# *KANEx*: Translating Kolmogorov-Arnold Networks' Interpretability to Medical Explainability

Krithi Shailya✉[1], Ananya Lakshmi Ravi[1], Venkatanathan K. V.[1], Sowmya S. Sundaram[1], Gokul S. Krishnan[1], Aditi Anand[1,2⋆], and Balaraman Ravindran[1]

[1] Centre for Responsible AI, Wadhwani School of Data Science and AI, Indian Institute of Technology Madras, Chennai, India
krithishailya01@gmail.com
[2] Vanderbilt University School of Medicine, Nashville, USA

**Abstract.** Computer vision models have become highly effective for medical applications, yet their black-box nature continues to undermine clinician trust. In clinical workflows, chest X-ray classifiers are increasingly paired with Vision-Language Models (VLMs) to generate natural-language explanations. However, these systems add linguistic fluency without addressing the underlying opacity of the visual model. With the emergence of Kolmogorov-Arnold Networks (KANs), whose spline-based components provide inherently interpretable functional units, we investigate whether this architectural transparency can be leveraged to produce more trustworthy textual explanations. We introduce *KANEx*, the first ever framework that leverages the symbolic transparency of KANs to ground VLM reasoning. This interpretability also made it possible to design *KAN-Map*, a novel heatmap generation method derived directly from KAN models rather than gradient approximations. We feed these grounded contexts into downstream VLMs for enhanced explainability. Benchmarked on the MIMIC-CXR dataset, we demonstrate that KAN-based architectures with ResNet/ViT baselines demonstrate improved semantic similarity while producing significantly more faithful saliency maps. KAN architectures improve visual localization and downstream reasoning quality by ∼10%. Our findings suggest that grounding linguistic explanations and visual attributions in mathematically interpretable units is a necessary step toward trustworthy medical AI.

**Keywords:** Explainability · Kolmogrov Arnold Networks · Vision Language Models · Trustworthy AI

## 1 Introduction & Background

Artificial Intelligence (AI) models, particularly computer vision models, are being deployed to assist clinicians, helping with tasks such as triage, prioritization, and diagnosis in medical settings [14]. Although these systems often achieve strong performance on benchmark datasets [17, 18], their internal

⋆ Work done while the author was at the Centre for Responsible AI, IIT Madras

decision-making processes are typically inaccessible and erodes clinician trust and sustained deployment. As a result, explainability is a pivotal requirement for medical AI models where transparency, and more importantly accountability are essential [25].

In this context, there has been a rise in the research landscape on explainable AI models for medicine [23]. Prominent techniques include CAM [28], Grad-CAM [19], and occlusion sensitivity [27], which are widely used to verify attention to clinically relevant pathology [21, 1]. Vision-Language Models (VLMs) extend this approach by providing textual explanations, but they remain prone to hallucination and opaque reasoning [7].

To improve transparency, we leverage interpretable Kolmogorov–Arnold Networks (KANs) [16], whose spline-based functions reveal learned nonlinear mechanisms [5]. We introduce a novel approach, **KAN-Map**, a heatmap derived from KAN activations rather than gradients. Unlike Grad-CAM, which relies on a linear approximation of feature importance through gradient backpropagation, KAN-Map directly analyzes forward activations of spline functions to assess the contribution of each spatial patch. This approach eliminates the need for backward passes, yielding higher computational efficiency and capturing higher-order importance cues inherent in KAN's nonlinear representations. Applied to chest X-ray classification, KAN-Map produces functionally grounded and faithful spatial attributions that better guide VLMs toward clinically consistent explanations. Our work presents the first concerted effort, to the best of our knowledge, to enhance the explainability of radiology reports on two fronts: (a) improved visual explainability using KAN-Map and (b) more robust explanations using VLMs that process these enhanced heatmaps.

KAN-based variants improve localization IoU by ∼10–15%, and improve semantic explanation quality (LExT [20]) by up to ∼20% over ResNet/ViT baselines. Our *KAN-Map* further yields ∼25% higher IoU, >20% faithfulness gains, and ∼23% LExT improvement over gradient-based methods.
Our contributions are threefold:

- **KANEx**: A practical X-ray explainability pipeline which provides *visual* heatmap localization, and *textual* explanations in a single system.
- **KAN-Map** A novel heatmap generation method built on the interpretable KAN components.
- A first of a kind systematic empirical comparison of KAN variants of ResNet / ViT hybrid backbones within KANEx.

