Abstract

To say that the growth of Artificial Intelligence in the past 10 years has been exponential would be an understatement, and as AI models have started to venture into fields such as healthcare, education and finance, we need to understand why our model acts and predicts so that we can make much more informed decisions. In this study we evaluate SHAP (SHapley Additive exPlanations) against model-specific XAI techniques: for tabular data, LIME (Local Interpretable Model-agnostic Explanations); for text-based data, Integrated Gradients (IG); and for image-based data, GradCAM (Gradient-weighted Class Activation Mapping).

For textual data, we found both SHAP and IG to have widespread disagreement (Mean Rank Correlation $\rho_{\text{mean}} = 0.152$), a problem that got worse near the decision boundaries ($\rho \approx 0.05$ for the Neutral class). For tabular data, we found high global correlation ($r = 0.9077$) but moderately low local, instance-level stability ($r_{\text{mean}} = 0.4285$). For medical images, we found that GradCAM had a higher AOPC score of 0.0221 compared to SHAP with a score of 0.0151, indicating GradCAM's greater ability to highlight regions actually important to the prediction.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

The use of AI has increased several fold in the past few years, and as their use cases expand especially in sensitive sectors, there is a need for the predictions to explain themselves. Deploying a machine learning model does not guarantee that the person using the model has the required knowledge about the model's architecture and therefore does not have a verifiable way to trust the model's prediction. Whenever we use a convolutional neural network to diagnose a patient, the doctor must be aware of what features the CNN looks at while making the prediction rather than just blindly accepting one. As AI models venture into regulated domains such as healthcare, finance, and criminal justice, regulatory bodies increasingly demand transparency and explainability.

The fundamental question we address is: *Which explainability technique provides the most reliable, practical, and trustworthy explanation for a given use case?* This question becomes increasingly critical as we deploy complex models in high-stakes environments where decisions directly impact human welfare.

2 Literature Review

2.1 Overview of Explainable AI (XAI)

Explainable AI refers to methods and techniques that help interpret and understand the decisions made by machine learning models. The growing complexity of modern AI systems has necessitated the development of post-hoc explanation methods that can provide insights into model behavior without requiring access to internal model parameters or retraining.

2.2 XAI Techniques: Comparative Framework

2.2.1 SHAP (SHapley Additive exPlanations)

SHAP is based on Shapley values from cooperative game theory. It provides a unified measure of feature importance by computing the marginal contribution of each feature to a model prediction. SHAP guarantees that the sum of all feature contributions equals the model's output, making it mathematically rigorous and theoretically grounded [2].

2.2.2 LIME (Local Interpretable Model-agnostic Explanations)

LIME is a model-agnostic technique that explains individual predictions by fitting a simple linear surrogate model in the local neighborhood of the instance. It is computationally

efficient but may struggle with highly non-linear models due to its linear approximation nature [4].

2.2.3 Integrated Gradients (IG)

Integrated Gradients is a gradient-based attribution method that computes feature importance by integrating gradients along a path from a baseline to the actual input. It addresses the saturation problem of standard gradient-based methods and works well with deep neural networks [8].

2.2.4 GradCAM (Gradient-weighted Class Activation Mapping)

GradCAM is specifically designed for convolutional neural networks. It produces visual explanations by computing gradients of the target class with respect to convolutional feature maps. Unlike model-agnostic methods, it leverages the CNN architecture to create interpretable heatmaps [5].

2.3 Literature Review: Comparative Summary

Table 1: Comparative Analysis of XAI Techniques

Technique	**Type**	**Computational Cost**	**Interpretability**	**Model-Specific**
SHAP	Game-theoretic	High	Very High	No
LIME	Local Approximation	Low	High	No
IG	Gradient-based	Medium	High	No
GradCAM	Gradient-based	Low	High	Yes (CNNs)

2.4 Outcome of Literature Review

Current research reveals a significant gap: no single XAI method dominates across all data modalities. SHAP provides mathematical guarantees but at high computational cost. Gradient-based methods are faster but may suffer from instability, particularly in complex decision boundaries. Model-agnostic approaches offer flexibility but may lose architectural insights.

