FunnyNodules: A Customizable Medical Dataset Tailored for Evaluating Explainable AI

Luisa Gallée [1,2] (iD)   LUISA.GALLEE@UNI-ULM.DE

[1] *Experimental Radiology, Ulm University Medical Center, Ulm, Germany*
[2] *XAIRAD - Cooperation for Artificial Intelligence in Experimental Radiology, Ulm, Germany*

Yiheng Xiong [1,2] (iD)

Meinrad Beer [2,3] (iD)

[3] *Department of Diagnostic and Interventional Radiology, Ulm University Medical Center, Ulm, Germany*

Michael Götz [1,2] (iD)

Abstract

Densely annotated medical image datasets that capture not only diagnostic labels but also the underlying reasoning behind these diagnoses are scarce. Such reasoning-related annotations are essential for developing and evaluating explainable AI (xAI) models that reason similarly to radiologists: making correct predictions for the right reasons. To address this gap, we introduce *FunnyNodules*, a fully parameterized synthetic dataset designed for systematic analysis of attribute-based reasoning in medical AI models. The dataset generates abstract, lung nodule–like shapes with controllable visual attributes such as roundness, margin sharpness, and spiculation. Target class is derived from a predefined attribute combination, allowing full control over the decision rule that links attributes to the diagnostic class. We demonstrate how FunnyNodules can be used in model-agnostic evaluations to assess whether models learn correct attribute–target relations, to interpret over- or underperformance in attribute prediction, and to analyze attention alignment with attribute-specific regions of interest. The framework is fully customizable, supporting variations in dataset complexity, target definitions, class balance, and beyond. With complete ground truth information, FunnyNodules provides a versatile foundation for developing, benchmarking, and conducting in-depth analyses of explainable AI methods in medical image analysis.

Keywords: Dataset, Explainable AI, Evaluation

1. Introduction

In medical image analysis, numerous machine learning models and explainable AI (xAI) methods have been proposed (Wang et al., 2022; Shi et al., 2023; Frasca et al., 2024; Bhati et al., 2024; Muhammad and Bendechache, 2024). While model performance is typically well evaluated, other aspects, such as the correctness of model reasoning, that is, whether a model makes the right decision for the right reasons, are often insufficiently assessed (Rudin, 2019). Even for xAI methods that are explicitly designed to improve interpretability, systematic evaluation remains a major challenge (Muhammad and Bendechache, 2024). A key reason for this limitation is the lack of comprehensive ground truth information. In addition to

annotations for the target classes, the evaluation of xAI methods also requires ground truth for the visual explanations at the sample level. Such annotations are rare, particularly in the medical domain where dataset sizes are inherently limited.

To address this gap, we introduce FunnyNodules, a synthetic image dataset specifically designed for evaluating AI and xAI methods in medical imaging. A core feature of FunnyNodules is its comprehensive annotation. Since the image generation process is explicitly controlled, all samples follow predefined appearance features (referred to as attributes), similar to the approach of Hesse et al. (Hesse et al., 2023), who applied this concept to natural images with discrete object parts (e.g., bird beak or wing) as features, and therefore did not include medical-specific aspects (e.g., lesion intensity or shape). The parameterized generative approach enables the automatic creation of complete ground truth information, including target class and attribute labels as well as region-of-interest (ROI) masks, without being affected by inter- or intra-rater variability.

Furthermore, the FunnyNodules framework provides a high degree of customization. Both the variability in image appearance and the decision rules for target classification can be adapted to different levels of complexity. This enables researchers to tailor the dataset to the specific evaluation needs of their models. In contrast to existing synthetic datasets generated using large generative models such as Diffusion Models (Pinaya et al., 2022; Khosravi et al., 2024; Gallée et al., 2025) or GAN (Frid-Adar et al., 2018), the goal of FunnyNodules is not to simulate realistic data. Instead, it focuses on simulating attribute relationships with clearly defined ground truth and full controllability. In addition, this approach does not require any real training data and is free from data-driven biases.

FunnyNodules provides a controlled environment for evaluating AI models and xAI methods in medical image analysis, allowing the investigation of model reasoning behavior without the limitations of real-world datasets. In this paper, we

- introduce FunnyNodules, a synthetic vision dataset inspired by medical image interpretation and designed for systematic evaluation of AI models as explainability methods.

- demonstrate how FunnyNodules can be used to assess model behavior with respect to attribute sensitivity, reasoning correctness, and trustworthiness.

The dataset generation code and all presented experiments are publicly available at https://github.com/XRad-Ulm/FunnyNodules.

