Persistent Misclassification Analysis for Improving Thyroid Cancer Classification from Ultrasound Images

Mariusz Rafało
Warsaw School of Economics
Warsaw, Poland *mrafalo@sgh.waw.pl*

Agnieszka Żyłka
Maria Skłodowska-Curie National Research Institute of Oncology
Warsaw, Poland *agnieszka.zylka.edu@gmail.com*

Abstract

We present a novel approach for identifying persistently misclassified images in real-world thyroid ultrasound data. Using 484 images of thyroid nodules, we evaluated four different convolutional neural network architectures. Persistent misclassification is defined as images repeatedly misclassified across models and cross-validation folds. These cases are validated by an experienced radiologist and subjected to Grad-CAM analysis. Results confirm that images, that have negative impact on model results, often exhibit atypical or ambiguous features. We emphasize that persistent misclassification is an important source of diagnostic error, independent of model choice. Recognizing misleading cases is crucial for dataset quality, model robustness and the trustworthiness of AI systems in clinical applications. This work highlights the need for incorporation data validation strategies alongside standard performance metrics in the development of deep learning models.

Keywords: persistent misclassification; thyroid cancer; deep learning; convolutional neural networks; Grad-CAM; medical imaging; cross-validation, impactful images, ultrasound data.

1. Introduction

Modern medical diagnostics increasingly rely on statistical analysis and, more recently, on AI and machine learning techniques. In the field of medical image diagnostics, the commonly used techniques are deep neural networks, in particular algorithms designed for image analysis - convolutional neural networks (CNN). CNN are a class of deep learning models designed to process data with a grid-like topology. Convolutional layers are used to capture spatial hierarchies and detect important image features. CNN models have been widely adopted in a variety of image and video-related tasks, including object detection, scene recognition and image classification. The introduction of increasingly sophisticated neural network architectures has consistently resulted in enhanced performance outcomes [23]. Deep learning has been widely adopted by researchers for diagnosing a broad spectrum of diseases, such as breast cancer [1], melanoma and lung cancer [2]. Among deep learning models, convolutional neural networks CNN have become a standard approach in medical image analysis. Numerous studies have validated the effectiveness of CNN-based methods, which have been successfully applied to the analysis of MRI [4], X-ray, CT scans [30], ultrasound images [13], [20], skin cancer [15], brain tumors [9] and liver lesions [21].

However CNN models can be significantly affected by certain images within a dataset. Some images may be inherently ambiguous, mislabeled or simply too difficult to classify consistently, even for state-of-the-art models. These problematic samples can erode overall model performance by introducing noise during training or bias during evaluation. A question arises: what if we could systematically identify and exclude such images from training and evaluation process? Removing them from the training set may improve generalization and reduce noise. Understanding how each individual image impacts model performance can help to locate these misleading samples.

One of the most popular approaches for assessing model performance is cross-validation. This technique ensures that the data used for testing are separate from those used in model training. A machine learning model learns patterns from a training set and is then evaluated on a distinct test dataset, which contains known classification labels. While cross-validation is a widely used method for model evaluation, it has limitations in medical imaging [25]. The assumption of independent and identically distributed samples is often violated due to correlated images from the same patient or similar acquisition conditions, leading to biased results. Additionally, feature imbalance within images can cause uneven representation across folds. In small datasets, common in rare diseases or specialized imaging, cross-validation can produce high variance and hinder deep learning models from learning stable, generalizable patterns.

The goal of this paper is to propose a novel, systematic approach for identifying images that negatively influence the performance of thyroid cancer classification models during cross-validation. We focus on detecting images that contain feature patterns which, when included in the training process, consistently lead to poor predictive outcomes. These so-called harmful images may either introduce noise, reflect underrepresented classes or embody ambiguous visual characteristics that confuse the learning process. By identifying and analyzing these images, we can to improve model robustness and enhance the reliability of the classification.

We introduce a concept of persistent misclassification analysis. For each image in the dataset, we track its classification outcomes across multiple cross-validation runs and different CNN architectures. We identify images that are repeatedly misclassified across runs and label them as systematically misclassified or harmful. For each model and iteration, we record how often each image appears in the training and test sets, whether it was correctly classified and which ultrasound features it exhibit. This approach allows us to analyze the consistency of model errors and explore how excluding problematic images improves classification performance.

We focus on the classification of thyroid cancer using ultrasound images, which are commonly employed in clinical practice for the evaluation of thyroid nodules. Distinguishing between malignant and benign thyroid nodule is crucial, because accurate identification not only avoids over treatment but also facilitates timely cancer detection [10]. A major challenge in this diagnostic process is the identification of robust features such as echogenicity or irregular shape as well as ensuring the appropriate quality of images.

Once misleading images have been identified, we use the proven Grad-CAM algorithm to investigate them. Grad-CAM was originally introduced by Selvaraju et al. [17] to provide visual explanations for CNN-based decisions. The algorithm is able to highlight regions of input images that influence predictions most strongly. It has been adopted in medical imaging for interpretability, particularly in domains where understanding model reasoning is important [26].

