# XAI in Medical Image Report Generation: Unlocking Stress Testing as a Responsible AI Practice for Multimodal Models

Flávia Carvalhido[1,*], Henrique Lopes Cardoso[1], Vítor Cerqueira[1] and Carlos Soares[1,2]

[1]LIACC, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias s/n, 4200-465, Porto, PORTUGAL
[2]Fraunhofer Portugal AICOS, Rua Alfredo Allen 455/461, 4200-135 Porto, PORTUGAL

**Abstract**

State-of-the-art multimodal models, namely Vision-Language Transformers, have revolutionized automatic image and text generation through the fusion of these two information media. Multimodal generation tasks, such as Medical Image Report Generation (MIRG), benefit greatly from these architectures, which are increasingly being adopted. However, MIRG is a safety-critical task, raising pressing concerns when it comes to Responsible AI development practices being adhered to. Any models that support clinical decision-making — in which MIRG can play an important role — must be highly reliable, interpretable, and trustworthy. As such, Explainable AI (XAI) methodologies are crucial for the correct development and usage of these models. We argue that XAI can help unlock further Responsible AI practices for MIRG multimodal model development, namely by serving as a guide for **Stress Testing** — a testing procedure that requires systematic probing of AI models with specific inputs to derive application and robustness limitations. By leveraging visual, example, and textual-based explanations, it is possible to better understand what input characteristics reduce the generation quality in MIRG models, thus guiding the stress testing process towards those worse-performing scenarios. We believe that XAI contributes both to the interpretability and reliability of MIRG approaches by offering tools that help develop more complex Responsible AI practices.

**Keywords**

Stress Testing, XAI, Medical Image Report Generation, Responsible AI

## 1. Introduction

State-of-the-art generative transformer models have revolutionized the landscape of image-text multimodal Artificial Intelligence (AI), achieving superior capabilities than previously seen by comprehensively fusing information from these two media [1]. One domain that benefits greatly from this is healthcare, namely medical imaging diagnostics, an inherently multimodal domain.

For every medical imaging exam, there is a complete report with a textual analysis and diagnosis. Typically, a radiologist must analyze the medical image and, based on their observations and acquired medical knowledge, compose a complete report with their findings. However, given that most hospital patients have to perform at least one type of imaging exam to aid their diagnosis, thousands of reports are written every day. As it is a time-consuming and fundamental process, an increased research effort in Medical Image Report Generation (MIRG) as a multimodal generation task has spurred the development and deployment of specialized AI models to aid healthcare professionals [2].

Recently, several concerns have been raised about Responsible AI usage, namely in safety-critical domains and tasks that involve sensitive information, such as MIRG. It is of the utmost importance that AI models used for any healthcare application be transparent, interpretable, reliable, privacy-preserving, and, all in all, trustworthy. Responsible AI practices should be followed when developing models for safety-critical applications, and special attention is given to this issue when analyzing AI regulation documents, such as the European Union AI Act. [1]

*MAI-XAI'25: Multimodal, Affective and Interactive eXplainable AI part of the 28th European Conference on Artificial Intelligence, October 25-26, 2025, Bologna, Italy*

*Corresponding author.
fcarvalhido@fe.up.pt (F. Carvalhido); hlc@fe.up.pt (H. Lopes Cardoso); vcerqueira@fe.up.pt (V. Cerqueira); csoares@fe.up.pt (C. Soares)
0009-0005-2681-8454 (F. Carvalhido); 0000-0003-1252-7515 (H. Lopes Cardoso); 0000-0002-9694-8423 (V. Cerqueira); 0000-0003-4549-8917 (C. Soares)

Explainable AI (XAI) methods are crucial when striving for interpretable and thus Responsible AI models. A clear issue with state-of-the-art Transformer-based multimodal models is their inherent complexity, which is what grants the model a higher degree of encoding of information and better performance, but requires billions of parameters. This large model size makes model interpretation difficult, and tracing the model's decision-making process nearly impossible, so Post-Hoc XAI techniques are necessary to safeguard interpretability. In MIRG, model interpretation also serves a bigger purpose, which is to boost model reliability and trust. The ability to correctly describe a diagnosis in a report, using the necessary information to justify a decision, is crucial for healthcare professionals and, ultimately, patients to effectively trust an AI model and its decisions [3].