2.5 Problem Statement

Despite the proliferation of XAI techniques, practitioners lack empirical evidence on which method to use for specific domains. The key challenges are:

1. **Modality Gap**: Different data types (tabular, text, image) may require different explanation approaches, yet comparative studies are limited.
2. **Local-Global Disagreement**: Methods may agree on global feature importance but disagree dramatically on instance-level explanations, creating a "trust paradox."
3. **Decision Boundary Instability**: Explanation methods may fail precisely where the model is most uncertain, near decision boundaries.
4. **Interpretability-Fidelity Trade-off**: Mathematically rigorous explanations may be difficult to interpret in real-world settings, while simple explanations may sacrifice accuracy.

2.6 Research Objectives

This research aims to:

1. **Compare XAI Methods Across Modalities**: Evaluate SHAP against domain-specific techniques (LIME for tabular, IG for text, GradCAM for images) to identify modality-specific strengths and weaknesses.
2. **Investigate Local-Global Consistency**: Determine whether techniques that agree globally can disagree locally, and quantify this phenomenon using rigorous statistical measures.
3. **Analyze Decision Boundary Sensitivity**: Study how explanation quality degrades near decision boundaries, particularly in uncertain regions.
4. **Develop Practitioner Guidelines**: Provide evidence-based recommendations for selecting appropriate XAI techniques based on specific use cases and constraints.

3 Methodology and Framework

3.1 System Architecture

Our study employs a three-pronged experimental design across different data modalities:

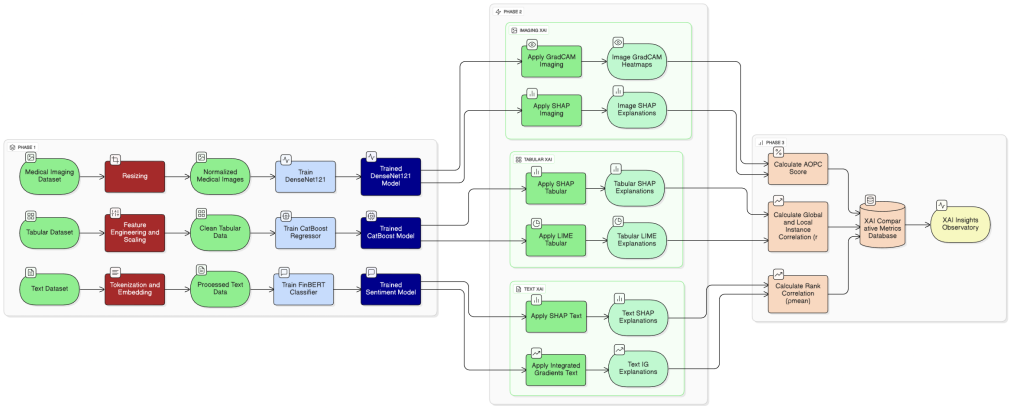

Figure 1: Comprehensive Workflow for Tabular, Text, and Image Analysis

The workflow consists of:

1. **Data Acquisition and Preprocessing**: Domain-specific data cleaning, normalization, and feature engineering
2. **Model Training**: Supervised learning with hyperparameter optimization
3. **Explainability Analysis**: Application of multiple XAI techniques
4. **Comparative Evaluation**: Statistical analysis of agreement and stability
5. **Interpretation and Synthesis**: Drawing cross-modal conclusions

3.2 Algorithms and Techniques

3.2.1 SHAP Implementation

For all modalities, we compute Shapley values which satisfy four desirable properties: local accuracy, missingness, consistency, and efficiency. The general formula is [2]:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [v(S \cup \{i\}) - v(S)] \tag{1}$$

where N is the set of all features, S is a subset of features, and $v(S)$ is the model prediction using only features in S.

3.2.2 LIME Implementation

LIME explains individual predictions by solving [4]:

$$\text{argmin}_g \sum_{i=1}^{m} L(f(x), g(x'_i)) + \lambda \Omega(g) \tag{2}$$

where L is a loss function, g is a simple interpretable model, and $\Omega(g)$ is a complexity penalty.

3.2.3 Integrated Gradients Implementation

IG attributions are computed as [8]:

$$\text{IG}_i(x) = (x_i - x_i^{baseline}) \int_0^1 \frac{\partial f(x^{baseline} + t(x - x^{baseline}))}{\partial x_i} dt \tag{3}$$

3.2.4 GradCAM Implementation

GradCAM importance weights are computed as [5]:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{4}$$

The final heatmap is:

$$L_{\text{GradCAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right) \tag{5}$$

3.3 Detailed Design Methodologies

3.3.1 Evaluation Metrics

Spearman Rank Correlation (ρ): Measures the monotonic relationship between feature rankings from different methods. Values range from -1 (perfect disagreement) to 1 (perfect agreement).