2. Related work and motivation

Recent studies have demonstrated that CNN can achieve high accuracy in differentiating benign from malignant thyroid nodules, often exceeding 80% accuracy and reaching area under the curve (AUC) values of up to 85% [8]. Several well-known CNN architectures have been applied to thyroid nodule classification tasks, including models from the VGG family [7], ResNet [6], InceptionNet [20] and DenseNet architectures [24].

These studies have explored multiple CNN architectures and methods for robust training. However, less attention has been paid to identifying specific training samples that harm model generalization. This issue is visible when trying to recreate the results of previous studies by running CNN models on other data sets. The low quality of such models results from the large dependence of the model parameters on the specific features of the training images (also including artifacts or low quality images). In these models, certain training samples (images) can negatively influence the model by introducing noise, spurious correlations or ambiguous patterns that degrade generalization performance. Previous studies confirm the existence of these limitations. E.g. Recht et al. [14] demonstrated that model performance can degrade on new test sets drawn from the same distribution, emphasizing the sensitivity of deep learning models. Zech et al. [27] showed that CNNs trained on chest X-ray data generalized poorly across institutions, partly due to confounding variables. Ribeiro et al. [16] proposed LIME for model interpretability, indirectly supporting case-level error analysis. Oakden-Rayner [12] highlighted how high-performing models can still fail on edge cases due to spurious correlations.

Several studies have explored methods for detecting and mitigating the effects of such impactful data. Influence functions, as introduced by Koh and Liang [5], offer a theoretical framework to estimate the impact of individual training points on model predictions. This approach has been used to trace mispredictions back to specific harmful training examples. However, influence functions are computationally expensive for large models like CNN and may not scale well with high-dimensional image data. Moreover, Toneva et al. [22] proposed the concept of forgetting events during training as a way to identify potentially noisy or unlearnable examples. They found that some data points are repeatedly forgotten across training epochs, and that removing such samples can improve model performance. This is particularly relevant in our context, where persistently misclassified images may reflect similar underlying issues.

Moreover, recent studies have demonstrated the utility of Grad-CAM method in analyzing correct and incorrect classifications [28], [19]. In thyroid ultrasound analysis, Grad-CAM is used to identify misleading images that may cause erroneous model decisions, by revealing whether the network focuses on medically irrelevant features [31]. Furthermore, by visualizing activation regions, Grad-CAM can assess how CNN behave on specific cases, helping to understand model biases [18].

Another line of research focuses on memorization in deep networks. Zhang et al. [29] demonstrated that deep neural networks are capable of memorizing random labels, raising concerns about their ability to generalize when trained on datasets with ambiguous samples.

In medical imaging Oakden-Rayner et al. [12] emphasized the risks of hidden stratification—where model performance appears high on aggregate metrics but fails on clinically important subgroups. This phenomenon can be driven by underrepresented samples that systematically bias the model. Similarly, Northcutt et al. [11] developed confident learning, a technique for detecting label errors in datasets, which is especially useful in domains with subjective labeling such as histopathology or ultrasound.

These studies collectively highlight the critical need for methods that can identify and manage misleading training data. Our work builds on this foundation by applying a concept of persistent misclassification analysis across CNN models and cross-validation iterations.

3. Method

This study is a retrospective analysis of US images of focal thyroid lesions in patients of Maria Skłodowska-Curie National Research Institute of Oncology in Warsaw (Poland). The dataset consist of 270 thyroid nodules from 270 patients who underwent diagnostic ultrasound examinations. Nodules with initial non-diagnostic or indeterminate cytologic findings that were not confirmed by subsequent histologic evaluation were excluded from the dataset. After this removal, the dataset consists of 242 nodules from 242 patients, resulting in a total of 484 ultrasound images.

Each nodule is captured in two orthogonal planes, providing paired images of the same focal change. Nodules are classified into two categories: malignant or benign (based on histopathological examination). Furthermore all images are manually reviewed by an experienced diagnostician. The expert evaluated each image for 20 known sonographic features that affect nodule classification, including: echogenicity (the relative brightness of the nodule compared to surrounding tissue), shape (round, oval, or irregular forms), borders (whether the edges were well-defined or blurred), presence of calcification, cystic components, shadowing artifacts, etc. This analysis of US features helped correlate image-level classification errors with clinically relevant visual patterns, providing insight into the limitations of CNN models.

Images were then cropped and resized to a fixed resolution of 140×140 pixels and 3 color channels, to ensure consistency across the dataset and compatibility with standard CNN architectures.

To assess the effect of individual images on model performance, we constructed a diverse set of 4 convolutional neural network (CNN) models. These included established architectures such as ResNet (ResNet152) DenseNet (DenseNet121) and two relatively simple (in terms of number of convolution layers) custom architectures designed and trained from scratch (cnn1 and cnn2). The custom models varied in depth and convolutional structure, as well as number of trainable parameters (132775 in cnn1 and 314230 in cnn2 respectively).