Another important feature for MIRG models is their high degree of reliability, which can ultimately only be achieved by safeguarding model robustness. If a model performs correctly and as expected in all scenarios, then the user can rely on the model more easily. Most MIRG model deployment information only reports on performance metrics and model characteristics, yet lacks details about the model's limitations. To fully rely on a model, the user should know the scope of its application and the scenarios in which the model might function incorrectly or output a prediction with a low degree of confidence, in order to avoid those scenarios and use the model in a responsible fashion.

In our perspective, XAI methods serve as a guide to finding model limitations and employing an essential Responsible AI development practice for MIRG: **Stress Testing**. By understanding the model's behavior on each prediction, it is possible to single out the inputs that cause the model to generate incorrect or low-quality reports and further localize its limitations. For example, understanding which visual features have the most impact on report generation can guide the stress testing process to focus on those features and further study the degradation of model performance when changes or perturbations are introduced. Ultimately, XAI will also help interpretation when stress testing is employed, by providing insight into the model's decision-making process when exploring underperforming scenarios. Thus, XAI methods can aid model explainability in two ways: explanations can be used to guide stress testing and to provide insight into the model's worst-performing areas.

Our stance is structured around three positions. The first is all-encompassing, while the second and third are complementary:

1. *XAI can unlock Stress Testing as a Responsible AI practice for multimodal models in MIRG.*
2. *XAI can guide the stress testing process by providing insight into the model's inner workings and finding the most impactful changes that might degrade model performance.*
3. *XAI can aid the interpretation of model limitations by explaining model behavior when used in a worst-performing scenario.*

This paper is structured as follows. Section 2 gives a brief overview of the related work on multimodal models for the MIRG task. Section 3 relates to our previously stated first and main position, describing existing works in the areas of XAI, robustness, and stress testing. Section 4 details existing works that support the development of stress testing approaches using XAI, justifying our second and third positions. Finally, Section 5 concludes this paper with a short overview of the discussed topics.

## 2. Related Work

MIRG is an image-to-text generation task, with the most used multimodal model architecture being the encoder-decoder stack, composed of a visual encoder and a language decoder [1]. Mostly used in a pre-trained fashion [4], state-of-the-art MIRG models often leverage the Transformer architecture and adopt the nomenclature of Vision-Language (VL) Transformers, generating a detailed medical textual report based solely on a medical imaging exam and, optionally, a prompt.

MIRG models are often trained on large task-specific datasets or generalist medical domain datasets. Some representative radiology-specific models include R2Gen-CMN [5] and RGRG [6]. MAIRA-1 [7], LLM-RG4 [8], and DART [9] are three of the newest MIRG-specific models proposed for this task.

[1] Full-text available at https://artificialintelligenceact.eu/chapter/3/

Among the generalist biomedical models, which include MIRG as a sub-task, it is worth highlighting Med-PaLM Multimodal [10] and BiomedGPT [11]. These models are often trained and tested on large, anonymized, and publicly available medical image and textual report datasets, with MIMIC-CXR [12], CheXpert [13], IU-CX [14], and ROCOv2 [15] among the most commonly used.

There are also some challenges in the evaluation of MIRG, given its domain-specific requirements for clinical validity, which surpass the common evaluation schemes for text generation. Recently, task-specific evaluation schemes such as RadGraph-F1 [16] and CheXprompt [17] have been proposed to further evaluate report quality, completeness, and diagnosis accuracy, leveraging large language models as evaluators, entity labeling schemes and other more advanced methods to encompass task-specific requirements.

## 3. MIRG: a Responsible AI perspective and approaches

As a safety-critical task, MIRG approaches have to adhere to Responsible AI practices and values. Among those values, the ones we will focus on are reliability and interoperability, inherently connected to robustness and XAI approaches. The following section will mention key concepts and works directly related to our positions on XAI and stress testing.