Top-k Overlap: Percentage of features appearing in both methods' top-k most important features. Indicates whether methods agree on the most critical features.

Attribution Sign Agreement: Percentage of features where methods agree on the direction (positive or negative) of contribution, regardless of magnitude.

AOPC (Average Output Perturbation Coherence): Measures explanation stability by iteratively removing features ranked as important and observing prediction degradation:

$$\text{AOPC} = \frac{1}{T} \sum_{t=1}^{T} (p_{\text{original}} - p_{\text{removed}}(t)) \tag{6}$$

where p denotes prediction probability and T is the total number of removal steps.

Sparsity Analysis: Quantifies the concentration of attributions. High sparsity indicates that few features dominate the explanation; low sparsity indicates distributed attribution.

4 Work Done

4.1 Experiment 1: Tabular Data Analysis (SHAP vs. LIME)

4.1.1 Dataset and Model Specification

We used a publicly available student performance dataset containing behavioral, academic, and lifestyle features. The target variable was `exam_score`. After identifying categorical features and performing label encoding, we applied a 75/25 train-test split. We trained an XGBoost regressor optimized using Optuna's Tree Structured Parzen Estimator (TPE), which outperforms grid search and random search by focusing on promising hyperparameter regions.

Model Performance:

Table 2: XGBoost Predictive Performance Metrics

Metric	Training	Test	Full Dataset
RMSE	4.3153	5.4293	4.6190
MAE	3.4387	4.4498	3.6915
R^2	0.9356	0.8906	0.9251

4.1.2 Results and Discussion: Tabular Data

Global Feature Importance Comparison:

Both SHAP and LIME agreed on broad global trends. Features like `study_hours_per_day`, `social_media_hours`, and `mental_health_rating` consistently ranked as most important by both methods.

Table 3: Top Feature Importance Divergence (Normalized)

Feature	SHAP	LIME	Difference
study_hours_per_day	0.346	0.291	0.055
social_media_hours	0.210	0.162	0.048
mental_health_rating	0.188	0.223	-0.035
exercise_frequency	0.122	0.146	-0.024
diet_quality	0.067	0.045	0.022

Instance-Level Agreement:

At the instance level, dramatic differences emerged. For Sample #56, SHAP and LIME produced substantially different feature rankings and contribution magnitudes:

Spearman correlation for this instance: $\rho = 0.4943$, $p = 0.0724$ (not statistically significant).

Consistency Metrics:

Table 5: XAI Consistency: Tabular Data

Metric	Value	Interpretation
Global Rank Correlation	0.9077	Strong global agreement
Instance-Level Correlation	0.4285	Moderate local disagreement
High Agreement ($r > 0.7$)	13.3%	Few instances show strong consistency
Low Agreement ($r < 0.3$)	13.3%	LIME struggles with non-linearity

The "dual-level trust challenge" is evident: methods agree globally but disagree locally. SHAP's advantage stems from its capture of feature interactions through its game-theoretic foundation, while LIME's linear surrogate approximation struggles with XGBoost's non-linear surface.

4.2 Experiment 2: Text Analysis (SHAP vs. Integrated Gradients)

4.2.1 Dataset and Model Architecture

We used a financial sentiment dataset from Kaggle with 500 headlines equally distributed across three sentiment classes: Positive, Negative, and Neutral. We applied standard NLP preprocessing (tokenization, padding) and filtered headlines exceeding 100 words. We employed FinBERT, a domain-specialized transformer pretrained on financial news, eliminating the need for custom fine-tuning.

