VPNet: Tractable and Explainable Classification with Probabilistic Circuits on ViT Features

Taishi Kusumoto

September 30, 2025

Abstract

In this paper, we propose VPNet, where we combine two parameter-frozen vision transformers (DINO and CLIP) with probabilistic circuits for a specific classification task. We demonstrated that this model is theoretically tractable for visualizing its decision processes and can achieve outstanding performance in classification tasks. In the future, it is expected that scientists will be able to discover new laws in the real world by asking an explainable AI to reveal the basis of its decision processes. This innovative model may serve as a next-generation framework to enable such achievements.

This manuscript is a draft; the analyses and results presented here are preliminary and may be revised in future versions.

1 Introduction

Medical or biological image classification with Artificial Intelligence (AI) has great potential to help innovative medical discoveries by uncovering hidden relationships between the target classes while extracting important feature representations from the images. However, most of such classification models lack interpretability due to their black-box characteristics. This complicates the introduction of such classification models for both practical and research purposes.

Explainable AI (XAI) models, which aim to clarify their decision processes, have been an active area of research, with various methods and algorithms proposed,[1] including Grad-CAM-based approaches[2], LRP-based approaches[3][4], and attention map–based approaches[5][6]. However, these methods merely visualize the complex feature representations of the model, without providing theoretical evidence that the visualization directly corresponds to the actual decision-making process of the model.

To address this limitation, we introduce an innovative classification model, VPNet, in which patch embedding vectors from parameter-frozen vision transformers (ViTs)[7] are directly connected to probabilistic circuits (PCs)[8][9][10][11]. The parameters of the PCs are then optimized for a specific classification task under supervised learning. This approach simultaneously maintains probabilistic interpretability and representation ability. In this paper, we demonstrate that VPNet achieves outstanding performance in image classification tasks and can visualize the probabilistic contribution of each patch in the image and/or each vision transformer connected to the PCs. We argue that this innovative approach represents a next-generation framework, enabling scientists to leverage artificial intelligence models to uncover new relationships in the real world.

2 Methods

2.1 Tractable Probabilistic Circuit for Classification

A probabilistic circuit is a neural network whose outputs represent the joint probability distribution over all input leaf nodes, provided that the circuit satisfies the following conditions: (a) each scope corresponds to a single (univariate) random variable, and (b) smoothness and decomposability are preserved.[8] We connected each patch embedding vector from a parameter-frozen vision transformer to leaf nodes as a Gaussian mixture, i.e., a convex combination of univariate Gaussian models over the scalars of the vector:

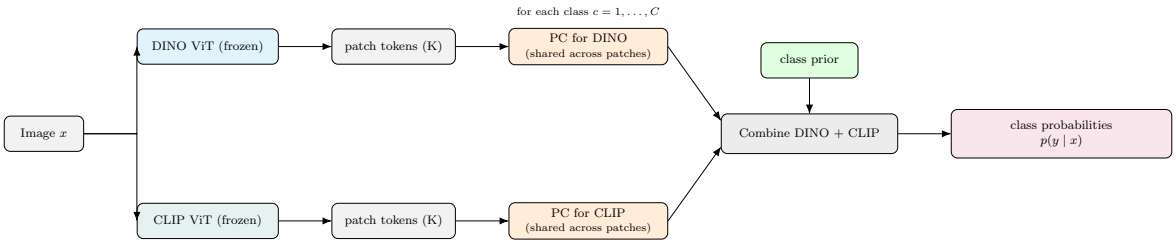


Figure 1: Simplified overview of the class-conditional PC classifier.

$$\mathbf{z}_i = g_{\theta^\star}(x)_i \in R^d, \quad i = 1, \ldots, N.$$

For each coordinate $j \in \{1, \ldots, d\}$, the mixture realized by the sum node s_j is

$$p_{s_j}(z_{ij}) \;=\; \sum_{u \in \mathrm{ch}(s_j)} w_{j,u}\, \mathcal{N}\big(z_{ij};\, \mu_{j,u},\, \sigma^2_{j,u}\big), \qquad w_{j,u} \geq 0, \; \sum_{u \in \mathrm{ch}(s_j)} w_{j,u} = 1.$$

where z_{ij} denotes the j-th scalar of the i-th patch embedding $\mathbf{z}_i$; s_j is the sum node associated with coordinate j; $\mathrm{ch}(s_j)$ is the set of Gaussian leaf children of s_j; $w_{j,u}$ are the convex edge weights of s_j; and $\mu_{j,u} \in R$ and $\sigma^2_{j,u} > 0$ are the mean and variance parameters of the univariate Gaussian leaf u. We assume that each scalar derived from a patch embedding vector defines a univariate scope, and the Gaussian leaves under its associated sum node share that scope. Because the Vision Transformer (ViT) parameters are frozen, the encoder does not learn relationships across patches during optimization for the classification task, so no additional dependencies are introduced outside the circuit, and the independence assumption is maintained.

While strictly observing the rules of smoothness and decomposability, we increase the complexity of the probabilistic circuits to reinforce their representational ability. Probabilistic densities from a leaf node are allowed for the following two mathematical operations:

$$\textbf{Sum node:} \quad p_v(\mathbf{z}) \;=\; \sum_{u \in \mathrm{ch}(v)} w_{v,u}\, p_u(\mathbf{z}), \qquad w_{v,u} \geq 0, \; \sum_{u \in \mathrm{ch}(v)} w_{v,u} = 1,$$

$$\textbf{Product node:} \quad p_v(\mathbf{z}) \;=\; \prod_{u \in \mathrm{ch}(v)} p_u(\mathbf{z}),$$

where $\mathbf{z}$ denotes the input variables, v is an internal node of the circuit, $\mathrm{ch}(v)$ is the set of its children, $w_{v,u}$ are the convex weights associated with edges from a sum node to its children, and $p_u(\mathbf{z})$ denotes the probability distribution computed at child node u.

As long as sum nodes are applied between nodes with the same scope, and product nodes are applied between nodes with independent scopes, the smoothness and decomposability are preserved, so the probabilistic interpretability of the model can be maintained.The output from the probabilistic circuit is the joint probability of all input random variables or patch embedding vectors:

$$p(\mathbf{z}_1, \ldots, \mathbf{z}_N) \;=\; p_{\mathrm{root}}(\mathbf{z}_1, \ldots, \mathbf{z}_N),$$

where $\mathbf{z}_i \in R^d$ denotes the embedding vector of the i-th patch, and p_{root} is the probability distribution computed at the root node of the circuit.