Table 1. Mean, minimum, and maximum AUC scores for selected models

Model	**Conv. layers**	**AUC (mean) ± std. dev.**	**AUC (min)**	**AUC (max)**
ResNet152	151	0.81 ± 0.05	0.76	0.93
cnn1	6	0.80 ± 0.03	0.74	0.86
cnn2	7	0.83 ± 0.04	0.74	0.90
denseNet121	120	0.86 ± 0.03	0.80	0.91

For each model, we performed 100 randomized dataset splits into training (80%), validation (10%), and test (10%) subsets. We performed 100 iterations to increase the variability of the data division into training and test sets as much as possible. The splits were stratified to preserve target class distribution across folds. During each training run, model predictions were logged for all test samples, along with the cut-off threshold values. This setup enabled detailed tracking of individual image behavior across multiple architectures and training scenarios. Aggregated AUC results of these iterations are presented in table 1

The framework was implemented in Python programming language. The source code and detailed architecture of cnn1 and cnn2 models are published in the author's public GitHub repository[1].

We define a *persistent misclassification* as the phenomenon where a given image is misclassified in a high proportion of cross-validation iterations, despite variation in model initialization, architecture, and data splits. We focus on images misclassified in more than 90% (we assume threshold $\theta = 0.9$) of 100 cross-validation iterations across 4 CNN model architectures.

For each image in the dataset, we track its classification outcome across cross-validation iterations for each model. We focus on test set predictions but we control train and validation dataset structure as well. For each image j, we define its misclassification frequency F_j as: $F_j = \frac{E_j}{O_j}$, where E_j is a number of incorrect classifications of image j and O_j is a total number of times image j appeared in the test dataset. Images with $F_j > \theta$ are labeled as systematically misclassified (harmful) images. For each CNN model and for each iteration, we record a number of times each image appeared in the training and test datasets. Then, we log a number of times the image was correctly classified. Moreover, we record US image features described by an experienced diagnostician, along with an indication of the percentage of nodules that have a given feature (in the entire data set and in the training set). In the experiment, we used the κ value as the threshold for improving AUC after excluding impactful images. Initially, we used

[1] https://github.com/mrafalo/persistent-misclassification

$\kappa = 0.05$. If in a given iteration of the algorithm we do not have adequate AUC improvement, then we slightly decrease (by 0.01) θ value to reduce the rigor of the image search. The persistent misclassification evaluation algorithm is presented in the algorithm diagram 1.

Algorithm 1 Image misclassification evaluation algorithm

```
Set $\theta = 0.9$
Set $\kappa = 0.05$
for each model $M$ in model list do
    for each iteration $i = 1$ to $100$ do
        Randomly split dataset into train (80%), validation (10%) and test (10%) datasets
        Train model $M$ on training set
        Validate model on validation set
        for each image $k$ from test dataset do
            Predict labels on image $k$
            Save model prediction $p$, actual value $a$ and threshold value $t$
        end for
    end for
    Compute mean AUC before exclusion, denote as $AUC_{\text{before}}$
end for
for each image $j$ in dataset do
    Count number appearances in test dataset: $O_j$
    Count number of incorrect predictions: $E_j$
    Compute misclassification frequency: $F_j = \frac{E_j}{O_j}$
end for
Flag images where $F_j > \theta$
Recompute AUC after excluding flagged images, denote as $AUC_{\text{after}}$
Compute improvement: $\Delta AUC = AUC_{\text{after}} - AUC_{\text{before}}$
if $\Delta AUC < \kappa$ then
    Decrease $\theta$ by $0.01$
    Run the function again
end if
```

In addition to misclassification frequency, we employ Grad-CAM technique to explore the underlying causes. Grad-CAM is applied to visualize the regions influencing model predictions. We use Grad-CAM to assess whether models attend to clinically relevant features, as opposed to artefactual or background regions. This method enables the identification of cases where the model focuses on artifacts or irrelevant structures, which can compromise diagnostic reliability. Highly misclassified samples are further analyzed through experienced diagnostician re-inspection to identify potential sources of error, including annotation ambiguity, atypical histopathology or suboptimal image quality. This comparison allows for a systematic assessment of whether model failures are attributable to model limitations or underlying data issues.

4. Persistent Misclassification Analysis

Through a systematic analysis across cross-validation iterations and 4 CNN architectures, we identified a subset of 25 images that were consistently misclassified with high frequency. These images, which exhibited misclassification rates exceeding 90%, were labeled as persistently misclassified cases. To assess their impact on overall model performance, we conducted an experiment in which these problematic samples were excluded from the training process. The exclusion resulted in a measurable improvement in model performance. Figure 1 illustrates the mean AUC of 100 iterations for CNN architectures (ResNet152, DenseNet121, cnn1 and cnn2) as a function of the number of excluded images identified as harmful. The exclusion of systematically misclassified images results in a consistent improvement in model performance across all architectures. DenseNet121 and ResNet152, which achieved the highest baseline AUC values, exhibited the most significant gains, reaching approximately 0.87 and 0.865 respectively. Although cnn1 and cnn2 demonstrated performance improvements, their final AUC values remained lower relative to the more advanced models.

