

Enhancing the Trustworthiness of Deep Ensemble via Explainability in Medical Imaging

Gbègninougbo Aurel Davy Tchokponhoue[1] and Ali Idri[2(✉)]

[1] Faculty of Medical Science, UM6P Hospitals, Mohammed VI Polytechnic University, Ben Guerir 43150, Morocco
aurel.tchokponhoue@um6p.ma

[2] Data and Software Sciences, ENSIAS, Mohammed V University, Rabat, Morocco
ali.idri@ensias.um5.ac.ma

Abstract. Deep Ensembles are among the best methods for uncertainty quantification in machine learning, especially in high-stakes areas like medical imaging. However, explaining how they work is still a challenge. To address this, we present two algorithms for their explainability: *Ensemble Conditioned Class Activation Map (ECAM)* and *Selective Ensemble Class Activation Map (SECAM)*. ECAM uses adaptive, sample-specific weighting based on predictive entropy to combine explainability maps that match ensemble predictions, which improves interpretability. SECAM uses the ADCC metric to choose the best explainability map from ensemble members whose predictions are closest to the output of the ensemble. This provides a model-specific view. We tested the suggested methods on a breast cancer ultrasound dataset (BUSI), which includes ground truth masks for localization assessment. The results show that ECAM and SECAM highlight important clinical areas and align well with the ground truth masks. These findings suggest that the proposed methods can improve the transparency of deep ensemble models and contribute to more trustworthy AI systems for medical diagnosis.

Keywords: Explainability · Deep ensemble · Uncertainty quantification · Trustworthy AI

1 Introduction

Deep learning (DL) has become a game-changing tool in medical diagnostics, showing unprecedented accuracy in various clinical tasks like imaging-based disease detection, predicting outcomes, and treatment planning [18]. However, the deployment of DL models in real-world clinical contexts remains limited due to concerns about their reliability, interpretability, and trustworthiness [7]. In high-stakes fields like healthcare, uncertainty quantification (UQ) is very important

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 162–176, 2027.
https://doi.org/10.1007/978-3-032-30494-0_11

for making reliable models because it shows how sure the model's predictions are.

In recent years, Monte Carlo Dropout, Bayesian Neural Networks, and Deep Ensembles are some of the UQ methods that have been investigated in the healthcare domain [7]. Among these, Deep Ensemble models have become the state-of-the-art, providing great results in both predicting accuracy and estimating uncertainty [2,6]. However, the explainability of these models continues to pose a challenge and remains largely unexamined in UQ studies for imaging applications, resulting in a significant gap.

In the wider area of ensemble learning applied to imaging, attempts to improve explainability often centered on combining interpretation maps created by individual ensemble members [4,11,15,19]. These methods were inspired by research that combined maps from different interpretability techniques for a single network [5,9,13,16]. They used techniques like averaging, summation, median calculation, or attention mechanisms. These strategies showed that they could produce good explainability maps for the prediction of a single model, setting the stage for further research in ensemble learning [13].

Building on this foundation, some studies on ensemble methods used similar aggregation strategies while others extended them to new ones [21]. Specifically, they introduced weighted aggregation schemes, where the weights assigned to individual model maps are based on the validation performance of each ensemble member [11]. Although these approaches usually achieve acceptable results, they have some limitations. First, they ignore the sample-specific variability of each model performance, assuming that each model works the same way across the dataset. However, a model member that performs well overall may struggle with certain cases, leading to unreliable or misleading explainability maps for those instances. Second, these methods produced an explainability map that is independent of the final prediction of the ensemble. This is an important drawback because the predictions from individual models can sometimes contradict the ensemble's overall decision. As a result, the combined maps may include information that does not match the ensemble's logic, adding noise and reducing the overall reliability of the explainability.

Class Activation Mapping (CAM) techniques [22] are widely used for explainability in DL, as they enable visualization of image regions that most influence a model's predictions. This paper describes two novel explainability algorithms developed for Deep Ensemble models, both leveraging CAM techniques. The first algorithm, Ensemble Conditioned Class Activation Map (ECAM), combines the explainability maps of individual ensemble members and weights their contributions according to predictive uncertainty, rather than using uniform or validation-based weighting. This method conditions the explainability on the ensemble's collective decision, producing a coherent and sample-specific representation. The second algorithm, Selective Ensemble Class Activation Map (SECAM), uses a selective approach, identifying and using only the most relevant explainability maps from the individual models that match the ensemble's overall prediction, reducing noise and inconsistencies. This strategy is inspired

by the use of surrogate models, which are commonly used to interpret complex black-box models.

ECAM and SECAM address an important gap in the explainability of Deep Ensemble models, improving the transparency of this UQ method. We show the effectiveness of these algorithms in breast cancer screening using ultrasound data. By comparing the explainability maps generated by the ensemble model to the ground truth masks, we evaluate their ability to capture clinically relevant features and accurately reflect the ensemble's decision-making process.

