Revealing Hidden Failure Modes in Chest X-ray Classification via Spectral Domain Analysis

Samuel Halimi SAMUEL@AZMED.CO
Loic Themyr LOIC@AZMED.CO
Arnaud Abreu ARNAUD@AZMED.CO
10 rue d'Uzès 75002 Paris

Editors: Under Review for MIDL 2026

Abstract

Deep learning models for chest X-ray anomaly detection remain vulnerable to subtle distributional shifts (e.g., acquisition technique, patient-related factors, and preprocessing). Traditional error analysis often relies on semantic metadata or model embeddings, which can mask low-level signal variations that degrade performance. In this work, we propose a data-centric framework for automated failure mode discovery using spectral analysis. We project images into the frequency domain and extract a compact profile summarizing the distribution of signal energy across frequency bands. By performing unsupervised clustering on these spectral profiles, we demonstrate that model failures are not randomly distributed, but are strongly concentrated within specific spectral clusters. This method effectively isolates "blind spots", enabling the prediction of model reliability and the discovery of performance-degrading data slices without requiring ground-truth failure annotations.

Keywords: Chest X-ray, Failure Mode Discovery, Spectral Analysis, Model Robustness, Unsupervised Clustering.

1. Introduction

Deep learning has achieved remarkable success in the interpretation of chest X-rays (CXRs), where algorithms now frequently match or exceed radiologist performance in controlled settings (Katzman et al., 2024; Bettinger et al., 2024). However, in the context of clinical deployment, models often fail to generalize to data that deviates statistically from their training distribution, resulting in significant and unanticipated performance degradation (Zech et al., 2018; Yu et al., 2022). Ensuring the reliability of these systems requires the ability to anticipate and identify failure modes before they impact patient care.

Two main approaches have been explored to isolate underperforming samples from unseen data. Metadata stratification uses sensitive categorical attributes to spot biases in the system (Gichoya et al., 2023), while model-centric techniques examine the embedding of models, either to spot out-of-distribution samples (Hong et al., 2024) or to cluster coherent underperfoming slices in the datasets (Olesen et al., 2024). However, both metadata stratification and latent space analysis operate on high-level semantic abstractions and fail to capture subtle signal-level shifts and irregularities that drive performance degradation.

Therefore, in this work, we propose a *data-centric* framework that prioritizes signal characteristics over semantic features or metadata labels. Inspired by recent studies in the field of domain adaptation and generalization (Wang et al., 2020; Xu et al., 2021), we hypothesize that some systematic model failures correlate with specific profiles in the frequency domain.

From the *Fourier transform* of an image, we compute a compact representation that summarizes the distribution of signal energy across frequency bands. In that representation space, we apply an unsupervised clustering method to partition validation data into spectrally coherent subgroups. Our experiments demonstrate that these spectral clusters might act as predictors of model reliability. We show that model failures are not uniformly distributed, but are concentrated within specific spectral clusters and effectively reveal "blind spots" in the model's generalization capability. This approach allows for the identification of performance-degrading data slices rooted in image physics, enabling predictive reliability estimation without the need for ground-truth failure annotations.

2. Related work

2.1. Slice Discovery in medical image datasets

To identify systematic failures, the standard approach for medical applications consists in slicing validation data by categorical attributes (Seyyed-Kalantari et al., 2020; Larrazabal et al., 2020; Ahluwalia et al., 2023). Although essential for fairness auditing, this approach is entirely based on the availability of tabular data. It remains blind to "hidden stratification" (Oakden-Rayner et al., 2020), where performance degrades due to signal-level characteristics that are not recorded in clinical logs.

Beyond categorical attributes, Out-of-Distribution (OOD) detection uses feature space distances to spot anomalies (Lee et al., 2018; Liu et al., 2020) and is also widely explored in medical image applications (Tardy et al., 2019; Berger et al., 2021; Roy et al., 2022; Araujo et al., 2023). Although effective in identifying stark outliers, these methods struggle with subtle variations in image acquisition that degrade performance without triggering distance-based alarms (Wiles et al., 2021).

To uncover coherent high-error subsets of data, recent unsupervised methods apply clustering algorithms in the latent embedding space of models (Eyuboglu et al., 2022; d'Eon et al., 2022; Olesen et al., 2024). However, this approach faces a technical paradox: deep networks are explicitly optimized to be invariant to nuisance variables (Achille and Soatto, 2018). Consequently, the latent space often suppresses the irregularities that cause model failure, rendering these failure modes invisible to latent space clustering strategies.

