USEFUL NONROBUST FEATURES ARE UBIQUITOUS IN BIOMEDICAL IMAGES

Coenraad Mouton[⋆†] *Randle Rabe*[†] *Niklas C Koser*[⋆]
Nicolai Krekiehn[⋆] *Christopher Hansen*[⋆] *Jan-Bernd Hövener*[⋆] *Claus-C. Glüer*[⋆]

[⋆] Section Biomedical Imaging, Department of Radiology and Neuroradiology,
University Hospital Schleswig-Holstein, Kiel University
[†] Faculty of Engineering, North-West University, South Africa

2026 IEEE 23rd International Symposium on Biomedical Imaging (ISBI) | 979-8-3315-7763-6/26/$31.00 ©2026 IEEE | DOI: 10.1109/ISBI61048.2026.11515306

ABSTRACT

We study whether deep networks for medical imaging learn useful nonrobust features - predictive input patterns that are not human interpretable and highly susceptible to small adversarial perturbations - and how these features impact test performance. We show that models trained only on nonrobust features achieve well-above-chance accuracy across five MedMNIST classification tasks, confirming their predictive value in-distribution. Conversely, adversarially trained models that primarily rely on robust features sacrifice in-distribution accuracy but yield markedly better performance under controlled distribution shifts (MedMNIST-C). Overall, nonrobust features boost standard accuracy yet degrade out-of-distribution performance, revealing a practical robustness-accuracy trade-off in medical imaging classification tasks that should be tailored to the requirements of the deployment setting.

***Index Terms*—** Nonrobust Features, Adversarial examples, Adversarial Robustness, MedMNIST

1. INTRODUCTION

Deep neural networks (DNNs) are increasingly being deployed in medical imaging domains such as radiology, digital pathology, and ophthalmology. Given the high-stakes of these application domains, it is critical that these models are reliable, trustworthy and (ideally) explainable. Despite this, prior work on natural images has shown that neural networks learn *nonrobust features*: input patterns which are extremely susceptible to small perturbations and completely uninterpretable to humans, yet highly predictive [1, 2]. It is further shown that the existence of adversarial examples - tiny, virtually imperceptible perturbations to images that reliably cause misclassifications - can be ascribed (at least in part) to a model's reliance on nonrobust features (see Fig. 1). Such features are naturally troublesome should they also occur in the medical domain: not only would they hinder any attempts at model interpretation [3] and increase susceptibility to adversarial examples, but one must then also consider whether a user should *trust* a prediction that was made based on uninterpretable and fragile patterns.

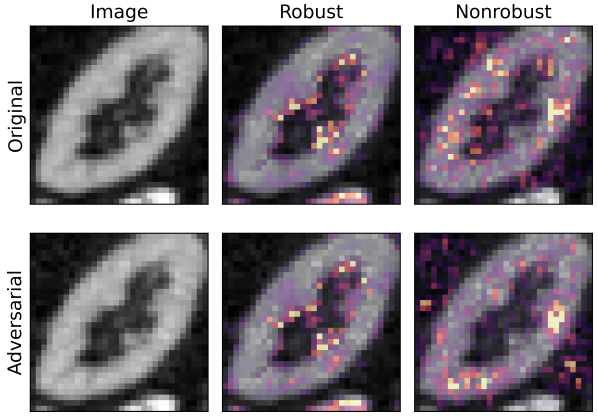

Fig. 1. Saliency map comparison between robust and nonrobust organ-classification models for a right-kidney CT slice (top) and a slightly perturbed adversarial version (bottom). The robust model's attributions remain stable between the original and adversarial images and concentrate on kidney-relevant anatomy, whereas the nonrobust model's attributions change markedly under perturbation and appear almost noise-like. Overlays are normalized to $[0, 1]$; brighter colors indicate larger attribution.

To approach this matter, we consider two fundamental questions: (1) Do DNNs learn useful, well-generalizing nonrobust features from various biomedical imaging modalities, and (2) How essential are such nonrobust features for high performance.

Accordingly, we consider the role that robust and nonrobust features play in several classification tasks spanning diverse imaging modalities such as computed tomography (CT), radiography, ultrasound, and histopathology images. Furthermore, we conduct our investigation in both an in-distribution and an out-of-distribution (OOD) setting.

Our primary contribution is that we provide the first systematic investigation of the generalization properties of robust and nonrobust features in medical image classification tasks.