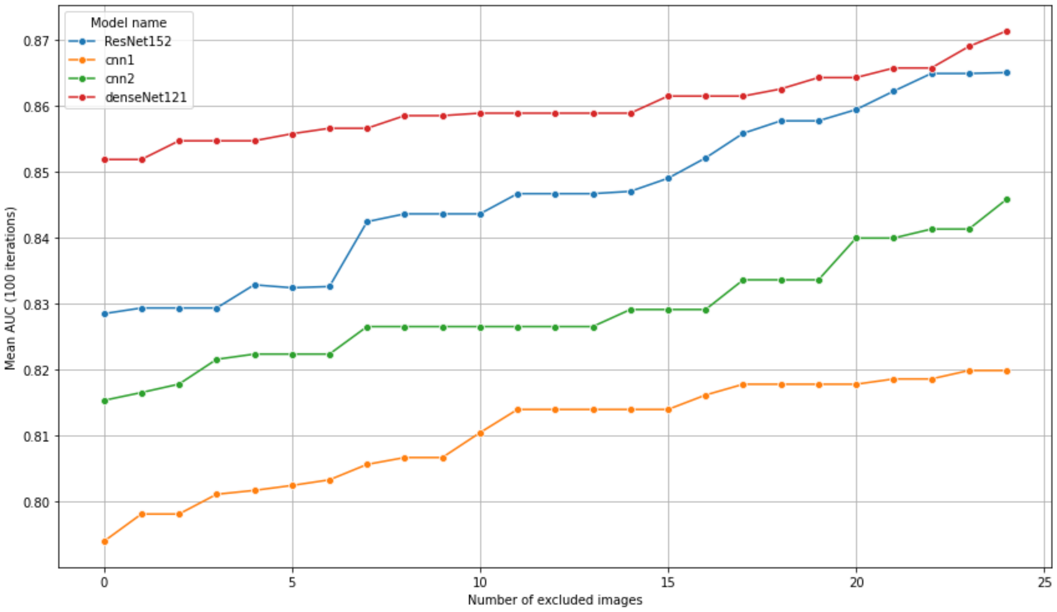


Fig. 1. AUC improvement by excluding persistently misclassified images

Table 2 summarizes the most common features identified in the misleading images. We identify these features in order to verify how often they are present in the dataset and in the train dataset. The presence of rare features may explain that the image containing this feature is often misclassified. The descriptions of individual thyroid nodules were, as previously mentioned, performed manually by an experienced radiologist. The results demonstrate that these images are characterized by features that are relatively rare within the overall dataset and sometimes even less common within the training dataset.

Two distinct groups of uncommon features were identified. The first group corresponds to rare histological types of thyroid cancer, including follicular thyroid carcinoma (FTC), follicular tumor of uncertain malignant potential (FTUMP), nodular overgrowth, Hurthle cell carcinoma and medullary thyroid carcinoma (MTC). These cancer types are infrequent in clinical practice and thus underrepresented in the dataset. The majority of malignant nodules in the dataset are papillary thyroid carcinoma (74% of all malignant nodules). These rare cancer types are markedly underrepresented in the dataset, consistent with their low incidence in clinical practice. Their scarcity within the training set limits the model's ability to generalize to such cases. Furthermore, the analysis suggests that these rare cancers exhibit imaging characteristics that differ significantly from PTC, which dominates the malignant cases in the dataset. This distinction may contribute to the observed misclassifications.

The second group covers US characteristics of thyroid nodules, taking into account features related to shape, echogenicity, calcification, halo, etc. In addition, the radiological description includes inflammation of the lymph nodes and the thyroid gland itself. The low frequency of these features further suggests that their under-representation may contribute to the difficulty in correct model classification. The analysis of features associated with harmful images is consistent with clinical expectations. For example, the presence of thyroid disorders, such as chronic

Table 2. US feature occurrence grouped by type with summarized number of images

Group	Feature	[%] of dataset	[%] of train dataset	Images
Rare cancer type	FTC	5%	3%	2
	FTUMP	6%	6%	1
	Nodular overgrowth	15%	12%	4
	Hurthle	3%	2%	1
	MTC	5%	2%	3
	Total			**11**
US features	Isoechoic	18%	19%	1
	Smooth margins	40%	39%	1
	Halo	5%	4%	1
	Round shape	2%	2%	3
	Isthmus	5%	5%	1
	Capsular invasion	2%	1%	1
	Pat. lymph nodes	8%	7%	3
	Macrocalcifications	13%	12%	1
	Thyroid disorder	26%	24%	2
	Total			**14**

