Comparative Evaluation of Explainable AI Techniques for Deep Learning in Image Recognition

Angelos Tzirtis, Christos Troussas, Akrivi Krouska, Phivos Mylonas, Cleo Sgouropoulou
Department of Informatics and Computer Engineering
University of West Attica
Egaleo, 12243, Greece
{cs171152, ctrouss, akrouska, mylonasf, csgouro}@uniwa.gr

Abstract. Deep learning models achieve high accuracy in image recognition but often function as "black boxes", making their decision-making processes difficult to interpret. Explainable AI (XAI) techniques aim to enhance transparency by providing insights into how deep neural networks reach their conclusions. This study presents a comparative evaluation of prominent XAI methods used in convolutional neural networks (CNNs), specifically Gradient-weighted Class Activation Mapping (Grad-CAM) and Saliency Maps. The techniques were applied to image classification tasks using benchmark datasets (ImageNet and CIFAR-10) and evaluated based on clarity, completeness, and trustworthiness. Our experimental results, conducted using VGG16 and ResNet50 architectures, demonstrate that Grad-CAM produces interpretable heatmaps that highlight relevant image regions, whereas Saliency Maps offer pixel-level feature importance with higher granularity but increased noise. The findings provide guidance for selecting suitable XAI methods depending on interpretability requirements, and we propose future research directions, including hybrid XAI approaches for improved transparency.

Keywords: Explainable AI, Deep Learning, Grad-CAM, Saliency Maps, Image Recognition, CNNs

1 Introduction

Deep learning has achieved significant success in image recognition applications, ranging from medical diagnostics to autonomous driving [1]. However, despite their high accuracy, convolutional neural networks (CNNs) lack transparency, making it challenging to understand the factors influencing their predictions. This "black-box" nature raises concerns in critical applications where trust, fairness, and accountability are necessary [2].

Explainable Artificial Intelligence (XAI) has emerged as a crucial research area aimed at improving the interpretability of deep learning models. By offering visual or numerical explanations, XAI techniques help users comprehend the decision-making

processes of neural networks. Among the most commonly used post-hoc explanation methods are Grad-CAM and Saliency Maps, which generate visual representations highlighting important features in an image [3].

Recent literature emphasizes the increasing demand for interpretability in AI models, particularly in domains such as healthcare, law, and finance, where explainability is essential for ethical and practical deployment [4][5]. Furthermore, regulatory frameworks such as the EU's AI Act highlight the need for transparency and accountability in AI systems [6]. Researchers have developed a wide array of techniques to address the opacity of deep learning models, with studies showing that visual explanations not only improve user trust but also help detect biases and errors in models [7][8].

Despite this progress, many existing techniques differ in their approach, performance, and applicability to real-world scenarios. While some methods focus on feature importance or perturbation-based strategies, others employ gradient-based visualizations for interpretability. Among these, Grad-CAM and Saliency Maps have received significant attention due to their visual nature and compatibility with popular CNN architectures [9][10].

This study aims to compare Grad-CAM and Saliency Maps in terms of their ability to provide meaningful explanations for CNN predictions. We evaluate these methods using standardized criteria—clarity, completeness, and trustworthiness—on two widely used architectures, VGG16 and ResNet50. By conducting experiments on benchmark datasets (ImageNet and CIFAR-10), we aim to offer insights into the strengths and weaknesses of each technique and provide recommendations for their practical application.

2 Related Work

The field of Explainable AI has grown substantially in recent years, with numerous techniques proposed to address the interpretability gap in deep learning models. Among these, Gradient-weighted Class Activation Mapping (Grad-CAM) and Saliency Maps have emerged as two of the most widely used visual explanation methods. Grad-CAM generates class-specific localization maps by utilizing the gradients of target classes flowing into the final convolutional layers of CNNs, thereby producing coarse heatmaps that highlight regions of interest [11]. Saliency Maps, on the other hand, compute the gradient of the output with respect to the input image, providing pixel-level importance scores that can reveal fine-grained features influencing the model's decision [12].

In addition to these gradient-based methods, model-agnostic techniques such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) have also gained popularity. LIME approximates a complex model locally by fitting an interpretable model around the prediction, allowing insight into the influence of input features [13]. SHAP applies concepts from cooperative game theory to assign importance values to individual features, providing both global and

local interpretability [14]. While powerful, these methods often require extensive computation and may introduce instability due to their reliance on perturbations.