In Sect. 2, we provide an overview of related work on explainability in deep learning, with a focus on imaging applications. Section 3 offers background information on the deep ensemble method and the Class Activation Mapping (CAM) strategy employed in this study. In Sect. 4, we introduce two novel Deep ensemble explainability algorithms. Section 5 presents the experimental results, including the generated maps produced by the algorithms. Finally, Sect. 6 discusses the limitations of the study, and Sect. 7 concludes the paper, offering suggestions for future research directions.

2 Explainability of Deep Learning-Based Imaging Applications

Explainability methods are crucial in deep learning, especially for image-based tasks, where understanding the decision-making process of black-box convolutional model is vital for sensitive domains like healthcare and autonomous systems. Local Interpretable Model-Agnostic Explanations (LIME) addresses this by creating local surrogate models to explain predictions, identifying important features through perturbations while balancing interpretability using a complexity penalty [12]. Similarly, SHapley Additive exPlanations (SHAP), based on game theory, computes Shapley values to attribute feature contributions to predictions, providing detailed insights into individual and interactive feature effects [8]. CAM techniques identify the important regions of the image that drive the model decision by combining weights from the final fully connected layer with activations from the last convolutional layer, generating class-specific heatmaps for CNNs [22]. Grad-CAM extends CAM by using backpropagated gradients with feature maps to highlight the regions most relevant to specific predictions and improving visual explanations [14]. Enhanced variants like Grad-CAM++, XGradCAM, and ScoreCAM give stronger and visually refined explanations. In the context of ensemble models, Rajani and Raymond introduced Weighted Average Ensemble Explanation and Penalized Weighted Average Ensemble Explanation for Visual Question Answering (VQA), where they used validation score as a global weight for model aggregation [11]. In contrast, we propose a new approach that uses sample-level weights derived from the predictive entropy of each ensemble member for the particular prediction. This led to adaptive explanations that moved beyond naive aggregation to better reflect the behavior of the Deep Ensemble method.

3 Background

3.1 Deep Ensemble

Deep ensembles [6] have become a strong method for estimating predictive uncertainty in DL. Unlike BNN, which need complex changes and are heavy on computation, Deep Ensembles provide a simpler and more scalable option. The idea is to train multiple neural networks independently, with each network initialized with different weights, and then combine their predictions to assess uncertainty. This method captures the differences in model outputs that result from different weight settings. It effectively measures both epistemic and aleatoric uncertainty.

Let $p(y \mid \mathbf{x})$ represent the predicted posterior distribution over the classes for a given input $\mathbf{x}$. In the case of an ensemble comprising M models, the predictive distribution is calculated by averaging the outputs of the individual models:

$$p(y \mid \mathbf{x}) = \frac{1}{M} \sum_{m=1}^{M} p_{\theta_m}(y \mid \mathbf{x}), \tag{1}$$

where θ_m denotes the parameters of the m-th model.

Epistemic uncertainty is quantified by calculating the variance of ensemble predictions:

$$\sigma^2 \big[p(y \mid \mathbf{x})\big] = \frac{1}{M} \sum_{m=1}^{M} \left(p_{\theta_m}(y \mid \mathbf{x}) - p(y \mid \mathbf{x})\right)^2. \tag{2}$$

The aleatoric uncertainty is measured using the entropy of the predictive distribution:

$$H\big[p(y \mid \mathbf{x})\big] = -\sum_{y \in \mathcal{Y}} p(y \mid \mathbf{x}) \log p(y \mid \mathbf{x}). \tag{3}$$

3.2 Score-CAM

Score-CAM [20] is a method for explaining the decisions of convolutional neural networks (CNNs). It highlights the areas of an image that have a big impact on the model's predictions. Unlike methods that rely on gradients, Score-CAM avoids using gradients, which makes the explanations less sensitive to gradient noise and improves their robustness. To start, the method gets activation maps from a convolutional layer that has been chosen. After that, each activation map is resized to fit the size of the input image and normalized to make a mask. This mask is applied to the original image, and the modified image is fed back into the network to observe changes in the target class score. The increase in the target class score from the masked image acts as a weight for the corresponding activation map. The final class activation map is created by calculating a weighted sum of these activation maps. This method highlights the most important areas

of the input image. Score-CAM has shown better performance in both recognition and localization tasks compared to earlier methods. It offers clearer visual explanations without needing gradient information.

Given a convolutional layer l of a model f and a target class c, the Score-CAM localization map is defined as:

$$L^{c}_{\text{Score-CAM}} = \text{ReLU}\left(\sum_{k} \alpha^{c}_{k} A^{k}_{l}\right), \tag{4}$$

where A^{k}_{l} denotes the k-th activation map at layer l, and α^{c}_{k} is the importance weight defined as:

$$\alpha^{c}_{k} = \text{softmax}_{k}\left(Y^{(c)}(M_k) - Y^{(c)}(\mathbf{X}_b)\right), \tag{5}$$

with $Y^{(c)}(\cdot)$ denoting the pre-softmax score for class c, $\mathbf{X}_b$ a baseline input, and M_k defined as:

$$M_k = \left(\frac{U(A^{k}_{l}) - \min_m U(A^{m}_{l})}{\max_m U(A^{m}_{l}) - \min_m U(A^{m}_{l})}\right) \odot \mathbf{X}_b, \tag{6}$$

where $U(\cdot)$ denotes upsampling and $\odot$ element-wise multiplication.

