] offers explainability catered for individual user preferences. Decision trees [36] is also a interpretable model that provide global explanations by providing the split criteria at each and every node that can reveal the feature importance. Correlation based attribution [23] offers global explanations by identifying the features that influence most in the model's output through correlation analysis.

Table B.4: XAI Strategies

XAI Type	**Specific Method**	**Reach**	**Strengths**	**Limitations**
Post-Hoc	LIME, SHAP, B-LRP, DALEX, AL-DLIME	Local, Global	Model agnostic, flexible for black-box models.	High computational costs, lacks faithfulness.
Visual based	GradCAM, Saliency maps, Bar charts.	Local	Effective for domain experts and image-based tasks.	Subjective interpretation, require visual literacy.
Intrinsic	Decision Trees, Logistic Regression, GPT-XAI.	Global	In-built interpretability, post-processing not needed.	Limited to simple models, poor scalability.

B.3.4 ML and DL Algorithms in XAL & Performance Comparison of XAL & AL

A diverse machine learning (ML) and deep learning (DL) algorithms have been employed that include interpretable models and black-box models. Traditional ML models like logistic regression[16][5][50] and support vector machines[5][50] are the popular models used for their simplicity and inherent intrinsic explainability. Logistic regression is also used alongside other models for a hybrid approach such as, AdaBoost[45] and Semi-supervised learning[7] to improve efficiency. Ensemble methods that include Random Forest[36][28][21] compared against Decision Tree [21], LightGBM[36][28], and XGBoost[24] (used Pearson correlation for feature selection) models are used for more accuracy and to mitigate the black-box problem.

Deep learning in XAL plays a major role for vision and sequential tasks. Transformer architecture approaches such as FLAN-T5[34], and PETER[56] are deployed in the NLP tasks for recommendation and classification. Convolutional Neural Networks[27] include U-Net[35][31], MiDas and DINOv2[35], ResNet-50[19], and fine-tuned VGG16[2] are being used for feature extraction and segmentation in IPCV tasks. Recurring models like UniLSTM with Attentional networks[42], and Active Graph Recurrent Network(Ac-GRN)[23] improves sequence modeling. There are few Hybrid models namely ResNet-18 with DRAEM along with GAN for image generation [46], Linear SVM combined with CNN and L2-Logistic regression [5]. These diverse models demonstrate the challenge of interpretability of the model along with optimizing performance.

Explainable Active Learning (XAL) enhances the performance by reducing reliance on superficial patterns and adds casual explanations by human validation[34]. It outperforms traditional AL methods like core-set, DEAL, Entropy AL[35], and CAIPI[42] while improving the quality and accuracy. Explanation correction further improve the model interpretability compared to standard AL[50]. Integrating the interpretability metrics like class ambiguity indices improves the model understanding found by comparing various AL strategies[31]. Some papers discuss about personalization and user trust[7] in intrusion detection[3]. Direct comparison of interpretability and performance between AL and XAL remains limited.

B.3.5 Evaluation Metrics and Frameworks

Two types of evaluation metrics have been identified: (1) performance evaluation metrics assessing XAL after the model training. (2) The evaluation metrics for evaluating the quality of the explanations by XAI methods within the XAL framework. They are further discussed in Table B.5.

The evaluation of XAL models involves diverse metrics which assess classification performance, uncertainty estimation, annotation quality, and assessing human-in-the-loop interactions. There are popular classification metrics like Accuracy, and F1-score [16][7][2][21][24][31], Matthews correlation coefficient (MCC)[45][7] and Area Under Curve (AUC)[36][42][5][19][28]. Along with the variants like AUC-ROC, and AUT-F1[3][46][27]. Precision, Recall, and sensitivity-specificity further refine the assessment of classification and annotations [2][19][46][24]. Prediction and segmentation accuracy are evaluated by Ranking loss[34], Mean Intersection over Union (mIoU)[35], and Root Mean Squared Error (RMSE)[56][23]. Uncertainty bias[45] is a novel metric developed to detect model overconfidence and mitigate uncertainty estimation. Human centered evaluation metrics such as confusion matrix[36][5], Human labeling performance[36][5], user trust questions[50],Degree of agreement between user-selected, and model-selected features[42] are used.