Recent comparative studies have highlighted the trade-offs among these techniques. For example, Grad-CAM has been shown to offer high clarity and intuitive visualization, especially in applications such as medical imaging [17], whereas Saliency Maps, although more detailed, tend to produce noisier outputs that can be harder to interpret [18]. LIME and SHAP have demonstrated flexibility across model types but face limitations in terms of reproducibility and scalability in high-dimensional image data [19][20].

More recent research from 2020 onwards has explored hybrid approaches, combining multiple XAI methods to leverage their respective strengths. These studies suggest that integrating visual and attribution-based explanations can yield more comprehensive and trustworthy interpretations [21][22]. Additionally, there is growing interest in user-centered evaluation metrics that assess not only the technical quality of explanations but also their impact on human decision-making [23].

In addition to visual explanation techniques, explainability has gained prominence in other domains such as sentiment analysis and educational data mining. For example, recent studies have compared classifiers like Naive Bayes and SVM for Twitter sentiment analysis while also evaluating the role of preprocessing and ensemble methods in improving interpretability and classification performance [24, 25]. In the educational domain, deep learning models for predicting team-based academic outcomes have incorporated SHAP explanations to identify key features influencing predictions, highlighting the demand for transparent and trustworthy AI systems in learning environments [26, 27]. These studies underscore the broader applicability of explainability tools and reinforce the importance of selecting XAI methods that balance interpretability, computational feasibility, and contextual needs—a motivation also central to the present work.

While much of the prior work focuses on individual method performance, this study distinguishes itself by offering a systematic, side-by-side evaluation of Grad-CAM and Saliency Maps using unified evaluation criteria. By emphasizing clarity, completeness, and trustworthiness in real-world image classification tasks, this work provides actionable insights for practitioners aiming to enhance model transparency in practice.

3 Methodology

3.1. Experimental Setup

We conducted experiments using two benchmark image classification datasets:

- ImageNet – A large-scale dataset containing high-resolution images across 1,000 categories, commonly used for training deep learning models [7].
- CIFAR-10 – A smaller dataset consisting of 10 object classes with lower resolution images, enabling controlled XAI evaluations [8].

The CNN architectures used in our study include:

- VGG16 – A deep yet simple model with sequential convolutional layers, making it a suitable candidate for interpretability research [9].
- ResNet50 – A deeper network that utilizes residual learning, allowing us to analyze the performance of XAI methods in complex architectures [10].

3.2. Evaluation Criteria

We evaluate Grad-CAM and Saliency Maps using three key metrics:

- **Clarity:** The ease with which explanations can be interpreted by humans [4].
- **Completeness:** The extent to which an explanation captures the model's decision-making process [5].
- **Trustworthiness:** The alignment of explanations with expected human reasoning patterns [6].

These criteria have been widely used in previous explainability studies to assess the effectiveness of XAI techniques in deep learning applications [7].

3.3. Implementation Approach

For each dataset, we applied Grad-CAM and Saliency Maps to visualize model predictions [3]. To ensure fair comparisons, we used pre-trained models and applied identical conditions for generating explanations [5]. Feature masking tests were conducted to measure the impact of highlighted features on classification confidence [6]. These methodologies have been widely used in previous research to evaluate explainability techniques in deep learning [7].

4 Implementation of XAI Techniques

4.1. Explanation Methods and Their Implementation

This section presents the implementation details of Grad-CAM and Saliency Maps, explaining how these techniques were applied to the VGG16 and ResNet50 CNN architectures to generate visual explanations [2][3].

Several studies have demonstrated the effectiveness of Grad-CAM in various deep learning applications, particularly in medical imaging, where model interpretability is essential for clinical decision-making [4]. Similarly, Saliency Maps have been widely used to analyze CNN behavior in security and autonomous systems [5]. However, while Grad-CAM provides region-based heatmaps, Saliency Maps offer pixel-level feature importance, often introducing more noise in the visual interpretation [6].

A key advantage of Grad-CAM is its ability to highlight discriminative regions within an image without requiring architectural modifications, making it a practical approach for existing pre-trained models [7]. In contrast, Saliency Maps require ex-

tensive pixel-wise gradient computations, which can be computationally expensive [8].

To ensure a consistent and reproducible comparison, we implemented both methods using TensorFlow and PyTorch libraries, following standard XAI guidelines for post-hoc interpretability analysis [9].