## 2 Method

We formalize *multi-label chest X-ray diagnosis* with unified explanations and introduce ***KANEx***, a pipeline that realizes this framework.

Let $x \in \mathbb{R}^{C\times H\times W}$, where $C$ is the number of channels, $H \times W$ the spatial dimensions, be an input chest X-ray image, $f(x)$ be a vision model that produces both classification and localization, $y \in \{0,1\}^K$ be a multi-label vector over $K$ clinical findings (e.g., pneumonia, effusion, cardiomegaly etc.)

Our goal is to develop a pipeline (Figure 1) that, for each image $x$, produces: (a) A set of predicted probabilities $f(x) = \hat{y} = (\hat{y}_1, \ldots, \hat{y}_K)$ over the $K$ labels, (b) Spatial explanations in the form of novel heatmaps derived from KAN importance (***KAN-Map***) $h_k(x)$ that localize image regions supporting each predicted label $k$. (c) A *single, holistic textual explanation* $t(x, \{h_k\}_k)$ formed by attention-weighted aggregation $\sum_k \hat{y}_k \cdot h_k(x)$ or multi-channel VLM input, synthesizing evidence across all labels into one practitioner-friendly description.

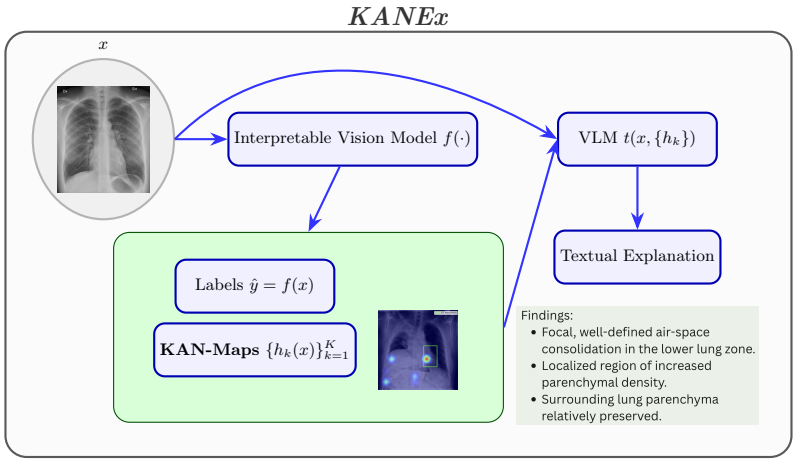


Fig. 1: *Overview of the proposed pipeline KANEx*: An input image $x$ is processed by a KAN backbone $f(\cdot)$ to produce labels $\hat{y}$ and heatmaps $\{h_k(x)\}_{k=1}^K$. These outputs are provided to a VLM which generates textual explanations.

### 2.1 Interpretable Models: Kolmogorov-Arnold Networks

Standard vision models for chest X-ray analysis include ResNets and vision transformers (ViTs) [26, 18]. ResNets [9] use residual connections with convolutional filters to capture local textures and patterns. Vision transformers (ViTs) [6] treat images as sequences of patches, applying self-attention to capture long-range dependencies. We integrate the popular Kolmogorov–Arnold Networks (KANs) [16] into these architectures, replacing neural layers with spline-based edge functions for providing *intrinsic* interpretability through mathematically structured representations. Unlike MLPs that apply fixed nonlinearities to linear combinations at each node, $x_i^{(l+1)} = \sigma\left(\sum_j w_{ij}^{(l)} x_j^{(l)}\right)$, KANs replace learnable weights $w_{ij}$ with univariate learnable functions $\phi_{i,j}$ on network edges: $x_i^{(l+1)} = \Phi_i\left(\sum_{j=1}^{n_l} \phi_{i,j}^{(l)}\left(x_j^{(l)}\right)\right)$, where $\Phi_i$ is a fixed nonlinearity and each $\phi_{i,j}$ is typically parameterized as a B-spline. This structure enables visualization of spline functions and symbolic simplification, exposing interpretable decision

rules [22]. We explore several KAN variants integrated into ResNet and ViT architectures. We consider three KAN variants: **VanillaKAN**, which uses spline-based edge functions [16]; **GroupKAN**, which groups splines for improved high-dimensional efficiency [13]; and **RationalKAN**, which replaces splines with rational activation functions to enhance expressivity [2].