The evaluation of XAL explanation incorporates both subjective and objective metrics for comprehensive assessment. Annotator agreement[16] measure the generated explanations alignment with the ground truth. Human evaluation and Ranking accuracy[34] evaluate the quality of the explanations as good or bad. Explanation Accuracy[7] is calculation with the benchmark data and estimating user trust by taking the survey to evaluate perceived reliability. Metrics include BLEU and ROUGE

scores measures the overlap between the generated explanations with the ground truth[56]. The Degree of agreement indicates the intuitiveness of the explanations with high agreement signaling more user friendly explanations[42]. F1-score can also be used in this context by comparing consistency between model generated and human provided explanations[50]. To evaluate the stability and faithfulness, Jaccard's distance can be used that assess consistency under single and incremental deletion[21]. Faithfulness, Monotonicity and Max Sensitivity measure the adaption of explanations to changes in model[24]. Finally, Spearman Rank Correlation Coefficient evaluates the ranking consistency across different scenarios[31].

Table B.5: Evaluation Metrics

Metric Type	**Specific Metrics**	**Purpose**
Performance Metrics	Accuracy, F1-Score, AUC-ROC, MCC.	Assess overall model performance.
	mIoU, RMSE, Ranking Loss.	Evaluate segmentation accuracy (mIoU), prediction accuracy (RMSE).
	Uncertainty bias	Detects model over-confidence and potential bias in uncertainty estimation.
Explanation Metrics	Faithfulness, Stability, Monotonicity , Max Sensitivity.	Ensures explanations reflects the actual model behavior under modifications.
	BLEU, ROUGE, Jaccard Distance, Spearman Correlation.	Quantify overlap, feature attribution stability, and ranking consistency of explanations.
Human-centric Metrics	Degree of Agreement, User Trust Survey.	Measures the alignment between model and user reasoning, assess trust in AI.
	Annotator Agreement , Explanation Accuracy.	Evaluates human-AI collaboration efficiency and correctness of explanations.

B.4 Discussion

The rapid growth of XAL research in the recent years from 2018 (fig:B.2) shows the increasing demand for explanations of AL in high stake domains namely healthcare and cybersecurity. Dominance of IPCV (fig: B.3) as the most researched areas reflects the need of interpretable and explainable models where transparent decision making is essential. Marginalized representation of NLP (19% of studies) in

XAL framework suggests a major gap in the areas like sentiment analysis and hate speech detection. It need more of contextual explanations which post-hoc methods fail to provide. Similarly the limited representation of TS (9.5% of studies) could also restrict the potential of XAL in finance, energy sectors. Where the need of explainability and interpretability is much needed. The imbalance in XAL research states while real-time and high stake processes demand a higher transparency. But obtaining the labeled training data in these domains is still a huge challenge.

B.4.1 Active Learning Strategies

Uncertainty sampling domination in the studies (75% of papers) underscores its effectiveness in reducing the labeling costs. In particular, medical imaging which is expensive and time consuming for labeling. Entropy based Uncertainty sampling prioritize the labeling of ambiguous samples which may not align with the intuitive labeling of the samples. Medical imaging focus must be on critical regions rather than just mathematical model uncertainty. To address this issues, novel methods like GradCAM [27] and DICE [35] may incorporate activation maps to identify image regions that can be domain specific. Even so their adoption is quite limited. Few studies suggests enhanced AL interpretability by using classing ambiguity indices. Others find that the human feedback on the model introduces more computation and slows down the AL processes. As the need for efficiency and explainability suggests adaptive AL strategies that uses uncertainty sampling in beginning and then,transitioned into explanation based sampling for further refinement. These hybrid AL strategies would provide balance between uncertainty, diversity, and explainability for optimal annotation efficiency.