2. Dataset

The aim of this synthetic dataset is to enable comprehensive evaluation of AI models for medical images. Full control over the image generation process allows the creation of depictions that follow defined visual attributes. This approach offers high customizability, scalability, and targeted adjustments, providing strong flexibility for model-agnostic evaluations.

The FunnyNodules framework is based on the idea of representing abstract lung nodules through controllable visual attributes. Rating the malignancy of lung nodules is a clinically relevant and well-studied task in radiology (Furuya et al., 1999; El-Baz et al., 2013). The

diagnostic evaluation strongly depends on quantifiable visual features, such as intensity, roundness, and margin sharpness (Armato III et al., 2015, 2011), which directly correspond to the concept of attributes in explainable AI.

2.1. FunnyNodules

The FunnyNodules dataset, employed in the experiments of this study, comprises abstract nodules described by six visual attributes:

- roundness   (r)   1-*round*, 5-*oval*
- spiculation   (sp)   1-*none*, 5-*marked*
- edge sharpness   (es)   1-*sharp*, 5-*soft*
- size   (s)   1-*small*, 5-*big*
- intensity   (i)   1-*dark*, 5-*bright*
- internal structure (is)   0-*absent*, 1-*present*

The target class is defined based on combinations of these attributes (see Algorithm 1) and is ordinal, ranging from 1 to 5, the same as all attributes except internal structure, which is binary. However, a major advantage of the synthetic FunnyNodules framework is the ability to implement different scales and rules as desired.

For performance evaluation, we use the Within-1-Accuracy metric, following prior attribute-based studies (LaLonde et al., 2020; Gallée et al., 2023). Predictions are considered correct if they deviate by at most ±1 from the ground truth score for ordinal labels, whereas internal structure requires exact agreement.
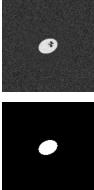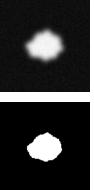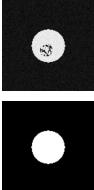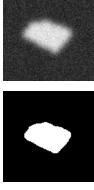
Histogram analysis of FunnyNodules can be found in Appendix A.1.

2.1.1. Image Generation

Each image in the FunnyNodules dataset is synthetically generated through a parameterized algorithm that constructs an abstract nodule as a grayscale image, see samples in Table 1. The function models nodules as elliptical shapes whose geometry, boundary, and intensity are determined by the six attributes. Spiculation is simulated via angular contour perturbations, edge sharpness by Gaussian blurring, and the optional central structure by adding a small textured subregion. Random perturbations of rotation and background noise ensure slight natural variation across samples while preserving exact attribute control.

2.2. Customizability

The dataset framework is highly customizable, allowing systematic investigation of the impact of various factors on model performance. Target definitions can follow simple linear rules or be based on difficult attribute correlated conditions, enabling analysis of model behavior under different levels of task complexity. The set of visual attributes can be varied in number, type, and scale. Experiments can be conducted with or without explicitly defined regions of interest (ROIs), and the presence or absence of background introduces additional complexity in nodule detection and segmentation. Furthermore, image size can be adjusted to approximate realistic scaling, rather than relying on interpolation. Finally, input channels can be selected according to the simulated application domain, e.g., grayscale for chest X-ray

target: 1	target: 2
r: 3	r: 3
sp: 1	sp: 4
es: 1	es: 4
s: 1	s: 3
i: 1	i: 3
is: 1	is: 0

target: 4	target: 4
r: 1	r: 2
sp: 1	sp: 1
es: 1	es: 2
s: 5	s: 4
i: 3	i: 5
is: 1	is: 1

target: 4	target: 5
r: 3	r: 4
sp: 5	sp: 4
es: 4	es: 4
s: 5	s: 4
i: 2	i: 1
is: 0	is: 0

(a)

Algorithm 1: target

$score \leftarrow 0$;
if $is = 0$ **then**
    if $r \geq 4$ **then** $score \leftarrow +2$;
    else if $r \leq 2$ **then** $score \leftarrow -2$;
else if $is = 1$ **then**
    if $r \geq 4$ **then** $score \leftarrow -2$;
    else if $r \leq 2$ **then** $score \leftarrow +2$;
if $sp \geq 4$ **then** $score \leftarrow +2$;
else if $sp \leq 2$ **then** $score \leftarrow -2$;
if $es \geq 4$ **then** $score \leftarrow -2$;
else if $es \leq 2$ **then** $score \leftarrow +2$;
if $s \geq 4$ **then** $score \leftarrow +2$;
else if $s \leq 2$ **then** $score \leftarrow -2$;
if $i = 5$ **then** $score \leftarrow -1$;
else if $i \leq 2$ **then** $score \leftarrow +1$;

if $score \leq -1$ **then** $target \leftarrow 1$;
else if $score = 0$ **then** $target \leftarrow 2$;
else if $score = [1, 2]$ **then** $target \leftarrow 3$;
else if $score = [3, 4]$ **then** $target \leftarrow 4$;
else $target \leftarrow 5$;