2. ROBUST AND NONROBUST FEATURES

In this section we introduce key notation and formalize robust and nonrobust features. Let f denote a deep neural network (DNN) with parameters θ trained on a dataset of sample-label pairs $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ using a suitable loss function $\mathcal{L}$. We define a *useful feature* learned by f for a class c as a function $g_c : \mathbb{R}^d \to \mathbb{R}$ which is correlated with $y = c$ under $\mathcal{D}$. Intuitively, a feature can be considered a single learned property or concept of the input data, for example 'cat ears' or 'human face' which aids in classification for a given class c.

We then further distinguish between *robust* and *nonrobust* features. Consider an adversarial perturbation $\boldsymbol{\delta}$ such that

$$\boldsymbol{\delta}^\star = \underset{\|\boldsymbol{\delta}\|_\infty \leq \epsilon}{\arg\max}\, \mathcal{L}\left(y, f_\theta(\mathbf{x} + \boldsymbol{\delta})\right) \tag{1}$$

where ϵ is some small perturbation bound, for example $\frac{4}{255}$ is commonly used for image data [4]. A feature g_c is then defined as *robust* if $g_c(\mathbf{x} + \boldsymbol{\delta}^\star)$ remains correlated with c, but *nonrobust* if $g_c(\mathbf{x}+\boldsymbol{\delta}^\star)$ is instead correlated with some $j \neq c$. See [1] for formal definitions.

More intuitively, we aim to convey that features are robust if they remain readily recognizable and correlated with a class label after some small adversarial perturbation. For example, a human face is still easily identifiable after slight distortion of the underlying pixel values. Conversely, nonrobust features are easily switched from being correlated with one class to another by tiny perturbations. A side-effect of this brittle nature is that these features are generally not recognizable, or interpretable, by humans, and therefore appear as noise-like patterns when isolated in the image domain.

To illustrate, Fig. 1 compares the saliency (gradient of the predicted class score w.r.t. the input) of an adversarially robust and an adversarially nonrobust classification model trained on OrganSMNIST [5]. This is done for a test image and an adversarially perturbed version that is misclassified by the nonrobust model. The visual contrast between these saliency maps mirrors our definitions: robust features are stable and more anatomically meaningful, whereas nonrobust features are brittle and easily altered by small perturbations.

3. EXPERIMENTAL SETUP

In this section we describe the datasets, models, and procedures used to isolate and evaluate robust vs. nonrobust features.

3.1. Isolating nonrobust features

Our first avenue of inquiry concerns determining whether well generalizing nonrobust features can be found in the various medical imaging datasets we consider. More specifically, we aim to establish whether nonrobust features alone suffice for non-trivial performance (better than random guessing) on these classification tasks.

To investigate this, we borrow an experimental setup from Ilyas et al. [1] and proceed as follows. Consider a medical imaging dataset $\mathcal{D}$, split into two subsets, $\mathcal{D}^{\text{train}}$ and $\mathcal{D}^{\text{test}}$. Furthermore, let f correspond to a DNN trained using $\mathcal{D}^{\text{train}}$ such that it performs well on $\mathcal{D}^{\text{test}}$. This model should presumably rely on both robust and (should they exist) nonrobust features in order to make predictions.

Our goal is to construct a new dataset, $\hat{\mathcal{D}}^{\text{train}}$, where the only correlation between the labels and inputs is the nonrobust features learned by f. We construct such a dataset as follows. Consider all m samples $(\mathbf{x}, y) \in \mathcal{D}^{\text{train}}$ that are correctly classified by f such that $f(\mathbf{x}) = y$. For each of these samples, we randomly select a target class t and then create an adversarial example $\hat{\mathbf{x}} = \mathbf{x} + \boldsymbol{\delta}$ (per equation 1) such that $f(\hat{\mathbf{x}}) = t$. The dataset $\hat{\mathcal{D}}^{\text{train}}$ is then constructed from these adversarial examples and labels, i.e. $\hat{\mathcal{D}}^{\text{train}} = \left\{ (\hat{\mathbf{x}}_i,\ t_i) \right\}_{i=1}^m$. We then train a new classifier on $\hat{\mathcal{D}}^{\text{train}}$ and evaluate it on the original, unadulterated test set $\mathcal{D}^{\text{test}}$. This allows us to determine to which extent the nonrobust features learned by f alone can suffice for classification.