B.4.2 Explainable AI Strategies

Post-hoc methods such as LIME and SHAP dominates XAL research because of their model agnostic approach. A novel adaptation like AL-DLIME [21] enhances the reproducibility with the use of the ridge regression as the surrogate model specifically catered for AL provides local explanations limiting its global insights. In contrast to these methods the usage of intrinsic algorithms such as decision trees[36], Logistic Regression[7] provide a inherent transparency but struggles on complex tasks. The use of visual explanations such as GradCam, saliency maps, anomaly maps, and pixel level segmentation provide significant explanations in IPCV. But they rely on the annotator‘s visual literacy which favor experts over laymen. The use of conversational XAI based on GPT-based explanations[34] can enable interactive querying that answers queries like "why are we using these data points?". SHAP, and LIME usage is 35% in the literature because of their popularity for explanations. DL explanations such as PETER, INNvestigate, and Encoder-Decoder shows that explanations for DL can be enhanced by use of niche algorithms. SHAP and LIME are not used in IPCV explanations, and a wide range of methods that are used in the papers states that explainability methods in XAL are very domain and use case-specific approaches.

B.4.3 ML & DL Algorithms with Performance Comparison

The integration of intrinsic models like Decision Trees, Logistic Regression, and black-box models into XAL highlights the challenge of balancing transparency and performance. The hybrid methods like LR with CNN and MLP attempts to bridge the gap. But they often create gray boxes in which interpretability can differ. Models must be designed carefully to avoid them becoming closed gray boxes. Traditional ML models are still relevant because of their transparency. Although they cannot provide good performance for complex tasks like medical image analysis. Whereas DL algorithms provide a better performance in complex tasks but are not transparent. In this literature review, 30% ML, 60% DL, and 10% Hybrid algorithms are used. DL algorithms have highest percentage of usage as they tend to achieve high performance which is important for AL applications. Although, XAL have a greater advantage over traditional AL for reducing labeling costs. While only 35% of the studies directly compare them and don't provide a robust answer if their is an increased performance.

B.4.4 Evaluation Metrics

In this literature review, a diverse range of evaluation metrics have been used to emphasize the technical performance alongside human centric explainability. While most of the studies use traditional methods like Accuracy (40%), F1-Score (55%), AUC-ROC (40%) for model performance assessment. These methods provide a good evaluation but they might not be able to comply with the nuances of XAL framework. Novel metrics like Uncertainty Bias and mIoU provide the evaluation for domain specific applications. Although the inconsistent usage in the studies complicate the cross domain comparisons. Human centric metrics like Degree of Agreement[16][42] and User Trust Survey[7] reflect the AI interpretability. Human opinions vary and lack of standardized tests which make it challenging to compare different studies. In contrast to the evaluation of performance of XAL, evaluating the quality of XAL explanations is challenging. As it involves balancing technical metrics alongside subjective and real-world usefulness of the explanations. Technical metrics like Faithfulness[21][24], Stability[21], BLEU, ROUGE scores [56], and Monotonicity[24] will ensure that explanations aligns with the AI's action. However they might not provide good explanation in real-time scenarios. For example, a faithful explanation can be provided by AI for medical imaging domain that may not align with the doctor's preconceived diagnosis.

A more comprehensive evaluation framework is needed to evaluate the explanations. The SAFE framework [44] integrates Sustainability, Accuracy, Fairness, and Explainability providing a model-agnostic framework to evaluate trustworthiness of AI models. Particularly black-box models to balance performance with interpretable decision making. Group Shapley [55] with robust significant testing framework extends traditional shapely values for the feature importance assessment in XAI. From this literature review, it can be inferred that a single evaluation metric will not suffice for XAL evaluation. Rather, a multifaceted approach of quantitative performance, explanation quality, and human centered interpretability metrics are needed. They

capture both technical and real-world explanations across diverse domains and user levels.