We prepared two parameter-frozen Vision Transformers, DINO[12] and CLIP[13]. Each ViT is connected separately to a probabilistic circuit, and the patch embedding vectors produced by the two models are treated as distinct independent random variables, even if they are derived from the same patch region. The final joint probability is obtained by a product node that combines the DINO–PC and the CLIP–PC:

$$p_{\mathrm{root}}\big(\{\mathbf{z}_i^{\mathrm{D}}\}_{i=1}^N, \{\mathbf{z}_i^{\mathrm{C}}\}_{i=1}^N\big) \;=\; \prod_{i=1}^{N} \Big[\, p_{\mathrm{root}}^{\mathrm{DINO}}\big(\mathbf{z}_i^{\mathrm{D}}\big)\; p_{\mathrm{root}}^{\mathrm{CLIP}}\big(\mathbf{z}_i^{\mathrm{C}}\big) \Big],$$

where $\mathbf{z}_i^{\mathrm{D}} \in R^{d_{\mathrm{D}}}$ and $\mathbf{z}_i^{\mathrm{C}} \in R^{d_{\mathrm{C}}}$ denote the patch embedding vectors from DINO and CLIP for patch i, respectively, and $p_{\mathrm{root}}^{\mathrm{DINO}}$ and $p_{\mathrm{root}}^{\mathrm{CLIP}}$ are the distributions computed at the roots of the DINO–PC and CLIP–PC.

The above concept is represented in Figure 1.

2.2 Optimization for classification

The above circuit is prepared for the number of defined classes, and one circuit is assigned to each class. The joint probability distribution from the root node of the circuit assigned to class c is regarded as an approximation of the class-conditional distribution:

$$p\big(\mathbf{z}_1, \ldots, \mathbf{z}_N \mid y = c\big) \;\approx\; p_{\text{root}}^{(c)}\big(\mathbf{z}_1, \ldots, \mathbf{z}_N\big),$$

where y denotes the class label, $\{\mathbf{z}_i\}_{i=1}^{N}$ are the patch embedding vectors, and $p_{\text{root}}^{(c)}$ is the distribution computed at the root node of the probabilistic circuit assigned to class c.In addition, we set a prior probability distribution as a categorical distribution,

$$p(y = c;\, \pi) \;=\; \pi_c, \qquad \pi_c \geq 0,\;\; \sum_{c=1}^{C} \pi_c = 1,$$

where $\pi = (\pi_1, \ldots, \pi_C)$ denotes the learnable class prior parameters.The joint probability distribution from a class and all of the input random variables is calculated from the joint rule,

$$p(z, y = c;\, \theta, \pi) \;=\; p(y = c;\, \pi)\, p(z \mid y = c;\, \theta_c),$$

where $\pi = (\pi_1, \ldots, \pi_C)$ denotes the categorical prior parameters over classes, $\theta = \{\theta_1, \ldots, \theta_C\}$ represents the collection of all class-conditional parameters, and $\theta_c \in \theta$ denotes the parameters of the probabilistic circuit specific to class c. The conditional probabilistic distribution of a class given input random variables is calculated by the following Bayes rule,

$$p(y = c \mid z;\, \theta, \pi) \;=\; \frac{p(y = c;\, \pi)\, p(z \mid y = c;\, \theta_c)}{\sum_{c'=1}^{C} p(y = c';\, \pi)\, p(z \mid y = c';\, \theta_{c'})}.$$

We defined a cross entropy loss function[14] to optimize the parameters,

$$\mathcal{L}_{\text{CE}}(X, Y; \theta, \pi) = -\frac{1}{B} \sum_{i=1}^{B} \log p(y_i \,|\, X_i;\, \theta, \pi)\,,$$

and a Shannon entropy loss function to avoid overfitting [15],

$$\mathcal{L}_{\text{ENT}}(X; \theta, \pi) = \frac{1}{B} \sum_{i=1}^{B} \Big(-\sum_{c=1}^{C} p(y = c \mid X_i; \theta, \pi)\, \log p(y = c \mid X_i; \theta, \pi) \Big),$$

where $X = \{X_1, \ldots, X_B\}$ denotes a mini-batch of B input images and $Y = \{y_1, \ldots, y_B\}$ are the corresponding class labels.

2.3 Interpretability via Patch and ViT Marginalization

In our implementation, instead of preparing a single large PC that jointly models all patch embedding vectors, we adopt a class-conditional design in which all patch embedding vectors from the same class and the same backbone ViT share one smaller PC. Each of these PCs accepts a single patch embedding vector as input. This follows the assumption of conditional independence between patch embedding vectors, which is consistent with the smoothness and decomposability properties of probabilistic circuits. Consequently, the joint probability over all patch embedding vectors reduces to the product of the per-patch likelihoods, i.e.,

$$p(z_1, \ldots, z_K \mid y = c;\, \theta_c) = \prod_{k=1}^{K} f_c(z_k; \theta_c),$$

where $z_k \in R^{qd}$ denotes the k-th patch embedding vector, $f_c(\cdot; \theta_c)$ is the class-c PC over a single patch embedding vector with parameters θ_c, and K is the number of patches extracted by the ViT. In practice, for numerical stability, we compute the average log-likelihood,

$$\log p(z_1, \ldots, z_K \mid y = c; \theta_c) \;\approx\; \frac{1}{K} \sum_{k=1}^{K} \log f_c(z_k; \theta_c).$$

The most important concept in this paper is that we can change patches or a ViT (or both) into the unobserved state and then calculate the marginal probability distribution that excludes such patches or the ViT. Formally, let $g \in \mathcal{G}$ index backbone groups (e.g., CLIP, DINO), and let $z_{g,k} \in R^{qd}$ denote the k-th patch embedding from group g ($k = 1, \ldots, K_g$). For each class $c \in \{1, \ldots, C\}$ and group g, let $f_{g,c}(\cdot; \theta_{g,c})$ be the class-conditional PC over a single patch embedding. Under the conditional independence assumption across patches within a group, the class-conditional likelihood for an observed patch index set $O_g \subseteq \{1, \ldots, K_g\}$ is

$$p\big(\{z_{g,k}\}_{k \in O_g} \mid y = c; \theta_{g,c}\big) = \prod_{k \in O_g} f_{g,c}(z_{g,k}; \theta_{g,c}),$$

where the variables in $U_g = \{1, \ldots, K_g\} \setminus O_g$ are marginalized out.