More simply, we construct a new training set consisting solely of adversarial examples and assign each example the label of its randomly chosen adversarial target class. Intuitively, this procedure switches the image's nonrobust features from the source class to the target class (as learned by the original classifier) while leaving the robust, more human-interpretable features largely intact. Therefore, the robust features are now uncorrelated with the new label, and the entire dataset appears (to a trained radiologist or specialist) unchanged but mislabeled. The only systematic signal that remains aligned with the (new) train set labels are the switched nonrobust features. Consequently, if a model trained on this dataset achieves non-trivial (better than chance) accuracy on the original test set with the original labels, it must be doing so by exploiting nonrobust features alone.

We perform this experiment using 5 medical imaging classification datasets of distinct imaging modalities from the MedMNIST [5] collection. For each dataset, we (1) train a WRN-16-8 (WideResNet [6]) model on the original dataset, (2) use this trained model to construct a nonrobust feature dataset, then (3) train a new WRN-16-8 model on the constructed dataset, and finally (4) evaluate both models on the original test set. We perform an extensive hyperparameter search for both steps (1) and (3) to ensure that each model is well optimized. Furthermore, we use 32×32 sized images in all cases (resized from the original 64×64) with the predefined train/val/test splits, and we explicitly do not use any pretrained weights, as this would introduce additional features. Finally, we rely on a ℓ_∞ perturbation bound of $\epsilon = \frac{4}{255}$ for step (2), where Equation 1 is approximated using 100-step PGD (projected gradient descent) [7].

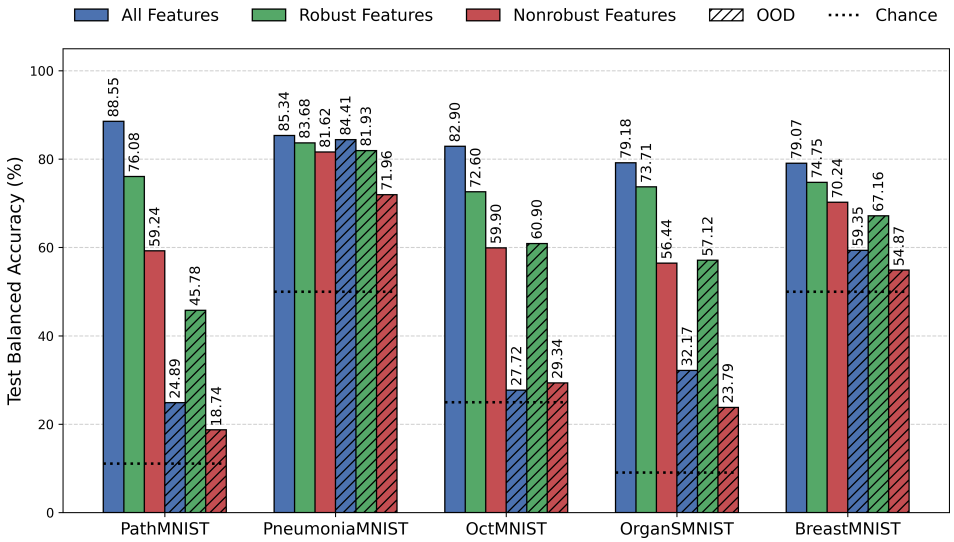

Fig. 2. Test accuracy comparison of DNNs using different features for 5 MedMNIST datasets. Solid bars: in-distribution performance. Dashed bars: average out-of-distribution performance.

3.2. Isolating robust features

We now ask the converse question: How well do models that rely primarily on *robust* features generalize? To this end, we make use of adversarial training [7], implemented via the TRADES [8] loss function:

$$\mathcal{L}_{\text{TR}}(\mathbf{x}, y; \theta) = \ell_{\text{CE}}\big(y, f_\theta(\mathbf{x})\big) + \beta\, \text{KL}\big(f_\theta(\mathbf{x}) \,\|\, f_\theta(\mathbf{x} + \boldsymbol{\delta}^\star)\big) \tag{2}$$

where ℓ_{CE} is the cross-entropy loss, KL is the Kullback-Leibler divergence, and we assume here that the output of f_θ is a probability vector (not logits or a scalar class prediction). The TRADES loss essentially balances the performance on clean data (left term) with the performance on adversarial data (right term) according to a hyperparameter β. This encourages the model to ignore nonrobust cues, and allows us to determine how performance changes as the reliance on nonrobust features decreases.