3.3 Average DCC (ADCC)

The Average DCC (ADCC) explainability map evaluation metric [10], was employed to develop our second algorithm, which adopts a selective CAM strategy. This strategy identifies the explainability map that best matches the output of the Deep Ensemble model. ADCC combines three key scores: Maximum Coherency, Minimum Complexity, and Minimum Average Drop, using a harmonic mean, as shown in (7):

$$\text{ADCC}(\mathbf{x}) = 3\left(\frac{1}{\text{Coherency}(\mathbf{x})} + \frac{1}{1 - \text{Complexity}(\mathbf{x})} + \frac{1}{1 - \text{AverageDrop}(\mathbf{x})}\right)^{-1}. \tag{7}$$

The Coherency score component makes sure that the CAM captures all relevant features that explain a prediction. It also removes any irrelevant or redundant features in a clear and logical manner, as defined in (8).

$$\text{Coherency}(\mathbf{x}) = \frac{\text{Cov}\big(\text{CAM}_c(\mathbf{x}) \odot \text{CAM}_c(\mathbf{x}), \text{CAM}_c(\mathbf{x})\big)}{\sigma_{\text{CAM}_c(\mathbf{x}) \odot \text{CAM}_c(\mathbf{x})}\, \sigma_{\text{CAM}_c(\mathbf{x})}}. \tag{8}$$

The complexity score promotes simplicity by requiring the CAM to show the smallest set of pixels that sufficiently explains the prediction, as defined in (9).

$$\text{Complexity}(\mathbf{x}) = \|\text{CAM}_c(\mathbf{x})\|_1. \tag{9}$$

The Average Drop part means that a good explanation map should make the confidence drop as little as possible when the masked input, based on the

explanation map, is used instead of the original image. A smaller drop shows that the explanation map retains the most important information for the prediction.

Recently, new metrics have come up to improve the evaluation of these maps, especially by using segmentation-based ground truth annotations. For example, [17] proposed the Explanation Significance Assessment (ESA) and its WESA extension.These methods use pixel-level spatial analysis and distance measurements to evaluate how well explanation maps match with annotated regions. In contrast, the ADCC metric we use in our study does not depend on ground-truth annotations. This makes it suitable for fully weakly-supervised settings without needing extra labeling.

4 Novel Deep Ensemble Explainability Algorithms

In this section, we present two new explainability algorithms, ECAM and SECAM, created for Deep Ensemble models. These algorithms make the model predictions easier to understand by providing clear explanations that reflect the decision-making process of the ensemble. We give a detailed description of each algorithm and explain their basic principles.

4.1 Ensemble Conditioned Class Activation Map (ECAM)

Algorithm 1 introduces *Ensemble Conditioned Class Activation Map (ECAM)*, a novel method to build an explainability map for deep ensemble by integrating the outputs of the individual models in the deep ensemble. Unlike existing approaches that combine class activation maps (CAMs) using uniform averaging or validation-based weighting, ECAM leverages *uncertainty-aware metrics* derived from the *normalized inverse entropy* of each model's predictive distribution.

Moreover, while traditional methods compute model-specific CAMs independently, Algorithm 1 ensures that the CAM of each member of the ensemble M_t is conditioned on the global prediction of the ensemble $\mathbf{y}_{\text{ens}}$. This design matches each explainability map with the group's behavior of the ensemble.

Given an input image $\mathbf{X}$, each model M_t, for $t = 1, \ldots, T$, produces a CAM, noted as $\text{CAM}_t(\mathbf{X})$, based on $\mathbf{y}_{\text{ens}}$. The reliability weight for each model is calculated using its predictive entropy $H(\mathbf{p}_t)$ as:

$$w_t = 1 - \frac{H(\mathbf{p}_t)}{\log C}, \tag{10}$$

where $\mathbf{p}_t \in \mathbb{R}^C$ is the predicted probability distribution of the model M_t.$H(\mathbf{p}_t)$ is the predicted entropy and C is the number of output classes.

The weights $\{w_t\}_{t=1}^T$ are normalized so that $\sum_{t=1}^T w_t = 1$. The final ECAM output is computed as the weighted sum of individual CAMs:

$$\text{ECAM}(\mathbf{X}) = \sum_{t=1}^{T} w_t \cdot \text{CAM}_t(\mathbf{X}). \tag{11}$$

This formulation makes sure that the resulting explanation focuses on the more reliable members of the group. This improves both the interpretability and how well it matches the overall decision of the group.