### 3.1. Robustness for Reliability

Chander et al. [18] define robustness with respect to technical robustness and safety of healthcare systems, stating that "Healthcare is compassionate, and unconditionally, the AI schemes need to produce consistent and reliable results since human lives could be on the line if anything goes wrong. [...] AI systems, while executing, may generate fault outcomes, so designers must build methods that efficiently handle issues and conflicts if any arise". Technical robustness, in general, refers to the model's ability to provide accurate results while withstanding attacks, threats, or generally unforeseen scenarios/inputs.

Performance benchmarking on a task-specific dataset is the most common way to measure model robustness, but it is commonly limited to general or ideal application scenarios. Authors commonly report model performance on these datasets – such as MIMIC-CXR [12], CheXpert [13] and IU-CX [14], – to better position their work and compare it to other existing approaches. As benchmark datasets for MIRG do not often contain adversarial, perturbed, or out-of-distribution examples, which provide a more complete insight into the technical robustness of a model, there is a need for curated datasets that probe MIRG models under such conditions.

MedMNIST-C [19] aims to tackle this gap by offering a robustness benchmark on common image corruptions for medical imaging, leveraging 5 different corruption categories with several intensity levels. There are also a multitude of adversarial attack methods to probe the adversarial robustness of medical imaging analysis models and perform adversarial training, but only a few have been applied to MIRG. Dong et al. [20] report on diverse adversarial attacks for medical image analysis and further explore adversarial defense mechanisms for MIRG, showcasing their applicability potential. Among these adversarial attacks, the usage of Generative Adversarial Networks (GANs) is highlighted as a preferred attack method that allows for the conditioned generation of adversarial images.

Technical robustness and model explainability suffer from a trade-off. Complex models are shown to obtain good performance and high robustness across several domains, which comes at the cost of explainability. On the other hand, by decreasing the complexity of models to maintain a high degree of explainability, technical robustness decreases significantly. A balance must be achieved to prioritize these two values, both of which are crucial to the responsible development of MIRG models.

### 3.2. XAI for Interpretability

MIRG models must prioritize a high degree of explainability, namely by providing clinically meaningful explanations. A plurality of surveys on XAI methods for medical image analysis and MIRG have been published, with a particular focus on Post-Hoc methods that leverage visual, example, and textual-based explanations [18, 21, 22].

Perturbation XAI approaches, namely SHAP [23] and its variations, focus on observing the effect that input perturbations have on model behavior, further studying feature importance and ”cooperation” in the model’s final prediction. Van der Velden et al. [24] use Deep SHAP to explain the image regions that contributed to the volumetric density estimation on breast MRI images.

Saliency map XAI methods are commonly used to explain which regions or features of a medical image are most relevant to produce the model output, providing a visual explanation. Approaches such as Grad-CAM [25], Integrated Gradients [26], Deep Taylor Decomposition [27], among others, can provide an attention map as an explanation that gives insight into the most important regions and features of the image. Raghavan et al. [28] propose an attention-guided Grad-CAM, focusing on both channel and spatial attention, to extract saliency maps for breast cancer detection that leverage the attention weights from the model. Sayres et al. [29] leverage Integrated Gradients to explain diabetic retinopathy predictions from retinal fundus images, but found that the explanations suffered from a bias for positive features. Yoon et al. [30] use Deep Taylor Decomposition on MRI images to explain the model’s predictions regarding temporomandibular joint anterior disk displacement.

Example-based explanations can commonly be obtained using three categories of XAI techniques: prototypes and criticisms, counterfactuals, and adversarial examples [31]. MMD-Critic [32] is a ”prototypes and criticisms” approach, using maximum mean discrepancy to obtain representative class example images that both support and contrast the model’s prediction. cGAN [33], meaning conditional GAN for Counterfactual image generation, incorporates object detection and image segmentation information into a gradual generation process to produce interpretable images that preserve clinically relevant radiographic features. TraCE [34], which stands for “Training Calibration-based Explainers”, is an uncertainty-based interval calibration counterfactual image generation model, showcasing good results for the task of chest x-ray anomaly detection.