(b)

Table 1: **FunnyNodules** is a parametrized nodule-generation framework that enables full control and annotation of samples (a), as well as customizable image complexity. (b) Target rules are fully configurable and can represent complex attribute-correlated rules (is, r) or simpler rules (sp, es, s, i).

or CT images and RGB for dermatological images, providing flexibility for diverse imaging tasks. While these examples illustrate key customization options, the framework is flexible and can accommodate additional variations.

3. Evaluation Methods

FunnyNodules offers broad opportunities for evaluating explainable AI methods, with the following examples illustrating the dataset's key capabilities. Building upon the definitions introduced by Nauta et al. (Nauta et al., 2023), we focus primarily on two complementary aspects of explanations, while acknowledging that others are also possible:

- Correctness: Measures how truthful an explanation is about the model's actual decision process, i.e. high correctness means the explanation reflects accurately what the model is doing.

- Contrastivity: Measures how well an explanation distinguishes the target outcome from other outcomes or events, i.e. high contrastivity means explanations highlight what makes the predicted class different from alternatives.

We used various models for our experiments, including standard models and prototype-based approaches. We trained ResNet-50 (He et al., 2016) and DenseNet-121 (Huang et al., 2017) in a multitask classification setting, where attributes and target predictions are concurrently predicted in the final layer. Models with hierarchical structures, such as Proto-Caps, HierViT, and Concept Bottleneck Networks (joint setting), naturally reflect the attribute-to-target relationship. We note that the models were not extensively optimized for the dataset, because the goal of this section is to demonstrate how FunnyNodules can be used for evaluation rather than to compare model performance.

3.1. Evaluation of a Model's Reasoning

The FunnyNodules dataset enables systematic assessment of whether a model correctly learns the relationships between diagnostic attributes and target classes through controlled variation of attribute values. By keeping the random seed and all other attributes constant, one can analyze how changes in a single attribute affect the target prediction, e.g., *If this nodule were more round, how would the model's prediction change?* This approach allows for the identification of incorrectly learned relationships or biases in the model's reasoning. As shown in Figure 1, 100 samples are generated for each attribute variation.
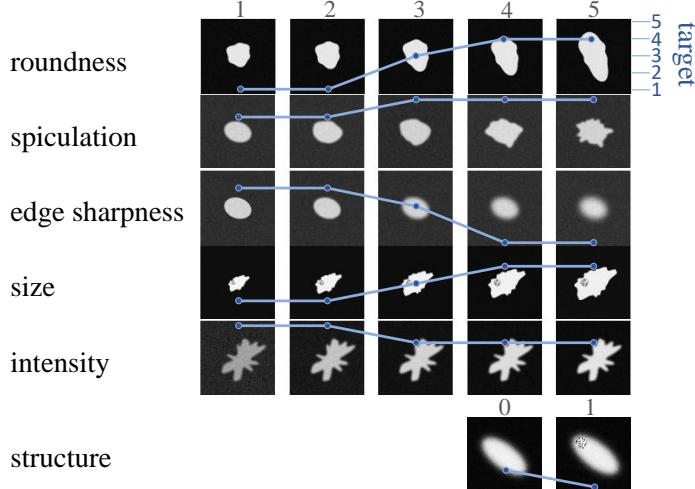

Figure 1: The controlled generative framework allows the generation of images differing in exactly one attribute, which facilitates analyzing how attribute changes influence the target class (blue line).