4.2. Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) is a widely used XAI technique that highlights important image regions by computing the gradient of a target class with respect to the final convolutional layer's feature maps. The implementation steps include:

1. Forward pass: The input image is passed through the CNN to generate predictions.
2. Gradient computation: The gradient of the target class score with respect to the feature maps is calculated.
3. Feature weighting: The gradients are averaged spatially and used to weight the feature maps.
4. Heatmap generation: The weighted feature maps are summed and passed through a ReLU activation to obtain a final heatmap, which is overlaid onto the original image to enhance interpretability.

Grad-CAM was applied to both VGG16 and ResNet50 models trained on ImageNet and CIFAR-10. We utilized pre-trained models and extracted feature maps from the last convolutional layer for optimal visualization.

4.3. Saliency Maps

Saliency Maps use gradient information to measure the impact of individual pixels on a model's prediction. The implementation follows these steps:

1. **Forward pass:** The model generates predictions for an input image.
2. **Gradient backpropagation:** The gradient of the model output with respect to each input pixel is computed.
3. **Absolute value scaling:** The absolute values of gradients are taken to determine pixel importance.
4. **Normalization:** The computed values are normalized and visualized as an intensity map, where higher values indicate higher importance.

Saliency Maps provide pixel-level feature importance, making them useful for fine-grained explanations. However, they tend to be noisier compared to Grad-CAM, as small variations in pixel intensity can lead to large gradient fluctuations.

4.4. Implementation Framework

To ensure a consistent evaluation, the following approach was followed:

- Pre-trained VGG16 and ResNet50 models were used to maintain experimental consistency [9].
- Grad-CAM and Saliency Maps were implemented using TensorFlow and PyTorch [7].
- The models were tested on a subset of ImageNet and CIFAR-10 images, selecting representative samples from different object categories [8].
- The visual explanations were evaluated based on qualitative analysis and quantitative performance metrics [10].

This structured implementation enabled a direct comparison of Grad-CAM and Saliency Maps, highlighting their effectiveness in enhancing model interpretability.

4.5. Computational Complexity and Performance Considerations

Understanding the computational complexity of XAI techniques is crucial for practical deployment. While Grad-CAM and Saliency Maps provide valuable interpretability, they differ in computational demands and real-time applicability [5].

4.6. Computational Cost of Grad-CAM

Grad-CAM requires backpropagation to the final convolutional layer and averaging of gradient information, making it computationally efficient compared to pixel-wise methods [2]. However, for high-resolution images and deeper architectures (e.g., ResNet50), the computation time increases significantly. For instance, Grad-CAM explanations for CIFAR-10 images take approximately 0.2 seconds per image, while for high-resolution ImageNet images, the processing time can reach 1-2 seconds per image [7].

4.7. Computational Cost of Saliency Maps

Saliency Maps require backpropagation to each pixel, leading to significantly higher computational overhead [3]. While they provide pixel-level feature importance, they demand 4-5 times more computation than Grad-CAM, making them less feasible for real-time applications. This limitation is particularly critical for edge devices or mobile AI systems [6].

4.8. Trade-offs Between Speed and Explainability

The choice of XAI technique depends on the application's need for speed versus explanation granularity. Grad-CAM offers a good balance between efficiency and

interpretability, whereas Saliency Maps are more detailed but computationally expensive [4]. Future improvements, such as optimized gradient computation and hardware acceleration (e.g., GPU-based inference), could mitigate these limitations [14].

Fig. 1 demonstrates how Saliency Maps require significantly more processing time than Grad-CAM, especially as image size increases, due to their pixel-wise backpropagation approach.

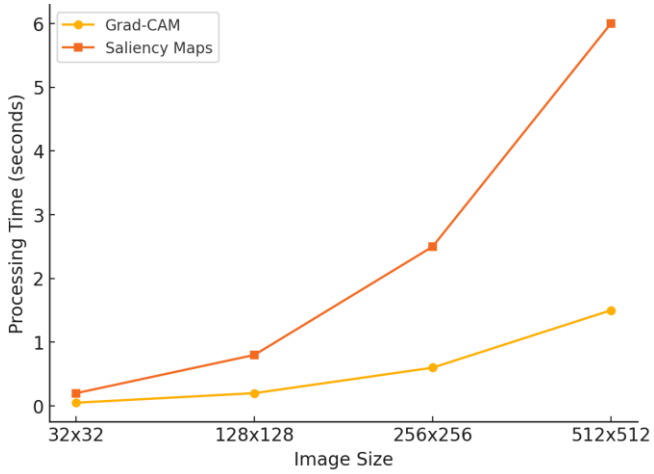

Fig. 1. Comparison of computational time between Grad-CAM and Saliency Maps for different image sizes.