Textual explanations fulfill a different role in MIRG. MIRG is inherently multimodal, and models generate a textual output comprised of the radiology report based on visual information. One could argue that a textual explanation for MIRG is an approximation to the generated medical image report, a position that several authors adopt. Van der Velden et al. [22] and Patrício et al. [21] divide textual explanation methodologies into three different categories with regard to medical imaging: image captioning/reporting, image captioning/reporting with visual explanations, and concept attribution. As such, the usage of MIRG models – included in Section 2 – can provide an explanation in itself. A high-quality medical image report must be descriptive and contain all relevant information for a diagnosis, thus giving insight into what the model perceives in the image and derives from it.

All the aforementioned XAI techniques provide insights into multimodal models’ mode of functioning and justify their generation process, leveraging different kinds of explanations, which can also be combined.

### 3.3. Our Position: XAI to Unlock Stress Testing

Stress testing for AI is a recent field, although this concept has been explored in different contexts/domains, such as software engineering, finance, and materials science, among many others [35, 36, 37]. In the specific case of software engineering, stress testing involves targeted tests on a software application, exploring its edge-case usage scenarios to probe if there are any deviations from its programmed behavior under non-ideal circumstances. While transferring this notion to generative AI, a different set of challenges is imposed, as generative AI behavior is not deterministic, and edge-case usage scenarios are more difficult to define.

Some initial attempts at defining stress testing in AI mention the following: “evaluations that probe the properties of a predictor by observing its outputs on specifically designed inputs” [38]; ”Large Language Model (LLM) stress testing includes identifying logical gaps within the LLM by giving it inferential rules for it to discern some general compositional or primitive rule.” [39]; and, ”quantify model robustness [...], complementing more qualitative tools for explainable AI” [40].

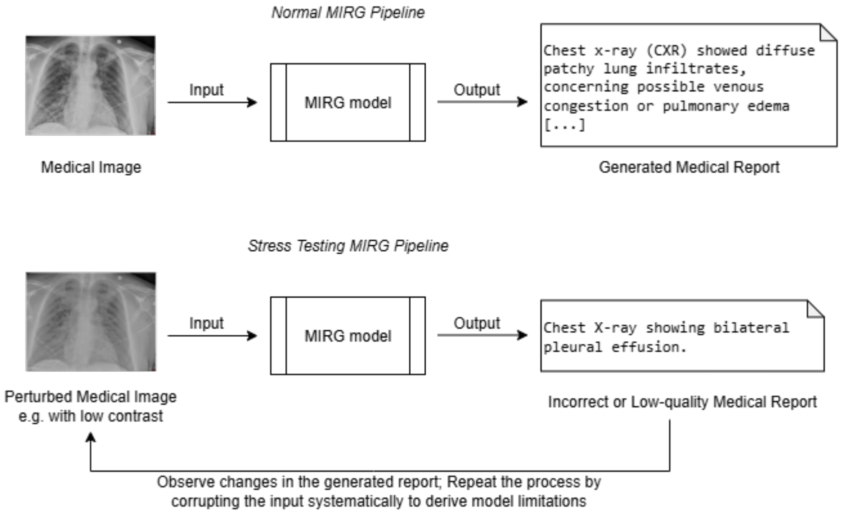
Normal MIRG Pipeline
Input
MIRG model
Output
Chest x-ray (CXR) showed diffuse patchy lung infiltrates, concerning possible venous congestion or pulmonary edema [...]
Medical Image
Generated Medical Report
Stress Testing MIRG Pipeline
Input
MIRG model
Output
Chest X-ray showing bilateral pleural effusion.
Perturbed Medical Image e.g. with low contrast
Incorrect or Low-quality Medical Report
Observe changes in the generated report; Repeat the process by corrupting the input systematically to derive model limitations

**Figure 1:** Normal versus Stress Testing MIRG pipelines. Images and report in the Normal MIRG pipeline from the ROCOv2 [15] dataset.