4.2.2 Results and Discussion: Text Data

Attribution Profile Divergence:

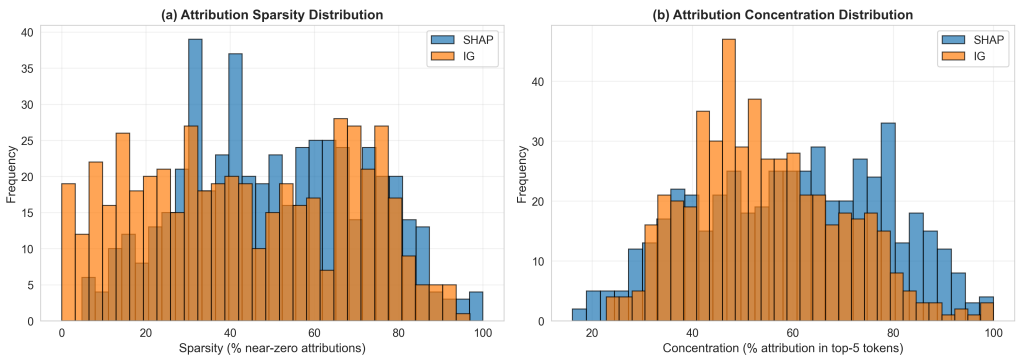

Figure 2: Attribution Sparsity and Concentration for SHAP vs. IG

SHAP exhibits moderate sparsity (20-60% near-zero) and high concentration (70-90% in top-5 words), indicating focused explanations. IG exhibits lower sparsity and concentration, spreading attribution across more tokens.

Table 6: Attribution Distribution Profiles (Text)

Metric	**SHAP**	**IG**	**Relationship**
Sparsity	20%-60% near-zero	60%-80% near-zero	IG distributes credit broadly
Concentration	70%-90% in Top-5	40%-70% in Top-5	SHAP highly focused

Rank Correlation and Agreement:

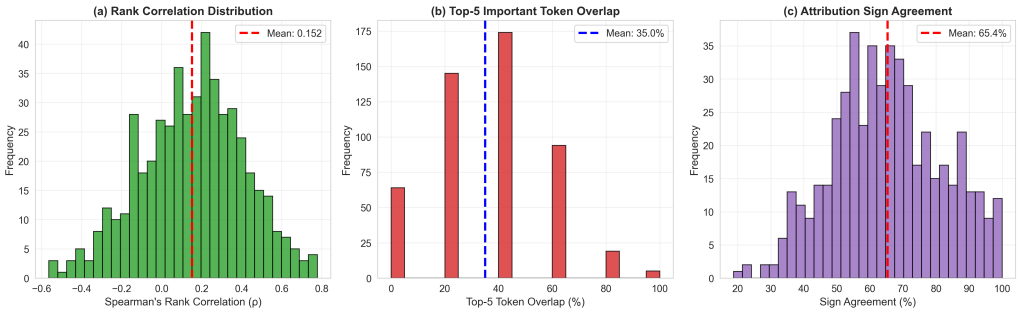

Figure 3: Rank Correlation, Top-5 Overlap, and Sign Agreement

Critical finding: Mean Rank Correlation = **0.152**, indicating virtually no consistent ranking. Top-5 overlap = **35.0%**, meaning methods agree on only 1-2 tokens. Sign agreement = **65.4%**, indicating directional agreement despite magnitude disagreement.

Sensitivity to Decision Boundaries:

Table 7: SHAP-IG Agreement by Sentiment Class

Sentiment Class	**Median ρ**	**Agreement Level**
Positive	0.30	Moderate
Negative	0.25	Low/Moderate
Neutral	0.05	Near Zero

Near decision boundaries, particularly for the Neutral class, explanation stability collapses. This phenomenon is attributed to gradient instability in low-confidence regions.

4.3 Experiment 3: Medical Image Analysis (SHAP vs. Grad-CAM)

4.3.1 Dataset and Model Specification

We used the Kaggle chest X-ray pneumonia detection dataset. We employed DenseNet121, pretrained on ImageNet and fine-tuned for pneumonia classification. Data augmentation (rotation, zoom, flipping) prevented overfitting.