2.2. Frequency domain analysis in deep learning

In this work, we look for failure modes in the frequency domain of the images, to remove metadata supervision and dependence on subjective model embeddings to focus only on the intrinsic irregularities of the signal. Important studies on image frequencies have been conducted in domain generalization (DG) (Huang et al., 2021a; Zhao et al., 2022) and domain adaptation (DA) (Huang et al., 2021b; Yang et al., 2022) to increase the robustness of deep learning models with respect to frequency perturbations. These methods alter either the low frequencies (Guo et al., 2018) or the high frequencies (Wang et al., 2020) of the images to train models on adversarial examples in a data augmentation scheme. Although not directly related to our approach, the success of these techniques brings evidence that some failure modes of deep learning models are explained by characteristics of the Fourier spectrum of input images.

Unlike DG and DA approaches that manipulate spectra to train robust models, we utilize spectral analysis as a *post-hoc* diagnostic tool. This allows us to cluster data based on image physics rather than semantics, exposing signal-driven failures that escape standard monitoring.

3. Methods

3.1. Radially Averaged Power Spectrum (RAPS)

We analyze signal-level variations by projecting each image into the frequency domain. For an image x, we compute its 2-D discrete Fourier transform $F = \mathcal{F}(x)$ and magnitude spectrum $S = |F|$. To obtain a 1-D orientation-invariant descriptor P, we compute the Radially Averaged Power Spectrum (RAPS) of the image following the extraction protocol by Torralba and Oliva (2003). The computation of the RAPS is illustrated in Figure 1.
For a given discrete radius $r \in \mathbb{N}$, let $\Omega(r)$ be the set of frequency coordinates such as:

$$\Omega(r) = \{(u, v) \in \mathbb{Z}^2 \mid \sqrt{u^2 + v^2} = r\} \tag{1}$$

Then, for a given sampling of N discrete radii $[r_1, \ldots, r_N]$, the n-th component $P(n)$, of the RAPS P, is computed as the average of the spectral magnitude S over $\Omega(r_n)$:

$$P(n) = \frac{1}{|\Omega(r_n)|} \sum_{(u,v)\in\Omega(r_n)} S(u, v) \tag{2}$$

We employ the RAPS profile to compress images into compact, rotation-invariant vectors, ensuring that downstream clustering tasks focus on intrinsic statistics rather than superficial variations in object pose.

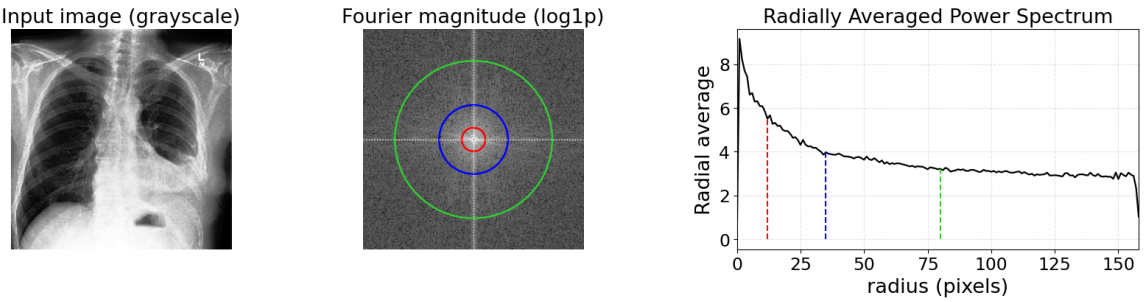


Figure 1: Computation of the RAPS of an image. We highlight the contribution of three frequency bands (red, green, blue) in the spectrum to the RAPS representation (dashed lines with corresponding colors).

To ensure comparability across images, all Fourier spectra must lie on the same frequency grid. We therefore restrict the analysis to images of similar native resolution and enforce identical spatial dimensions using minimal spatial cropping. This avoids spectral distortions: resizing modifies high-frequency content through interpolation, while padding artificially

increases low-frequency energy. A minimal uniform crop preserves the relevant spectral structure and maintains the physical interpretation of Fourier frequencies.