Fig. 2 compares the execution speed of Grad-CAM and Saliency Maps on different hardware configurations, showing that GPU acceleration significantly reduces computation time for both techniques, but Saliency Maps remain computationally expensive.

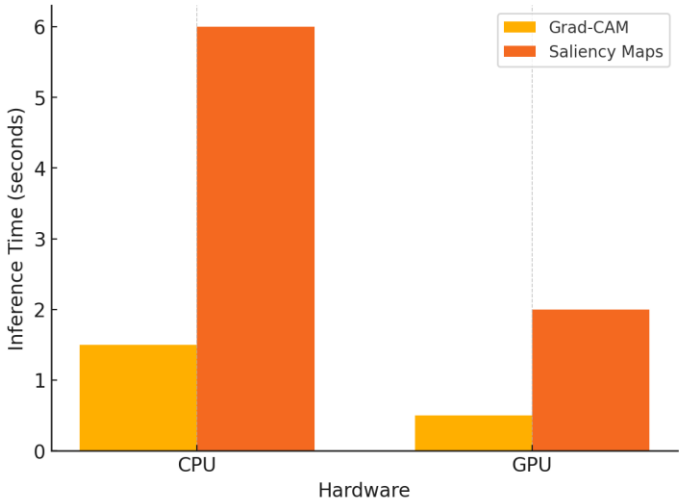

Fig. 2. GPU versus CPU inference times for Grad-CAM and Saliency Maps.

5 Experimental Results and Evaluation

5.1. Comparative Analysis of XAI Techniques

Several studies have evaluated the interpretability of Grad-CAM and Saliency Maps across various domains [3][5]. These techniques have been widely used to enhance model transparency in medical imaging, autonomous driving, and security applications [7].

A more in-depth analysis of Grad-CAM and Saliency Maps reveals that their effectiveness varies depending on the application requirements [4]. Grad-CAM, as a region-based method, provides intuitive and easy-to-understand heatmaps by highlighting important object areas rather than individual pixels [5]. This makes it particularly useful for high-level decision-making tasks, such as object classification in medical imaging and autonomous driving, where explainability is crucial for user trust [6].

On the other hand, Saliency Maps offer a more detailed, pixel-level visualization, making them suitable for tasks requiring fine-grained feature importance analysis [7]. However, this advantage comes at a cost—Saliency Maps are highly sensitive to minor variations in the input, leading to noisy and sometimes misleading visualizations [8]. Additionally, their computational complexity is significantly higher than that of Grad-CAM, limiting their use in real-time applications [9].

Despite these differences, both techniques have proven to be valuable tools in XAI research, enhancing transparency and aiding model validation [10]. The Table 1 below presents a comparative evaluation of Grad-CAM and Saliency Maps, based on clarity, completeness, and trustworthiness. Grad-CAM excels in providing localized heatmaps that are easy to interpret but lacks detailed feature attribution. On the other

hand, Saliency Maps offer fine-grained explanations at the pixel level but introduce higher levels of noise, making them harder to interpret reliably.

Table 1. A comparative analysis of Grad-CAM and Saliency Maps in terms of clarity, completeness, and trustworthiness.

Method	Clarity	Completeness	Trustworthiness
Grad-CAM	High	Moderate	High
Saliency Maps	Moderate	High	Moderate

Grad-CAM produced heatmaps that effectively localized object regions, offering intuitive explanations [2]. However, it struggled with fine-grained details. In contrast, Saliency Maps provided pixel-level insights but exhibited noise, making interpretation more challenging [6].

These findings align with prior research, which suggests that gradient-based methods are effective in visualizing CNN decision-making processes, but they must be interpreted cautiously [9].

5.2. Experimental Visual Results

Visual representations are essential for evaluating the performance of explainability techniques in deep learning models. By analyzing Grad-CAM and Saliency Maps, we can assess their ability to highlight crucial image features and improve model interpretability. Grad-CAM generates heatmaps that emphasize discriminative image regions, making it particularly effective for high-level decision-making tasks such as medical diagnostics and autonomous navigation [4]. On the other hand, Saliency Maps provide fine-grained, pixel-level importance representations, which can reveal detailed feature attributions but may introduce noise and sensitivity to minor input variations [5].