Algorithm 1. Ensemble Conditioned Class Activation Map (ECAM)

Require: Input image $\mathbf{X}$; Ensemble models $\{M_1, \ldots, M_T\}$; Number of classes C; Ensemble prediction $\mathbf{y}_{\text{ens}}$
Ensure: Explanation map ECAM($\mathbf{X}$)
Initialize empty lists $\mathcal{W} \leftarrow [\,]$ and $\mathcal{C} \leftarrow [\,]$
for $t = 1$ to T **do**
  $\mathbf{p}_t \leftarrow \text{softmax}(M_t(\mathbf{X}))$
  $\text{CAM}_t(\mathbf{X}) \leftarrow \text{ScoreCAM}(M_t, \mathbf{X}, \mathbf{y}_{\text{ens}})$
  $H_t \leftarrow -\sum_{c=1}^{C} \mathbf{p}_t[c] \log \mathbf{p}_t[c]$
  $\tilde{w}_t \leftarrow 1 - \frac{H_t}{\log C}$
  Append $\tilde{w}_t$ to $\mathcal{W}$
  Append $\text{CAM}_t(\mathbf{X})$ to $\mathcal{C}$
end for
for $t = 1$ to T **do**
  $w_t \leftarrow \frac{\mathcal{W}[t]}{\sum_{j=1}^{T} \mathcal{W}[j]}$
end for
Compute $\text{ECAM}(\mathbf{X}) \leftarrow \sum_{t=1}^{T} w_t \cdot \mathcal{C}[t]$
return $\text{ECAM}(\mathbf{X})$

4.2 Selective Ensemble Class Activation Map (SECAM)

Algorithm 2, *Selective Ensemble Class Activation Map (SECAM)*, presents a focused method for creating explainability maps. It picks only the most relevant ensemble models, specifically those whose predictions match the overall output of the ensemble. Unlike current methods that combine all ensemble member CAMs or calculate them without considering the ensemble's output, SECAM filters ensemble members. This process ensures that the final explanation comes from the most trustworthy and semantically aligned model.

Given an input image $\mathbf{X}$, each model M_t in the ensemble is checked for eligibility based on whether its prediction $\mathbf{y}_t$ matches the ensemble's global prediction $\mathbf{y}_{\text{ens}}$. The set of eligible models is defined as:

$$\mathcal{M}_{\text{eligible}} = \{M_t \mid t \in \{1, \ldots, T\},\ \mathbf{y}_t = \mathbf{y}_{\text{ens}}\},$$

For each eligible model $M_t \in \mathcal{M}_{\text{eligible}}$, a class activation map $\text{CAM}_t(\mathbf{X})$ and an associated alignment score $\text{ADCC}_t(\mathbf{X})$ are computed. The alignment score $\text{ADCC}_t(\mathbf{X})$ serves as a measure to assess the quality of the explanation provided by CAM_t.

The final SECAM explanation comes from the CAM linked to the eligible model with the highest alignment score. By using both an eligibility criterion and the alignment score, SECAM makes sure that the final explanation comes from

the most relevant and reliable model in the group. This improves robustness and prevents noisy contributions from misaligned predictions.

Algorithm 2. Selective Ensemble Class Activation Map (SECAM)

Require: Input image $\mathbf{X}$; Ensemble models $\{M_1, \ldots, M_T\}$; Ensemble prediction $\mathbf{y}_{\text{ens}}$
Ensure: Explanation map SECAM($\mathbf{X}$)
Initialize $s^* \leftarrow -\infty$ and SECAM($\mathbf{X}$) $\leftarrow \varnothing$
for $t = 1$ to T **do**
  $\mathbf{p}_t \leftarrow \text{softmax}(M_t(\mathbf{X}))$
  $\mathbf{y}_t \leftarrow \arg\max_c \mathbf{p}_t[c]$
  if $\mathbf{y}_t = \mathbf{y}_{\text{ens}}$ **then**
    $\text{CAM}_t \leftarrow \text{ScoreCAM}(M_t, \mathbf{X}, \mathbf{y}_{\text{ens}})$
    $s_t \leftarrow \text{ADCC}(\mathbf{X}, \text{CAM}_t)$
    if $s_t > s^*$ **then**
      $s^* \leftarrow s_t$
      $\text{SECAM}(\mathbf{X}) \leftarrow \text{CAM}_t$
    end if
  end if
end for
return SECAM($\mathbf{X}$)

5 Experiments

Dataset: The Breast Ultrasound Dataset [1] used in this study contains breast ultrasound images from 600 female patients aged 25 to 75 years. The dataset was put together in 2018 and has 780 images, each with an average resolution of 500×500 pixels in PNG format. These images are divided into three classes: normal, benign, and malignant, with ground truth annotations provided for each image. This dataset provides a robust foundation for training and evaluating machine learning models for breast cancer classification, detection, and segmentation.