Figure 2 shows the mean target values for each attribute alongside the model predictions, which enables an assessment of how consistently the model follows the ground truth trends.
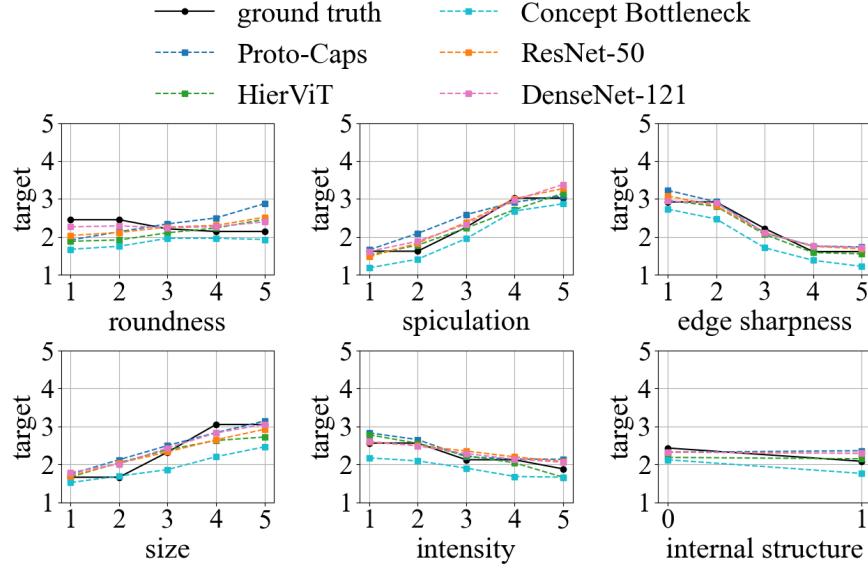
Figure 2: The sensitivity of the models' target predictions to varying attributes reflects whether the target rule was captured correctly, which is mostly the case except for the complex notion of roundness.

FunnyNodules also allows the investigation of more complex rules, such as the attribute roundness (r). Its effect on the target depends on the presence of an internal structure (is), as defined in Algorithm 1:

$$\text{target score} = \begin{cases} +2, & \text{if } (is = 0 \wedge r \geq 4) \text{ or } (is = 1 \wedge r \leq 2), \\ -2, & \text{if } (is = 0 \wedge r \leq 2) \text{ or } (is = 1 \wedge r \geq 4). \end{cases} \quad (1)$$

Using the parameterized generative algorithm of the FunnyNodules framework, we can systematically analyze attribute dependencies by isolating the effects of roundness and internal structure on the target, while fixing all other attribute values at 3 (see Figure 3). Identifying attribute-dependent performance weaknesses is crucial for taking targeted measures, such as adapting model architectures or training procedures to increase sensitivity to specific attributes.

3.2. Investigating the Trustworthiness of Attribute-based Explanations

Attributes form the basis for explaining the model's target class prediction. Their correctness is an important measure for the truthfulness of the explanations and can be quantified in the same way as the target correctness. The relationship between these two performances provides insights into the model's reasoning process. To quantify it, we define a Trust Index

$$TI = P_{\text{target}} - \frac{\frac{1}{N} \sum_{i=1}^{N} A_i}{P_{\text{target}}}, \quad (2)$$
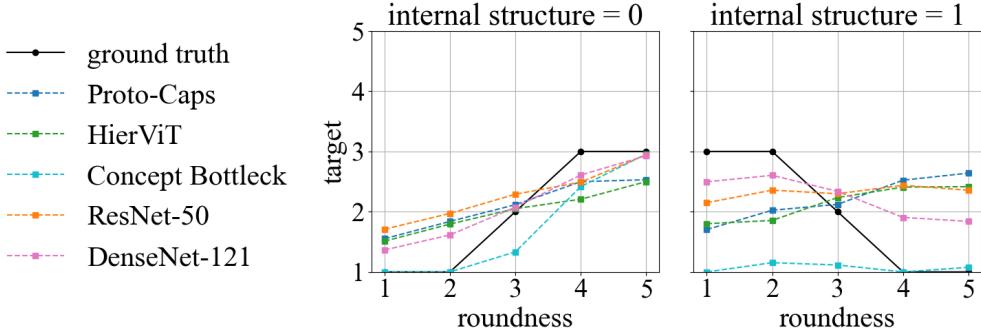
Figure 3: FunnyNodules allows in-depth evaluation of complex decision rules, such as correlated attributes. For example, the effect of roundness on the target depends on the presence of an internal structure. This conditional relation was captured correctly only for one value of internal structure = 0, indicating a general weakness in handling correlated rules across all tested models.

where P_{target} is the target performance, A_i denotes the performance for attribute i, and N is the total number of attributes. The objective is to achieve a Trust Index (TI) close to zero. If $TI \gg 0$, the model exhibits strong target prediction performance, but the decisive attributes are not adequately learned. This suggests that the model's predictions are misguided and therefore should not be trusted. Conversely, if $TI \ll 0$, the model demonstrates strong attribute extraction capabilities, yet the mapping from attributes to the target class is insufficiently learned.