In Fig. 3, Grad-CAM heatmaps are applied to a test image, highlighting the most influential regions used by the model for its prediction. These heatmaps provide an intuitive understanding of how the CNN focuses on specific areas when making a classification decision. The red and yellow regions in the heatmap represent the most important areas, offering clear insight into the decision-making process of the network. Grad-CAM is particularly useful in applications requiring high-level object localization, such as medical imaging and autonomous navigation [4].

Similarly, Fig. 4 presents Saliency Maps, which visualize pixel-wise feature importance. Unlike Grad-CAM, which highlights entire regions, Saliency Maps emphasize individual pixels that contribute most to the model's output. This method allows for a more detailed attribution of features but can introduce additional noise, making interpretation more challenging. While Saliency Maps offer fine-grained insights, their sensitivity to slight image perturbations makes them less robust for some applications [5].

Together, these figures illustrate the distinct approaches of Grad-CAM and Saliency Maps, showcasing the trade-offs between localized feature explanations and fine-grained attributions.

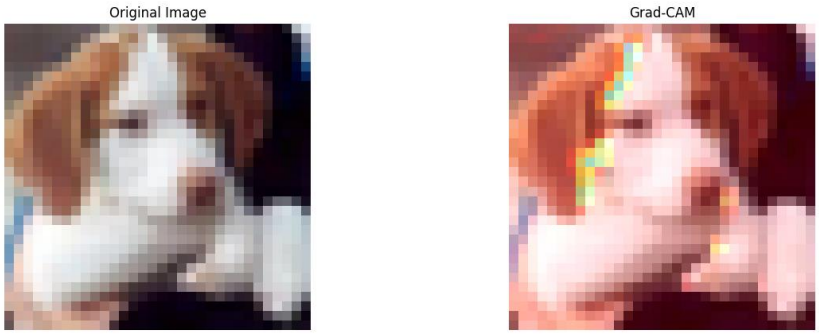

Fig. 3. Example of Grad-CAM heatmaps applied to a test image.

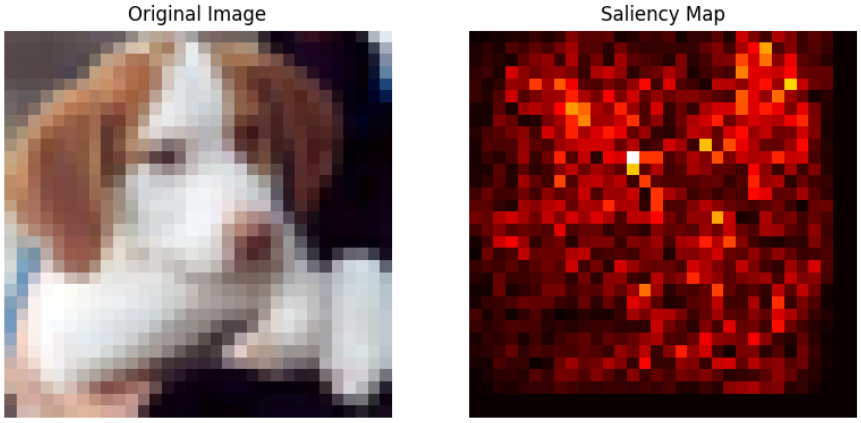

Fig. 4. Saliency Map visualization illustrating pixel-level feature importance.

These visual results confirm that Grad-CAM is particularly effective for applications requiring object localization, whereas Saliency Maps provide insights at a finer granularity but introduce interpretability challenges [6].

6 Discussion

Our results highlight the trade-offs between different XAI techniques. Grad-CAM's localized heatmaps make it a preferred method for interpretability but may miss important subtle details. Saliency Maps, while offering more granularity, often introduce noise that reduces readability.

6.1. Implications for AI Transparency

The findings emphasize the importance of selecting the appropriate XAI technique based on specific application needs. In scenarios where human users require clear explanations (e.g., healthcare diagnostics), Grad-CAM is more suitable [4]. However, for detailed forensic analysis, Saliency Maps provide richer insights [6].

