Human-Centered Contrastive Explanations for Medical Imaging using VAE-AC-WGAN

Mirtha Lucas[a], Miguel Lerma[b], Jacob Furst[a], Daniela Raicu[a]
[a]DePaul University, Chicago, IL USA; [b]Northwestern University, Evanston, IL USA

ABSTRACT

We introduce a novel contrastive explanation framework for deep learning models in the medical domain based on Variational AutoEncoder Auxiliary Classifier Wasserstein Generative Adversarial Networks (VAE-AC-WGANs). This framework is designed to generate synthetic medical images with and without pathological features, thereby enhancing interpretability and enabling clearer insights into model behavior. Unlike traditional attribution methods, VAE-AC-WGAN supports more faithful reconstructions and targeted perturbations of key image features. Building upon previous work, our approach addresses limitations in image quality and computational efficiency. We evaluate our method qualitatively and quantitative on brain MRI and the Lung Image Database Consortium (LIDC) datasets and present comparisons with prior techniques, aiming to contribute toward more transparent and trustworthy AI-assisted clinical decision-making systems.

Keywords: Explainable AI, Contrastive Explanations, Medical Imaging, GAN, Variational Autoencoder, Counterfactuals.

1. DESCRIPTION OF PURPOSE

The goal of this research is to improve model explainability in medical imaging by providing contrastive explanations. Rather than localizing attention, our method reconstructs and contrasts images with and without target features (e.g., presence vs. absence of pathology), thereby aligning explanations with human reasoning.

2. METHODOLOGY

We introduce the VAE-AC-WGAN architecture combining: 1) Variational AutoEncoder (VAE) to encode images and bypass latent search; 2) Auxiliary Classifier GAN structure, using class labels as input; and 3) Wasserstein GAN with gradient penalty for training stability.

Our framework is shown in Fig. 1. The generator G receives a latent vector and a label to synthesize images. A classifier C, encoder E, and discriminator D collaborate through backpropagation using a combination of loss terms: reconstruction loss (mean squared error, MSE), classification loss (binary cross-entropy, BCE), and Kullback–Leibler (KL) divergence.

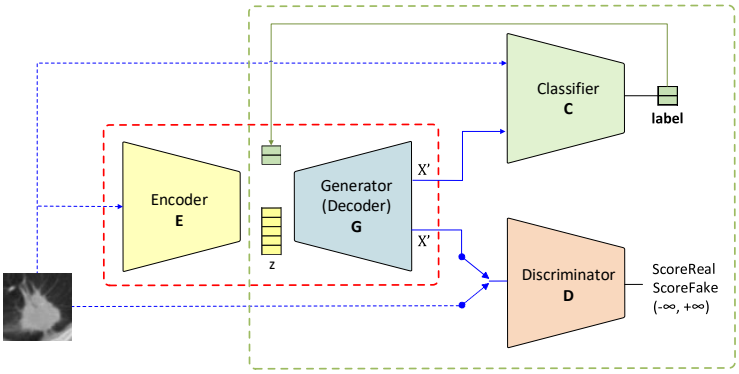

Figure 1: Architecture VAE-AC-WGAN

A key mechanism in our framework is Feature Inversion, which enables the generation of contrastive images. This is achieved by toggling the class label at the decoding stage to reconstruct either a pathological or a non-pathological version of the input image. For example, given an image of a healthy lung or brain, our system can reconstruct an alternative version that simulates the presence of a lesion or tumor, and vice versa. This inversion mechanism does not rely on manually editing the image or searching the latent space instead, it directly switches the output condition during decoding, producing

semantically opposite versions of the same anatomical structure. This facilitates the creation of contrastive explanations that align with human reasoning. This feature inversion capability allows us to highlight what specific visual features the model associates with a particular pathology, thereby offering deeper interpretability compared to traditional saliency methods like Grad-CAM [6].

To quantify explanation effectiveness, we employ the Area Under the Receiver Operating Characteristic Curve (AUROC). Following [7], a continuous anomaly score is computed from the anomaly map (mean of differences between input and negative reconstructed image), and a threshold is used to make a binary decision to distinguish between images with and without anomaly.

3. RESULTS

To show generalizability of our proposed framework, we conducted our experiments for two anatomical structures: brain and lung. We used the Brain MRI Dataset (BRATS2020 [2]) consisting of 95 tumor and 144 non-tumor axial brain MRI images (64×64), with heads approximately aligned, and the LIDC Dataset [1], which includes 228 non-spiculated and 31 markedly spiculated lung nodule images (64×64, max slice per nodule). Using Feature Inversion, we were able to reconstruct not only faithful versions of the original brain MRIs, but also their contrastive images that simulate the opposite class (e.g., converting a tumor image into a tumor-free version, and vice versa), highlighting pathological differences in a controlled and interpretable way. For lung images, Feature Inversion also revealed pathological differences, although with some entanglement. Data augmentation was done to balance the dataset - including random ±180° rotations and horizontal flips - to address class imbalance, particularly between non-spiculated and markedly spiculated nodules.