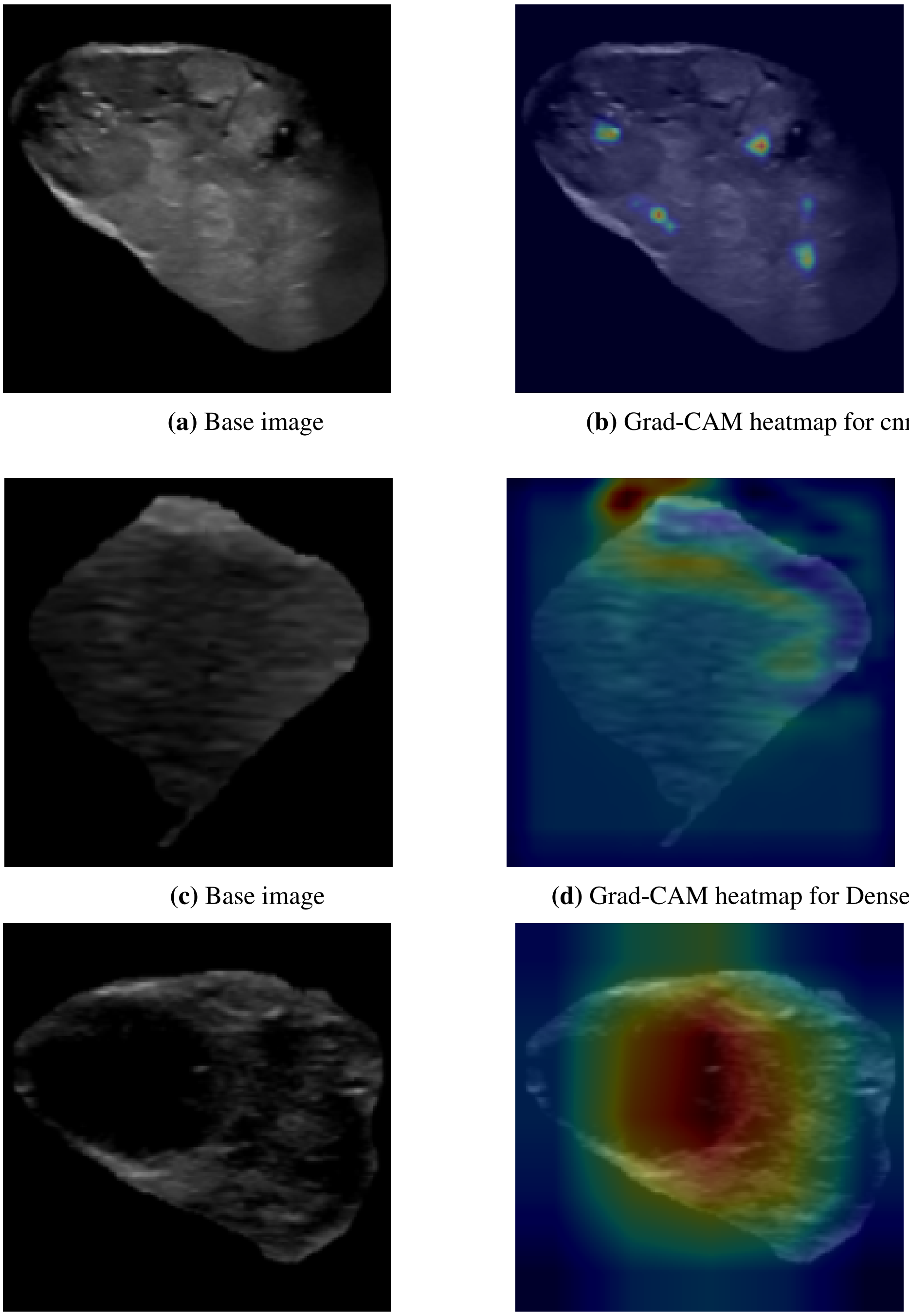
lymphocytosis thyroiditis (Hashimoto's thyroiditis), significantly impairs the ability to accurately identify malignancy in ultrasound images [3].

The conclusions were drawn following in-depth manual radiological assessment of each ultrasound image. Such detailed evaluation is challenging or even impossible when working with multiple models and large volumes of US images. Therefore, performing a persistent misclassification analysis is essential to systematically identify patterns of recurrent errors.

In the second stage of the analysis, Grad-CAM was applied to verify how individual CNN models focus on specific image areas. This approach allows for the assessment of whether the models attend to clinically relevant regions, thereby providing insight into model behavior and the reliability of predictions. Moreover, this analysis can be performed automatically in each cross-validation iteration, without the need for manual work or expert evaluation.

Figure 2 shows examples of misleading images of thyroid nodules. Images on the left side are the baseline images (these images were fed into models and classified). Images on the right are baseline images overlaid with a Grad-CAM heatmap derived from the final convolutional layer of the CNN model. The heatmap 2b highlights regions the network considers important for its classification. In this case, the model wrongly focuses on image artifacts rather than on the nodule itself. These artifacts, visible as small concentrated bright spots, are unrelated to pathological structures and suggest poor feature localization, which leads to an incorrect classification. The heatmap 2d shows that the model fails to accurately detect the shape and borders of the nodule. Instead, the model's attention is dispersed across surrounding tissue or irrelevant structures. This lack of focus on the lesion boundaries contributes directly to misclassification. Similar behavior can be observed on heatmap 2f where the dark image caused difficulty in identifying nodule boundaries.