Table 2: **Trust Index (TI)** reflects both the reliability of a prediction and the correctness of its underlying decision rule. $TI \gg 0$ indicates low trustworthiness (e.g., orange), whereas $TI \ll 0$ suggests an insufficiently learned target rule (e.g., pink).

[train,val,test]	[100, 50, 500]		[500, 50, 500]		[1800, 200, 500]	
	P_{target}	TI	P_{target}	TI	P_{target}	TI
ResNet-50	0.828	-0.225	1.0	0.002	1.0	0.001
DenseNet-121	0.824	-0.242	1.0	0.083	0.998	0.009
HierViT	0.848	-0.240	0.982	-0.024	0.997	-0.003
Proto-Caps	0.744	-0.299	1.0	0.001	1.0	0.000
Concept Bottleneck	0.478	-1.341	0.498	-1.222	0.952	-0.083

The Trust Index provides an immediate overview of the relationship between attribute and target prediction performance, see Table 2. By varying the amount of training data, it can also be used to simulate different real-world data availability scenarios. For $TI \gg 0$, attribute extraction should be improved, for example by using differently weighted losses,

whereas for $TI \ll 0$, the mapping from attributes to the target classes should be enhanced, for instance by employing more complex target layers.

3.3. Attribute ROI Masks for Attribute Attention Assessment

For models that perform attribute-based reasoning for diagnostic classification, evaluating attribute prediction becomes a key aspect. While standard evaluation metrics, such as the prediction scores presented above, provide an initial insight into model performance, a more in-depth analysis is required, analogous to the evaluation of target class predictions. One approach is to analyze the model's attention region most relevant for a prediction (Simonyan et al., 2014; Zeiler and Fergus, 2014). Applied at the attribute level, this allows the model to highlight the image regions that most strongly influenced its prediction for each attribute.

Assessing whether this attention is accurate requires ground truth annotations of the relevant regions of interest (ROIs). Creating real-world datasets with such attribute ROI annotations is highly challenging, particularly for medical images, as it requires manual labeling by experts. Although previous work (Choi et al., 2022) attempted automatic post-hoc annotation of attribute-specific regions in real lung nodule datasets, it was limited to just two of the eight attributes.

In contrast, the procedurally generated FunnyNodules dataset enables fully controlled, parametric synthesis of attribute appearances and their corresponding annotations. This allows the generation of precise, attribute-specific ROIs directly during image creation, rather than relying on post-hoc segmentation, and thus provides exact and scalable ground truth for evaluating attribute-level attention, as illustrated in Figure 4.
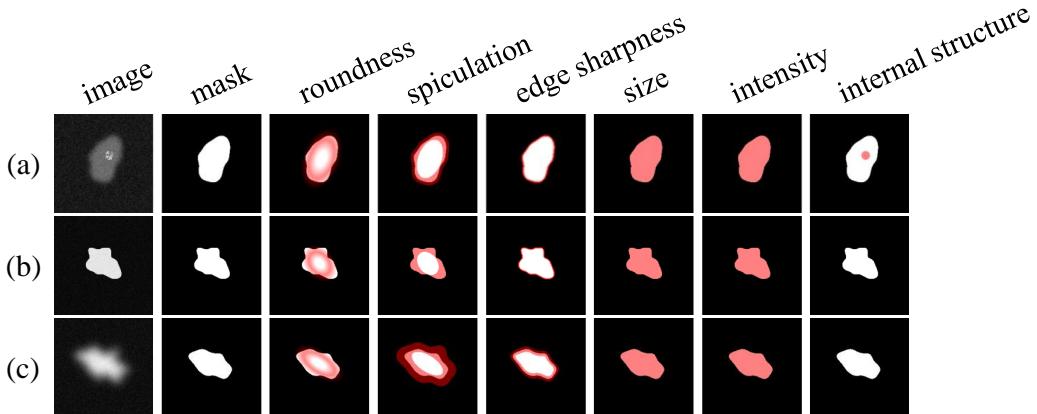

Figure 4: **Attribute ROIs** Ground-truth masks are being created during image generation and enable evaluation of attention in attribute prediction.

For roundness, the nodule's elliptical shape is the key structural feature. Spiculation is defined by the protruding spikes of the main nodule (samples b and c). Edge sharpness depends on the nodule border, with the attention ring narrowing for sharp edges (sample b, $es = 1$) and widening for soft edges (sample c, $es = 5$). Internal structure corresponds to the

internal texture area (sample a). For attributes such as size and intensity, we hypothesize that the entire nodule is relevant for the model's prediction.