Let $\pi_c = p(y = c)$ be the categorical class prior. Then the marginal posterior that excludes all unobserved patches is

$$p\big(y = c \mid \{z_{g,k}\}_{g,\, k \in O_g};\, \theta, \pi\big) = \frac{\pi_c \prod_{g \in \mathcal{G}} \prod_{k \in O_g} f_{g,c}(z_{g,k}; \theta_{g,c})}{\sum_{c'=1}^{C} \pi_{c'} \prod_{g \in \mathcal{G}} \prod_{k \in O_g} f_{g,c'}(z_{g,k}; \theta_{g,c'})}.$$

Special cases include:

$$\begin{aligned}
\text{(drop-one patch } k^\star \text{ in group } g^\star) \quad & O_{g^\star} = \{1, \ldots, K_{g^\star}\} \setminus \{k^\star\}, \quad O_g = \{1, \ldots, K_g\} \ (g \neq g^\star), \\
\text{(keep-only patch } k^\star \text{ in group } g^\star) \quad & O_{g^\star} = \{k^\star\}, \quad O_g = \ (g \neq g^\star), \\
\text{(drop-one ViT group } g^\star) \quad & O_{g^\star} =, \quad O_g = \{1, \ldots, K_g\} \ (g \neq g^\star).
\end{aligned}$$

After that, we can understand which patches or a ViT are more valuable for the classification task by comparing the resulting posterior distributions. For example, for a target class t, we can use the change in log-posterior:

$$\Delta_{\text{patch}}(g^\star, k^\star) = \log p(y = t \mid \{z_{g,k}\}_{\text{full}}) - \log p\big(y = t \mid \{z_{g,k}\}_{\text{drop } (g^\star, k^\star)}\big).$$

2.4 Datasets

We prepared a brain tumor classification dataset from MRI images, BRISC2025[16], and a pneumonia classification dataset from chest X-ray images, ChestX-ray8[17], to demonstrate the tractability and classification ability of our model.

BRISC2025 has 4 specific classes, including no tumor, glioma, meningioma, and pituitary, with 1067, 1147, 1329, and 1457 training images, and 140, 254, 306, and 300 validation images, respectively. We integrated all the training and validation images except the no-tumor class into a single tumor class. Then, the training samples in the no-tumor class were oversampled so that the ratio of the two classes is nearly equal. This dataset has ground truth segmentation masks corresponding to tumor regions for every image in the tumor class.

ChestX-ray8 comprises 2 specific classes, normal and pneumonia, with 1341 and 3875 training images, 8 and 8 validation images, and 234 and 350 test images, respectively. For this dataset, we did not make any modifications and use it directly.

In addition, we prepared a simpler dataset for anomaly detection, the MVTec dataset.[18] We used two of its categories, zipper and tile, comprising two specific classes, good and NG.

Mvtec-zipper includes 240 training images for the good class, and 32 and 119 test images for the good and NG classes, respectively. We selected the first 21 images from the NG test set and used them as training images for the NG class. The training NG images were oversampled so that the ratio of the two classes was nearly equal. All the remaining test images were used as validation images.

MVTec Tile contains 230 training images for the good class, and 33 and 84 test images for the good and NG classes, respectively. We selected 26 images and 5 images from the NG test dataset and 5 images from the good training dataset, and used them as training images and validation images, respectively. The training NG images were oversampled so that the ratio of the two classes was nearly equal.

Table 1: Validation/Test metrics for each dataset. Values are AUROC, Average Precision (AP), and Max Accuracy with the threshold used.

2*Dataset	Validation (Val)			Test		
	AUROC	AP	Max Acc. @ thr	AUROC	AP	Max Acc. @ thr
BRISC2025 (2-class)	0.997251	0.997393	0.978235 @ 0.5370	—	—	—
MVTec (zipper)	0.991709	0.997337	0.953846 @ 0.2690	—	—	—
MVTec (tile)	—	—	—	**1.000000**	**1.000000**	**1.000000 @ 0.5103**
ChestX-ray8	—	—	—	0.971828	0.980382	0.923077 @ 0.7880

2.5 Experimental setting

In all experiments, the probabilistic circuits (PCs) were initialized with the following default configuration: a depth of 5, 64 latent variables, 3 repetitions, 2 mixture pieces per leaf, unless explicitly stated otherwise. Two vision transformers, DINO ViT-B/16 and CLIP ViT-B/16, were used as the backbones of the circuits. Since these ViTs are optimized for 224×224 images, all input images were resized to this resolution. As a result, each input image was divided into 196 patches. The Adam optimizer [19] was adopted for training.

During training, we set a threshold of 0.5 for the anomaly class, then calculated the accuracy on the validation dataset, and saved the parameters that achieved the maximum accuracy.

3 Results

3.1 Overall Classification Performance

The joint probability without any unobserved state of patch or ViT is used for the classification task. Table 1 shows the AUROC, Average Precision (AP), and maximum accuracy from the experimental datasets. We calculated the evaluation metrics for MVTec-zipper and BRISC2025 using the validation dataset, and the test dataset for ChestX-ray8 and MVTec-tile. AUROC, Average Precision, and Accuracy are defined as follows:

$$\text{AUROC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR}), \tag{1}$$

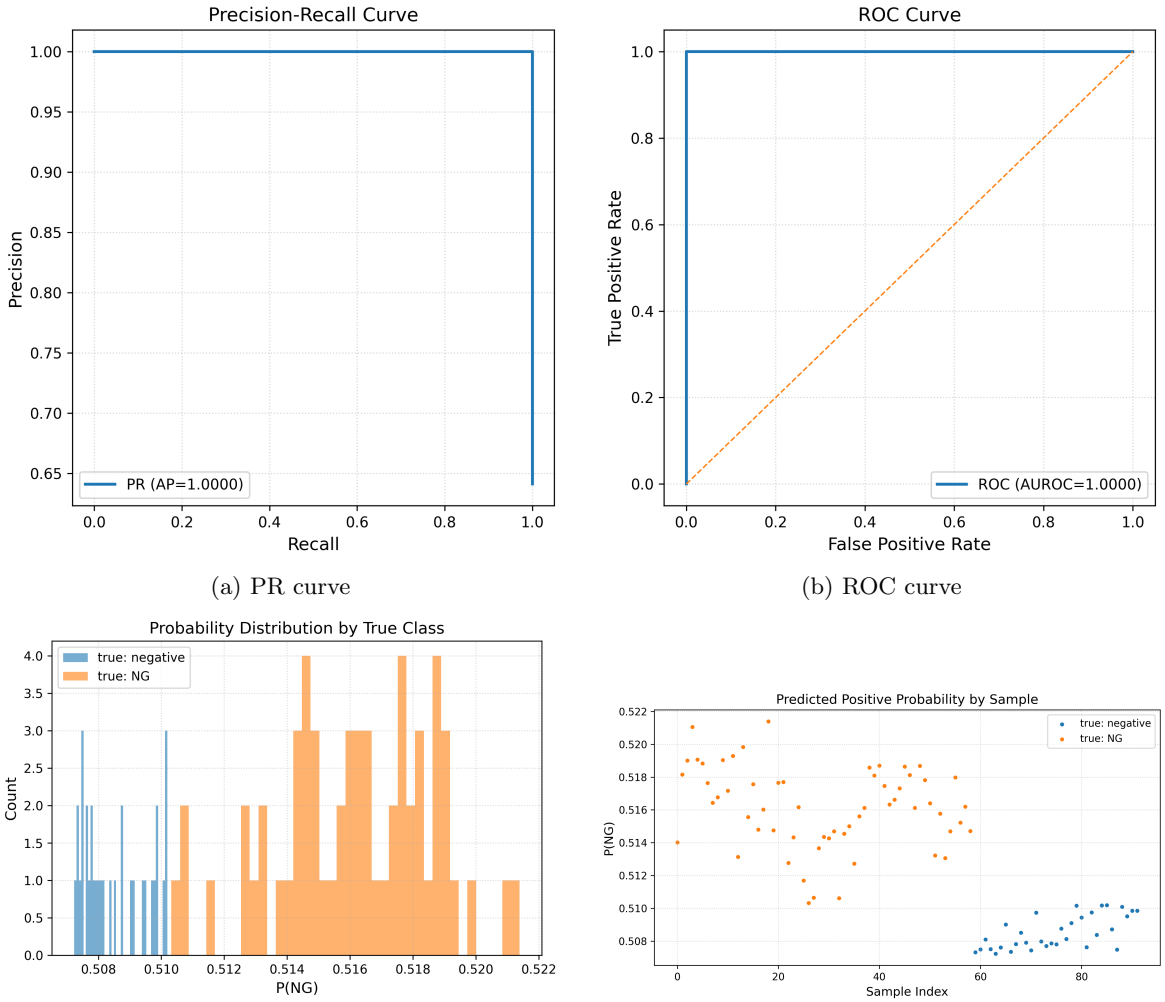
$$\text{AP} = \sum_{n=1}^{N} \big(R_n - R_{n-1}\big) P_n, \tag{2}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{3}$$