Model Architecture: The model architecture is based on the ResNet50 [3] network pre-trained on ImageNet. A custom classification head is added on top, consisting of three fully connected layers with 512, 128, and 3 neurons, respectively. A dropout layer with a probability of 0.2 is included between these layers to reduce overfitting. The Deep Ensemble model comprises 10 instances of this architecture, where each model is initialized with different random weights. During training, only the final convolutional layer of the ResNet50 backbone is fine-tuned, while the remaining layers are frozen.

Evaluation Protocol: As the primary focus of this study is on explainability within deep ensemble uncertainty quantification (UQ) methods, optimizing classification performance was not the main objective. Nevertheless, we evaluated

the classification performance of the model using standard metrics, including accuracy, precision, recall, and F1-score, to ensure a reasonably strong baseline. We then analyzed the relationship between classification performance and uncertainty estimates, specifically predictive entropy, in order to assess the ability of the UQ method to identify misclassified samples through elevated uncertainty. In addition, we evaluated the quality of the explainability maps generated by ECAM and SECAM by analyzing their alignment with ground truth masks, thereby assessing the reliability of the model's explanations. As a baseline, we considered the VCAM ensemble explainability method [11], which aggregates explanation maps from individual ensemble members using validation performance as weighting. This comparison allows us to qualitatively assess the effectiveness of the proposed methods.

5.1 Relationship Between Predictive Uncertainty and Model Performance

In our experiment, the Deep Ensemble model shows an accuracy of 88.61%, a precision of 89.46%, a recall of 88.61%, and an F1-score of 88.75%. These numbers suggest that the model's performance is good and similar to the best results on the Breast Ultrasound dataset. Additionally, we examine the relationship between predictive uncertainty and model performance (see Fig. 1).This helps us understand the practical value of predictive entropy derived from Deep Ensemble as a measure of uncertainty.

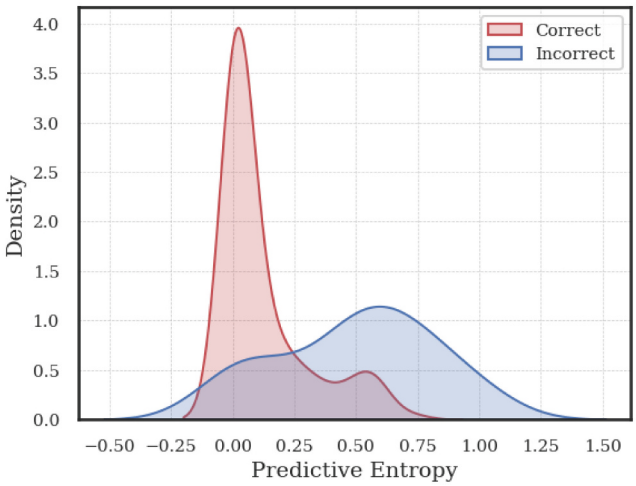

Fig. 1. Distribution of predictive entropy given classification performance.

A key observation is the difference between correctly classified and misclassified samples when examining predictive entropy. Correctly classified samples

have low predictive entropy, showing high confidence in predictions. In contrast, misclassified samples have high entropy, indicating greater uncertainty. This pattern shown through the density plots supports the value of predictive entropy as a measure of model confidence. These findings have several important implications. First, the link between high entropy and misclassified samples highlights the potential of predictive entropy as a tool for identifying unclear or difficult cases. In clinical settings, these cases could be identified for secondary review. This improves patient safety by reducing the dependence on uncertain predictions. This supports the broader goal of using risk-aware AI systems in sensitive areas like healthcare. Additionally, the results highlight opportunities for selective prediction. In this approach, the model avoids making predictions on samples with high uncertainty, referring them to human experts or other diagnostic tools.

Finally, misclassified samples with high uncertainty may reveal issues in the dataset, like edge cases or classes that are not well represented. Examining these situations further could result in specific improvements in both the dataset and the model training process.

5.2 Explanation Map Analysis

The qualitative analysis of the explainability maps presented in Fig. 2 and Fig. 3 provides detailed insight on how ensemble-based interpretability methods work in both correct and failure scenarios. In Fig. 2, where predictions are correct and the regions of interest (ROIs) are correctly identified, both ECAM and SECAM create activation maps that match the ground truth masks. When compared to the baseline VCAM, the proposed methods show more spatially focused and semantically consistent activations. This indicates that ECAM and SECAM better gather ensemble information to highlight clinically important areas, thereby improving how we interpret deep ensemble models.

Figure 3 on the other hand, shows two different types of failure modes. First, in the examples with red frames (correct classification but wrong localization), the model predicts the right label while concentrating on areas that don't match the annotated ROI. This behavior suggests that the model may depend on spurious or weakly correlated features that are predictive of the class but lack clinical significance. Alternatively, it may show that there are important visual cues that the ground truth annotations didn't pick up on. This indicates that the model can sometimes find complementary patterns that aren't in the masks that people defined. These situations emphasize the importance of explainability in understanding model behavior and improving UQ. Notably, ECAM usually generates more diffuse and less overconfident activation maps, similar to VCAM. In contrast, SECAM often produces sharper but possibly misleading explanations because it relies on just one ensemble member. This observation points out a trade-off between localization and predictive performance. It also promotes including uncertainty estimation alongside the model explanation. This helps assess how confident the model is in its interpretability results.