3.2. Similarity metric

For the subsequent clustering phase, we advocate the use of the correlation distance as a similarity metric between RAPS descriptors. Given the RAPS of two different unseen images, P_i and P_j, their correlation distance is given by:

$$d_{\text{corr}}(i,j) = 1 - \text{corr}(P_i, P_j) \tag{3}$$

Where the correlation between P_i and P_j is computed across the N sampled radii as:

$$\text{corr}(P_i, P_j) = \frac{\sum_{n=1}^{N} \left(P_i(n) - \bar{P}_i\right)\left(P_j(n) - \bar{P}_j\right)}{\sqrt{\sum_{n=1}^{N} \left(P_i(n) - \bar{P}_i\right)^2}\sqrt{\sum_{n=1}^{N} \left(P_j(n) - \bar{P}_j\right)^2}} \tag{4}$$

Unlike the Euclidean distance, which is sensitive to absolute magnitude, the correlation distance is invariant to additive and multiplicative shifts in signal intensity. This ensures that vectors are grouped based on their intrinsic spectral morphology rather than physically irrelevant variations in global energy or sensor gain. Consequently, the clustering process effectively isolates stable structural features, such as relative sharpness (shape of the high-frequency tail), sensor or reconstruction noise patterns (boosting of high frequencies), as well as characteristic spectral slopes of specific imaging devices, while remaining robust to acquisition-level scaling and offsets.

3.3. K-means clustering

In a first exploratory phase, we use the standard K-means clustering to validate the relevance of RAPS descriptors for performance analysis. As it is not specifically optimized for slice discovery, it allows unbiased exploration of the relationship between the frequency domain and model failures. The possibility of varying the number of clusters is also convenient to confirm an observed tendency across partitions of different granularities.

3.4. Hierarchical Agglomerative Clustering (HAC)

To optimize the slice discovery process, we resort to a hierarchical clustering technique. In a bottom-up fashion, each radial profile starts as an individual cluster and we iteratively merge the closest pairs of clusters according to the average linkage rule. To partition the resulting dendrogram we set the 0.5-quantile of the distribution of pairwise distances as a threshold to stop the merging procedure. Adopting this strategy gives three key advantages. First, the partition scale naturally adapts to the inherent variability of the spectral representations. Second, it circumvents the limitation of pre-specifying an arbitrary cluster count. Finally, this flexibility effectively optimizes the isolation and discovery of specific underperforming slices.

4. Experimental setup

4.1. Datasets and preprocessing

We conducted all experiments using four publicly available chest X-ray datasets commonly employed for disease classification:
CheXpert (Irvin et al., 2019). A large-scale dataset from Stanford containing frontal and lateral views, labeled for 14 thoracic findings using report-derived NLP. **MIMIC-CXR** (Johnson et al., 2019). A de-identified dataset from Beth Israel Deaconess Medical Center with paired radiology reports and substantial acquisition and population diversity. **PadChest** (Bustos et al., 2020). A Spanish dataset with multi-view radiographs and detailed labels covering findings, diagnoses, and anatomical locations. **NIH ChestX-ray14** (Wang et al., 2017). A frontal-view dataset labeled for 14 conditions via report mining, notable for heterogeneous acquisition and label noise, making it a common robustness benchmark.

To ensure that Fourier transforms and spectral profiles were comparable across datasets, we standardized image resolution with minimal geometric distortion. We selected 264×224 as the target size, as it was the most common resolution across all sources. Images were first resized while preserving aspect ratio and then center-cropped to the exact target dimensions. Samples with incompatible shapes (e.g., near-square images) were discarded to avoid excessive rescaling.

We focus on three clinically common findings: Atelectasis, Consolidation, and Pleural Effusion. We trained our models in a mono-pathology setting. For each pathology, we constructed separate training and validation splits. The training sets are exclusively based on CheXpert images. Table 1 details the datasets sizes. The test sets are built from each dataset, so MIMIC-CXR, PadChest, and NIH are unseen-domain in our evaluation. This enables the analysis of cross-dataset generalization under distributional shift.

Table 1: The size of each mono-pathology dataset used in the study is detailed, broken down by the number of images in the training set, the seen-domain test set, and the unseen-seen test set. Note that all datasets are balanced, containing an equal number of positive and negative examples.

Dataset	Type	Atelectasis	Consolidation	Pleural Effusion
CheXpert	Train	39K	39K	39K
CheXpert	Test	256	256	256
MIMIC	Test	476	400	520
NIH	Test	1308	642	1386
Padchest	Test	400	400	816

4.2. Models, training and performance assessment

For all experiments, we trained a DenseNet-121 (Huang et al., 2018) classifier using binary cross-entropy loss and the Adam (Kingma and Ba, 2017) optimizer (learning rate 1×10^{-4}, weight decay 1×10^{-3}). A Reduce-on-Plateau scheduler (minimum learning rate 1×10^{-5}) and a batch size of 16 were used. Models were trained for up to 30 epochs with early stopping

based on validation performance. All training was performed on an NVIDIA GeForce RTX 3080 GPU.