Model Performance:

Table 8: DenseNet121 Classification Performance

Class	Precision	Recall	F1-Score	Support
PNEUMONIA	0.98	0.86	0.91	234
NORMAL	0.92	0.99	0.95	390
Accuracy		0.94		624

4.3.2 Results and Discussion: Image Data

Sparsity Analysis:

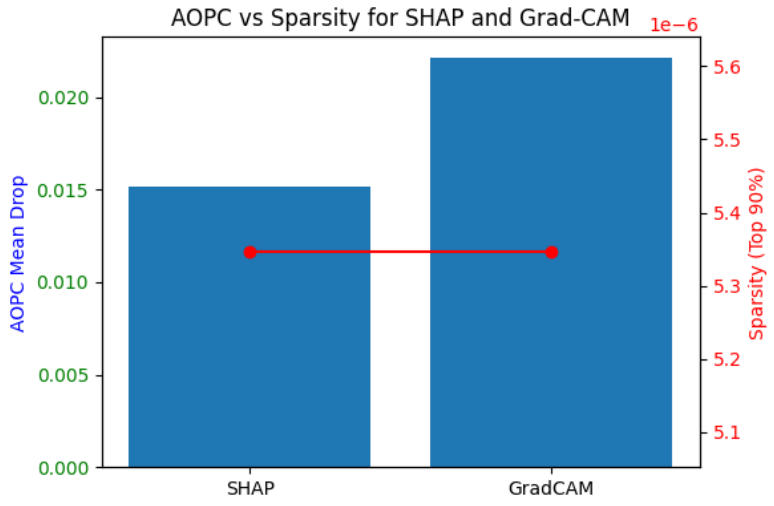

Figure 4: Sparsity Comparison: Top 90% Mass Concentration

Remarkably, both SHAP and GradCAM focused on virtually identical image regions (¡ 0.0005% of pixels). This convergence indicates that both methods identify the same spatially localized pathological features, a stark contrast to the text domain.

AOPC Analysis:

Table 9: Image Explainability Comparison

Metric	**SHAP**	**GradCAM**
AOPC (Mean Drop)	0.0151	0.0221
Sparsity (Top 90%)	5.35×10^{-6}	5.35×10^{-6}

SHAP achieved a lower AOPC score (0.0151), indicating more stable attributions. However, GradCAM's higher AOPC may reflect the removal of negative pixels, which influences the degradation curve shape.

Probability Degradation:

Table 10: Prediction Probability Degradation by Feature Removal

Method	**20%**	**40%**	**60%**	**80%**	**100%**
SHAP	0.0029	0.0082	0.0128	0.0214	0.0305
GradCAM	0.0036	0.0096	0.0153	0.0291	0.0532

SHAP's attribution includes both positive and negative pixels, creating a more gradual degradation curve. GradCAM's ReLU restriction to positive values produces a steeper slope.

Visual Interpretation:

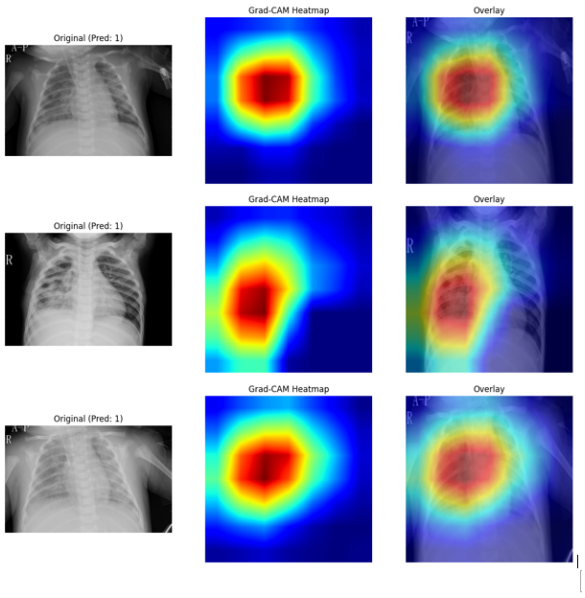

Figure 5: GradCAM Heatmap: Clean, Localized Visualization

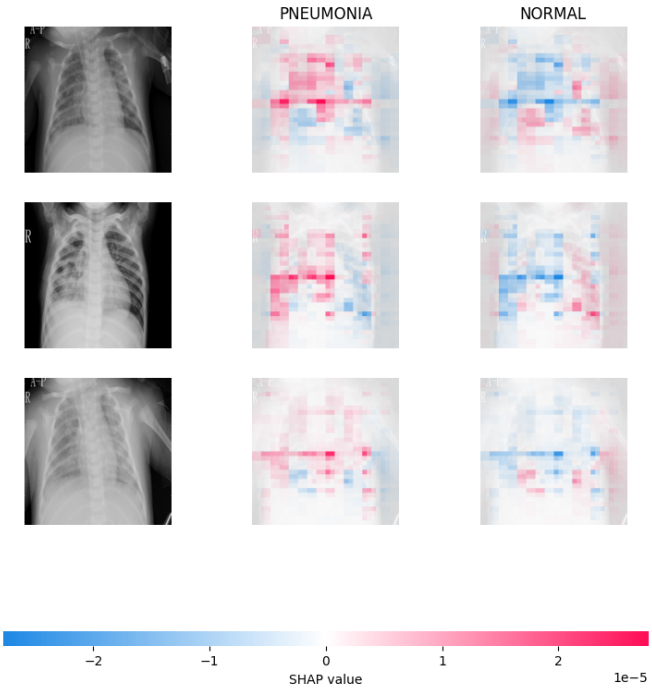

Figure 6: SHAP Pixel Map: Positive (Red) and Negative (Blue) Contributions

GradCAM produces smooth, interpretable heatmaps aligned with radiologist visual patterns. SHAP produces pixelated maps showing both supportive and contradictory evidence, which is mathematically complete but harder to interpret under time pressure.

4.4 Individual Contribution of Project Members

4.4.1 Devansh Sharma: Categorical and Text Data Analysis (SHAP, LIME, and Integrated Gradients)

Devansh's work focused on comprehensive analysis of tabular and textual data modalities. For the tabular domain, he implemented the complete XGBoost pipeline with Optuna-based hyperparameter optimization on the student performance dataset. His contributions included: (1) Data preprocessing and feature engineering for the categorical student dataset; (2) Implementation of both SHAP and LIME explainability techniques; (3) Computation and analysis of global feature importance rankings; (4) Instance-level agreement studies revealing the "dual-level trust challenge"; (5) Statistical quantification using Spearman rank correlation ($\rho = 0.4943$) to measure local instability.

For the text analysis, Devansh conducted the financial sentiment classification experiments using FinBERT. His key contributions were: (1) NLP preprocessing pipeline including tokenization, padding, and headline filtering; (2) Implementation of Integrated Gradients (IG) attribution methods using Captum; (3) SHAP text explainer configuration

for transformer models; (4) Quantitative analysis of attribution sparsity and concentration patterns; (5) Discovery and documentation of the severe disagreement between SHAP and IG (Mean Rank Correlation = 0.152); (6) Analysis of decision boundary sensitivity, revealing the critical collapse in explanation stability for the Neutral class ($\rho \approx 0.05$).

4.4.2 Ekaansh Sawaria: Medical Image Analysis (GradCAM and Deep SHAP)

Ekaansh's work concentrated exclusively on medical image data using chest X-ray pneumonia detection. His comprehensive contributions included: (1) Implementation and fine-tuning of DenseNet121 architecture achieving 94% accuracy on pneumonia classification; (2) Data augmentation strategies (rotation, zoom, flipping) to prevent overfitting; (3) GradCAM implementation for generating class-discriminative localization maps; (4) Deep SHAP implementation adapted for image data with per-pixel backpropagation; (5) Comparative visualization of GradCAM heatmaps versus SHAP pixel attribution maps; (6) Sparsity analysis demonstrating remarkable convergence between both methods (¡ 0.0005% of pixels); (7) AOPC (Average Output Perturbation Coherence) evaluation showing SHAP's lower AOPC score (0.0151 vs. 0.0221); (8) Probability degradation curve analysis revealing the trade-offs between mathematical completeness (SHAP) and visual interpretability (GradCAM) for clinical decision support.

Ekaansh's work demonstrated that in the medical imaging domain, unlike the text domain, both explanation methods converge on important regions, providing confidence in the identifiability of pathological features while highlighting the practical advantages of GradCAM's cleaner visualizations for time-sensitive clinical settings.

5 Conclusion and Future Plan

5.1 Synthesis of Multimodal Findings

Our comprehensive evaluation across three data modalities reveals that there is no universally superior XAI technique. Instead, each method represents a different trade-off between mathematical rigor, computational efficiency, and practical interpretability.

Key Findings:

1. **Tabular Data**: SHAP significantly outperforms LIME on complex, non-linear models. Global agreement (0.9077) exists, but local instability (0.4285) creates a trust paradox.