3.4. Assessing Prototype Reasoning Correctness

Some xAI models leverage prototypes for explanation, as they resemble human reasoning (e.g., *This bird is an eagle because its beak resembles the eagle prototype*, or *this lung nodule is malignant because its spiculation strongly resembles that of a learned prototype*). Compared to general-domain approaches that use object parts (Chen et al., 2019) or pixel values (Nauta et al., 2021) as explanatory features, this concept can be adapted to medical image diagnosis, with each prototype representing a single attribute as a decisive criterion (Gallée et al., 2025b,a). For an inference sample, the models present a prototype for each attribute, a training image closest to detected inference attribute value. The analysis of these prototype images provides insight into the performance of prototype learning and therefore the truthfulness of prototype reasoning. For more details about the prototype learning method, we refer to the respective works (Gallée et al., 2023, 2025a,b).

Table 3: **Prototype Correctness** is an important measure of the truthfulness of a models' prototype reasoning.

	attribute prototypes						attribute prototypes induced target
	r	sp	es	s	i	is	
HierViT	0.974	0.998	0.880	0.960	1.0	0.964	0.977
Proto-Caps	0.982	0.915	0.945	0.863	0.995	0.979	0.988

3.5. Scalable Dataset Size for Unlimited Evaluation

Dataset size is a fundamental limitation in medical imaging. The sensitive nature of patient data restricts the total number of images available, and the specialized expertise required for annotation further constrains the size of fully annotated datasets. While disease labels can be extracted automatically from clinical reports (Smit et al., 2020), key attributes for reasoning alignment in AI models must be annotated manually (Gallée et al., 2025). This process is costly and further limits the availability of real-world datasets.

Consequently, evaluating model robustness under dataset constraints using real data is inherently limited. In contrast, synthetic datasets, such as FunnyNodules, offer virtually unlimited scalability. As demonstrated in Table 2, the Concept Bottleneck model is highly sensitive to sample size, impacting both predictive performance and the quality of explanations. Such findings pose key considerations when selecting models.

3.6. Further Evaluation Ideas

Beyond the presented approaches for using the dataset in model evaluation, more evaluation strategies can be considered, such as:

Background independence In real medical images, background structures can interfere with nodule detection. The FunnyNodules framework allows the controlled addition of background structures, enabling systematic evaluation of models' robustness to such interference.

Model-specific evaluation Because image generation is fully controlled, differences in model activations can be systematically observed and analyzed, enabling identification of the internal representations affected by specific attribute variations.

4. Discussion and Conclusion

In this work, we introduce the FunnyNodules dataset, which provides extensive opportunities for evaluating diverse aspects of model behavior and explainability methods, with a particular focus on attribute-based reasoning models. The fully controlled parametric synthesis of attribute characteristics enables the creation of multiple types of ground truth annotations, such as attribute-level scores and region-of-interest masks. Moreover, the framework's customizability allows full control over the relationship between attributes and target classes, making it possible to simulate different levels of reasoning complexity and decision rules. We also provided example analyses highlighting the dataset's versatility for evaluating and interpreting AI models.

The controlled conditions of FunnyNodules offer clear advantages in terms of scalability, reproducibility, and precise ground truth. While it is not designed nor capable of replacing the evaluation on real world data, which is a clear limitation, it opens opportunities which are simply not possible for real-life data, with its inherent uncertainty and complexity. Although the obtained performance may not translate to real-life applications, the dataset can be used to obtain valuable insights into the internal mechanisms and tendencies of different model architectures and can serve as a complementary benchmark for comparing model performance. The flexibility of the framework also makes it possible to adapt FunnyNodules to systematically study specific properties of AI models, such as label imbalance, sparse attribute annotations, or attribute attention maps. By increasing data availability and complementing existing real medical datasets, it facilitates model development and enables broader participation in research, including by groups without access to large, comprehensively annotated datasets.

Comprehensive evaluation of explanations generated by xAI models is essential for their intended purpose. One key component is the inclusion of humans in the loop (Dieber and Kirrane, 2022; Rong et al., 2023; Gallée et al., 2024), as xAI targets human-centered aspects such as understanding, acceptance, and usability. Conducting user studies requires access to participants, which in medical applications means medical experts, and involves considerable effort and cost. Datasets such as FunnyNodules provide a complementary approach that reduces this burden, as no human study is needed, and can be used for objective evaluation of the correctness and truthfulness of explanations. The FunnyNodules dataset framework provides a versatile foundation for the systematic evaluation of AI models, particularly explainable approaches, and supports progress towards more transparent and trustworthy medical AI systems.