### 2.2 KAN-Map: Spline-Derived Heatmaps

Heatmaps provide visual explanations by highlighting regions that drive model predictions. Traditional Class Activation Mapping (CAM) [28] assumes linear classifiers, while Grad-CAM [19] relies on gradient approximations. We extend this idea to KAN classifiers by directly evaluating the learned spline functions over spatial feature vectors. We pass the visual features from each image region through the KAN's learned spline functions and measure how strongly that region supports the prediction, using these scores to generate the heatmap.

Let $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ denote the final convolutional feature map, where $C$ is the channel dimension and $H \times W$ are spatial dimensions. For each spatial location $(x, y)$, the feature vector $\mathbf{F}_{x,y} \in \mathbb{R}^C$ is passed through the trained KAN classifier. The class-$k$ heatmap value is obtained as

$$h_k(x, y) = \phi_{2,k}(\phi_1(\mathbf{F}_{x,y}))\,,$$

where $\phi_1 : \mathbb{R}^C \to \mathbb{R}^D$ and $\phi_2 : \mathbb{R}^D \to \mathbb{R}^K$ denote the first and second KAN spline layers, respectively. Each univariate spline unit is parameterized using a B-spline basis with grid size $G$=5 and order $p$=3:

$$\phi_{j,c}(z) = \sum_m w_{j,c,m}\, B_m(z),$$

where $B_m$ are B-spline basis functions and $w_{j,c,m}$ are learned coefficients.

**Implementation:** The feature map is reshaped to $(HW)\times C$ and forward-passed through the trained KAN head to obtain class logits for every spatial location. The logits corresponding to class $k$ are reshaped to $H \times W$, passed through ReLU, and normalized to $[0, 1]$ to obtain the final heatmap. This forward-only procedure requires no gradients or linear approximations and directly reflects the learned spline mappings.

### 2.3 Prompting the Vision Language Model (VLM)

The VLM receives saliency-enhanced images along with the predicted diagnosis as input and is prompted to generate a natural-language explanation justifying that diagnosis. We design a detailed prompt that encompasses the medical context, the X-ray, the heatmap and prompt radiology report generation.

**Prompt for extracting explanations from normal/enhanced images:**
**Input:** A chest X-ray image and the target diagnosis label.
**Instruction:** Analyze the provided chest X-ray and generate a structured radiology report using language appropriate for a physician. The diagnosis for this case is `{diagnosis}`.
**Output format:**
*Findings:* Describe the radiographic abnormalities that support the given diagnosis, or state if findings are subtle.
*Explanation:* Explain, using expert radiologic reasoning, how the imaging findings support the diagnosis and discuss relevant differential considerations if applicable.
**Example:** <Example taken from ground truth>
Now generate a report for the given chest X-ray using the same format

## 3 Experiments & Results

We describe further details of the dataset, experimental setup and metrics below.

***Custom Vision-Language Dataset for Explainability:*** We constructed a custom multimodal dataset to comprehensively evaluate visual explainability and explanation quality in our experiments. We use a subset of the chest radiographs from MIMIC-CXR [11, 12]. These were matched, via subject and study identifiers, to structured diagnosis codes from MIMIC-IV and narrative "Brief Hospital Course" notes from MIMIC-IV-Note. For each matched case, we extracted the X-ray and radiology findings, CheXpert labels for training, retrieved note passages referencing the X-ray, and synthesized these elements into a single clinically grounded natural language rationale. For evaluation, we used a subset of data points which were derived from MS-CXR [3] that provided bounding boxes for MIMIC data for heatmap evaluation. This process resulted in a total of 20k cases, each with an image, a diagnosis, a segmentation, and a comprehensive ground-truth explanation. From these cases, 30% were held out as an unseen test set for evaluation.