We report the Area Under the Receiver Operating Characteristic (AUROC) to evaluate the discriminative performance of the models. This metric was chosen for its invariance to decision thresholds and its robustness to the fluctuating prevalence in unsupervised clustering assignments.

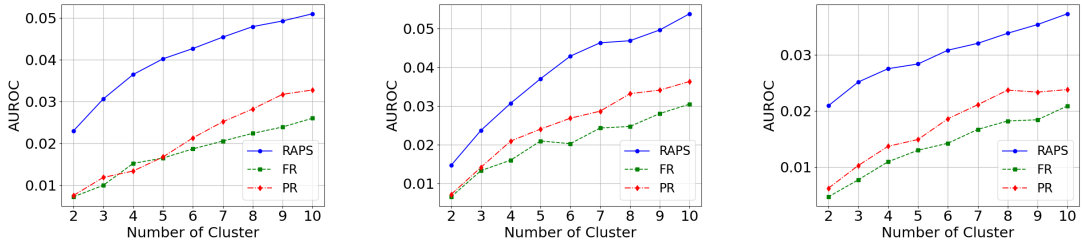
5. Experiments and results

5.1. Relevance of RAPS descriptors for performance analysis

Given a K-means partition of the RAPS descriptors (Section 3.3), we use the standard deviation of the AUROC across the K clusters to measure how effectively the partition stratifies performance. To isolate the effect of spectral coherence from artifacts of cluster size or class distribution, we implement two stochastic baselines: a *fully random* (FR) assignment and a *pseudo-random* (PR) control that preserves the specific cluster size and label statistics of the K-means solution. To study the impact of granularity, we repeat the above experiment for values of K ranging from 2 to 10. We then average the results over 50 seeds to reduce variability from K-means initialization and baseline stochasticity.

As shown in Figure 2, the dispersion of AUROC in the K-means partitions of RAPS profiles (blue) remains consistently above, sometimes twice as high, those of both random baselines (red and green). This persistent separation indicates that the partitions induced by the correlation distance between RAPS profiles capture genuine heterogeneity in the model behavior.

Because the correlation distance emphasizes shape differences, this gap suggests that differences in spectral composition, such as relative distribution of low versus high frequencies, decay rates, or local spectral slopes, are systematically associated with differences in model performance. The fact that this phenomenon appears for all pathologies (see Appendix A) and all values of K supports the hypothesis that the spectral representation explains, at least, part of the observed variability in the reliability of the model.



(a) Atelectasis (b) Consolidation (c) Pleural Effusion

Figure 2: AUROC variability across different number of clusters (K) made by the k-means method with the correlation metric vs. clusters made by fully and pseudo random methods. Results are averaged over 50 seeds.

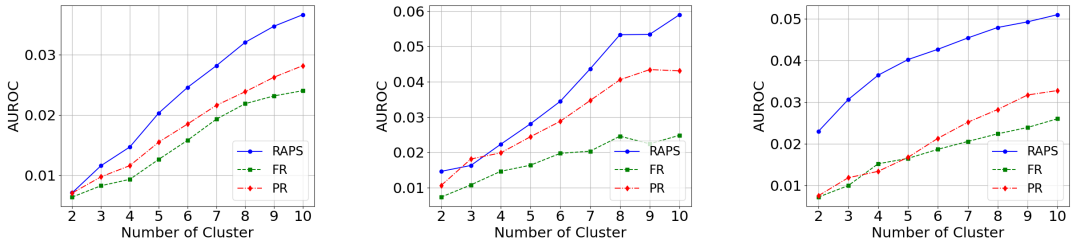
5.2. Influence of the similarity metric on performance stratification

To assess how the distance metric influences the K-means performance stratification, we compare three metrics on the radial profiles: Euclidean, Spearman, and correlation. The evaluation follows the same protocol as in Section 5.1, with AUROC variability compared to the FR and PR baselines and averaged over multiple seeds.

Across all values of K, the correlation distance produces the largest separation between the clustering curve and both baselines (Figure 3). The Spearman distance induces greater variability than the Euclidean distance, though still below the correlation distance.

By contrast, the L_2 distance shows only limited gains over the baselines, suggesting that amplitude-based similarities contribute little to the stratification of performance compared to differences in the *shape* of the spectral profiles. This ordering among metrics indicates that the performance variability is more closely related to the *relative distribution* of spectral energy than to its absolute magnitude. Metrics that are invariant to global scaling and affine transformations, such as the correlation distance and, to a lesser extent, the Spearman distance, capture spectral characteristics that are more informative for distinguishing model behavior. This supports the choice of the correlation distance in our clustering framework and is consistent with Section 5.1, where it produces the largest performance disparities across clusters.

