Reliable and Trustworty Learning Prototype: Insight from POCUS

Giacomo Ignesti
Institute of Science and Information Technologies
National Research Council
Pisa, Italy

Gennaro D'Angelo
Institute of Clinical Physiology
National Research Council
Pisa, Italy

Lorenza Pratali
Institute of Clinical Physiology
National Research Council
Pisa, Italy

Davide Moroni
Institute of Science and Information Technologies
National Research Council
Pisa, Italy

Massimo Martinelli
Institute of Science and Information Technologies
National Research Council
Pisa, Italy

***Abstract*—Deep learning models often lack the interpretability and trustworthiness required for clinical use. This paper proposes a prototype-regularised training method to analyse 1,208 lung ultrasound images, focusing on B-line artefacts. A ConvNeXt-Tiny architecture is used, adding a novel reconstruction loss to the standard classification loss. The model is guided to extract meaningful prototypes and uses them to classify the ultrasound images. To prevent these constraints from hindering generalisation, it is used in pairs with the proposed reconstruction loss, a set of plausible data augmentation of the ideal researched prototypes, and a geometry-aware network, a spatial transformer network, to measure which solutions help the network towards outputting the most reliable outcomes. The resulting models are precise, lightweight and interpretable, indicating that the proposed solution can be embedded in an ultrasound device to assist healthcare specialists in point-of-care applications.**

***Index Terms*—Ultrasounds, XAI, Pattern Recognition, Training Function, POCUS**

I. Introduction

Foundational deep learning (DL) models are spreading in the medical analysis domain, creating new types of medical applications. Starting from Resnet [1] and U-Net [2] architectures, artificial intelligence (A.I.) unlocked the potential to extract patterns and relations in medical imaging [3]. A substantial amount of annotated data often constrains the reliability and precision of these models, and the vast majority of the provided outputs lack clear connections with the state-of-the-art medical knowledge in the analysed domain. Ultrasound images are affected a little more by these problems, despite DL models performing accurately in histology and computed tomography (CT) domains.. In the sonography domain, certain factors influence the depth of the analysis that AI uses to make its decisions. Outside the breast cancer dataset and lung ultrasound data collected during the SARS-CoV-2 era, ultrasound images are rare and often lack detailed annotations. Sonography is well known, and the literature about it is massive. Still, the abundance of affine artefacts and significant operator-dependent variability generally impacts the quality of the analysis that can be automatically extracted from a dataset. To be clinically useful, an embedded DL model must not only be precise (accuracy) and efficient (lightweight) but also trustworthy (interpretability). Post-hoc interpretability methods like Class Activation Maps (CAMs) [4], Layer-Wise Relevance Propagation (LRP) [5], and other algorithms, such as LIME and SHAP [6] [7], enable interpretability but are strictly applied to already trained networks. While transparency-by-design networks, such as ProtoPNet [8] or Concept Bottleneck Models (CBMs) [9], force the model to adapt to defined concepts, they hinder the network's capacity to learn generalising concepts. This paper proposes an alternative method to analyse lung ultrasound images, utilising a previously explored training procedure that leverages CAM's ability to detect the features the network uses for classification decisions. The training methodology is tailored to detect a specific artefact, the B-Lines or comet tails, which are particularly used to identify pulmonary oedema, pneumonia, and pulmonary fibrosis. This paper is organised

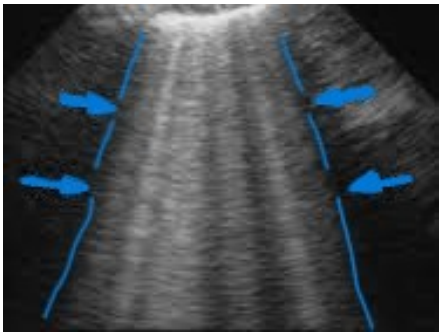

Fig. 1. Comet-tails, the limited area in the ultrasounds represents also the ideal prototype for B-Lines extracted as a feature map by the model for confluent B-lines

as follows. Section II first introduces the characteristics of

the employed dataset and details the training methods. It then presents our proposed adaptation of the Batch-CAM and prototype-informed training procedure [10] for the complex lung ultrasound domain. Following this, Section III presents the early results of this work. Finally, Sections IV and V provide a brief discussion of the methodology and explore possibilities for expanding the algorithms in future work.