B.4.5 Research Gap

Despite increasing research on XAL, a gap remains in evaluating explanation quality within AL workflows. Existing studies often treat AL, and XAI separately. This lacks insights into how explanations evolve across AL cycles. The absence of standardized evaluation framework makes it unclear whether explanations remains stable and reliable over iterations. Additionally, comparative analysis of different XAI methods within AL workflow is limited that leaves questions unanswered regarding effectiveness in explaining selected samples. Addressing these gaps will be essential for ensuring explanations in AL are interpretable, robust and informative across learning process.

B.5 Conclusion

This literature review emphasizes the significance of Explainable Active Learning (XAL) in enhancing the interpretability of ML models and optimizing the data labeling costs. The analysis showcase that uncertainty sampling mainly entropy based uncertainty sampling dominates the AL strategies used. While a variety of XAI methods are used, most of them are post-hoc especially SHAP and LIME which are widely used for explainability. A wide variety of ML and DL are used. Being consistent with post-hoc, DL models tops in the usage. However, a gap can be observed in evaluating the quality of explanations generated by XAL over each AL cycle. Existing metrics of XAI like monotonicity, stability along side of human evaluation provide a partial insights into XAL explanations. Furthermore the popularity of post-hoc methods raises the concerns on their applicability across diverse domains. Further research can focus on design and development of a comprehensive assessment methodology that secure the explanations of XAL to be reliable, interpretable and align with real-world requirements.

B.6 Appendix:A

TABLE B.6: Techniques, Strategies, Contributions observed in selected papers with focus on XAI, and AL methods.

Author	Main Contributions	AL Strategy	XAI Strategy	Algorithms Used	Evaluation Metrics
Ghai et al. [16] [2021]	The use of Explainable Active Learning as the interface for an Active Learning algorithm. Instead of requesting the labels without justification for selected instances, the model presents both the prediction and its explanation to request feedback from humans.	Entropy-Based Uncertainty sampling	Local feature importance visualization using Bar charts. Explanation Quality: Annotator Agreement, i.e. alignment with ground truth.	Logistic Regression(L2 Regularization)	Accuracy, F1- Score
Phillips et al. [45] [2018]	Integration of LIME with uncertainty sampling, and to generate explanations for queried data points. Introduces "Uncertainty Bias" to analyze subgroup-specific uncertainties and tracks the resolutions during AL process. MCC highlights improved interoperability and transparency in AL workflows.	Uncertainty Sampling	Local, Extrinsic explanations using LIME.	Logistic Regression & Ada-boost	Matthews correlation Constant(MCC) & Uncertainty Bias(Novel Metric)

Luo et al. [34] [2023]	Implementation of a novel framework of XAL in text classification that encourages classifiers to justify their conclusions to dive into unlabeled data for which they cannot provide any reasonable explanation. The use of pre-trained Bi-directional Encoder & Uni-Directional Decoder for both AL and XAI strategies is noteworthy.	Weighted sum of Entropy based uncertainty sampling and explanation score from decoder.	Intrinsic , Local explanations using Pre-Trained uni-directional decoder trained using GPT-based explanations generator. Explanations Quality: Human evaluation, Ranking Accuracy (good vs bad explanations).	Pre-Trained FLAN-T5-Large: encoder-decoder architecture, Transformer-based encoder for classification.	Macro F1-score, Ranking Loss.
Cantürk and Aydoğan [7] [2023]	This study proposes a preference elicitation framework that maximizes user information by integrating AL with user interaction. This framework presents the user with preference-based explanations and collects the feedback to personalize the model's decision making process.	Uncertainty-based Sampling, K-Medoids Clustering, Clustered Uncertainty-based Sampling, Uncertainty-based Clustering Sampling, & Most Uncertain Cluster Sampling	Intrinsic Explainability with feature weights of LR, Global Explanations, Explanation Quality: Explanation Accuracy, user trust(Survey)	Logistic Regression(LR), & Semi-Supervised Learning	F1-Score, & Matthews correlation coefficient(MCC)
Mandalika and Nambiar [35] [2024]	A novel XAL framework called SegXAL is introduced for semantic segmentation in self driving scenarios. It proposes a module Explainable Error Mask (EEU) which integrates uncertainty AL and proximity aware explainability to improve annotation efficiency.	Entropy based uncertainty & DICE- based selection	Technique: GradCAM, & Entropy Heat Map, Intrinsic, Local Explanations	Semantic segmentation Deep Neural Network: U-Net, Depth Estimation Models: MiDaS, DINOv2	Mean Intersection over Union (mIoU) DICE coefficient & Standard Deviation