These findings suggest that DG and DA strategies based on the frequency domain should not focus exclusively on matching absolute spectral energy distributions, as is often done in methods that impose target spectra or mix frequency components directly. Instead, our results indicate that the *relative* structure of the radial spectrum, such as the balance between low and high frequencies and the rate at which spectral energy decays, plays a more consequential role. Methods that explicitly account for these relative spectral characteristics may therefore offer a more effective direction for adaptation and robustness.

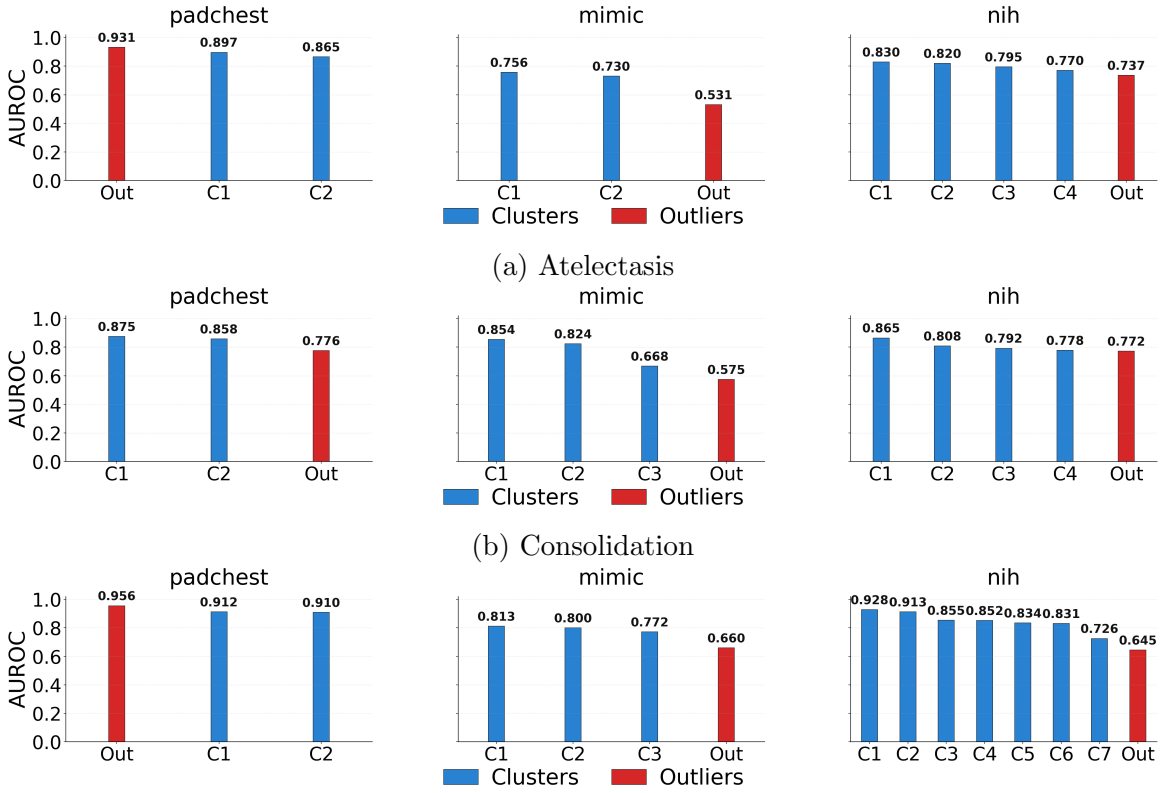


(a) L_2 (b) Spearman (c) Correlation

Figure 3: Ablation study evaluating the influence of the distance metric used in k-means clustering on AUROC variation for the Atelectasis pathology. We compare three metrics: Euclidean (L_2) distance, Spearman distance, and the correlation distance employed in our proposed method.

5.3. Slice discovery and outlier detection

To confirm the robustness of our data-centric framework, we apply HAC in the space of RAPS descriptors to isolate under-performing data slices on unseen domain data. After clustering, following the method from Section 3.4, we aggregate clusters with fewer than 10 samples into a single outlier group; this ensures sufficient sample sizes for the primary partitions while isolating spectrally atypical profiles. We then assess the AUROC on each cluster to pinpoint specific spectral regions where model performance deviates from the norm.