Our preliminary results show successful reconstruction and manipulation of tumor presence in brain MRIs and spiculated features in lung nodules (Fig.2). Unlike medXGAN [3], our architecture does not require latent optimization**,** meaning we do not need to perform iterative search in the latent space to generate a plausible reconstruction. Instead, our model reconstructs images directly via the encoder-decoder pathway conditioned on class labels, making it computationally more efficient and fully deterministic at inference time. Additionally, our method consistently generates sharper and more accurate reconstructions, especially when reconstructing fine-grained pathology features.

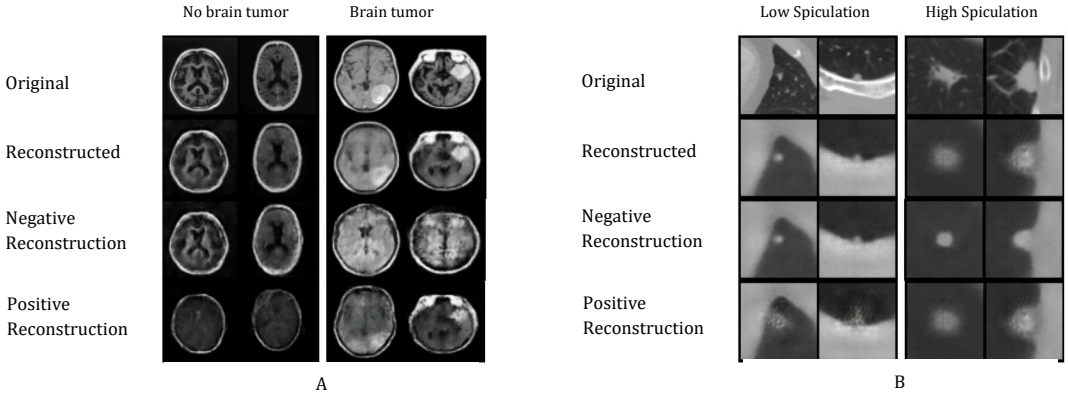

Figure 2A (left) shows brain MRI outputs with original, reconstructed, and tumor-removed images, illustrating our model's contrastive capabilities. Figure 2B (right) shows similar results for LIDC-IDRI scans.

Using AUROC metric to measure the separatation between images with pathology present versus without pathology (Table 1), we show that the VAE-AC-WGAN approach produces statitically significant better results compared to medXGAN across reconstruction variants and ground truth baselines (Dataset Label or Classifier Output) as illustrated in Figures 3 to 6.

Table 1: AUROC for medXGAN and VAE_AC-WGAN

		medXGAN (AC + GAN)	VAE + AC + WGAN
	Reference truth	**Metric: AUROC from Anomaly Maps**	**Metric: AUROC from Anomaly Maps**
Dataset: BRATS2020 (axial cross-sectional only)	Classifier Output	0.69	0.91
	Dataset Label	0.71	0.99
Dataset: LIDC (spiculation feature)	Classifier Output	0.82	1.00
	Dataset Label	0.72	0.80