Zhang et al. [56] [2024]	This study addresses the challenge of obtaining explanation ground truth under budget constraints. It proposes framework (ActiveEXR) which integrate two strategies to improve the data labeling for explainable recommendation system. It demonstrates that selection of key samples strategically enhance performance with reducing annotation costs.	Entropy-based Uncertainty sampling, Influence-based sampling, Acquisition Function: Linear combination of both samplings.	Intrinsic-explainability using PETER, Explanation Quality: BLEU & ROUGE scores.	PETER: Personalized Transformer for Explainable Recommendation.	RMSE (Root Mean Squared Error): Rating prediction accuracy.
Mark Chignell[36] [2020]	This paper proposes an Interactive ML framework using AL for cybersecurity threat detection. This framework integrates feature-relevance based explainability to improve trust and reduce false alarm rates via AL. It evaluates expert-AI compatibility in an iterative model for enhanced interpretability.	Pool-Based AL with Entropy based uncertainty sampling.	Intrinsic, Global Explanations, Feature importance scores from Decision Trees.	LightGBM (Gradient Boosting Trees) as learner, Random Forest & AdaBoost for comparison.	AUC, Confusion Matrix, Human-Labeling performance.
Adhane et al. [2] [2022]	This paper use Monte Carlo Dropout as adversary for uncertainty sampling for AL. It employs B-LRP for visual explanations, that enhances the interpretability of model, and improve classification accuracy by 4%.	Uncertainty-sampling using Monte Carlo Dropout.	Intrinsic, Local, Post-Hoc explanations using saliency map by Bayesian Layer-wise Relevance Propagation (B-LRP).	Fine Tuned VGG16: Deep Neural Network with 16 layers.	Non-Rejection Accuracy, Precision, Recall, F1-Score.

Nishimura et al. [42] [2023]	This paper proposes ALLFA, an active learning framework for integrating XAI-based feature selection. It uses SHAP for feature importance with feature augmentation for enhanced labeling. Improves classification accuracy over traditional AL for better interpretability.	Margin Sampling using uncertainty. Along with labels, feature information is also given.	Extrinsic, Local explanations using SHAP values for feature importance explanations. Explanation Quality: Degree of agreement, higher agreement: more intuitive explanations.	Uni-LSTM, Attentional Neural Network.	F-1 Score, AUC, Degree of agreement between user-selected and model-selected features.
Baur et al. [5] [2020]	This paper introduce Explainable Cooperative Machine Learning that speed up the annotation of the social signals in large multi-modal databases by incorporating AL (speed) and XAI(interpretability). It proposes a two staged confidence-based AL strategy with LIME and iNNvestigate for enhanced labeling.	Uncertainty sampling: Confidence Based.	Extrinsic, Local explanations using LIME, iNNvestigate: Deep Learning Explanation.	Linear SVM, CNN, L2- Logistic Regression.	Classification Accuracy, Confusion Matrix, Unweighted Average of AUC.
Hemelings et al. [19] [2020]	This study proposes automated glaucoma detection through deep learning by using AL to reduce labeling costs. It provides model decision insights via saliency maps, ultimately achieving high performance with a limited dataset.	Uncertainty sampling, Random sampling (Baseline).	Local and Extrinsic visual explanations using saliency maps.	ResNet-50 CNN	AUC, Sensitivity, Specificity.