(a) Atelectasis

(b) Consolidation

(c) Pleural Effusion

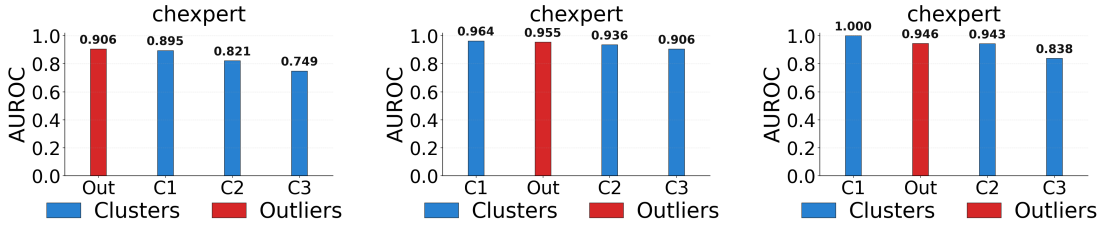
Figure 4: Comparative analysis of model performance (AUROC) across data clusters (blue) defined by HAC. The figure highlights the distinct model reliability on the identified clusters versus the significantly different performance (red) observed on the HAC-isolated outlier set. Results on the unseen-domain test datasets.

As depicted in Figure 4, the AUROC across HAC clusters exhibits distinct and variable performance levels for all pathologies and external validation centers. This finding mirrors the performance stratification observed with the K-means clustering approach. This convergence confirms that the spectral differences encoded through the correlation distance

reliably capture meaningful variations in model performance, regardless of the underlying clustering algorithm employed.

Importantly, we demonstrate that across almost all considered datasets and pathologies, the proposed method systematically isolates a cluster exhibiting significantly degraded AUROC (bars on the far right of the different plots). This consistency underscores the efficacy of applying HAC to RAPS descriptors for slice discovery, as it reliably uncovers latent subgroups where model performance is compromised.

Additionally, a particularly characteristic and instructive pattern emerged from the analysis of the outlier super-group (red), which was formally defined by merging all HAC clusters containing fewer than 10 samples. The AUROC for this super-group consistently exhibited a marked separation from the performance mean of the main clusters. Specifically, the performance was frequently significantly lower; as observed, for example, in Atelectasis and Pleural Effusion on the MIMIC and NIH datasets, and in Consolidation on Padchest and MIMIC; indicating substantial failure concentration. Conversely, in select cases, the AUROC was noticeably higher (e.g., Atelectasis and Pleural Effusion on Padchest), but critically, it was never comparable to the average cluster performance. This consistent performance separation provides strong quantitative evidence: slices possessing atypical spectral signatures reliably translate into performance-level outliers within the model.



(a) Atelectasis (b) Consolidation (c) Pleural Effusion

Figure 5: Comparative analysis of model performance (AUROC) across data clusters (blue) defined by HAC. The figure highlights the distinct model reliability on the identified clusters versus the significantly different performance (red) observed on the HAC-isolated outlier set. Results on seen-domain test datasets: CheXpert.

We also evaluated the model on seen-domain images with the CheXpert test dataset. As shown in Figure 5, the performance on the outlier super-group is not markedly different from that of the other clusters. This outcome is expected: because the model was trained exclusively on this domain, it learned to handle the full range of RAPS variations present in the training distribution. Nevertheless, we still observe clusters with distinct performance levels, indicating that HAC can partition the domain into meaningful RAPS subsets on which the model performs comparatively better or worse.

Finally, for Pleural Effusion cases from the NIH center, we compare the AUROC of the worst slice isolated by our method with traditional slicing by available metadata attributes, i.e. age, gender, and view. As demonstrated in Table 2, metadata slicing sometimes fails to identify performance-degrading subsets, resulting in only minor AUROC variations. In

Table 2: Comparison of model performance achieved by slicing the data using conventional semantic metadata against the proposed HAC of RAPS method. Results are presented for the Pleural Effusion pathology within the NIH dataset. (P: Posterior; A: Anterior)

Category	**Subgroup**	**AUROC**
Age	< 20	0.863
	20–60	0.832
	> 60	0.839
Gender	Female	0.856
	Male	0.824
View	AP	0.827
	PA	0.845
Proposed	**Our Slice**	**0.645**

contrast, our data-centric framework successfully isolates a highly vulnerable data slice that exhibits a substantial and critical performance drop. This result definitively establishes our method as an effective complementary solution for automated failure mode discovery.