II. Material and Method

A. Dataset and Pre-processing

The primary dataset used in this study consists of 1,208 lung ultrasound images, sourced from the TiAsssito telemedicine research project funded by the COVID-19 Tuscany Region call. Ultrasounds are classified into four severity scores based on the presence and confluence of B-line artefacts: '0' (A-lines, no B-lines), '1-3' (moderate B-lines), '4-5' (severe B-lines), and '6-8' (confluent B-lines/white lung).

The dataset was split into training (845 images, $\sim$70%), validation (181 images, $\sim$15%), and test (182 images, $\sim$15%) sets using a stratified random split, ensuring class distribution was maintained.

Before training, we calculated the per-channel mean and standard deviation of the training set for normalisation. All images were resized to 224×224 pixels and normalized using the calculated values (Mean: $[0.118, 0.118, 0.118]$, Std: $[0.115, 0.115, 0.115]$).

B. Model Architecture

The main model is a ConvNeXt-Tiny, pre-trained on ImageNet, as our feature extractor [11], the weights are used as a starting point since the whole network is trained. Given an input image of 224x224, the final stage of the model's feature extractor was hooked to retrieve the 7×7 feature maps. These maps, which have a shape of $[Batch, Channels, 7, 7]$, are the target for our Grad-CAM calculations. The resulting first model has 27.81 million parameters and an operational complexity of 4.46 GFLOPs, making it suitable for deployment on resource-constrained Point-of-Care Ultrasounds (POCUS) devices.

C. Geometric-Aware Model

The second architecture, a geometric-aware model, explicitly models and corrects for spatial variance by inserting a Spatial Transformer Network (STN) module [12] before the ConvNeXt backbone. This STN is a differentiable module that learns to apply an optimal 2D affine transformation to each input image before the primary classifier processes it. The STN operates sequentially: first, a small localisation sub-network analyses the input image and regresses the six parameters of a 2x3 affine transformation matrix, which is initially set to an identity transformation. Second, a grid generator component utilises this matrix to compute a sampling grid, which maps coordinates in the desired transformed output back to the corresponding coordinates in the original input. Finally, a sampler uses this grid to resample the input image, producing a new, geometrically normalised version. This transformed image is then fed into the ConvNeXt backbone for classification. A critical implication of this design is that the subsequent Grad-CAM explanations are generated from this geometrically normalised representation, potentially offering more consistent visual saliency maps that are invariant to the initial presentation of the anatomy. The resulting second model has 28.68 million parameters and an operational complexity of 4.54 GFLOPs.

D. Prototype-based Loss Functions

Our core hypothesis is that model interpretability and robustness can be improved by forcing the model's CAMs to resemble a prototype of each class.

1) Class Prototype Generation: We first compute a mean prototype image for each of the four classes by averaging all normalised images belonging to that class from both the training and validation sets. This results in four 224×224 prototype images, $\mathbf{P}_c$, where c is the class index.

2) CAM-Prototype Regularization: During training, the total loss $\mathcal{L}_{\text{total}}$ is a weighted sum of the standard classification loss and a novel reconstruction loss, Equation 1:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{recon}} \tag{1}$$

Where $\mathcal{L}_{\text{CE}}$ is the Cross-Entropy loss for the classification task and $\mathcal{L}_{\text{recon}}$ is our CAM-prototype loss (in our experiments, $\lambda = 1$).