In turn, Figure 3 presents the correct classification of the thyroid nodules. Grad-CAM visualizations reveal that the models concentrated their attention on specific morphological features within the nodules, particularly areas of altered echotexture and internal irregularities. This focused activation (heatmaps 3b, 3d and 3f)suggests that the networks learned clinically relevant features, rather than relying on artefactual patterns.



(a) Base image

(b) Grad-CAM heatmap for cnn1

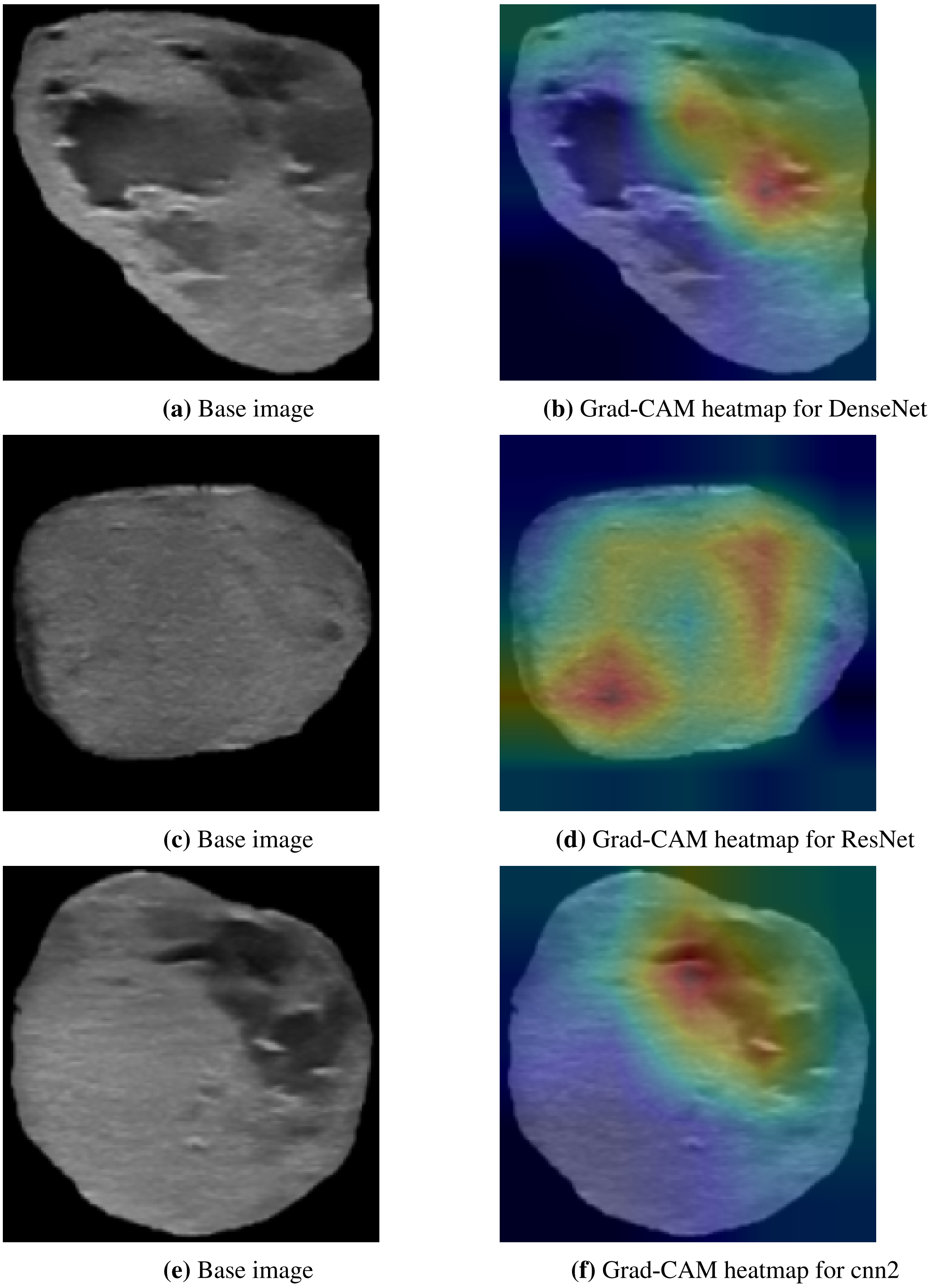
(c) Base image

(d) Grad-CAM heatmap for DenseNet

(e) Base image

(f) Grad-CAM heatmap for ResNet

Fig. 2. Example of misclassified images: base US image (left) and Grad-CAM heatmap (right)



(a) Base image

(b) Grad-CAM heatmap for DenseNet

(c) Base image

(d) Grad-CAM heatmap for ResNet

(e) Base image

(f) Grad-CAM heatmap for cnn2

Fig. 3. Example of properly classified images: base US image (left) and Grad-CAM heatmap (right)

5. Summary

We propose the concept of persistent misclassification analysis as a structured diagnostic tool during CNN training for medical image analysis. Using the algorithm we identified persistently misclassified samples. These were confirmed by experienced radiologists and validated using Grad-CAM visualization. The models consistently failed on these cases across different architectures, indicating that the observed errors were not incidental but reflected broader limitations in data representation and model learning. Grad-CAM analysis revealed that the networks often focused on non-diagnostic regions, such as image artifacts or surrounding tissue, rather than the lesion itself.