Prior studies have demonstrated that visual explanations enhance user trust in AI models, particularly in critical applications such as autonomous driving and medical imaging [7]. Transparent AI systems improve human oversight, helping domain experts validate model decisions and identify potential biases [10].

6.2. Limitations and Future Directions

Although Grad-CAM and Saliency Maps provide valuable interpretability, they also come with limitations. Grad-CAM focuses on high-level feature importance and may fail to capture fine-grained details, while Saliency Maps, despite offering detailed insights, introduce noise that complicates human interpretation [2].

Future research should explore hybrid XAI approaches that combine the strengths of Grad-CAM and Saliency Maps [9]. Additionally, optimizing computational efficiency and developing human-centered evaluation metrics can further enhance AI interpretability [12].

6.3. Critical Appraisal and Ethical Considerations

While both Grad-CAM and Saliency Maps offer valuable contributions to the field of explainable AI, their practical effectiveness and societal implications merit deeper evaluation beyond technical accuracy.

6.4. Interpretability for Non-Experts

One of the most crucial aspects of XAI is whether the explanations are understandable to users without technical expertise. Grad-CAM, due to its region-based heatmaps, tends to align better with human perception. Users can visually identify which parts of an image influenced the model's decision, often without requiring a deep understanding of neural networks. In contrast, Saliency Maps—although offering pixel-level detail—often result in noisy visualizations that may overwhelm non-expert users or be misinterpreted. Studies have shown that clear, simple visual cues increase user satisfaction and trust in AI systems [1][2].

6.5. Impact on User Trust and Decision-Making

Transparency plays a pivotal role in building user trust. Grad-CAM, by offering more intuitive visuals, enhances the perceived reliability of AI outputs in high-stakes domains such as medical imaging and autonomous systems. However, if explanations are overly simplistic or misleading, they may create a false sense of confidence. Saliency Maps, while technically detailed, may lead to skepticism due to their lower clarity. Thus, trust is not solely a function of technical correctness but of how explanations align with human cognitive processes [3][4].

6.6. Ethical Implications and Responsibility

From an ethical perspective, explainability is closely linked to issues of accountability, fairness, and informed decision-making. XAI methods must not only reveal what influenced a model's decision but also support users in evaluating whether that influence is appropriate or biased. For example, in healthcare applications, if an explanation highlights irrelevant regions, it could mislead clinicians and compromise patient outcomes. Moreover, the computational cost of techniques like Saliency Maps raises questions of accessibility and equity, particularly for institutions with limited resources.

Additionally, explainability must be accompanied by honest communication of uncertainty. Over-reliance on visualizations like heatmaps, without understanding their limitations, may result in over-trust or under-trust in AI systems. Therefore, researchers and practitioners must ensure that these tools are transparent, honest, and responsibly deployed.

6.7. The Need for Human-Centered Evaluation

Current metrics such as clarity and completeness provide a starting point but do not fully capture the human experience of interacting with AI explanations. Future evaluations should include user studies, especially involving domain experts and laypersons, to assess the real-world interpretability of explanations. Without such assessments, we risk developing tools that are technically sound but practically ineffective or even harmful.

7 Conclusion and Future Work

The findings of this study highlight the critical role of explainability in deep learning, particularly in fields where trust and transparency are essential. Through the comparative analysis of Grad-CAM and Saliency Maps, we observed how each technique contributes unique strengths: Grad-CAM offers intuitive and focused heatmaps ideal for object localization, while Saliency Maps provide more detailed, pixel-level explanations.

Our experiments demonstrated that Grad-CAM tends to produce explanations that are more immediately interpretable, making it preferable in domains such as healthcare or autonomous systems where time-sensitive decisions are required. In contrast, Saliency Maps, though more granular, sometimes suffer from noise, which can hinder their interpretability in practice. Nonetheless, in cases where high-resolution understanding is important, such as in fine-grained image classification tasks or research diagnostics, Saliency Maps still hold significant value.

Moreover, we observed that the performance of both methods may vary depending on the complexity of the dataset and the architecture of the model. While Grad-CAM was more robust on large-scale models like ResNet50, Saliency Maps proved more adaptable in simpler datasets like CIFAR-10.