We investigated several variants of $\mathcal{L}_{\text{recon}}$:

1) **BatchCAM Prototype Loss (BatchCAM):** Instead of calculating a CAM for every image, we compute a single, averaged CAM for each class present in the batch. This average CAM is then compared to the corresponding class prototype $\mathbf{P}_c$.
2) **Augmented Prototype Loss (Augmented):** This loss compares an *individual* image's CAM to a randomly *augmented* version of its class prototype. Augmentations included random rotations ($\pm 15°$), translations ($\pm 10\%$), and scaling ($0.9 \times -1.1\times$), specifically designed to simulate the affine artefacts and operator variability common in ultrasound imaging.
3) **Augmented BatchCAM Prototype Loss:** A hybrid approach combining both batch-averaging of CAMs and augmentation of the target prototypes.

For the comparison metric between the CAM (a 7×7 heatmap) and the prototype (downsampled and converted to grayscale), we tested two functions: L1 Loss (mean absolute error) and a structural loss based on SSIM ($1 - \text{SSIM}$).

E. Training

All models were trained for a maximum of 35 epochs with a batch size of 32, using the Adam-W optimiser with a learning rate of 1×10^{-4}. A scheduler monitored validation accuracy, reducing the learning rate on a plateau. Early stopping was employed with a patience of 10 epochs to prevent overfitting.

TABLE I
FINAL TEST SET PERFORMANCE OF EXPERIMENTAL MODELS. ALL MODELS SHARE THE SAME ARCHITECTURE (CONVNEXT-TINY) AND COMPLEXITY, EXCEPT FOR STN_PROTOTYPE

Experiment	Accuracy (%)
BatchCAM_Prototype_L1	**92.86**
STN_Prototype_L1	**92.31**
Augmented_Prototype_L1	91.76
BatchCAM_Prototype_SSIM	91.21
Augmented_BatchCAM_Prototype_L1	90.11
Augmented_Prototype_SSIM	89.56

III. RESULTS

The performance of each experimental configuration on the 182-image test set is summarised in Table I.

The baseline model, BatchCAM_Prototype_L1, which regularises the batch-averaged CAM against the static prototype using L1 loss, achieved the highest test accuracy of 92.86%.

The Augmented_Prototype_L1 model, designed to handle affine variations explicitly, performed slightly worse at 91.76%. This suggests that explicit augmentation of the prototype may not be as effective as the implicit regularisation provided by batch-averaging.

Furthermore, models using L1 loss for $\mathcal{L}_{\text{recon}}$ consistently outperformed their SSIM-based counterparts. For instance, BatchCAM_Prototype_L1 (92.86%) was superior to BatchCAM_Prototype_SSIM (91.21%).

Qualitative analysis of the generated CAMs, particularly for mislabeled test images, confirms that the models predominantly learn to focus on relevant sonographic features. The prototype-based loss appears to regularise these attention maps, making them more consistent and less noisy compared to what is typically seen in models trained with classification loss alone [13], [14].

IV. DISCUSSION

Our results demonstrate that a prototype-regularised training scheme can successfully balance precision, complexity, and interpretability for POCUS analysis.

The higher performance of the BatchCAM_Prototype_L1 model is the key finding. We hypothesise that averaging the CAMs within a batch before comparison (as done in BatchCAM) acts as an effective noise reduction and variance stabilisation mechanism, as shown in Figure 2. The geometry-aware model performs in a similar range to the BatchCAM_Prototype approach, and the feature extraction results have a comparable accuracy, Figure 3. The batch-averaging implicitly handles the minor affine variations present in different patient images, rendering the explicit $Augmented_Prototype_Loss$ redundant or even counter-productive. The random augmentations may not perfectly model real-world sonographic distortions, whereas batch-averaging leverages the real data variance itself. The preference for L1 loss over SSIM suggests that a direct, pixel-wise enforcement of similarity between the CAM and the prototype's grayscale representation is more beneficial than a purely structural comparison for this task. This may force the model to learn a more constrained and biased attention map for each class.