As defined by existing works and depicted in Figure 1, AI stress testing is a complex practice that requires exploring specifically engineered out-of-distribution, perturbed, or adversarial inputs [38], observing and documenting changes in model behavior and performance, and finally extracting a comprehensive list of model limitations that clearly establish robustness and application boundaries, thus enhancing model explainability [40].

As such, stress testing is highly connected to XAI and robustness research, relying on the key concepts of both and, more importantly, on their approaches. Yet, it is in XAI that stress testing finds its meaning, since XAI methods can guide the stress testing process into the models' worst-performing areas, and help interpret the results of stress testing as well.

Our first stated and main position is: **XAI can unlock Stress Testing as a Responsible AI practice for multimodal models in MIRG**. As a novel Responsible AI practice, stress testing still lacks a consensual definition and standard practices, yet the existing works concur in exploring model limitations with specifically built inputs [38, 40]. This procedure requires understanding a model's mode of functioning in order to guide the input construction, which is the reason why XAI methods can unlock stress testing. Instead of developing AI stress testing approaches from scratch, we argue that there already exist approaches within the XAI field that should act as the basis for the development of stress testing procedures, specifically for multimodal models, as described in the following section. We take MIRG as an example domain in which XAI methods are crucial, and there is an increased need for new Responsible AI practices, such as stress testing.

The usage of GANs, adversarial and perturbed examples, and counterfactuals aligns with the already perceived notions of stress testing. But while XAI's goal is limited to explaining a model prediction and mode of functioning, stress testing involves the continuous application of these techniques to approximate and derive model limitations with the aim of thoroughly defining the model's application and robustness boundaries.

## 4. Resources and Methods

In this section, some existing works that support our second and third positions are explored and detailed.

## 4.1. XAI to Guide Stress Testing

As previously stated, **XAI can guide the stress testing process by providing insight into the model's inner workings and finding the most impactful changes that might degrade model performance**. XAI methods that rely on feature perturbations [23, 24] and adversarial or counterfactual examples [34, 33] derive explanations by analyzing how modifications to the initial input might affect model performance or the final prediction itself. This perspective of identifying which features most impact the model prediction can be leveraged to understand the scenarios that negatively influence the model's performance and which feature changes will achieve this.

If applied systematically to MIRG's generative use case, these approaches can help identify increasingly more challenging or stress-inducing inputs in a progressive and guided manner, in order to effectively understand what impacts report generation the most. The usage of adversarial robustness benchmarks, such as MedMNIST-C [19], can also help by exposing the model to inherently perturbed inputs, approximating the idea of systematically probing the models but without the direction that XAI methods can provide.

Both XAI and robustness research fields often employ GANs for adversarial example generation, simulating inputs that target the model's weaker-performing cases. In an early attempt at stress-testing, GASTeN [41] (Generative Adversarial Stress Test Networks) aims to generate realistic adversarial data that will increasingly approximate the decision boundary of a model, allowing for better visualization of which data is classified with the lowest confidence by the model and, thus, derive model limitations.

Additionally, in the previously mentioned TraCE [34], Thiagarajan et al. argue that their counterfactual generation technique can be used to approximate model decision boundaries through the guided generation of counterfactual images.

RadEdit [40] is a stress testing approach specifically developed for medical images that generates synthetic examples using masks to protect regions of the image that should not be altered, while locating regions that should, so as to mimic plausible dataset shifts. This stress testing approach aims to complement XAI methods by probing models on their robustness against data distribution shifts. To guide the mask placement during generation, the authors used already existing masks from the data itself or trained separate segmentation models to generate the masks. One future work direction that the authors report is the need for quantitative evaluations for the introduced image changes when generating stress testing inputs, to better guide the generation process towards high-quality changes that can effectively showcase meaningful model limitations.