Second, in the green-framed examples (misclassification with correct localization), the explainability maps clearly show the pathological areas despite the

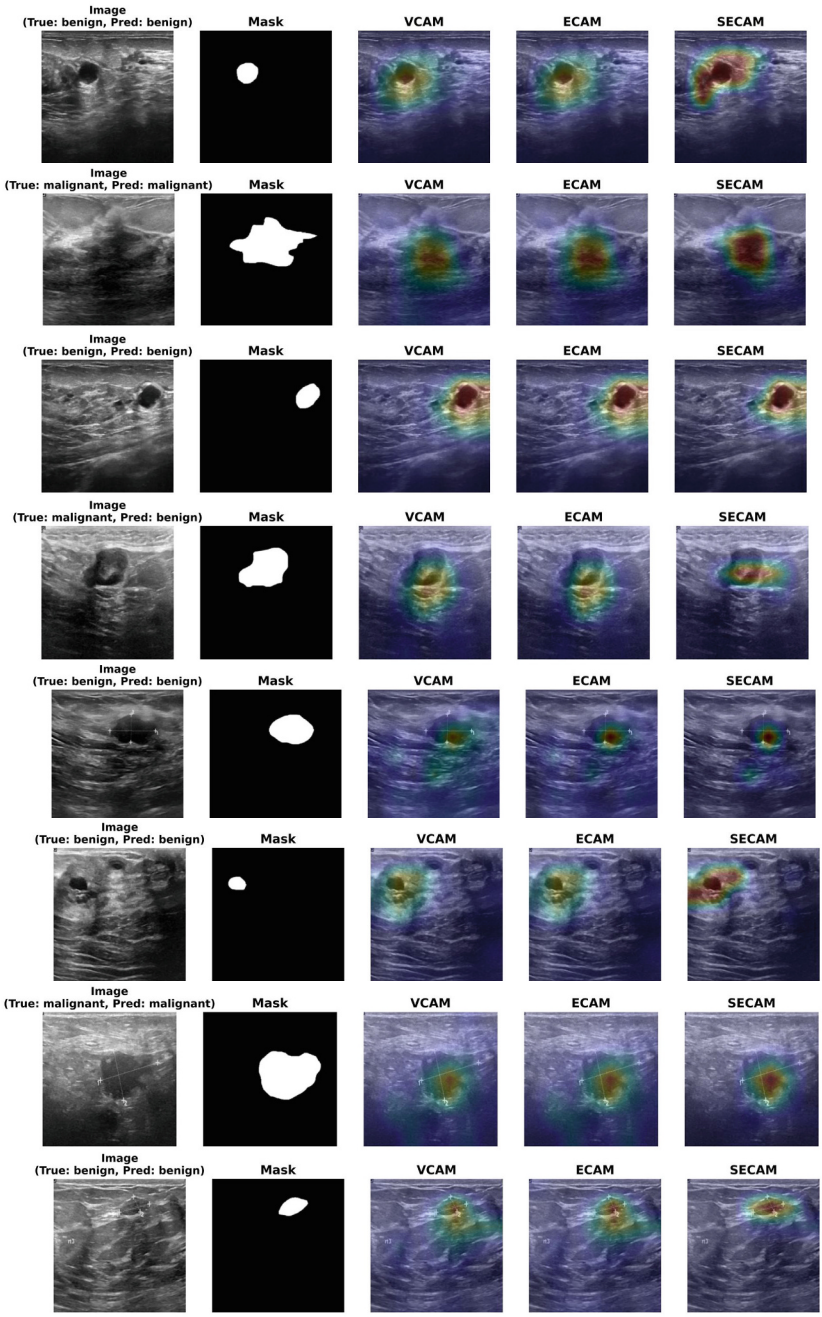

Fig. 2. Explainability map generated by ECAM and SECAM compared to VCAM on some correctly classified samples with accurate localization of regions of interest (ROIs).