Andresini et al. [3] [2021]	This paper presents INSOMNIA: a semi-supervised concept drift aware network intrusion detection system. It integrates AL, Explainability(feature importance tracking). It adapts DNN models incrementally that improves performance and insights in attack behavior.	Uncertainty sampling, Replaced human annotator (label estimator) with NC (Nearest Centroid Neighbor classifier).	Global, Extrinsic explanations with DALEX (moDel Agnostic Language for Exploration and eXplanation), tracks feature importance over time.	DNN (Deep Neural Network)	TESSERACT framework, F1-Score, AUT(F1) (Area under Time aware F1-curve)
Huang et al. [23] [2023]	This paper proposes a Ac-GRN that combines active deep learning and GNN for heat load forecasting. It selects most informative and representative samples to train GNN with bidirectional recurrent conditions. Extensive experiments have been conducted and compared with 11 best methods and gives the best performance.	Least confidence uncertainty sampling	Main Explainability: Global, Intrinsic Correlation based Attribution (identify influential feature), Supporting Explainability: Local, Intrinsic Attention Mechanism(assign importance to spatial temporal data)	Graph Neural Networks: Active Graph Recurrent Network(Ac-GRN).	RMSE,MAE
Teso and Kersting [50] [2019]	This paper introduces Explanatory Interactive Learning(XIL), and proposed CAIPI framework. It is the first XIL method that integrate explanation feedback into AL. It enables query-based explanations which opens black-box of AL and turns into cooperative learning process which increases user trust, performance of the mode.	Pool-based AL with uncertainty sampling.	Local, Extrinsic explanations using LIME. Explanation Quality: F1-score for model vs human explanations.	Multilayer Perceptron, Logistic Regression, Support Vector Machines (L1,L2 Regularized).	Classification Accuracy, User Trust through questionnaires.

Kuelzer et al. [28] ([2021])	This paper presents a DNN for latency prediction with active learning framework. It utilizes passive probing data that enhances classification accuracy through the use of SHAP values in autonomous driving applications.	Pool-based certainty sampling.	Local and Global, Extrinsic explanations using SHAP.	DNN with custom architecture, Random Forest (RF) and Light Gradient Boosted Machine (LGBM) as comparisons.	Metric Accuracy (ACC), AUC, F1-Score.
Kriznar et al. [27] [2023]	This paper introduces $GradCam_{avg}$ an active learning strategy to guide sample selection based on structural similarity index measure (SSIM) using GradCAM activation maps that enhances explainability in AL workflow and reduce annotation effort.	Novel sampling strategy $GradCam_{avg}$, compared with random, uncertainty sampling.	Intrinsic, Local explainability using GradCam which is Post-hoc technique that visualizes image region influencing predictions.	Not mentioned, but applicable for CNN-based models.	AUC ROC with different AL strategies and with, incremental, batch learning.
Holm and Macedo [21] [2023]	This paper presents a novel model, AL-DLIME (Active Learning- Deterministic Local Interpretable Model-Agnostic Explanations) that enhances explainability by incorporating AL in XAI algorithm. It provides stable, faithful explanations by outperforming traditional models in interpretability and accuracy in Healthcare applications.	Uncertainty based sampling using Logistic Regression.	Extrinsic, Local, Post-hoc, Deterministic explanations using AL-DLIME. Surrogate model in DLIME: Ridge Linear Regression, Explanations Quality: Faithfulness, Stability (Jaccard's Distance), Single and Incremental deletion.	Random Forest: black-box model, Decision Trees: Baseline (Transparent model), to find if black-box model needed for better accuracy.	Accuracy, F1-Score.