***Experimental Setup:*** For the vision backbones, we used both ResNet and Vision Transformer (ViT) architectures as baselines. To study the interaction between KANs and existing architectures in a controlled manner, we replaced the final MLP classification head with the KAN variants while freezing the pretrained backbone in all experiments. This design choice ensured that any changes in explanation quality or interpretability could be attributed specifically to the KAN-based head rather than differences in feature extraction or overall model capacity. All backbone weights, training configurations, and optimization settings were kept identical across models to maintain a fair and consistent comparison. To evaluate visual explainability, we compared our proposed **KAN-Map** with two other methods: Grad-CAM [19], gradient-weighted attention rollout (Attn-R)[10].

For generating explanations, we used LLaVA (Large Language and Vision Assistant) [15], an open-source instruction-tuned multimodal model built on the

LLaMA backbone, due to its strong image–text reasoning ability to generate structured long-form explanations aligned with physician reports, and because prior evidence indicates that general-purpose models often outperform finetuned medical models in generative explanation quality [20]. The model was prompted to respond as a physician interpreting a chest X-ray for a clinical audience, ensuring consistent tone and structure with the ground truth. The curated dataset, prompts used, the modified model architectures and training configurations are available in our code base[3].

***Evaluation Metrics:*** We evaluate model performance across two complementary dimensions (Table 1): *Visual Explainability* and *Explanation Quality*. We examine localizations (Table 2) using IoU, Energy@10, Area@50 and Faithfulness (Table 1) across Grad-CAM (G), and our KAN-Map (K) method. In the case of ViTs, we additionally compare with the Attn-R (A) method. We also present some of (Figure 2) the heatmaps. Finally, we compare the text explanation quality generated by the VLM directly against the various KAN types with the three different heatmap configurations (Grad-CAM, Attn-R, KAN-Map) using LExT-C, which provides a NER-based embedding overlap to align We also look at how the explanations differ when we use KAN-Map based images versus the baseline (Figure 3). Friedman test was run to compare model distributions on IoU. The distributions are significantly different with $p < 0.001$.

Table 1: Definitions of evaluation metrics used in the proposed fraemwork

| **Category** | **Metric** | **Definition (Range, Direction)** |
|---|---|---|
| Visual Explainability | IoU | Intersection-over-Union between predicted heatmaps and ground-truth bounding boxes |
| | Energy@10 | The fraction of total heatmap energy contained within the top 10% most salient pixels [19] |
| | Area@50 | The fraction of image area required to capture 50% of total heatmap energy [19] |
| | Faithfulness | We assess explanation faithfulness as the difference between *Deletion AUC*, which measures the drop in model confidence when salient regions are progressively removed and *Insertion AUC*, which measures confidence recovery when salient regions are gradually introduced [19]. |
| Explanation Quality | LExT-C | The correctness subset of LExT measures clinical alignment and correctness between generated explanations and reports through lexical and factual overlap [20]. |

[3] https://github.com/cerai-iitm/KANEx

Table 2: Visual explainability comparison across RN50, ViT, and their KAN variants. Bold values indicate the best performance within each backbone family and explanation method.

| | **IoU ↑** | | | **Area@50 ↓** | | | **Energy@10 ↑** | | | **Faithfulness ↑** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | G | A | K | G | A | K | G | A | K | G | A | K |
| RN50 | 0.062 | – | – | 0.184 | – | – | 0.268 | – | – | 3.910 | – | – |
| ViT | 0.066 | 0.067 | – | 0.011 | 0.016 | – | 0.392 | 0.381 | – | 1.050 | 2.450 | – |
| RN50(KAN) | 0.060 | – | 0.069 | 0.166 | – | 0.075 | 0.280 | – | 0.568 | 3.893 | – | 9.116 |
| RN50(rKAN) | 0.061 | – | **0.076** | 0.197 | – | 0.114 | 0.257 | – | 0.432 | **5.554** | – | **9.789** |
| RN50(gKAN) | **0.063** | – | 0.075 | **0.151** | – | **0.075** | **0.340** | – | **0.570** | 3.086 | – | 9.300 |
| ViT(KAN) | 0.067 | 0.068 | 0.079 | 0.010 | **0.017** | 0.023 | 0.388 | 0.376 | 0.365 | 0.960 | 3.330 | 6.200 |
| ViT(rKAN) | 0.068 | 0.070 | **0.079** | 0.010 | 0.019 | 0.025 | 0.386 | 0.372 | 0.358 | **1.010** | **4.850** | **7.500** |
| ViT(gKAN) | **0.070** | **0.074** | 0.077 | **0.008** | **0.017** | **0.021** | **0.390** | **0.379** | **0.367** | 0.840 | 4.050 | 6.270 |