To characterize these misleading cases further, we examined low-level US features. The analysis shows that these images contained features that were either rare or poorly represented in the training set, limiting the models' ability to generalize. Also, the types of malignant thyroid cancer that were misclassified were rare clinical events, underrepresented in the dataset.

This analysis is important for two key reasons. First, model auditing: average metrics such as accuracy, AUC and F1-score can obscure critical failure cases that persist across experimental variations. Second, dataset quality assessment: systematically misclassified images may indicate labeling errors, low-quality US scans or feature representations not well captured by the models.

Our findings have practical implications for model development process. Persistent misclassification analysis can serve as an early diagnostic step in the training process, helping to identify problematic data. Moreover, Grad-CAM-based interpretability supports error auditing by allowing practitioners to assess whether model decisions align with known diagnostic criteria.

This study is limited by the scope of the dataset, which, while well-curated, is relatively small and imbalanced in terms of cancer type representation and US features representation. Rare cancer types such as MTC or Hurthle cell carcinoma were underrepresented, which constrained the model's ability to learn their imaging characteristics.

Future work should explore systematic integration of persistent misclassification analysis into CNN training workflows, possibly through automated detection and targeted data exclusion. The results of Grad-CAM analysis can be automatically run to flag images deemed to be misleading.

References

[1] Ayana, G., Park, J., and Choe, S. W.: Patchless Multi-Stage Transfer Learning for Improved Mammographic Breast Mass Classification. In: *Cancers* 14.5 (2022).

[2] Brinker, T. J., Hekler, A., Enk, A. H., Klode, J., Hauschild, A., and Berking, C. et al.: Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. In: *European Journal of Cancer* 113 (2019), pp. 47–54.

[3] Jeon, E. J., Jeong, Y. J., Park, S. H., Cho, C. H., Shon, H. S., and Jung, E. D.: Ultrasonographic Characteristics of the Follicular Variant Papillary Thyroid Cancer According to the Tumor Size. In: *Journal of Korean Medical Science* 31.3 (2016), pp. 397–402. URL: https://doi.org/10.3346/jkms.2016.31.3.397.

[4] Jiang, Z., Salcudean, S. E., and Navab, N.: Robotic ultrasound imaging: State-of-the-art and future perspectives. In: *Medical Image Analysis* 89.July (2023), p. 102878. arXiv: 2307.05545. URL: https://doi.org/10.1016/j.media.2023.102878.

[5] Koh, P. W. and Liang, P.: Understanding Black-box Predictions via Influence Functions. In: *Proceedings of the 34th International Conference on Machine Learning (ICML)* (2017), pp. 1885–1894.

[6] Kunapinun, A., Dailey, M. N., Songsaeng, D., Parnichkun, M., Keatmanee, C., and Ekpanyapong, M.: Improving GAN Learning Dynamics for Thyroid Nodule Segmentation. In: *Ultrasound in Medicine and Biology* 49.2 (2023), pp. 416–430.

[7] Lee, J., Kim, Y., Ahn, Y., Park, S., et al.: Investigation of optimal convolutional neural network conditions for thyroid ultrasound image analysis. In: *Sci. Rep.* 13 (2023), pp. 1–9.

[8] Li, X., Zhang, S., Zhang, Q., Wei, X., Pan, Y., Zhao, J., Xin, X., Qin, C., Wang, X., Li, J., Yang, F., Zhao, Y., Yang, M., Wang, Q., Zheng, Z., Zheng, X., Yang, X., Whitlow, C. T., Gurcan, M. N., Zhang, L., Wang, X., Pasche, B. C., Gao, M., Zhang, W., and Chen, K.: Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. In: *The Lancet Oncology* 20.2 (2019), pp. 193–201. URL: http://dx.doi.org/10.1016/S1470-2045(18)30762-9.

[9] Liu, M., Zhang, J., Adeli, E., and Shen, D.: Deep multi-task multi-channel learning for joint classification and regression of brain status. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Cham: Springer, 2017, pp. 3–11.

[10] Moon, J. H., Kim, Y. I., Lim, J. A., Choi, H. S., and Cho, S. W.: Thyroglobulin in washout fluid from lymph node fine-needle aspiration biopsy in papillary thyroid cancer: large-scale validation of the cutoff value to determine malignancy and evaluation of discrepant results. In: *The Journal of Clinical Endocrinology and Metabolism* 98.3 (2013).