## 4.2. XAI to Interpret Model Limitations

Our final position is that **XAI can aid the interpretation of model limitations by explaining model behavior when used in a worst-performing scenario**. Deriving a definition of model limitations from the inputs used during stress testing is important in order to provide some interpretation to the stress testing findings. Thus, XAI methods' role in stress testing is twofold: to guide the stress testing process and to interpret its final results.

Building upon GASTeN [41], Gomes et al. [42] use this GAN-based stress testing technique and then apply XAI approaches to build prototypes that demonstrate which type of inputs fall under model limitations, highlighting the features that negatively affect the confidence of the models the most. They do so by leveraging prototype generation and GradientSHAP [23].

Moreover, the intrinsic explainability that the generated medical image reports offer can also provide some insight into model limitations. Although quantifying the model's performance drop based on the report quality can be quite challenging, by observing the changes in the generated outputs from the MIRG methodologies, we can also derive some in-domain explanations for model shortcomings. As this is a less explored area, there are only a few works that surround these ideas. Baia et al. [43] perform black-box attacks on multimodal models to understand the effect certain perturbations on the image can have on the generated explanation, introducing attacks that lead the models to provide unfaithful explanations, thus showcasing the model's susceptibility to certain changes in the input.

## 5. Conclusion

This paper offers a novel perspective on the role of XAI for multimodal models, with a particular focus on the MIRG task. Responsible AI practices are increasingly necessary to better understand model behavior and obtain information on these technologies, so XAI methods play a crucial role in developing novel Responsible AI practices, such as Stress Testing. Our expressed positions offer research perspectives yet to be explored, while placing XAI at the center of a new approach for reporting on multimodal model limitations and, ultimately, increasing model explainability and trust.

## Acknowledgments

This work is a result of the Agenda “Center for Responsible AI”, nr. C645008882-00000055, investment project nr. 62, financed by the Recovery and Resilience Plan (PRR) and by European Union - NextGeneration EU. Funded by the European Union – NextGenerationEU. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

Financially supported by UID/00027 - Artificial Intelligence and Computer Science Laboratory (LIACC), funded by Fundação para a Ciência e a Tecnologia, I.P./ MCTES through the national funds.

## Declaration on Generative AI

The authors have not employed any Generative AI tools.

## References

[1] R. Guo, J. Wei, L. Sun, B. Yu, G. Chang, D. Liu, S. Zhang, Z. Yao, M. Xu, L. Bu, A survey on image-text multimodal models, arXiv preprint arXiv:2309.15857 (2024).

[2] H. Ali, Z. Shah, T. Alam, P. Wijayatunga, E. Elyan, Recent advances in multimodal artificial intelligence for disease diagnosis, prognosis, and prevention, Frontiers in radiology 3 (2024) 1349830.

[3] O. Wysocki, J. K. Davies, M. Vigo, A. C. Armstrong, D. Landers, R. Lee, A. Freitas, Assessing the communication gap between ai models and healthcare professionals: Explainability, utility and trust in ai-driven clinical decision-making, Artificial Intelligence 316 (2023) 103839. URL: https://www.sciencedirect.com/science/article/pii/S0004370222001795. doi:https://doi.org/10.1016/j.artint.2022.103839.

[4] S. Long, F. Cao, S. C. Han, H. Yang, Vision-and-language pretrained models: A survey, arXiv preprint arXiv:2204.07356 (2022).

[5] Z. Chen, Y. Shen, Y. Song, X. Wan, Generating radiology reports via memory-driven transformer, in: Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021.

[6] T. Tanida, P. Müller, G. Kaissis, D. Rueckert, Interactive and explainable region-guided radiology report generation, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2023. URL: http://dx.doi.org/10.1109/CVPR52729.2023.00718. doi:10.1109/cvpr52729.2023.00718.

[7] S. L. Hyland, S. Bannur, K. Bouzid, D. C. Castro, M. Ranjit, A. Schwaighofer, F. Pérez-García, V. Salvatelli, S. Srivastav, A. Thieme, et al., MAIRA-1: A specialised large multimodal model for radiology report generation, arXiv preprint arXiv:2311.13668 (2024).