Here,

- TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively.
- $\text{TPR} = \dfrac{TP}{TP + FN}$ is the true positive rate (recall).
- $\text{FPR} = \dfrac{FP}{FP + TN}$ is the false positive rate.
- P_n and R_n denote the precision and recall at the n-th threshold, respectively.

Figures 2–5 present the ROC curve, PR curve, scatter plot, and histogram of the conditional probability of the positive class for each dataset: MVTec Tile (Figure 2), MVTec Zipper (Figure 3), BRISC2025 (Figure 4), and ChestX-ray8 (Figure 5).



(a) PR curve

(b) ROC curve

(c) $P(\mathrm{NG})$ histogram

(d) Per-sample $P(\mathrm{NG})$

Figure 2: Performance summary of Mvtec-tile.

3.2 Patch-level Probabilistic Attribution

We compared the conditional probabilistic distribution from each patch in images labeled as the positive class with respect to the negative and positive classes. We then visualized the difference using the inferno colormap.[20] Let $x = \{x_1, x_2, \ldots, x_K\}$ denote the set of patch features. For each class $c \in \{\mathrm{pos}, \mathrm{neg}\}$, the conditional probabilistic distribution of patch k is

$$p_c(x_k) = f_c(x_k), \tag{4}$$

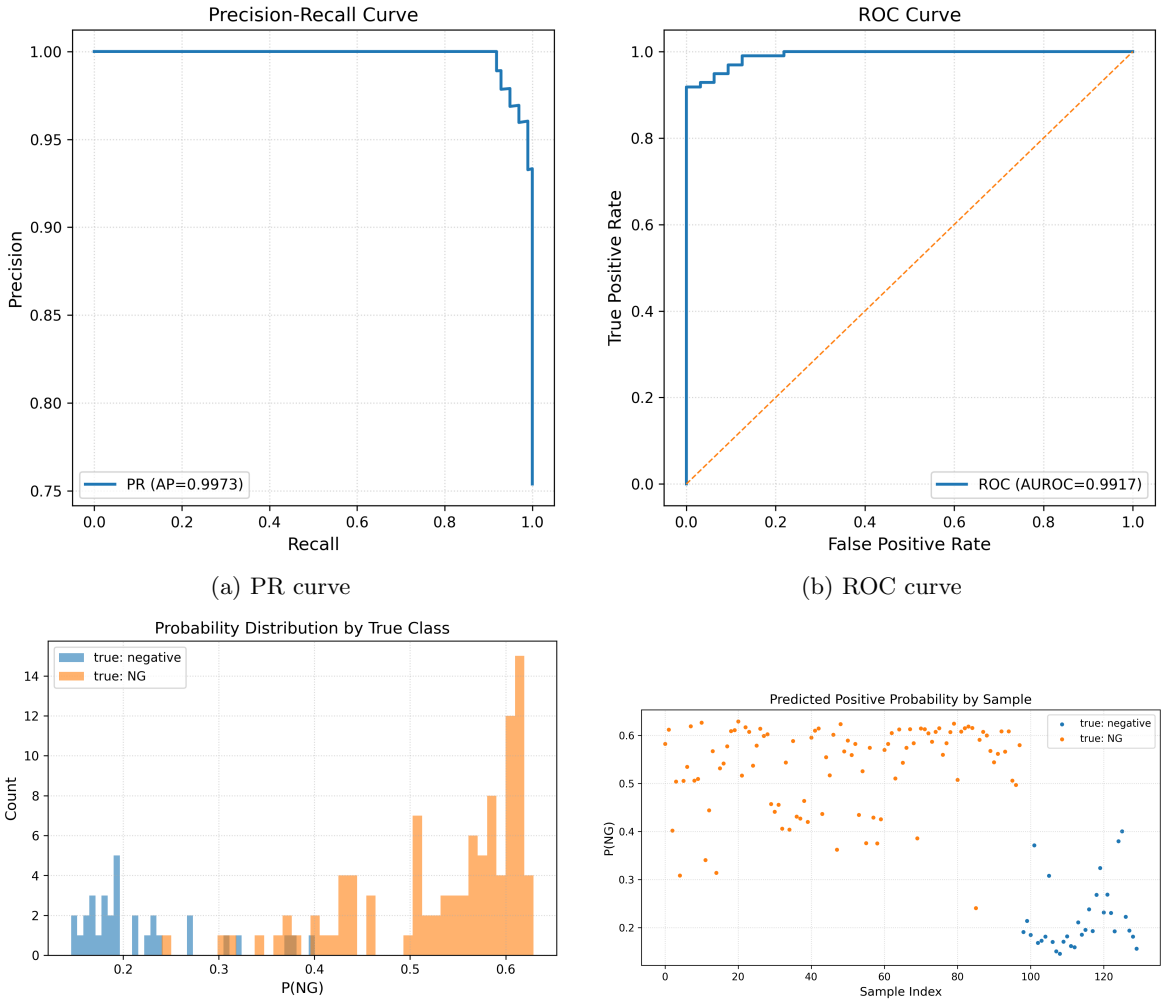
where $f_c(\cdot)$ denotes the class-conditional density modeled by the probabilistic circuit. The per-class joint log-likelihood is given by

$$\log p_c(x) = \frac{1}{K} \sum_{k=1}^{K} \log p_c(x_k). \tag{5}$$

The patch-level difference is defined as

$$\Delta_k = \log p_{\mathrm{pos}}(x_k) - \log p_{\mathrm{neg}}(x_k). \tag{6}$$

In the inferno heatmap visualization, dark red indicates patches with high Δ_k (i.e., a strong contribution to increasing the class-conditional joint probability for the class), yellow indicates medium Δ_k, and dark purple indicates low Δ_k. This means that in the heatmap for the positive class, dark red regions correspond to abnormal areas, whereas dark purple regions indicate normal areas. In the heatmap for the negative class, the interpretation is reversed.