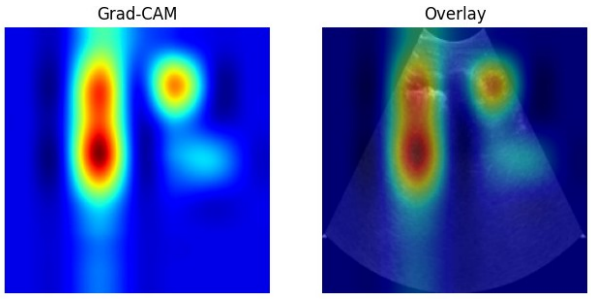

Fig. 2. Example of good prototype identification using the Batch-CAM prototype training approach

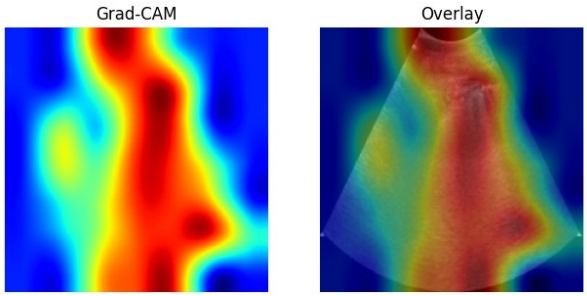

Fig. 3. Example of good prototype identification using the Batch-CAM prototype training approach with the geometric aware network

Ultimately, our approach successfully meets the three target criteria:

- **Precision:** Achieved 92.86% 4-class accuracy on a challenging LUS dataset.
- **Complexity:** The ConvNeXt-Tiny architecture is efficient and well-suited for embedding in POCUS devices.
- **Interpretability:** The training process optimises for interpretable CAM, guiding the model to a solution that mixes precision with a generalised identification approach.

V. FUTURE WORK

The following line of research lies in three main areas. First, we will apply this prototype-regularisation method to other lung ultrasound problems, such as disease classification, to validate its generalizability for tasks in which the desired objective is not well-defined in the image. Second, we will investigate more sophisticated prototype generation methods beyond simple averaging, such as using a variational autoencoder (VAE) to learn a richer latent space of class features. Furthermore, we believe integrating more complex geometry-aware DL networks could substantially improve model learning and generalisation capacity. By designing architectures that inherently respect the geometric properties of the data, such as rotational and translational symmetries common in medical imaging, we can instil a stronger inductive bias. This would help the model learn more fundamental anatomical representations rather than superficial, orientation-dependent artefacts. Such an approach should make the model more

robust and data-efficient, improving its ability to generalise to unseen data acquired from different devices or sonographers' scanning patterns. Finally, we will proceed with on-device testing of the BatchCAM-Prototype-L1 model to evaluate its real-world inference speed and diagnostic performance at the point of care.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," 2015.

[3] R. Obuchowicz, M. Strzelecki, and A. Piórkowski, "Clinical applications of artificial intelligence in medical imaging and image processing—a review," p. 1870, 2024.

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

[5] A. Binder, G. Montavon, S. Bach, K.-R. Müller, and W. Samek, "Layer-wise relevance propagation for neural networks with local renormalization layers," 2016. [Online]. Available: https://arxiv.org/abs/1604.00825

[6] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[7] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[8] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, "This looks like that: deep learning for interpretable image recognition," *Advances in neural information processing systems*, vol. 32, 2019.

[9] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, "Concept bottleneck models," in *International conference on machine learning*. PMLR, 2020, pp. 5338–5348.

[10] G. Ignesti, D. Moroni, and M. Martinelli, "Batch-cam: Introduction to better reasoning in convolutional deep learning models," 2025. [Online]. Available: https://arxiv.org/abs/2510.00664

[11] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," 2022. [Online]. Available: https://arxiv.org/abs/2201.03545

[12] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," *Advances in neural information processing systems*, vol. 28, 2015.

[13] G. Ignesti, D. Moroni, and M. Martinelli, "You've got the wrong number: evaluating deep learning training paradigms using handwritten digit recognition data," *Pattern Recognition and Image Analysis*, vol. 34, no. 4, 2024, in press.

[14] G. Ignesti, M. Martinelli, and D. Moroni, "You've got the wrong outfit: Evaluating deep learning paradigms on digit and fashion recognition," in *International Conference on Pattern Recognition*. Springer, 2024, pp. 202–215.