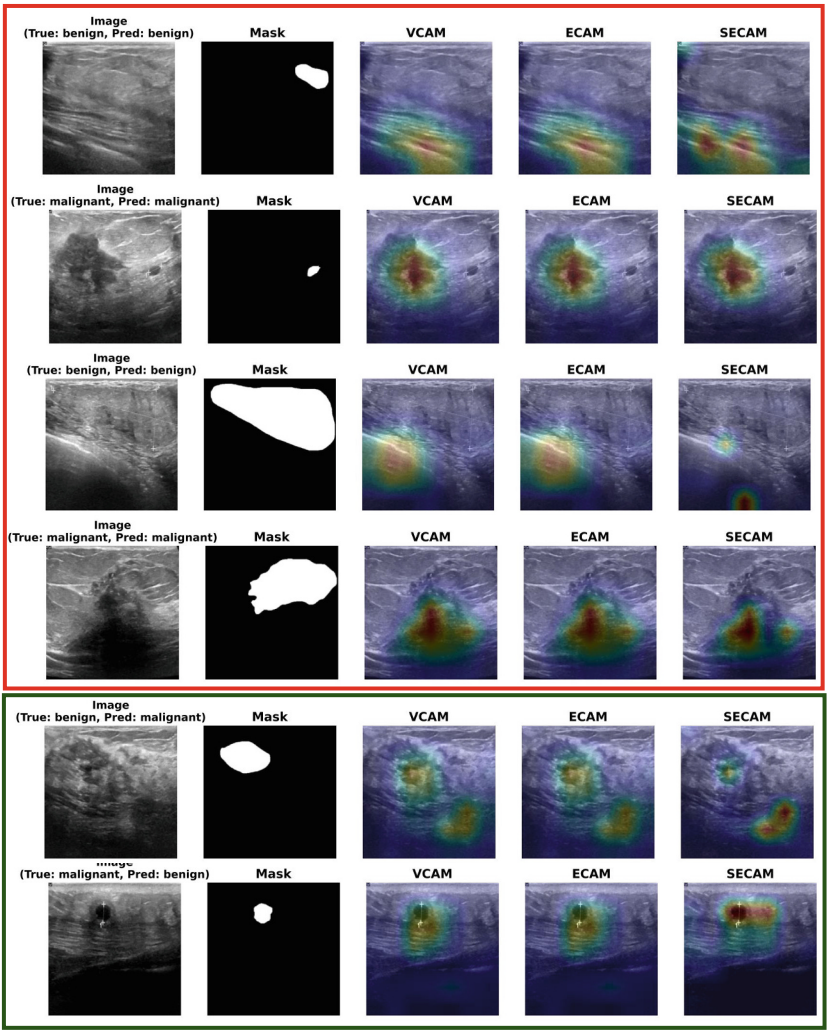

Fig. 3. Explainability failure modes. Red-framed examples show correct classifications with incorrect localization. Green-framed examples show misclassifications with correct localization. (Color figure online)

wrong predictions. This is an important observation because it suggests that the model can recognize key anatomical structures but struggles during the decision-making stage. This difficulty likely arises from confusion in feature representation or overlap among classes, like benign versus malignant. These cases show that classification errors do not always come from poor localization. Instead, they arise from limits in learning features that are discriminative. This difference is

important from a clinical point of view. Correct localization can still provide valuable diagnostic support, even if the predicted label is incorrect.

Overall, ECAM and SECAM show consistent and stable behavior, even though they use different methods to create ensemble-level explainability maps. The results show that adding explainability to deep ensemble frameworks makes the model more transparent and more informative. However, there is an important limitation: the quality of ensemble-level explanations depends on how reliable the individual ensemble members are. If the base models fail to provide reliable or meaningful explanations, the combined ensemble explanation cannot be completely trusted. This shows how important it is to improve model performance, estimate uncertainty, and enhance explainability. Doing so will help develop trustworthy AI systems for critical uses like medical imaging.

6 Limitations

Despite the promising results of this study, several limitations still exist. First, the algorithms were evaluated on a single dataset, and their generalizability to other imaging modalities and datasets has not yet been established. Second, the reliability of ensemble-level explanations is fundamentally dependent on the quality of individual ensemble members. This highlights a structural dependency that cannot be fully mitigated at the aggregation level alone and suggests that improving base model robustness is a prerequisite for reliable ensemble interpretability. Finally, the proposed methods rely on post-hoc explainability techniques, which inherit known limitations such as sensitivity to model architecture, gradient saturation, and lack of causal interpretability. As a result, the generated explanations should not be interpreted as causal evidence but rather as approximations of model attention, which may not always reflect true decision-making processes.

7 Conclusion

In this work, we introduced two novel ensemble-based explainability methods, ECAM and SECAM, designed to improve the explainability of Deep ensemble models used for uncertainty quantification in medical imaging. Our work was applied to the diagnosis of breast cancer using ultrasound images. By leveraging complementary aggregation strategies, the proposed methods provide insights into both collective ensemble behavior and individual model contributions, enabling a more comprehensive understanding of model decision processes. Through qualitative analysis, we demonstrated that ECAM and SECAM are capable of producing coherent and clinically relevant activation maps, while also revealing important failure modes such as the decoupling between localization and classification. Overall, this work contributes to the growing effort toward trustworthy AI by bridging ensemble learning, uncertainty quantification, and explainability. We believe that such integrated approaches are essential for enabling the safe and reliable deployment of deep learning systems in high-stakes domains such as medical imaging.