(a) PR curve

(b) ROC curve

(c) $P(\mathrm{NG})$ histogram

(d) Per-sample $P(\mathrm{NG})$

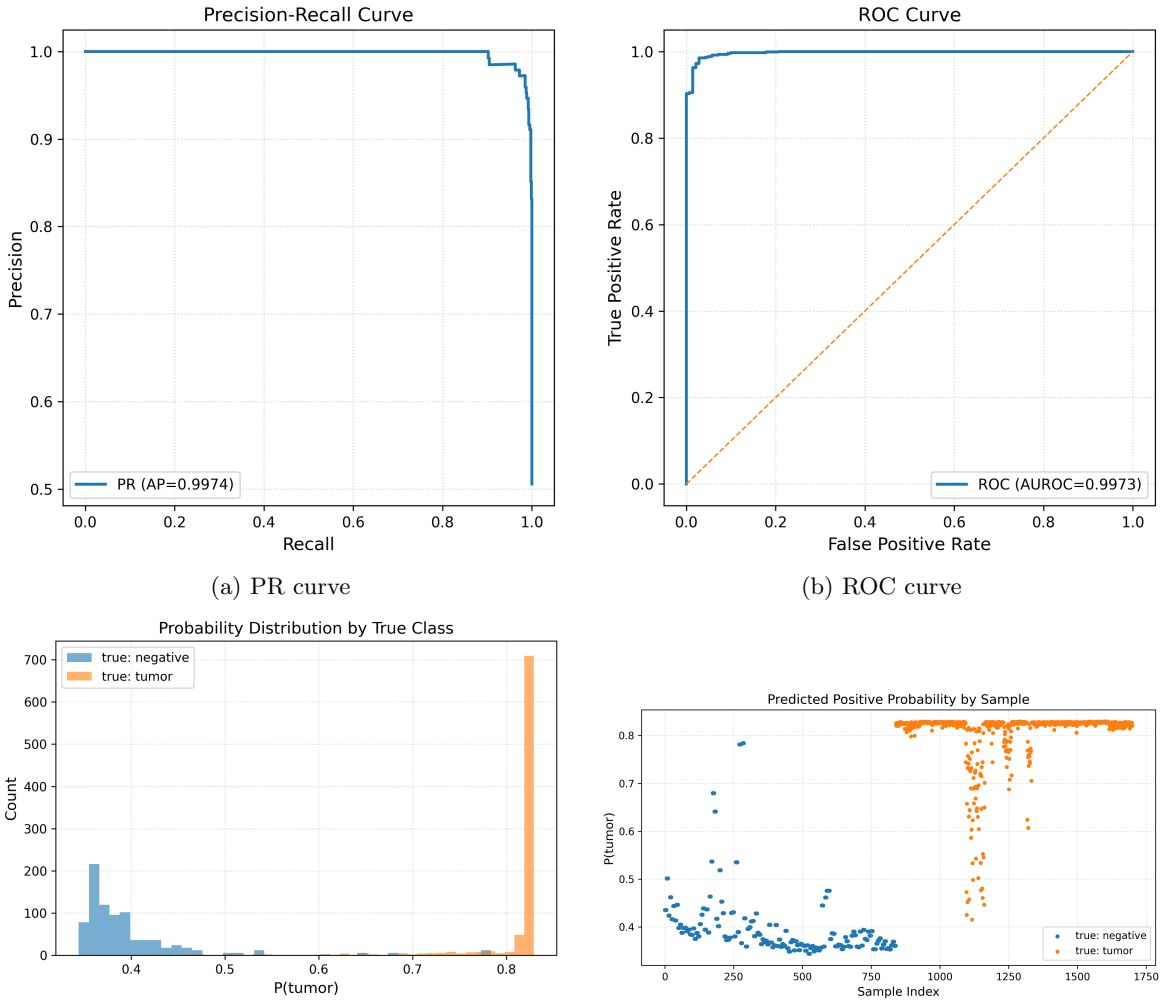
Figure 3: Performance summary of Mvtec-zipper.

Figure 6 illustrates two heatmaps for several anomaly images in the test dataset of MVTec Tile, together with the corresponding original images. It can be observed that both the dark purple regions in the heatmap from the negative class and the dark red regions in the heatmap from the positive class cover the defective parts in the originals.

Figure 7 depicts heatmaps for anomaly samples in the validation dataset of MVTec Zipper, along with the corresponding original images. Red squares are superimposed on the originals to mark defective regions. In this case, the dark purple areas in the negative-class heatmap successfully highlight the defects, and furthermore spread to adjacent defective and deformed areas. In contrast, although the dark red areas in the positive-class heatmap are theoretically expected to capture the defects, they fail to delineate them accurately.

In Figure 8, heatmaps for glioma images from the BRISC2025 validation dataset are presented for both the positive (tumor) and negative (no tumor) classes, together with the original images and the ground-truth masks. Unlike the industrial cases, the tumor regions are not clearly emphasized by the dark purple areas in the negative-class heatmap; instead, purple regions are more diffusely distributed throughout the brain.

Figure 9 shows examples of heatmaps for pneumonia cases from the ChestX-ray8 test dataset, corresponding to the positive (pneumonia) and negative (normal) classes, displayed alongside the original images. In this case, the distributions of the two heatmaps are notably different. The dark purple regions in the negative-class heatmap clearly concentrate over the entire chest, whereas the dark red regions in the positive-class heatmap accumulate in smaller areas around the neck, lower abdomen, and armpits.



(a) PR curve

(b) ROC curve

(c) P(Tumor) histogram

(d) Per-sample P(Tumor)