[11] Northcutt, C., Jiang, L., Chuang, J., Sagawa, S., Saunshi, N., and Lipton, Z. C.: Confident Learning: Estimating Uncertainty in Dataset Labels. In: *Journal of Artificial Intelligence Research* 70 (2021), pp. 1373–1411.

[12] Oakden-Rayner, L., Beam, A. L., and Palmer, L. J.: *Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging*. https://arxiv.org/abs/1909.12475. arXiv preprint arXiv:1909.12475. 2020.

[13] Peng, S., Liu, Y., Lv, W., Liu, L., Zhou, Q., Yang, H., et al.: Deep learning-based artificial intelligence model to assist thyroid nodule diagnosis and management: a multicentre diagnostic study. In: *Lancet Digit. Heal.* 3.4 (2021), e250–e259.

[14] Recht, B., Roelofs, R., Schmidt, L., and Shankar, V.: Do ImageNet Classifiers Generalize to ImageNet? In: *International Conference on Machine Learning (ICML)* (2019), pp. 5389–5400.

[15] Rguibi, Z., Hajami, A., Zitouni, D., Elqaraqoui, A., and Bedraoui, A.: CXAI: Explaining Convolutional Neural Networks for Medical. In: *Electronics* 11.11 (2022), pp. 1775–1794.

[16] Ribeiro, M. T., Singh, S., and 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*. ACM. 2016, pp. 1135–1144.

[17] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2017), pp. 618–626.

[18] Singh, A., Sengupta, S., and Lakshminarayanan, V.: *Explainable deep learning models in medical image analysis*. 2020. arXiv: 2005.13799 [cs.CV]. URL: https://arxiv.org/abs/2005.13799.

[19] Song, D., Yao, J., Jiang, Y., Shi, S., Cui, C., Wang, L., Wang, L., Wu, H., Tian, H., Ye, X., Ou, D., Li, W., Feng, N., Pan, W., Song, M., Xu, J., Xu, D., Wu, L., and Dong, F.: A new xAI framework with feature explainability for tumors decision-making in Ultrasound data: comparing with Grad-CAM. In: *Computer Methods and Programs in Biomedicine* 235 (2023), p. 107527. URL: https://www.sciencedirect.com/science/article/pii/S016926072300192X.

[20] Sorrenti, S., Dolcetti, V., Radzina, M., Bellini, M., Frezza, F., Munir, K., et al.: Artificial Intelligence for Thyroid Nodule Characterization: Where Are We Standing? In: *Cancers (Basel)* 14.14 (2022), pp. 1–15.

[21] Todoroki, Y., Iwamoto, Y., Lin, L., Hu, H., and Chen, Y.: Automatic Detection of Focal Liver Lesions in Multi-phase CT Images Using A Multi-channel & Multi-scale CNN. In: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2019, pp. 872–875.

[22] Toneva, M., Sordoni, A., Tarlow, D., Kiros, J., Furlanello, T., and Bengio, Y.: An Empirical Study of Example Forgetting During Deep Neural Network Learning. In: *International Conference on Learning Representations (ICLR)*. 2019.

[23] Wang, Y., Guan, Q., and Lao I., e. a.: Using deep convolutional neural networks for multi-classification of thyroid tumor by histopathology: a large-scale pilot study. In: *Annals of Translational Medicine* 7.18 (2019).

[24] Wei, X., Gao, M., Yu, R., Liu, Z., Gu, Q., Liu, X., Zheng, Z., Zheng, X., Zhu, J., and Zhang, S.: Ensemble Deep Learning Model for Multicenter Classification of Thyroid Nodules on Ultrasound Images. In: *Med. Sci. Monit.* 26 (2020), e926096.

[25] Xu, Q. S. and Liang, Y. Z.: Monte Carlo Cross Validation. In: *Chemometrics and Intelligent Laboratory Systems* 56.1 (2001), pp. 1–11.

[26] Xu, Y., Xu, M., Geng, Z., et al.: Thyroid nodule classification in ultrasound imaging using deep transfer learning. In: *BMC Cancer* 25 (2025), p. 544.

[27] Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., and Oermann, E. K.: Variable Generalization Performance of a Deep Learning Model to Detect Pneumonia in Chest Radiographs: A Cross-Sectional Study. In: *PLoS Medicine* 15.11 (2018), e1002683.

[28] Zhang, C., Liu, D., Huang, L., Zhao, Y., Chen, L., and Guo, Y.: Classification of Thyroid Nodules by Using Deep Learning Radiomics Based on Ultrasound Dynamic Video. In: *Journal of Ultrasound in Medicine* 41.12 (2022), pp. 2993–3002.

[29] Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O.: Understanding deep learning requires rethinking generalization. In: *International Conference on Learning Representations (ICLR)* (2017).

[30] Zhao, H., Liu, C., Ye, J., Chang, L., Xu, Q., Shi, B., et al.: A comparison between deep learning convolutional neural networks and radiologists in the differentiation of benign and malignant thyroid nodules on CT images. In: *Endokrynol. Pol.* 72.3 (2021), pp. 217–225.

[31] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A.: Learning Deep Features for Discriminative Localization. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 2921–2929.