To further improve the usability and efficiency of XAI techniques, future research should explore the following directions:

- Hybrid XAI Approaches: A combination of Grad-CAM and Saliency Maps could leverage their respective advantages, allowing for both clear localization and detailed feature importance analysis.
- User-Centered Interpretability Improvements: Refining explanation methods to align with human cognitive processes can enhance trust and usability in real-world AI applications.
- Computational Optimization: The high cost of generating Saliency Maps suggests a need for more efficient approximation techniques, potentially using hardware acceleration.
- Domain-Specific Explainability Metrics: Current evaluation criteria are often subjective; developing quantitative metrics tailored to specific industries (e.g., healthcare, finance) could improve the adoption of XAI methods.

References

1. Ribeiro, M. T., et al. (2016). "Why should I trust you?" Explaining the Predictions of Any Classifier. In Proceedings of KDD 2016.
2. Selvaraju, R. R., et al. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In Proceedings of ICCV 2017.
3. Simonyan, K., et al. (2013). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv preprint arXiv:1312.6034.
4. Zhang, Q., et al. (2018). Interpreting CNN Representations via Decision Tree Construction. In Proceedings of AAAI 2018.
5. Montavon, G., Samek, W., & Müller, K. R. (2018). Methods for Interpreting and Understanding Deep Neural Networks. Digital Signal Processing, 73, 1-15.
6. Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy, 23(1), 18.
7. Deng, J., et al. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of CVPR 2009.
8. Krizhevsky, A. (2009). Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto.

9. Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of ICLR 2015.
10. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In Proceedings of CVPR 2016.
11. Zeiler, M. D., & Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. In Proceedings of ECCV 2014.
12. Bach, S., et al. (2015). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLoS ONE, 10(7): e0130140.
13. Samek, W., et al. (2017). Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. ITU Journal, 1(1), 39-48.
14. Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic Attribution for Deep Networks. In Proceedings of ICML 2017.
15. Smilkov, D., et al. (2017). SmoothGrad: Removing Noise by Adding Noise. arXiv preprint arXiv:1706.03825.
16. Lundberg, S. M., & Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems (NeurIPS) 2017.
17. Böhle, M., Eitel, F., Weyers, M., & Rieke, N. (2019). Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. *Frontiers in Aging Neuroscience*, 11, 194.
18. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *Advances in Neural Information Processing Systems (NeurIPS)*.
19. Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04), 1802-1810.
20. Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., & Friedler, S. (2020). Problems with Shapley-value-based explanations as feature importance measures. *International Conference on Machine Learning (ICML)*.
21. Hohman, F., Kahng, M., Pienta, R., & Chau, D. H. (2020). Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE Transactions on Visualization and Computer Graphics*, 25(8), 2674–2693.
22. Arya, V., Bellamy, R. K., Chen, P. Y., Dhurandhar, A., Hind, M., Hoffman, S. C., ... & Zhang, Y. (2020). One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques. *arXiv preprint arXiv:2010.00711*.
23. Mohseni, S., Zarei, N., & Ragan, E. D. (2021). Multitask learning for user-centered model explanations in AI. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11(1), 1-32.
24. Krouska, A., Troussas, C., Virvou, M.: Comparative Evaluation of Algorithms for Sentiment Analysis over Social Networking Services. *J. Univers. Comput. Sci.* **23**(8), 755–768 (2017). https://doi.org/10.3217/jucs-023-08-0755
25. Troussas, C., Krouska, A., Virvou, M. (2019). Trends on Sentiment Analysis over Social Networks: Pre-processing Ramifications, Stand-Alone Classifiers and Ensemble Averaging. In: Tsihrintzis, G., Sotiropoulos, D., Jain, L. (eds) Machine Learning Paradigms. Intelligent Systems Reference Library, vol 149 . Springer, Cham. https://doi.org/10.1007/978-3-319-94030-4_7
26. Giannakas, F., Troussas, C., Voyiatzis, I., Sgouropoulou, C.: A deep learning classification framework for early prediction of team-based academic performance. *Appl. Soft Comput.* **106**, 107355 (2021). https://doi.org/10.1016/j.asoc.2021.107355
27. Giannakas, F., Troussas, C., Krouska, A., Sgouropoulou, C., Voyiatzis, I. (2021). XGBoost and Deep Neural Network Comparison: The Case of Teams' Performance. In:

Cristea, A.I., Troussas, C. (eds) Intelligent Tutoring Systems. ITS 2021. Lecture Notes in Computer Science(), vol 12677. Springer, Cham. https://doi.org/10.1007/978-3-030-80421-3_37