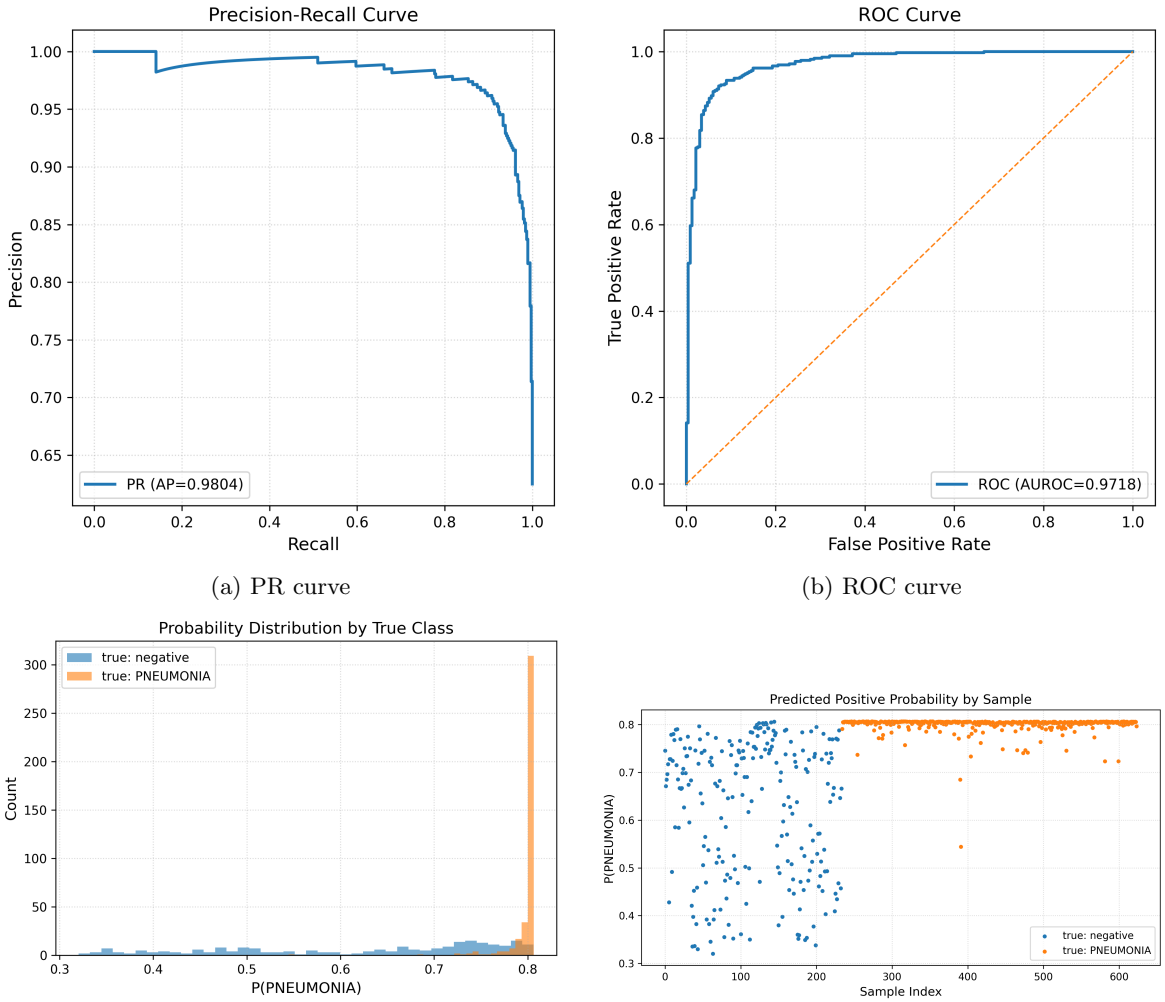
Figure 4: Performance summary of BRISC2025.

4 Discussion

One of the limitations of probabilistic circuits is their weak representational capacity for solving complex tasks.[21][22] This paper demonstrates that this limitation can be compensated by integrating high-performing ViT-based encoders such as CLIP or DINO. The evaluation results, including very high AUROC values, clearly indicate that our classification models achieve outstanding performance even on data domains outside the training distribution. In our experiments, we used ViT-B/16 models for both CLIP and DINO; however, ViTs with finer patch embeddings could also be employed as encoders. In this case, a more fine-grained visualization of the conditional probabilistic distributions would become possible.

Moreover, the input to the probabilistic circuit of VPNet is an encoded vector. This implies that other types of information (e.g., text vectors, audio vectors, or etc.) could be incorporated simultaneously with image-derived information for a specific classification task. Furthermore, joint probabilistic distributions as well as marginal distributions for each modality can still be extracted regardless of the input type. This avenue should be explored in future research.

The results show that at least one of the visualizations of the conditional probabilities of the negative and positive classes from MVTec Tile and MVTec Zipper clearly captures the defective parts of the images. This finding suggests that our probabilistic circuit–based approach is valid for visualizing the decision process of the classification model, i.e., the probabilistic contribution of each region in the image. However, although the AUROC for the validation dataset is quite high (0.9917), the dark red regions in the heatmap from the positive class of MVTec Zipper highlight areas irrelevant to the defective regions. This indicates that even if a model achieves outstanding performance on a



(a) PR curve

(b) ROC curve

(c) P(Pneumonia) histogram

(d) Per-sample P(Pneumonia)

Figure 5: Performance summary of ChestX-ray8.

specific classification task, this does not guarantee that all corresponding heatmaps clearly represent the information the model actually uses to solve the task. Nevertheless, it can be assumed that at least one of the heatmaps provides such information.

Based on this assumption, for the brain tumor classification task using the BRISC2025 dataset, the model may rely on information distributed across the brain rather than exclusively on the tumor regions. For the pneumonia classification task using the ChestX-ray8 dataset, it can be concluded that information outside the chest area, including the neck, lower abdomen, and armpits, may contribute to the classification, and/or that the entire chest region provides more useful information for the task.

One of the most important future objectives of this research is to achieve new scientific discoveries by extracting the model's decision processes for specific tasks. In the future, it may become possible to identify the causes of unexpected diseases or even to uncover new laws of nature by visualizing the complex rationales indicated by AI and analyzing them.

5 Conclusion

We implemented an explainable AI model for a classification task and demonstrated that this model can visualize the decision process by extracting the marginal probabilities from input patch embedding vectors. This model has the great potential to uncover new relationships in the real world. In future research, we plan to expand its functionalities to include self-supervised anomaly detection, zero-shot image segmentation, and knowledge distillation, as well as to incorporate multimodal inputs using different types of information such as text, signals, and audio.