During training, we incorporate most of the modern techniques that assist adversarial training, such as label smoothing, exponential weight averaging, swish activations, and cosine annealing[1], while early stopping is performed on an adversarial validation set [9, 10, 11]. Furthermore, the adversarial perturbations are crafted using 10-step PGD with $\epsilon = \frac{4}{255}$, and we set $\beta = 5$. As previously done, we perform an extensive hyperparameter search per model to ensure that it is well optimized.

3.3. Evaluating OOD performance

Nonrobust features are defined as 'nonrobust' because they are extremely susceptible to tiny *adversarial* perturbations. However, this does not necessarily imply that these features are brittle to other distribution shifts. To determine this, we compare the degree to which the various models are sensitive to controlled corruptions using the MedMNIST-C test sets [12]. These consist of the original MedMNIST test set altered by various transformations, such as contrast adjustments, Gaussian noise, blurring, JPEG compression, and others. We report on the average performance of each model across these different distribution shifts.

3.4. Evaluating adversarial performance

As a final experiment, we verify that our adversarially trained models are truly more robust to adversarial attacks. We measure this using the strong ensemble method AutoAttack [13] with $\epsilon = \frac{4}{255}$ perturbations.

4. RESULTS

In Fig. 2 we report the in-distribution (solid bars) and OOD performance (dashed bars) across five MedMNIST datasets, comparing base models (trained on all features; blue), robust models (green), and nonrobust-only models (red). Balanced accuracy is used throughout due to class imbalance.

4.1. Comparing in-distribution test performance

When considering the in-distribution results, we observe that in all cases the nonrobust-only models achieve well above chance performance (indicated by the dashed horizontal lines). For example, the nonrobust-only model achieves a balanced accuracy of 59.24% and 81.62% on PathMNIST and PneumoniaMNIST, respectively, which is substantially

[1] We also use these methods when training the base and nonrobust feature models, to ensure they are comparable.

higher than the 11% and 50% that would be achieved by random guessing. This is a strong indication that useful nonrobust features are present and learnable in these datasets.

It is also clear from these results that using all available features consistently yields the best test performance; relying primarily on robust features is second-best; and nonrobust-only performs worst. Furthermore, the performance difference between using all available features and only robust features can be significant, e.g. 88.55% versus 76.08% balanced accuracy for PathMNIST, and similarly 82.90% versus 72.60% for OctMNIST. The implication of this is that the addition of nonrobust features significantly improves performance in the standard classification setting.

4.2. Comparing OOD test performance

With respect to OOD performance, we find that adversarially trained models that primarily rely on robust features provide the best performance (except PneumoniaMNIST). Compared to using all features, the difference can be substantial, e.g. the balanced accuracy increases from 24.89% to 45.78% for PathMNIST and from 27.72% to 60.90% for OctMNIST when the model is adversarially trained.

By contrast, the nonrobust-only models generally mirror the OOD performance drop of the base models, which is often very large. We believe that in both cases this decrease in performance is tied to the models' reliance on nonrobust features. Despite this, we note that in all cases we still observe (marginally) higher than chance performance on the OOD test sets, and in one case strong performance (PneumoniaMNIST).

4.3. Comparing adversarial test performance

We find that all base and nonrobust-only models are highly susceptible to adversarial attacks, with an adversarial accuracy of $\leq 0.04\%$. Conversely, the adversarially trained models achieve between 57% and 74% depending on the dataset, confirming that they are (unsurprisingly) much more adversarially robust (these results are omitted from Fig. 2).

5. DISCUSSION

Let us consider our various observations and revisit the two originally posed questions:

1. **Do DNNs learn useful nonrobust features from medical images?** Yes - nonrobust features alone yield accurate classification on several datasets. While still inferior to robust features, they perform well above chance. By contrast, under OOD shifts, accuracy drops to only marginally above chance, and to zero under adversarial attacks.
2. **How essential are nonrobust features?** In-distribution, discouraging nonrobust features reduces classification performance (by $2-12\%$). Conversely, in the OOD setting, the presence of nonrobust features alongside robust ones lowers accuracy on most datasets (by $8-33\%$), whereas robust features alone are much more invariant to distribution shifts and adversarial attacks.