6. Conclusion

In this work, we introduced a data-centric framework that uses Fourier-domain analysis to examine performance variability in chest X-ray classifiers. Representing each image through its Radially Averaged Power Spectrum (RAPS) yielded model-agnostic descriptors capturing acquisition-level signal properties. Across experiments, K-means and hierarchical clustering consistently revealed meaningful differences in model behavior across groups defined by their spectral profiles, with hierarchical clustering further isolating small sets of images whose atypical frequency patterns were matched by equally atypical performance across evaluation domains.

Ablations over similarity metrics clarified which components of the spectral representation matter most. Correlation distance produced the strongest separation between clusters, indicating that performance differences relate primarily to the relative distribution of spectral energy, its shape and rate of decay, rather than to absolute magnitude. This identifies the spectral characteristics most relevant to model behavior and motivates approaches that explicitly account for such relative frequency patterns.

Together, these findings point to limitations of metadata- or embedding-based slice discovery. Signal-level irregularities rooted in acquisition physics, preprocessing, or frequency composition form predictive patterns that standard monitoring pipelines overlook. Spectral analysis therefore provides a principled and interpretable means of exposing these hidden strata and improving performance auditing. Beyond error analysis, our results suggest that robustness and adaptation strategies may benefit from targeting spectral structure directly, particularly the balance between low and high frequencies, offering a more physically grounded avenue for improving generalization in medical imaging.

Acknowledgments

The authors thank Alexandre Attia, Julien Vidal and Elie Zerbib for supporting this work at AZmed.

References

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50):1–34, 2018.

Monish Ahluwalia, Mohamed Abdalla, James Sanayei, Laleh Seyyed-Kalantari, Mohannad Hussain, Amna Ali, and Benjamin Fine. The subgroup imperative: chest radiograph classifier generalization gaps in patient, setting, and pathology subgroups. *Radiology: Artificial Intelligence*, 5(5):e220270, 2023.

Teresa Araujo, Guilherme Aresta, Ursula Schmidt-Erfurth, and Hrvoje Bogunović. Few-shot out-of-distribution detection for automated screening in retinal oct images using deep learning. *Scientific Reports*, 13(1):16231, 2023.

Christoph Berger, Magdalini Paschali, Ben Glocker, and Konstantinos Kamnitsas. Confidence-based out-of-distribution detection: a comparative study and analysis. In *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, pages 122–132. Springer, 2021.

Hubert Bettinger, Gregory Lenczner, Jean Guigui, Luc Rotenberg, Elie Zerbib, Alexandre Attia, Julien Vidal, and Pauline Beaumel. Evaluation of the performance of an artificial intelligence (ai) algorithm in detecting thoracic pathologies on chest radiographs. *Diagnostics*, 14(11):1183, 2024.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797, 2020.

Greg d'Eon, Jason d'Eon, James R Wright, and Kevin Leyton-Brown. The spotlight: A general method for discovering systematic errors in deep learning models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1962–1981, 2022.

Sabri Eyuboglu, Maya Varma, Khaled Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Ré. Domino: Discovering systematic errors with cross-modal embeddings. *arXiv preprint arXiv:2203.14960*, 2022.

Judy Wawira Gichoya, Kaesha Thomas, Leo Anthony Celi, Nabile Safdar, Imon Banerjee, John D Banja, Laleh Seyyed-Kalantari, Hari Trivedi, and Saptarshi Purkayastha. Ai pitfalls and what not to do: mitigating bias in ai. *The British Journal of Radiology*, 96 (1150):20230023, 2023.

Chuan Guo, Jared S Frank, and Kilian Q Weinberger. Low frequency adversarial perturbation. *arXiv preprint arXiv:1809.08758*, 2018.

Zesheng Hong, Yubiao Yue, Yubin Chen, Lele Cong, Huanjie Lin, Yuanmei Luo, Mini Han Wang, Weidong Wang, Jialong Xu, Xiaoqi Yang, et al. Out-of-distribution detection in medical image analysis: A survey. *arXiv preprint arXiv:2404.18279*, 2024.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2018. URL https://arxiv.org/abs/1608.06993.

Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6891–6902, 2021a.

Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Rda: Robust domain adaptation via fourier adversarial attacking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8988–8999, 2021b.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

Benjamin D Katzman, Mostafa Alabousi, Nabil Islam, Nanxi Zha, and Michael N Patlas. Deep learning for pneumothorax detection on chest radiograph: a diagnostic test accuracy systematic review and meta analysis. *Canadian Association of Radiologists Journal*, 75 (3):525–533, 2024.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23): 12592–12594, 2020.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, pages 151–159, 2020.