Kalakoti et al. [24] [2024]	This paper proposes an Explainable Active Learning (XAL) framework for Iot botnet detection using SHAP, LIME. It compares Uncertainty sampling using classification entropy, and Query-by-Committee using disagreement-based strategies. Namely vote entropy, consensus entropy, max disagreement. It evaluates the quality of the given explanations.	Uncertainty sampling, and Query-by-committee.	Local, Extrinsic, Post-Hoc explanations using SHAP and LIME. Explanation Quality: Faithfulness, Monotonicity, Max Sensitivity.	Extreme Gradient Boost (XGBOOST) classifier, Feature selection: Pearson correlation	F1-score, Accuracy, Precision, Recall
Lee and Li [31] [2024]	This paper proposes IDAL-FIM framework using interpretable deep learning for flood mapping. For interpretability, it proposes class ambiguity indices(BPR,MDF) along with five AL strategies. BALD, Entropy, Margin i.e. uncertainty, Density, Random sampling. Demonstrate their impact on model performance.	Uncertainty based: Predictive, Model uncertainty and Density-based sampling: PCA, K-Means, Random Sampling: Baseline.	Intrinsic, Local, Model specific explanations using Boundary Pixel Ratio (BPR), 2-D Density Plots. Explanation Quality: Spearman Rank Correlation Coefficient.	Deep Learning model: U-Net with MC-Dropout.	F1-Score
Rožanec et al. [46] [2022]	This paper integrates AL, XAI in visual inspection framework that enhances performance and interpretability with uncertainty management and user guidance.	Uncertainty sampling.	Intrinsic, Local explanations by GradCAM, Anomaly maps, Nearest labeled image.	ResNet-18, Unsupervised Anomaly detection: DRAEM, GAN: Defect image generation	AUC ROC, Precision, Recall, F1-score.

B.7 Appendix:B

ID	Research Domain	References	Remarks
1	Image Processing and Computer Vision (IPCV)	1. Baur et al. [5] 2. Hemelings et al. [19] 3. Adhane et al. [2] 4. Rožanec et al. [46] 5. Kriznar et al. [27] 6. Lee and Li [31] 7. Mandalika and Nambiar [35]	The papers mentioned here deals with Medical diagnosis, Autonomous driving, and Satellite Imagery that deals with high stake applications. All the applications mentioned above needs an enormous pool of labeled data, which is difficult to be obtained. AL with explainability will be the best option in this scenario.
2	Explainable Active Learning (XAL)	1. Phillips et al. [45] 2. Teso and Kersting [50] 3. Mark Chignell [36] 4. Ghai et al. [16] 5. Nishimura et al. [42] 6. Holm and Macedo [21]	The papers mentioned here does not deal with any research domain or application. They propose new frameworks for XAL applications that are not tied to a single domain.
3	Natural Language Processing (NLP)	1. Baur et al. [5] 2. Luo et al. [34] 3. Cantürk and Aydoğan [7] 4. Zhang et al. [56]	The papers mentioned here deals with the NLP applications with proposed frameworks that can label the data with limited resources. As NLP algorithms like BERT needs large amount of data for better performance.

ID	Research Domain	References	Remarks
4	Cyber Security (CS)	1. Mark Chignell [36] 2. Andresini et al. [3] 3. Kalakoti et al. [24]	The papers mentioned here deals with the implementation of the frameworks that can detect Network intrusion and Data Exfiltration attacks. It is difficult to label every record in a busy network, XAL provides a simple pathway for better performance.
5	Time Series (TS)	1. Kuelzer et al. [28] 2. Huang et al. [23]	The papers mentioned here deals with frameworks that deals with huge amount of historical data which forecasts the predication for real-time systems using DL modes. XAL provides a good explainability for closed black-box models.