These answers expose an uncomfortable truth: despite nonrobust features being uninterpretable and extremely vulnerable to both adversarial perturbations and distribution shifts, they are generally very useful in the standard classification setting. This conclusion echoes prior work on natural images which note a trade-off between adversarial robustness and accuracy [3, 8]. It is therefore difficult to determine whether the use of nonrobust features should be encouraged or dissuaded. We believe it depends on the expected deployment setting: if certain distribution shifts are expected and model safety and interpretation is paramount, nonrobust features are to be avoided. Conversely, if absolute in-distribution performance is the goal, they are likely best left included.

Besides this performance assessment, it is also crucial to consider how nonrobust features affect the trustworthiness of an imaging model. i.e. whether a user (e.g. a physician) can reasonably rely on the mechanism behind a prediction. Nonrobust features likely reduce trustworthiness: they are generally uninterpretable and fragile to small, plausible changes, making decisions harder to justify. That said, it is unclear whether users would prefer a more interpretable, shift-robust model over a more accurate one. A promising avenue for future work is to assess, in realistic clinical settings, which trade-off clinicians actually prefer for different use cases.

Despite our strong empirical results, we must also note several limitations. While we have compared OOD performance, we focus on artificial corruptions and believe that an evaluation on natural distribution shifts (e.g. independently sourced data) would also be insightful. Similarly, we did not explicitly train models to be robust to the MedMNIST-C shifts (e.g. with data augmentation), which could potentially shrink the gap between robust and nonrobust features. Furthermore, we solely focus on CNNs (WRN-16-8) trained on two-dimensional low resolution images. Although we expect similar trends for modern vision transformers and higher-resolution or three-dimensional imaging, this remains to be verified. Finally, we confine our study to the task of classification and we believe that an extension to segmentation or detection settings would also prove valuable.

6. CONCLUSION

In conclusion, we have shown that (1) nonrobust features are present in medical imaging datasets and exploited by DNNs, (2) these features are predictive in-distribution, and suppressing them harms in-distribution accuracy, yet (3) their presence degrades OOD performance, whereas robust features alone are markedly more invariant. This clarifies a robustness-accuracy trade-off that should be taken into account.

7. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by the MedMNIST and MedMNIST-C datasets [5, 12]. Ethical approval was not required as confirmed by the license attached with the open access data.

8. ACKNOWLEDGEMENTS

This project was supported by the modular AI Imaging Pipelines (mAIPipes) Grant, Application No. 22024025 KI-Förderrichtlinie Schleswig-Holstein, Germany, and supported in part by the National Research Foundation of South Africa (Ref Numbers PSTD23042898868, RCDL240215206999).

9. REFERENCES

[1] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry, "Adversarial examples are not bugs, they are features," in *Advances in Neural Information Processing Systems*, 2019.

[2] Nikolaos Tsilivis and Julia Kempe, "What can the neural tangent kernel tell us about adversarial robustness?," in *Advances in Neural Information Processing Systems*, 2022.

[3] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry, "Robustness may be at odds with accuracy," in *International Conference on Learning Representations*, 2019.

[4] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein, "Robustbench: a standardized adversarial robustness benchmark," in *Advances in Neural Information Processing Systems*, 2021.

[5] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni, "Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification," *Scientific Data*, vol. 10, 2023.

[6] Sergey Zagoruyko and Nikos Komodakis, "Wide residual networks," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2016.

[7] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[8] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan, "Theoretically principled trade-off between robustness and accuracy," in *International Conference on Machine Learning*, 2019.

[9] Leslie Rice, Eric Wong, and Zico Kolter, "Overfitting in adversarially robust deep learning," in *International Conference on Machine Learning*, 2020.

[10] Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan, "Better diffusion models further improve adversarial training," in *International Conference on Machine Learning*, 2023.

[11] Tianyu Pang, Xiao Yang, Yinpeng Dong, Hang Su, and Jun Zhu, "Bag of tricks for adversarial training," in *International Conference on Learning Representations*, 2021.

[12] Francesco Di Salvo, Sebastian Doerrich, and Christian Ledig, "Medmnist-c: Comprehensive benchmark and improved classifier robustness by simulating realistic image corruptions," *arXiv preprint arXiv:2406.17536*, 2024.

[13] Francesco Croce and Matthias Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *International Conference on Machine Learning*, 2020.