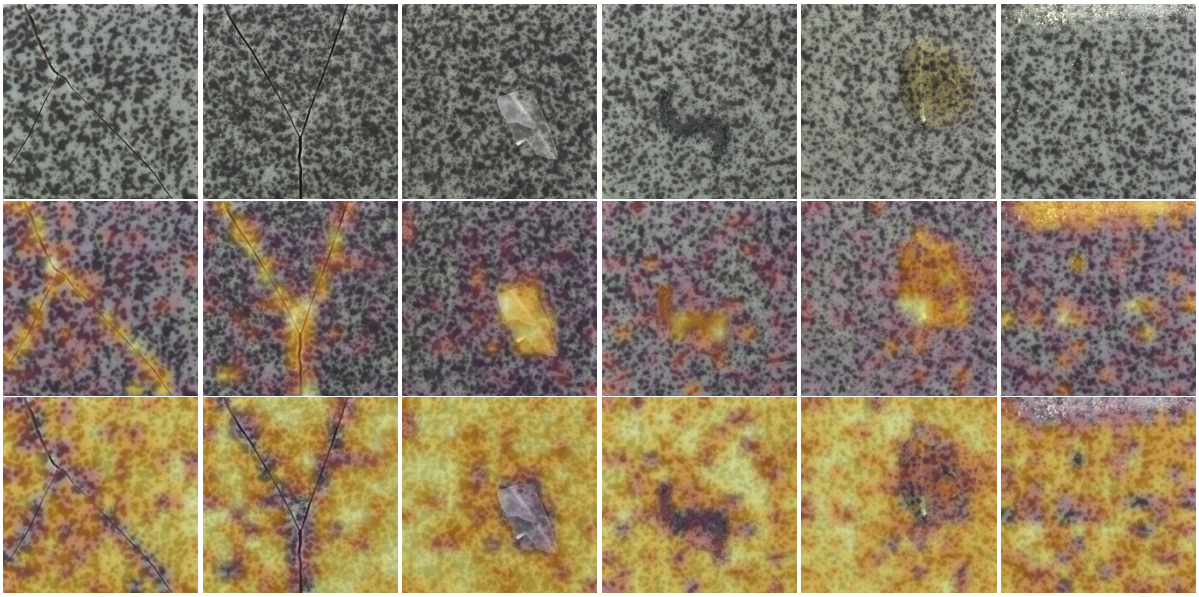


Figure 6: Three rows (top to bottom): original, heatmap (NG), heatmap (good). Six columns show different samples in MVTec-tile.

References

[1] C. Patrício, J. C. Neves, and L. F. Teixeira, "Explainable deep learning methods in medical image classification: A survey," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–41, 2023.

[2] P. Sharma, A. Singh, and S. Kumar, "Enhancing brain tumor detection in mri images through explainable ai using grad-cam with resnet50," *BMC Medical Imaging*, vol. 24, no. 1, p. 92, 2024.

[3] L. Arras, G. Montavon, K.-R. Müller, and W. Samek, "Explaining recurrent neural network predictions in sentiment analysis," in *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 159–168, 2017.

[4] M. Bohle, F. Eitel, M. Weygandt, and K. Ritter, "Layer-wise relevance propagation for explaining deep neural network decisions in mri-based alzheimer's disease classification," *Frontiers in Aging Neuroscience*, vol. 11, p. 194, 2019.

[5] M. Komorowski *et al.*, "Towards evaluating explanations of vision transformers for medical imaging," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 6133–6142, 2023.

[6] A. Rao, J. Park, S. Woo, J.-Y. Lee, and O. Aalami, "Studying the effects of self-attention for medical image analysis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3416–3425, 2021.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020. arXiv preprint arXiv:2010.11929.

[8] Y. Choi, A. Vergari, and G. Van den Broeck, "Probabilistic circuits: A unifying framework for tractable probabilistic models." http://starai.cs.ucla.edu/papers/ProbCirc20.pdf, 2020. UCLA Technical Report.

[9] S. Braun, S. Sidheekh, S. Natarajan, A. Vergari, M. Mundt, and K. Kersting, "Autoencoding probabilistic circuits," in *Eighth Workshop on Tractable Probabilistic Modeling*, 2025.

[10] W. Chen, S. Yu, H. Shao, L. Sha, and H. Zhao, "Neural probabilistic circuits: An overview," in *Eighth Workshop on Tractable Probabilistic Modeling*, 2024.

[11] A. Ciotinga and Y. Choi, "Optimal transport for probabilistic circuits," *arXiv preprint arXiv:2410.13061*, 2024.

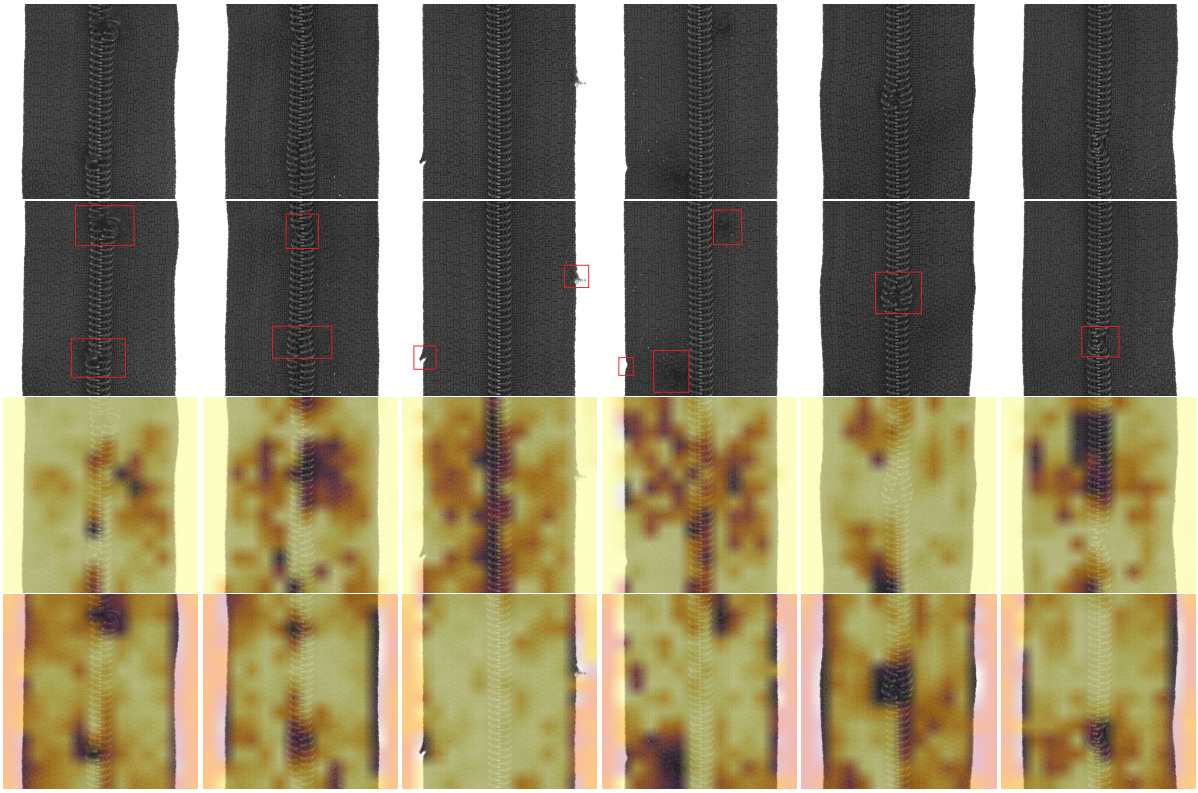


Figure 7: Four rows (top to bottom): original, original with defect boxes, heatmap (NG), heatmap (good). Six columns show different samples in MVTec-zipper.

[12] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. arXiv preprint arXiv:2104.14294.

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. arXiv preprint arXiv:2103.00020.

[14] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[15] A. Liu and G. Van den Broeck, "Tractable regularization of probabilistic circuits," *Advances in Neural Information Processing Systems*, vol. 34, pp. 3558–3570, 2021.