### 3.1 Discussion

**KAN-Map** consistently yielded improved localization and faithfulness compared with gradient-based baselines. Its evaluation of learned spline mappings at the spatial level produced heatmaps that were both more compact and more aligned with class-relevant behavior, demonstrating the benefit of incorporating KAN-aware spatial structure into interpretability evaluation.

The *Inter-KAN* analysis reveals distinctive performance patterns among the KAN variants. The GroupKAN model achieved the best overall classification and alignment with its internal function, attaining the highest KAN-Map Faithfulness and LExT scores. While RationalKAN reached the highest IoU with KAN-Map, the VanillaKAN and GroupKAN variants generated more compact and concentrated heatmaps, indicated by the lowest *Area@50* and highest *Energy@10* scores. This observation suggests a trade-off in interpretability: the most spatially accurate localization does not always correspond to the most focused explanatory pattern. Complementarily, it also helps translate theoretical interpretability: RationalKAN's rational spline functions provide smoother and more globally stable function approximations that improve attribution faithfulness, while GroupKAN's grouped basis decomposition preserves spatial feature separability, enabling sharper and more localized activations that translate into higher IoU and localization quality.

We also observe that KAN-Map based localisations (Figure 3) generally produce higher-quality textual explanations than GRAD-CAM and KAN variants achieve more plausible alignment between generated explanations and model reasoning, while the absolute localization accurazy indicates potential for enhancing the spatial alighment of explanations. Several interpretative factors merit consideration when assessing these findings. The evaluation of faithfulness depends on the chosen perturbation and inpainting strategies, and alternate methodologies could lead to different quantitative outcomes. Our work thus fits in with the

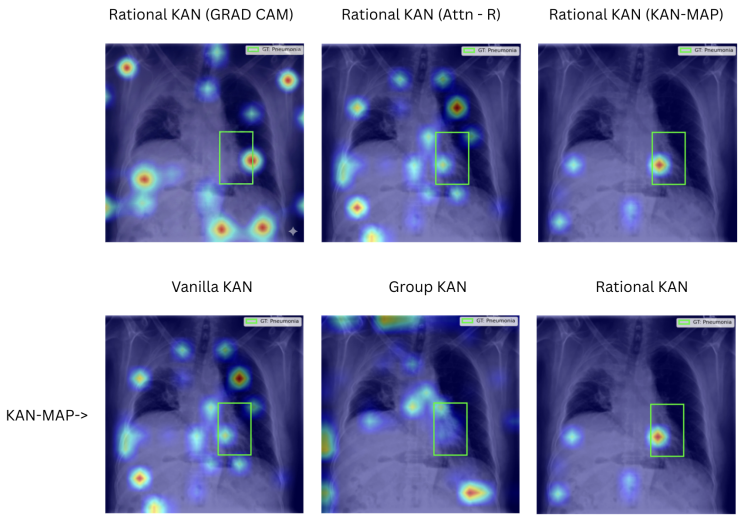


Fig. 2: Ablation Studies for KAN-Map: Localization methods for Rational KAN (Grad-CAM, Attn-R, KAN-Map) with ground truth bounding boxes (top) and KAN-Map across VanillaKAN, GroupKAN, RationalKAN (bottom)