[8] Z. Wang, Y. Sun, Z. Li, X. Yang, F. Chen, H. Liao, Llm-rg4: Flexible and factual radiology report generation across diverse input contexts, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, 2025, pp. 8250–8258.

[9] S.-J. Park, K.-S. Heo, D.-H. Shin, Y.-H. Son, J.-H. Oh, T.-E. Kam, Dart: Disease-aware image-text alignment and self-correcting re-alignment for trustworthy radiology report generation, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 15580–15589.

[10] T. Tu, S. Azizi, D. Driess, M. Schaekermann, M. Amin, P.-C. Chang, A. Carroll, C. Lau, R. Tanno, I. Ktena, et al., Towards generalist biomedical ai, arXiv preprint arXiv:2307.14334 (2023).

[11] K. Zhang, R. Zhou, E. Adhikarla, Z. Yan, Y. Liu, J. Yu, Z. Liu, X. Chen, B. D. Davison, H. Ren, et al., A generalist vision–language foundation model for diverse biomedical tasks, Nature Medicine (2024) 1–13.

[12] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, S. Horng, MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports, Scientific data 6 (2019) 317.

[13] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al., Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, in: Proceedings of the AAAI conference on artificial intelligence, volume 33, 2019, pp. 590–597.

[14] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, C. J. McDonald, Preparing a collection of radiology examinations for distribution and retrieval, Journal of the American Medical Informatics Association 23 (2016) 304–310.

[15] J. Rückert, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, S. Koitka, O. Pelka, A. B. Abacha, A. G. Seco de Herrera, H. Müller, P. A. Horn, F. Nensa, C. M. Friedrich, ROCOv2: Radiology objects in context version 2, an updated multimodal image dataset, Scientific Data 11 (2024). URL: http://dx.doi.org/10.1038/s41597-024-03496-6. doi:10.1038/s41597-024-03496-6.

[16] F. Yu, M. Endo, R. Krishnan, I. Pan, A. Tsai, E. P. Reis, E. K. U. N. Fonseca, H. M. H. Lee, Z. S. H. Abad, A. Y. Ng, et al., Evaluating progress in automatic chest x-ray radiology report generation, Patterns 4 (2023).

[17] J. Zambrano Chaves, S.-C. Huang, Y. Xu, H. Xu, N. Usuyama, e. a. Zhang, S, Towards a clinically accessible radiology foundation model: open-access and lightweight, with automated evaluation, arXiv preprint arXiv:2403.08002 (2024). URL: https://arxiv.org/pdf/2403.08002.

[18] B. Chander, C. John, L. Warrier, K. Gopalakrishnan, Toward trustworthy artificial intelligence (tai) in the context of explainability and robustness, ACM Computing Surveys 57 (2025) 1–49.

[19] F. Di Salvo, S. Doerrich, C. Ledig, MedMNIST-C: Comprehensive benchmark and improved classifier robustness by simulating realistic image corruptions, arXiv preprint arXiv:2406.17536 (2024).

[20] J. Dong, J. Chen, X. Xie, J. Lai, H. Chen, Survey on adversarial attack and defense for medical image analysis: Methods and challenges, ACM Computing Surveys 57 (2024) 1–38.

[21] C. Patrício, J. C. Neves, L. F. Teixeira, Explainable deep learning methods in medical image classification: A survey, ACM Computing Surveys 56 (2023) 1–41.

[22] B. H. van der Velden, H. J. Kuijf, K. G. Gilhuijs, M. A. Viergever, Explainable artificial intelligence (xai) in deep learning-based medical image analysis, Medical Image Analysis 79 (2022) 102470. URL: https://www.sciencedirect.com/science/article/pii/S1361841522001177. doi:https://doi.org/10.1016/j.media.2022.102470.