[16] A. Fateh, Y. Rezvani, S. Moayedi, S. Rezvani, F. Fateh, M. Fateh, and V. Abolghasemi, "Brisc: Annotated dataset for brain tumor segmentation and classification with swin-hafnet," *arXiv preprint arXiv:2506.14318*, 2025.

[17] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097–2106, 2017.

[18] P. Bergmann, K. Batzner, M. Fauser, D. Sattlegger, and C. Steger, "The mvtec anomaly detection dataset: A comprehensive real-world dataset for unsupervised anomaly detection," *International Journal of Computer Vision*, vol. 129, no. 4, pp. 1038–1059, 2021.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations (ICLR)*, 2015.

[20] K. Moreland, "Diverging color maps for scientific visualization," in *Advances in Visual Computing (ISVC)*, vol. 5876 of *Lecture Notes in Computer Science*, pp. 92–103, Springer, 2009.

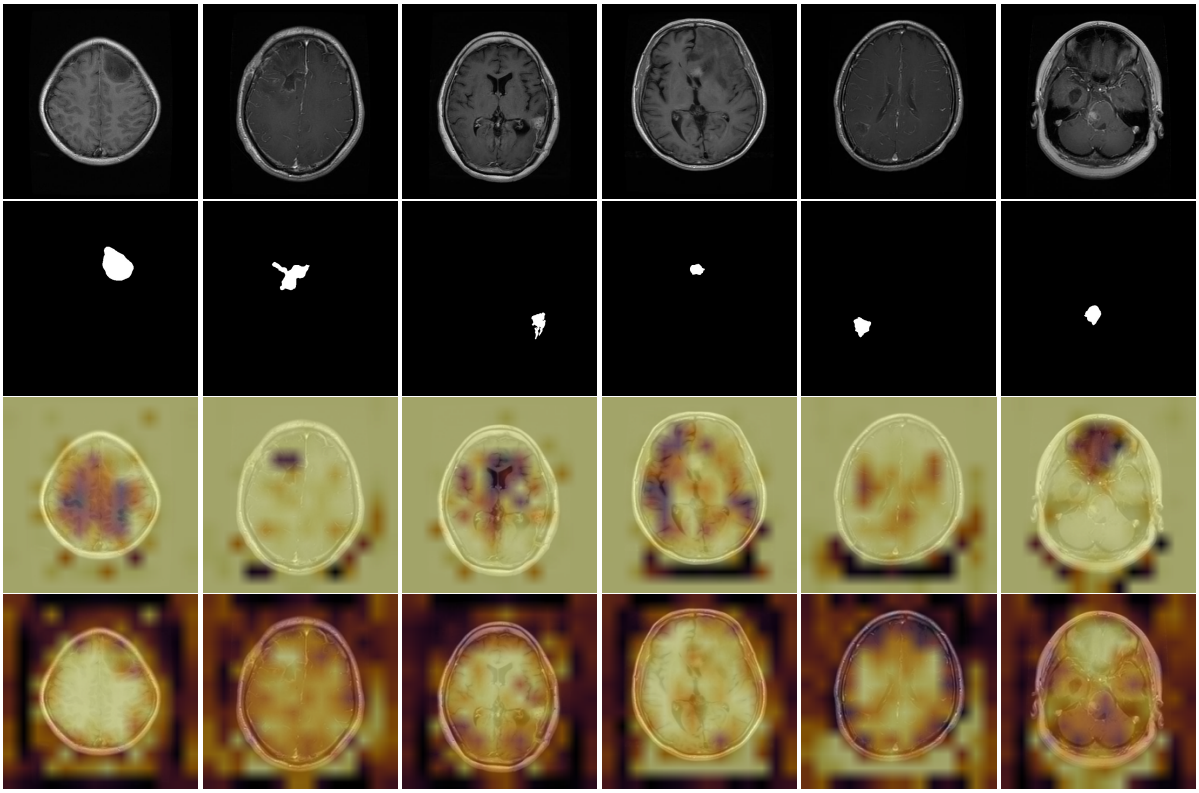


Figure 8: Four rows (top to bottom): original, original with defect boxes, heatmap (tumor), heatmap (no-tumor). Six columns show different samples in BRISC2025.

[21] H. Zhang, S. Holtzen, and G. Broeck, "On the relationship between probabilistic circuits and determinantal point processes," in *Conference on uncertainty in artificial intelligence*, pp. 1188–1197, PMLR, 2020.

[22] J. Leland and Y. Choi, "On the hardness of approximating distributions with probabilistic circuits," *arXiv preprint arXiv:2506.01281*, 2025.

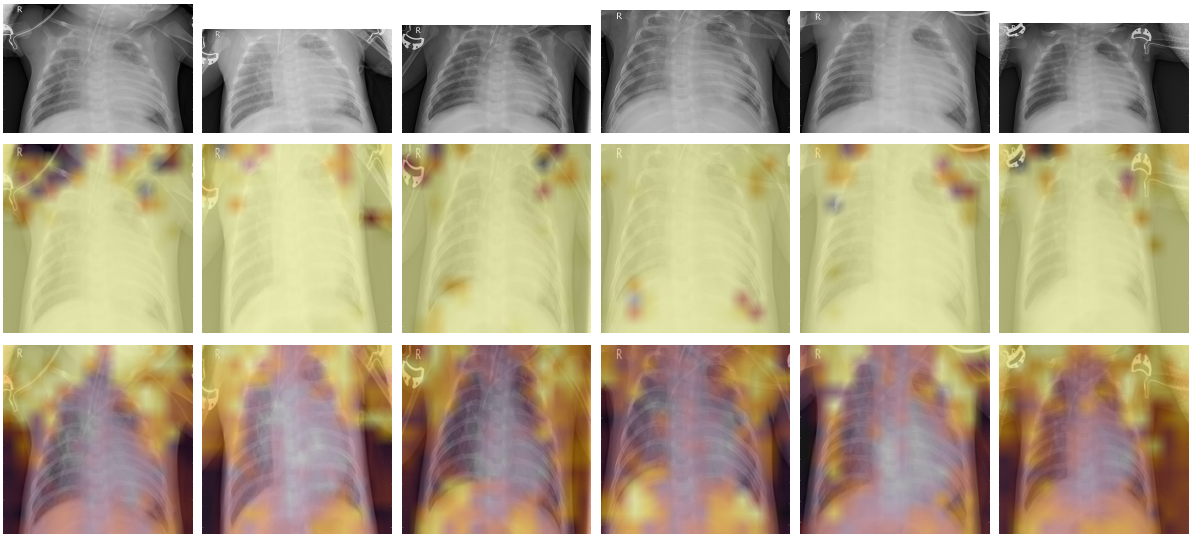


Figure 9: Three rows (top to bottom): original, heatmap (pneumoni), heatmap (normal). Six columns show different samples in ChestX-ray8.