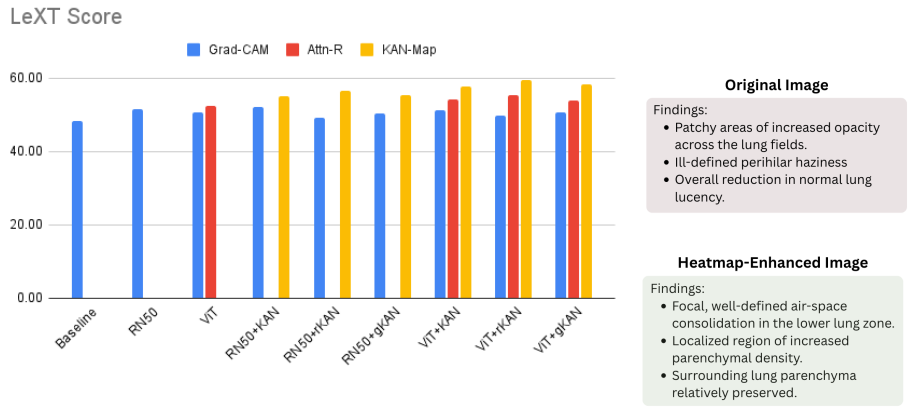


Fig. 3: LExT bar plot comparing baseline, RN/ViT, and KAN variants; corresponding explanations reveal sub-par baseline explanation (top-right) vs. improved KAN-Map explanation (bottom-right).

growing literature on semantic alignment and explainability, especially in the field of radiology reports [24, 8, 4].

To further underscore the trustworthiness and practical value of our pipeline, we plan to evaluate the proposed method with clinician oversight to assess its interpretability and relevance in real-world diagnostic contexts. Our findings encourage the adoption and further fine-tuning of KAN-based variants to more fully leverage their intrinsic functional interpretability, providing researchers, clinical-AI developers, and radiologists with a principled framework for explanation auditing, faithful VLM-grounded reporting, model debugging, and ultimately safer and more transparent clinical deployment.

## 4 Conclusion

In summary, we introduce a unified pipeline combining KANs with VLMs for multi-label chest X-ray diagnosis and practitioner-useful textual explanations. Evaluating VanillaKAN, GroupKAN, and RationalKAN variants across ResNet and ViT backbones, we show how intrinsic interpretability, exposed by our novel KAN-Map heatmaps, enables linking findings to localized image evidence. This work bridges model interpretability, visual grounding, and clinical utility, working towards providing a practical framework for trustworthy medical AI.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. p. 9525–9536. NIPS'18, Curran Associates Inc., Red Hook, NY, USA (2018)
2. Aghaei, A.A., Hosseinzadeh, M., Parand, K.: rkan: Rational kolmogorov-arnold networks. Neural Networks p. 108888 (2026)
3. Boecking, B., Usuyama, N., Bannur, S., Coelho de Castro, D., Schwaighofer, A., Hyland, S., Sharma, H., Wetscherek, M.T., Naumann, T., Nori, A., Alvarez Valle, J., Poon, H., Oktay, O.: MS-CXR: Making the Most of Text Semantics to Improve Biomedical Vision-Language Processing. PhysioNet (Nov 2024). https://doi.org/10.13026/9g2z-jg61, https://doi.org/10.13026/9g2z-jg61, version 1.1.0
4. Deperrois, N., Matsuo, H., Ruipérez-Campillo, S., Vandenhirtz, M., Laguna, S., Ryser, A., Fujimoto, K., Nishio, M., Sutter, T.M., Vogt, J.E., et al.: Radvlm: A multitask conversational vision-language model for radiology (2025)
5. Di Marino, A., Bevilacqua, V., Ciaramella, A., De Falco, I., Sannino, G.: Ante-hoc methods for interpretable deep models: A survey. ACM Comput. Surv. **57**(10) (May 2025). https://doi.org/10.1145/3728637
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021), https://openreview.net/forum?id=YicbFdNTTy