Future work will assess the generalizability of ECAM and SECAM across additional datasets and imaging modalities, including histopathology, cytology, mammography, and CT imaging. This assessment will help us evaluate their reliability in various clinical settings. We also plan to investigate uncertainty in explanations themselves, with the aim of further improving the reliability and transparency of model interpretation.

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data Brief **28**, 104863 (2020). https://doi.org/10.1016/j.dib.2019.104863
2. Gawlikowski, J., et al.: A survey of uncertainty in deep neural networks. Artif. Intell. Rev. **56**(1), 1513–1589 (2023). https://doi.org/10.1007/s10462-023-10562-9
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
4. Hosny, K.M., Mohammed, M.A., Salama, R.A., Elshewey, A.M.: Explainable ensemble deep learning-based model for brain tumor detection and classification. Neural Comput. Appl. (2024). https://doi.org/10.1007/s00521-024-10401-0
5. Hryniewska-Guzik, W., Longo, L., Biecek, P.: CNN-based explanation ensembling for dataset, representation and explanations evaluation. In: Longo, L., Lapuschkin, S., Seifert, C. (eds.) Explainable Artificial Intelligence, vol. 2154, pp. 346–368. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-63797-1_18
6. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017)
7. Lambert, B., Forbes, F., Doyle, S., Dehaene, H., Dojat, M.: Trustworthy clinical AI solutions: a unified review of uncertainty quantification in Deep Learning models for medical image analysis. Artif. Intell. Med. **150**, 102830 (2024). https://doi.org/10.1016/j.artmed.2024.102830
8. Lundberg, S.M., Lee, S.I.: A Unified Approach to Interpreting Model Predictions (2017)
9. Ntrougkas, M., Gkalelis, N., Mezaris, V.: TAME: attention mechanism based feature fusion for generating explanation maps of convolutional neural networks. In: 2022 IEEE International Symposium on Multimedia (ISM), pp. 58–65 (2022). https://doi.org/10.1109/ISM55400.2022.00014
10. Poppi, S., Cornia, M., Baraldi, L., Cucchiara, R.: Revisiting the evaluation of class activation mapping for explainability: a novel metric and experimental analysis. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2299–2304 (2021). https://doi.org/10.1109/CVPRW53098.2021.00260
11. Rajani, N.F., Mooney, R.J.: Ensembling visual explanations. In: Escalante, H.J., et al. (eds.) Explainable and Interpretable Models in Computer Vision and Machine Learning. TSSCML, pp. 155–172. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-98131-4_7
12. Ribeiro, M.T., Singh, S., Guestrin, C.: "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, pp. 1135–1144 (2016). https://doi.org/10.1145/2939672.2939778

13. Rieger, L., Hansen, L.K.: Aggregating explanation methods for stable and robust explainability. arXiv: Learning (2019)
14. 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: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 618–626 (2017). https://doi.org/10.1109/ICCV.2017.74
15. Shorfuzzaman, M., Hossain, M.S., El Saddik, A.: An explainable deep learning ensemble model for robust diagnosis of diabetic retinopathy grading. ACM Trans. Multimedia Comput. Commun. Appl. **17**(3s) (2021). https://doi.org/10.1145/3469841
16. Sritharan, N., Gnanavel, N., Inparaj, P., Meedeniya, D., Yogarajah, P.: Ensemble-CAM: unified visualization for explainable cervical cancer identification. In: 2024 International Research Conference on Smart Computing and Systems Engineering (SCSE), vol. 7, pp. 1–6 (2024). https://doi.org/10.1109/SCSE61872.2024.10550859
17. Stodt, J., Reich, C., Clarke, N.: A Novel metric for XAI evaluation incorporating pixel analysis and distance measurement. In: 2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 1–9. IEEE Computer Society (2023). https://doi.org/10.1109/ICTAI59109.2023.00009
18. Thakur, G.K., Thakur, A., Kulkarni, S., Khan, N., Khan, S.: Deep learning approaches for medical image analysis and diagnosis. Cureus **16**(5), e59507 (2024). https://doi.org/10.7759/cureus.59507
19. Vanitha, K., Mahesh T., R., Sree, S.S., Guluwadi, S.: Deep learning ensemble approach with explainable AI for lung and colon cancer classification using advanced hyperparameter tuning. BMC Med. Inform. Dec. Mak. **24**(1), 222 (2024). https://doi.org/10.1186/s12911-024-02628-7
20. Wang, H., et al.: Score-CAM: score-weighted visual explanations for convolutional neural networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 111–119 (2020). https://doi.org/10.1109/CVPRW50498.2020.00020
21. Zhao, Z., Alzubaidi, L., Zhang, J., Duan, Y., Naseem, U., Gu, Y.: Robust and explainable framework to address data scarcity in diagnostic imaging (2024). https://doi.org/10.48550/ARXIV.2407.06566
22. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2921–2929 (2016). https://doi.org/10.1109/CVPR.2016.319