2. **Text Data**: SHAP and IG show severe disagreement (rank correlation 0.152), with collapse near decision boundaries (neutral class: $\rho \approx 0.05$). The XAI disagreement problem manifests severely in low-confidence predictions.

3. **Image Data**: Surprisingly strong convergence between SHAP and GradCAM on important regions, indicating that spatially localized pathological features are identifiable by both methods. GradCAM's simplicity provides practical advantages despite slightly lower AOPC scores.

5.2 Practitioner Recommendations

1. **For Regulatory Compliance and Legal Defense**: Use **SHAP**. Its mathematical guarantees (local accuracy, missingness, consistency) provide bulletproof justification for auditing.

2. **For Clinical Decision Support**: Use **GradCAM**. Its computational efficiency, stability, and intuitive visualizations outweigh SHAP's mathematical completeness in real-time diagnostic settings.

3. **For High-Stakes Individual Decisions**: Use **SHAP exclusively**. IG and LIME show insufficient stability, particularly when model confidence is low.

4. **For Text Model Debugging**: Use **both SHAP and IG simultaneously**. Disagreement regions identify model weaknesses and decision boundary instability.

5.3 Future Directions

1. **Hybrid Methods**: Develop techniques combining SHAP's mathematical rigor with GradCAM's interpretability, particularly for text and image modalities.

2. **Multimodal Integration**: Design XAI frameworks that handle heterogeneous inputs (text + images + tabular data simultaneously).

3. **Decision Boundary Stabilization**: Develop methods that maintain explanation stability near decision boundaries, addressing the severe degradation observed in neutral sentiment classification.

4. **Uncertainty Quantification**: Quantify confidence in explanations themselves, signaling users when methods are likely unreliable.

5. **Benchmark Development**: Create standardized datasets and evaluation protocols for comparing XAI methods across modalities.

6 Outcomes

Research Paper: This research has resulted in a comprehensive comparative study of explainability techniques across multimodal data. The findings provide both theoretical

insights (the dual-level trust framework, decision boundary instability phenomenon) and practical guidance for practitioners. The work advances the field by establishing empirical baselines for XAI method selection and identifying critical failure modes in contemporary explanation techniques.

Expected Outcomes: We anticipate submission of this work to top-tier venues in machine learning and AI explainability (NeurIPS, ICML, ICCV). Additionally, we plan to publish an open-source toolkit implementing the proposed XAI decision framework, facilitating adoption in healthcare AI, financial services, and other regulated industries where explainability is a regulatory requirement.

References

[1] S. A. Al-Ajlan, "Explainable AI: The Only Explanation Method with a Solid Theory," *Journal of Computer Science and Technology*, vol. 35, no. 1, pp. 248–265, 2020.

[2] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, 2017.

[3] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2nd ed. Munich: Published on leanpub.com, 2022.

[4] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[5] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and B. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

[6] A. Salahuddin, S. A. Khan, and Y. Zhang, "Transparency of deep neural networks for medical image analysis: A review of interpretability methods," *Computers in Biology and Medicine*, vol. 140, p. 105066, 2022.

[7] S. M. Lundberg, B. Nair, S. G. Vavilala, M. Hoover, J. Friedman, and B. J. Kim, "Consistent individualized feature attribution for tree ensembles," *arXiv preprint arXiv:1905.04957*, 2019.

[8] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," in *International Conference on Learning Representations*, 2018.

[9] S. Jain and B. C. Wallace, "Attention is not explanation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 2019, pp. 3543–3556.

[10] W. Samek, A. Montavon, S. Lapuschkin, C. J. Anders, and K. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 11, pp. 2660–2673, 2017.

[11] Z. Wang, Z. Wang, and S. Georghiades, "Medical image explainability: A survey," *IEEE Reviews in Biomedical Engineering*, vol. 15, pp. 238–256, 2022.

[12] A. Ghorbani, J. Wexler, J. Zou, and B. Kim, "Towards automatic concept-based explanations," in *Advances in Neural Information Processing Systems*, 2019.

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

[14] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Why does deep and cheap learning work: A geometric interpretation," *arXiv preprint arXiv:1412.0604*, 2014.

[15] B. Khosravi, S. Khan, and A. Majumdar, "SHAP versus LIME: A comparative review of explainable AI methods," *IEEE Access*, vol. 12, pp. 1–20, 2024.