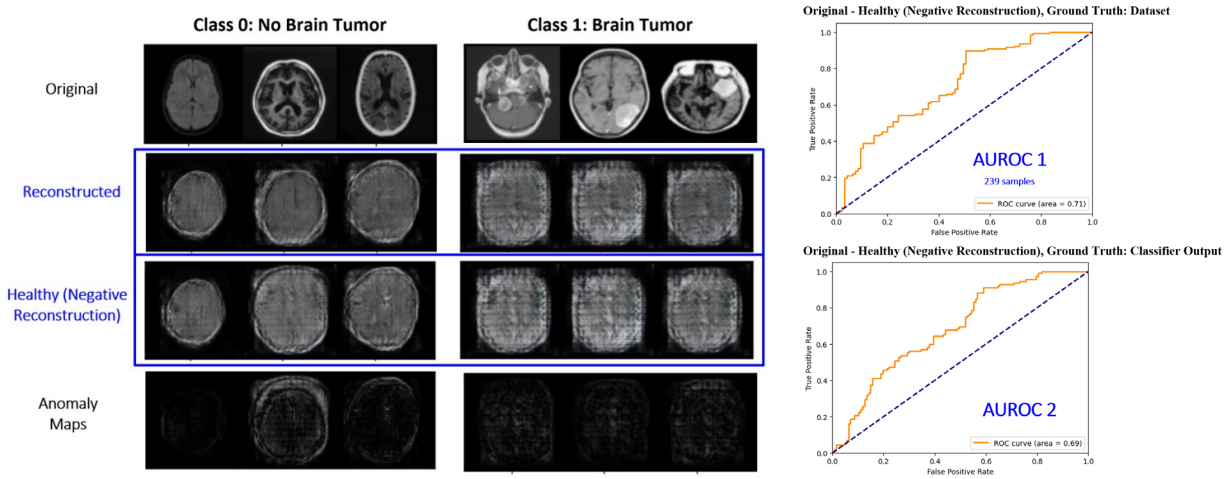

Figure 3: AUROC for medXGAN on brain dataset

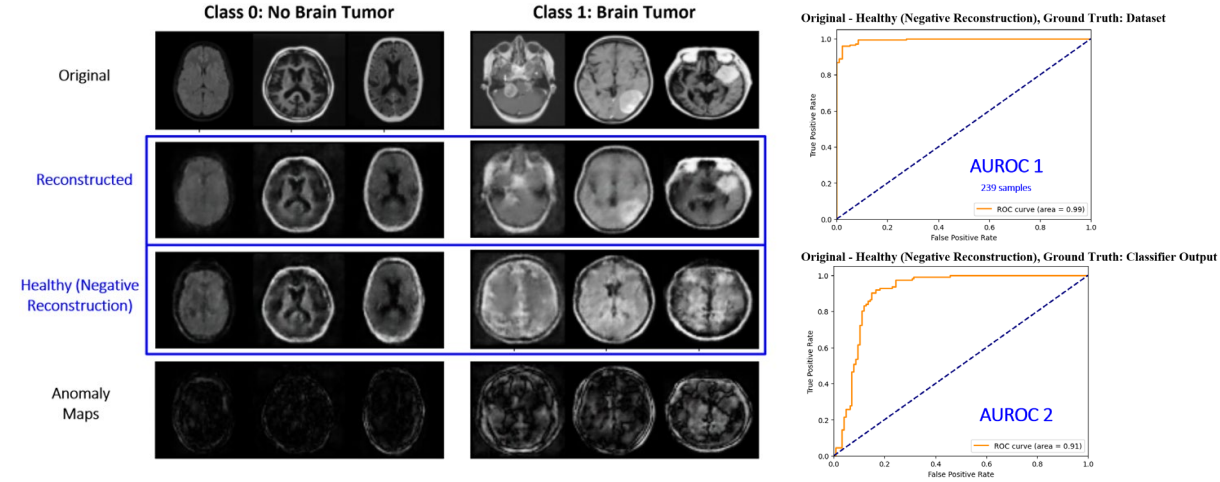

Figure 4: AUROC for VAE-AC-WGAN on brain dataset

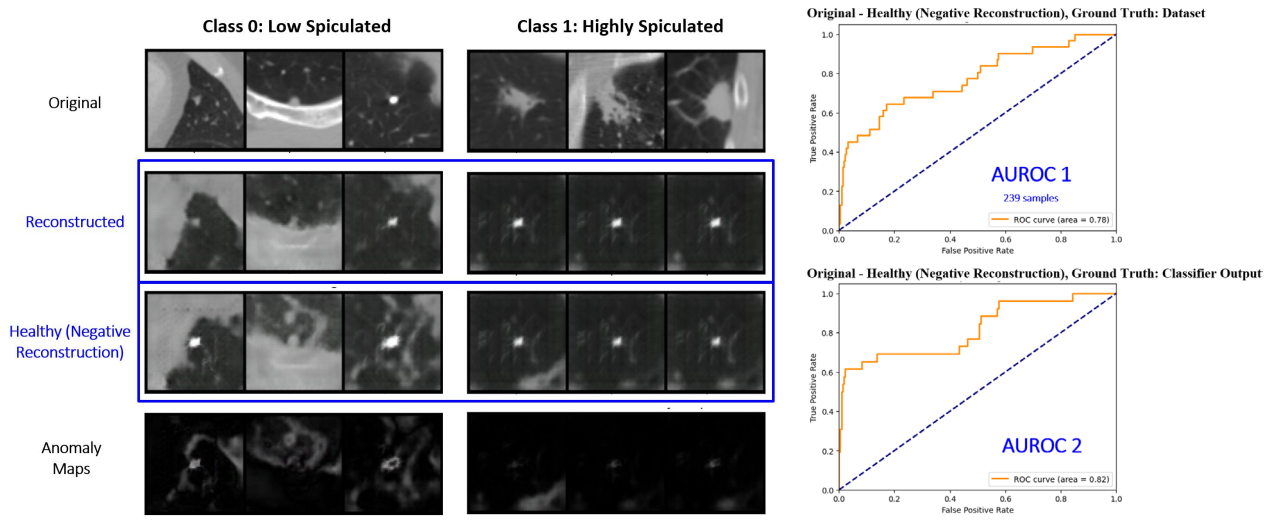

Figure 5: AUROC for medXGAN on the LIDC dataset

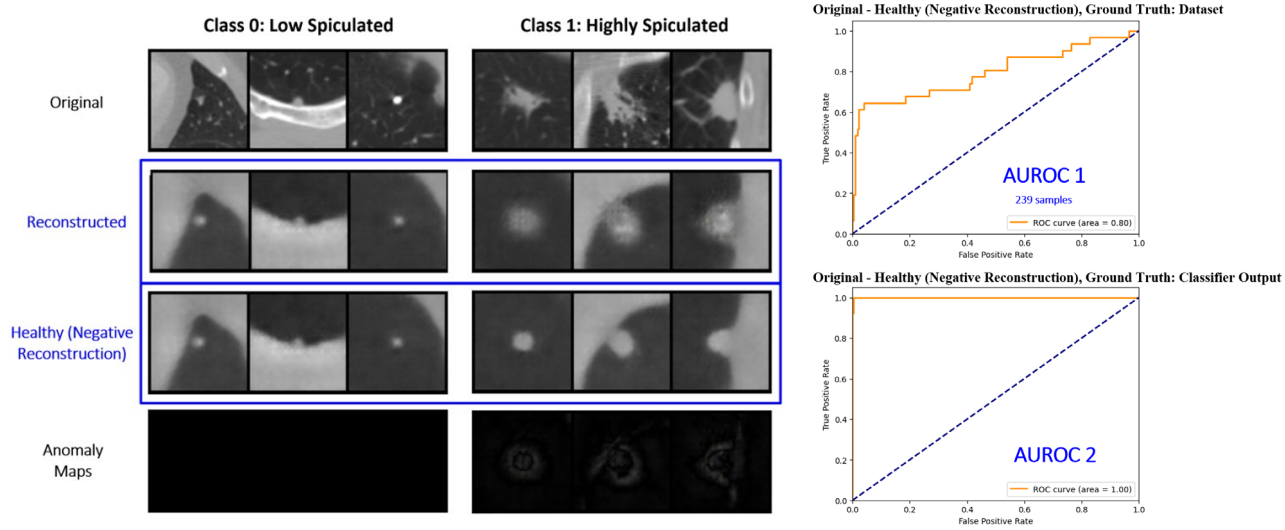

Figure 6: AUROC for VAE-AC-WGAN on the LIDC dataset

4. CONCLUSIONS AND FUTURE WORK

We introduced VAE-AC-WGAN as an improved contrastive explanation method for medical imaging, with promising results in generating meaningful visual contrasts between healthy and pathological cases. As part of future work, we plan to expand our validation to additional datasets such as CBIS-DDSM [4] for mammography and RadImageNet [5], and to apply the approach to a broader range of pathologies. These efforts aim to establish a benchmark for contrastive explanation quality in medical AI while also exploring enhanced disentanglement techniques for more robust and interpretable generative models.

BIBLIOGRAPHY

[1] Armato S. G. et al. (2011). The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans, Med Phys. 2011 Feb;38(2):915-31. doi: 10.1118/1.3528204.

[2] Bhuvaji S., Kadam A., Bhumkar P., Dedge S., and Swati Kanchan. Brain tumor classification (MRI), 2020. Kagle. DOI:10.34740/KAGGLE/DSV/1183165.

[3] Dravid A., Schiffers F., Gong B., and Katsaggelos A. K. (2022). medXGAN: Visual Explanations for Medical Classifiers through a Generative Latent Space. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 2935–2944, 2022.

[4] Lee R. et al. (2017), A Curated Mammography Data set for Use in Computer-aided Detection and Diagnosis Research. In Scientific Data (Nature's journal), volume 4, article number 170177, in 2017.

[5] Mei X. et al. (2022). RadImageNet: An Open Radiologic Deep Learning Research Dataset for Effective Transfer Learning. Radiology: Artificial Intelligence, 4, 2022.

[6] Selvaraju R. R. et al. (2019). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization (2019), International Journal of Computer Vision (IJCV), 2019.

[7] Wolleb et al. (2022). Diffusion Models for Medical Anomaly Detection. Medical Image Computing and Computer Assisted Intervention -MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII, Pages 35–45 https://doi.org/10.1007/978-3-031-16452-1 4