Acknowledgments

References

Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011. doi: 10.1118/1.3528204.

Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Christian R Meyer, Anthony P Reeves, Bo Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. Data from lidc-idri [data set]. https://www.cancerimagingarchive.net/collection/lidc-idri/, 2015. The Cancer Imaging Archive. doi: 10.7937/K9/TCIA.2015.LO9QL9SX.

Deepshikha Bhati, Fnu Neha, and Md Amiruzzaman. A survey on explainable artificial intelligence (xai) techniques for visualizing deep learning models in medical imaging. *Journal of Imaging*, 10(10):239, 2024. doi: 10.3390/jimaging10100239.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

Wookjin Choi, Navdeep Dahiya, and Saad Nadeem. Cirdataset: a large-scale dataset for clinically-interpretable lung nodule radiomics and malignancy prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2022. doi: 10.1007/978-3-031-16443-9_2.

Jürgen Dieber and Sabrina Kirrane. A novel model usability evaluation framework (muse) for explainable artificial intelligence. *Information Fusion*, 81:143–153, 2022. doi: 10.1016/j.inffus.2021.11.017.

Ayman El-Baz, Garth M Beache, Georgy Gimel'farb, Kenji Suzuki, Kazunori Okada, Ahmed Elnakib, Ahmed Soliman, and Behnoush Abdollahi. Computer-aided diagnosis systems for lung cancer: Challenges and methodologies. *International journal of biomedical imaging*, 2013(1):942353, 2013. doi: 10.1155/2013/942353.

Maria Frasca, Davide La Torre, Gabriella Pravettoni, and Ilaria Cutica. Explainable and interpretable artificial intelligence in medicine: a systematic bibliometric review. *Discover Artificial Intelligence*, 4(1):15, 2024. doi: 10.1007/s44163-024-00114-7.

Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn

performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018. doi: 10.1016/j.neucom.2018.09.013.

Kiyomi Furuya, S Murayama, H Soeda, J Murakami, Y Ichinose, H Yauuchi, Y Katsuda, M Koga, and K Masuda. New classification of small pulmonary nodules by margin characteristics on highresolution ct. *Acta Radiologica*, 40(5):496–504, 1999. doi: 10.3109/02841859909175574.

Luisa Gallée, Catharina Silvia Lisson, Christoph Gerhard Lisson, Daniela Drees, Felix Weig, Daniel Vogele, Meinrad Beer, and Michael Götz. Minimum data, maximum impact: 20 annotated samples for explainable lung nodule classification. *iMIMIC Workshop on Interpretability of Machine Intelligence in Medical Image Computing at MICCAI*, 2025. doi: 10.48550/arXiv.2508.00639.

Luisa Gallée, Meinrad Beer, and Michael Götz. Interpretable medical image classification using prototype learning and privileged information. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2023. doi: 10.1007/978-3-031-43895-0_41.

Luisa Gallée, Catharina Lisson, Christoph Lisson, Daniela Drees, Felix Weig, Daniel Vogele, Meinrad Beer, and Michael Götz. Evaluating the explainability of attributes and prototypes for a medical classification model. In *World Conference on Explainable Artificial Intelligence*, 2024. doi: 10.1007/978-3-031-63787-2_3.

Luisa Gallée, Catharina Lisson, Meinrad Beer, and Michael Götz. Hierarchical vision transformer with prototypes for interpretable medical image classification. Eprint arXiv:2502.08997, 2025a.

Luisa Gallée, Catharina Lisson, Timo Ropinski, Meinrad Beer, and Michael Götz. Protocaps: interpretable medical image classification using prototype learning and privileged information. *PeerJ Computer Science*, 2025b. doi: 10.7717/peerj-cs.2908.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Robin Hesse, Simone Schaub-Meyer, and Stefan Roth. Funnybirds: A synthetic vision dataset for a part-based analysis of explainable ai methods. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. doi: 10.1109/ICCV51070.2023.00368.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Bardia Khosravi, Frank Li, Theo Dapamede, Pouria Rouzrokh, Cooper U Gamble, Hari M Trivedi, Cody C Wyles, Andrew B Sellergren, Saptarshi Purkayastha, Bradley J Erickson, et al. Synthetically enhanced: unveiling synthetic data's potential in medical imaging research. *EBioMedicine*, 104, 2024. doi: 10.1016/j.ebiom.2024.105174.