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

[24] B. H. van der Velden, M. H. Janse, M. A. Ragusi, C. E. Loo, K. G. Gilhuijs, Volumetric breast density estimation on mri using explainable deep learning regression, Scientific reports 10 (2020) 18095.

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

[26] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International

conference on machine learning, PMLR, 2017, pp. 3319–3328.

[27] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, K.-R. Müller, Explaining nonlinear classification decisions with deep taylor decomposition, Pattern recognition 65 (2017) 211–222.

[28] K. Raghavan, Attention guided grad-CAM: an improved explainable artificial intelligence model for infrared breast cancer detection, Multimedia Tools and Applications 83 (2024) 57551–57578.

[29] R. Sayres, A. Taly, E. Rahimy, K. Blumer, D. Coz, N. Hammel, J. Krause, A. Narayanaswamy, Z. Rastegar, D. Wu, et al., Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy, Ophthalmology 126 (2019) 552–564.

[30] K. Yoon, J.-Y. Kim, S.-J. Kim, J.-K. Huh, J.-W. Kim, J. Choi, Explainable deep learning-based clinical decision support engine for MRI-based automated diagnosis of temporomandibular joint anterior disk displacement, Computer Methods and Programs in Biomedicine 233 (2023) 107465.

[31] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, F. Herrera, Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence, Information fusion 99 (2023) 101805.

[32] B. Kim, R. Khanna, O. O. Koyejo, Examples are not enough, learn to criticize! criticism for interpretability, Advances in neural information processing systems 29 (2016).

[33] S. Singla, M. Eslami, B. Pollack, S. Wallace, K. Batmanghelich, Explaining the black-box smoothly—a counterfactual approach, Medical image analysis 84 (2023) 102721.

[34] J. J. Thiagarajan, K. Thopalli, D. Rajan, P. Turaga, Training calibration-based counterfactual explainers for deep learning models in medical image analysis, Scientific reports 12 (2022) 597.

[35] A. J. Maâlej, M. Krichen, M. Jmaïel, A comparative evaluation of state-of-the-art load and stress testing approaches, International Journal of Computer Applications in Technology 51 (2015) 283–293.

[36] M. Barberio, M. Scisciò, S. Vallières, F. Cardelli, S. Chen, G. Famulari, T. Gangolf, G. Revet, A. Schiavi, M. Senzacqua, et al., Laser-accelerated particle beams for stress testing of materials, Nature communications 9 (2018) 372.

[37] M. Sorge, Stress-testing financial systems: an overview of current methodologies (2004).

[38] A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, et al., Underspecification presents challenges for credibility in modern machine learning, Journal of Machine Learning Research 23 (2022) 1–61.

[39] T. Elvira, T. T. Procko, L. Vonderhaar, O. Ochoa, Exploring testing methods for large language models, in: 2024 International Conference on Machine Learning and Applications (ICMLA), 2024, pp. 1152–1157. doi:10.1109/ICMLA61862.2024.00177.

[40] F. Pérez-García, S. Bond-Taylor, P. P. Sanchez, B. van Breugel, D. C. Castro, H. Sharma, V. Salvatelli, M. T. A. Wetscherek, H. Richardson, M. P. Lungren, A. Nori, J. Alvarez-Valle, O. Oktay, M. Ilse, Radedit: Stress-testing biomedical vision models via diffusion image editing, in: A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, G. Varol (Eds.), Computer Vision – ECCV 2024, Springer Nature Switzerland, Cham, 2025, pp. 358–376.

[41] L. Cunha, C. Soares, A. Restivo, L. F. Teixeira, GASTeN: Generative adversarial stress test networks, in: International Symposium on Intelligent Data Analysis, Springer, 2023, pp. 91–102.

[42] I. Gomes, L. F. Teixeira, J. N. Van Rijn, C. Soares, A. Restivo, L. Cunha, M. Santos, Finding patterns in ambiguity: Interpretable stress testing in the decision boundary, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 8316–8321.

[43] A. E. Baia, V. Poggioni, A. Cavallaro, Black-box attacks on image activity prediction and its natural language explanations, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 3686–3695.