7. Ghassemi, M., Oakden-Rayner, L., Beam, A.L.: The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health **3**(11), e745–e750 (Nov 2021). https://doi.org/10.1016/S2589-7500(21)00208-9
8. Gu, D., Gao, Y., Zhou, Y., Zhou, M., Metaxas, D.: Radalign: Advancing radiology report generation with vision-language concept alignment. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 484–494. Springer (2025)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2016)
10. Jo, S., Jang, G., Park, H.: Gmar: gradient-driven multi-head attention rollout for vision transformer interpretability. In: 2025 IEEE International Conference on Image Processing (ICIP). pp. 582–587. IEEE (2025)
11. Johnson, A., Pollard, T., Mark, R., Berkowitz, S., Horng, S.: MIMIC-CXR Database. PhysioNet (Jul 2024). https://doi.org/10.13026/4jqj-jw95, version 2.1.0
12. Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data **6**(1), 317 (Dec 2019). https://doi.org/10.1038/s41597-019-0322-0
13. Li, G., Majeed, A.P., Ateeq, M., Nguyen, A., Zhang, F.: Groupkan: Rethinking nonlinearity with grouped spline-based kan modeling for efficient medical image segmentation. arXiv preprint arXiv:2511.05477 (2025)
14. Linguraru, M.G., Bakas, S., Aboian, M., Chang, P.D., Flanders, A.E., Kalpathy-Cramer, J., Kitamura, F.C., Lungren, M.P., Mongan, J., Prevedello, L.M., Summers, R.M., Wu, C.C., Adewole, M., Kahn, C.E.: Clinical, cultural, computational, and regulatory considerations to deploy ai in radiology: Perspectives of rsna and miccai experts. Radiology: Artificial Intelligence **6**(4), e240225 (2024). https://doi.org/10.1148/ryai.240225
15. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. vol. 36, pp. 34892–34916 (2023)
16. Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljacic, M., Hou, T., Tegmark, M.: Kan: Kolmogorov–arnold networks. In: Yue, Y., Garg, A., Peng, N., Sha, F., Yu, R. (eds.) International Conference on Learning Representations. vol. 2025, pp. 70367–70413 (2025)
17. Manzari, O.N., Ahmadabadi, H., Kashiani, H., Shokouhi, S.B., Ayatollahi, A.: Medvit: a robust vision transformer for generalized medical image classification. Computers in Biology and Medicine **157**, 106791 (2023)
18. Manzari, O.N., Asgariandehkordi, H., Koleilat, T., Xiao, Y., Rivaz, H.: Medical image classification with kan-integrated transformers and dilated neighborhood attention. Applied Soft Computing p. 114045 (2025)
19. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision (2017)
20. Shailya, K., Rajpal, S., Krishnan, G.S., Ravindran, B.: Lext: Towards evaluating trustworthiness of natural language explanations. In: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency. FAccT '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3715275.3732104

21. Shi, C., Rezai, R., Yang, J., Dou, Q., Li, X.: A Survey on Trustworthiness in Foundation Models for Medical Image Analysis (Oct 2024). https://doi.org/10.48550/arXiv.2407.15851, http://arxiv.org/abs/2407.15851, arXiv:2407.15851 [cs]
22. Somvanshi, S., Javed, S.A., Islam, M.M., Pandit, D., Das, S.: A survey on kolmogorov-arnold network. ACM Computing Surveys **58**(2) (2025)
23. Sun, Q., Akman, A., Schuller, B.W.: Explainable artificial intelligence for medical applications: A review. ACM Trans. Comput. Healthcare **6**(2) (Feb 2025). https://doi.org/10.1145/3709367
24. Thawakar, O.C., Shaker, A.M., Mullappilly, S.S., Cholakkal, H., Anwer, R.M., Khan, S., Laaksonen, J., Khan, F.: Xraygpt: Chest radiographs summarization using large medical vision-language models. In: Proceedings of the 23rd workshop on biomedical natural language processing. pp. 440–448 (2024)
25. Wong, A., Roslan, N.L., McDonald, R., Noor, J., Hutchings, S., D'Costa, P., Via, G., Corradi, F.: Clinical obstacles to machine-learning pocus adoption and system-wide ai implementation (the compass-ai survey). The Ultrasound Journal **17**(1), 32 (Jul 2025). https://doi.org/10.1186/s13089-025-00436-2
26. Xu, W., Fu, Y.L., Zhu, D.: Resnet and its application to medical image processing: Research progress and challenges. Computer Methods and Programs in Biomedicine **240**, 107660 (Oct 2023), https://www.sciencedirect.com/science/article/pii/S0169260723003255
27. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 818–833. Springer International Publishing, Cham (2014)
28. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2016)