Vincent Olesen, Nina Weng, Aasa Feragen, and Eike Petersen. Slicing through bias: explaining performance gaps in medical image analysis using slice discovery methods. In *MICCAI Workshop on Fairness of AI in Medical Imaging*, pages 3–13. Springer, 2024.

Abhijit Guha Roy, Jie Ren, Shekoofeh Azizi, Aaron Loh, Vivek Natarajan, Basil Mustafa, Nick Pawlowski, Jan Freyberg, Yuan Liu, Zach Beaver, et al. Does your dermatology classifier know what it doesn't know? detecting the long-tail of unseen conditions. *Medical Image Analysis*, 75:102274, 2022.

Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. In *BIOCOMPUTING 2021: proceedings of the Pacific symposium*, pages 232–243. World Scientific, 2020.

Mickael Tardy, Bruno Scheffer, and Diana Mateus. Uncertainty measurements for the reliable classification of mammograms. In *International conference on medical image computing and computer-assisted intervention*, pages 495–503. Springer, 2019.

Antonio Torralba and Aude Oliva. Statistics of natural image categories. *Network: computation in neural systems*, 14(3):391, 2003.

Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8684–8694, 2020.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre Alvise-Rebuffi, Ira Ktena, Krishnamurthy Dvijotham, and Taylan Cemgil. A fine-grained analysis on distribution shift. *arXiv preprint arXiv:2110.11328*, 2021.

Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14383–14392, 2021.

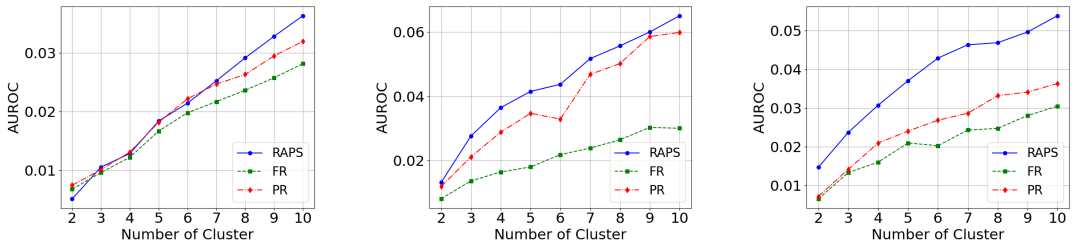
Chen Yang, Xiaoqing Guo, Zhen Chen, and Yixuan Yuan. Source free domain adaptation for medical image segmentation with fourier style mining. *Medical Image Analysis*, 79: 102457, 2022.

Alice C Yu, Bahram Mohajer, and John Eng. External validation of deep learning algorithms for radiologic diagnosis: a systematic review. *Radiology: Artificial Intelligence*, 4(3): e210064, 2022.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11): e1002683, 2018.

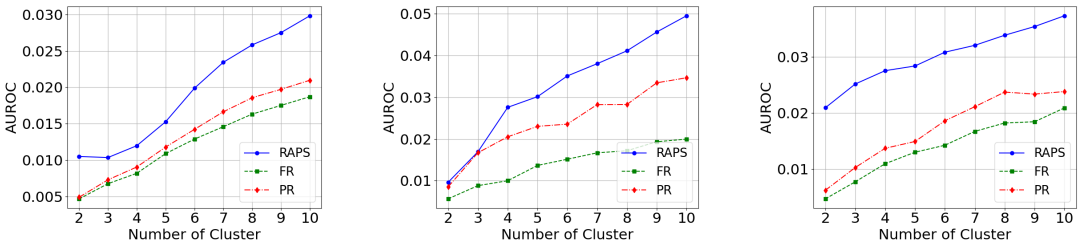
Xingchen Zhao, Chang Liu, Anthony Sicilia, Seong Jae Hwang, and Yun Fu. Test-time fourier style calibration for domain generalization. *arXiv preprint arXiv:2205.06427*, 2022.

Appendix A. Ablation extension



(a) L2 (b) Spearman (c) Correlation

Figure 6: Ablation study evaluating the influence of the distance metric used in k-means clustering on AUROC variation for the Consolidation pathology. We compare three metrics: Euclidean (L_2) distance, Spearman distance, and the correlation distance employed in our proposed method.



(a) L2 (b) Spearman (c) Correlation

Figure 7: Ablation study evaluating the influence of the distance metric used in k-means clustering on AUROC variation for the Pleural Effusion pathology. We compare three metrics: Euclidean (L_2) distance, Spearman distance, and the correlation distance employed in our proposed method.