Rodney LaLonde, Drew Torigian, and Ulas Bagci. Encoding visual attributes in capsules for explainable medical diagnoses. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 294–304. Springer, 2020. doi: 10.1007/978-3-030-59710-8_29.

Dost Muhammad and Malika Bendechache. Unveiling the black box: A systematic review of explainable artificial intelligence in medical image analysis. *Computational and structural biotechnology journal*, 24:542–560, 2024. doi: 10.1016/j.csbj.2024.08.005.

Meike Nauta, Annemarie Jutte, Jesper Provoost, and Christin Seifert. This looks like that, because... explaining prototypes for interpretable image recognition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 441–456. Springer, 2021. doi: 10.1007/978-3-030-93736-2_34.

Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice Van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42, 2023. doi: 10.1145/3583558.

Walter HL Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F Da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Brain imaging generation with latent diffusion models. In *MICCAI workshop on deep generative models*, pages 117–126. Springer, 2022. doi: 10.1007/978-3-031-18576-2_12.

Yao Rong, Tobias Leemann, Thai-Trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. Towards human-centered explainable ai: A survey of user studies for model explanations. *IEEE transactions on pattern analysis and machine intelligence*, 46(4):2104–2122, 2023. doi: 10.1109/TPAMI.2023.3331846.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019. doi: 10.1038/s42256-019-0048-x.

Jin Shi, David Bendig, Horst Christian Vollmar, and Peter Rasche. Mapping the bibliometrics landscape of ai in medicine: methodological study. *Journal of Medical Internet Research*, 25:e45815, 2023. doi: 10.2196/45815.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *2nd International Conference on Learning Representations, ICLR, Workshop Track Proceedings*, 2014. doi: 10.48550/arXiv.1312.6034.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. 2020. doi: 10.48550/arXiv.2004.09167.

Lu Wang, Hairui Wang, Yingna Huang, Baihui Yan, Zhihui Chang, Zhaoyu Liu, Mingfang Zhao, Lei Cui, Jiangdian Song, and Fan Li. Trends in the application of deep learning

networks in medical image analysis: Evolution between 2012 and 2020. *European journal of radiology*, 146:110069, 2022. doi: 10.1016/j.ejrad.2021.110069.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. doi: 10.1007/978-3-319-10590-1_53.

Appendix A. FunnyNodules

A.1. Histogram
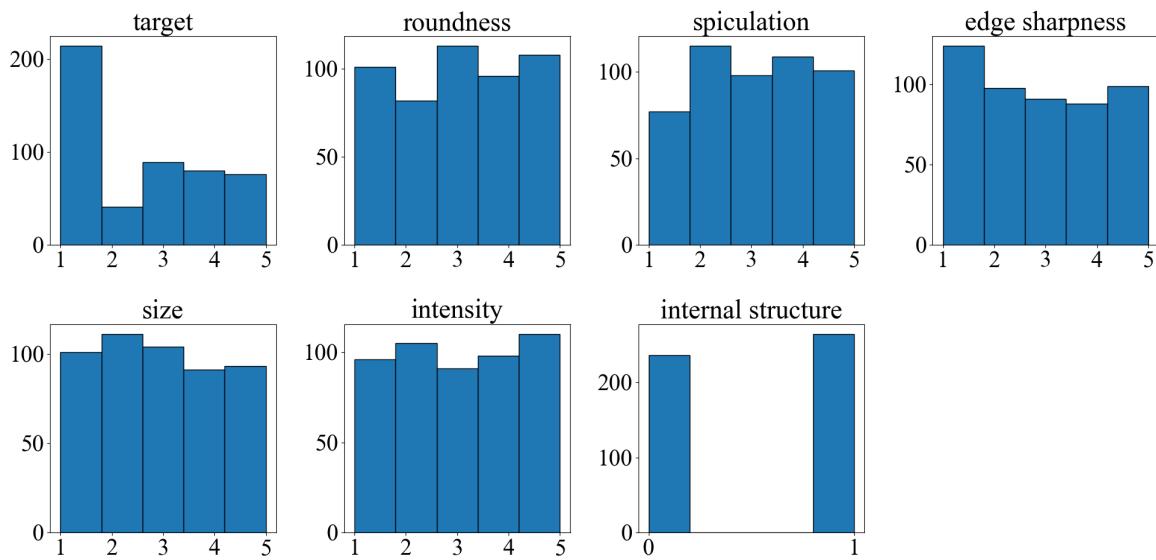
Figure 5: Histogram of 500 randomly generated FunnyNodules images.