GRASP-PsONet: Gradient-based Removal of Spurious Patterns for PsOriasis Severity Classification

Basudha Pal[1,2†], Sharif Amit Kamran[1], Brendon Lutnick[1], Molly Lucas[1], Chaitanya Parmar[1], Asha Patel Shah[1], David Apfel[1], Steven Fakharzadeh[1], Lloyd Miller[1], Gabriela Cula[1], and Kristopher Standish[1]

[1] Johnson&Johnson Innovative Medicine, New Brunswick, NJ, USA
[2] Johns Hopkins University, Baltimore, MD, USA
bpal5@jhu.edu

Abstract. Psoriasis (PsO) severity scoring is vital for clinical trials but is hindered by inter-rater variability and the burden of in-person clinical evaluation. Remote imaging utilizing patient-captured mobile photos offers scalability but introduces challenges, such as variations in lighting, background, and device quality that are often imperceptible to humans but may impact model performance. These factors, coupled with inconsistencies in dermatologist annotations, reduce the reliability of automated severity scoring. We propose a framework to automatically flag problematic training images that introduce biases and reinforce spurious correlations which degrade model generalization by using a gradient-based interpretability approach. By tracing the gradients of misclassified validation images, we detect training samples where model errors align with inconsistently rated examples or are affected by subtle, non-clinical artifacts. We apply this method to a ConvNeXT-based weakly supervised model designed to classify PsO severity from phone images. Removing 8.2% of flagged images improves model AUC-ROC by 5% (85% to 90%) on a held-out test set. Commonly, multiple annotators and an adjudication process ensure annotation accuracy, which is expensive and time-consuming. Our method correctly detects training images with annotation inconsistencies, potentially eliminating the need for manual reviews. When applied to a subset of training images rated by two dermatologists, the method accurately identifies over 90% of cases with inter-rater disagreement by rank-ordering and reviewing only the top 30% of training data. This framework improves automated scoring for remote assessments, ensuring robustness and scalability despite variability in data collection. Our method handles both inconsistencies in image conditions and annotations, making it ideal for applications lacking standardization of controlled clinical environments.

Keywords: Dermatology · Psoriasis · Multi-instance Learning · Explainability · Gradient Tracing · Spurious Correlations

[†] This work was done during an internship at J&J Innovative Medicine

1 Introduction

Psoriasis (PsO) is a chronic systemic inflammatory disease that affects 2%-3% of the global population and is associated with comorbidities such as psoriatic arthritis (PsA), diabetes, and cardiovascular diseases [28, 2]. PsO severity assessment is essential for decision-making during clinical trials, as treatment selection is based on disease severity. Dermatologists use the Psoriasis Area and Severity Index (PASI) as the gold standard for quantifying PsO, scoring lesion extent and severity on a scale of 0–72 [2]. Cases are then classified as mild, moderate, or severe based on predefined thresholds. However, in-clinic evaluations impose a logistical burden on both patients and physicians. Combining remote imaging, where patients capture images using mobile devices, with deep learning-based automated PASI scoring offers a scalable alternative. This approach reduces the need for in-person visits, minimizes subjectivity, and streamlines disease monitoring [12, 8, 23, 9]. Despite advancements in deep learning-based PsO assessment, existing models face several limitations, such as extensive pre-processing using bounding boxes [12], background removal [12, 8, 23], and exclusion of clinically relevant regions [25, 23] which are difficult to detect. Some approaches use separate models for regional scoring [9] or distinct computations for erythema, induration, desquamation, and lesion area ratio [17, 18] adding to complexity.

Despite being scalable, remote imaging can introduce significant variability in illumination, background, and device quality, which can lead to spurious correlations and degrade model performance. Furthermore, annotation inconsistencies among dermatologists contribute to unreliable training data, making it difficult for models to generalize [21, 13, 26]. A promising direction for addressing these challenges is data attribution, which aims to identify training samples that negatively impact model generalization. Koh *et al.* [11] and Yeh *et al.* [27] extended the concept of influence functions from robust statistics to deep learning. Researchers have explored feature-level influence by estimating how specific input features impact individual predictions [24, 14, 22], as well as training sample influence by assessing how data points contribute to overall model performance [16, 15]. Gradient-tracing methods, such as TracIn [20] and its practical adaptation TracInCP, estimate training sample influence by tracking gradient updates across minibatches and checkpoints. While [4], a closely related work, leverages Hessians to measure influence in the absence of a sample, TracInCP replays training via stored checkpoints to approximate its effect on test predictions. To emphasize on model explainability in medical imaging, some researchers have leveraged influence functions to analyze and interpret decision-making [7]. Recent work in medical imaging has leveraged model pruning to improve performance and reduce computation [1, 6]. The Dynamic Average Dice score by He *et al.* [5] on the other hand, focuses on data pruning which dynamically quantifies the importance of each training sample by assessing its contribution to the Dice coefficient, allowing a score guided identification and removal of non-informative training samples. To the best of our knowledge, such score-based data pruning methods have been less explored in medical imaging.

In this study, we introduce GRASP-PsONet, a gradient-tracing based influence estimation method for efficient data pruning in PsO severity classification. Given the challenge that a training image may be detrimental to one validation instance yet beneficial to another, GRASP-PsONet optimizes data selection. Our key contributions are as follows:

- GRASP-PsONet is built on an existing weakly supervised multi-instance learning (MIL) framework, provides an end-to-end solution which eliminates extensive pre-processing (e.g., bounding boxes, region-specific models) while remaining resilient to spurious correlations and annotation inconsistencies. Using gradient tracing, our method enhances model interpretability and generalizability by identifying and removing influential training samples that contribute to misclassifications in the validation set. Specifically, we compute influence scores from misclassified validation examples and systematically prune 2.8%-13.3% of the training dataset, improving overall robustness.
- We use self-influence scores to identify potential mislabeled examples [20]. Self-influence quantifies how removing a training example affects its own prediction, with high scores indicating mislabeled or atypical samples. We rank-order training images based on self-influence and demonstrate that this method effectively detects annotation inconsistencies. When applied to a subset rated by two dermatologists, reviewing only the top 30% of ranked data correctly identified 90.3% of inter-observer disagreements.
- Using data attribution based method boosts performance on the multi-class PsO classification task by improving AUC for both readers by approximately 5% and 10% for ConvNeXT and ViT-based encoders, respectively.

2 Methodology

2.1 Problem Setup

The data was divided into a training set, $S_{\text{train}} = \{(x_1^{\text{trainpv}}, y_1^{\text{trainpv}}), \ldots, (x_N^{\text{trainpv}}, y_N^{\text{trainpv}})\}$, with $N = 46$ images per patient visit (pv) and 610 unique patient visits. The validation set was $S_{\text{val}} = \{(x_1^{\text{valpv}}, y_1^{\text{valpv}}), \ldots, (x_N^{\text{valpv}}, y_N^{\text{valpv}})\}$, where each label y denotes PsO severity: mild $(\text{PASI} : 0 - 5)$, moderate $(\text{PASI} : 5 - 10)$, or severe $(\text{PASI} > 10)$ based on all 46 images per visit. Influence is the reduction in loss on a validation example $z' \in S_{\text{val}}$ caused by using a training example $z \in S_{\text{train}}$. The aim is to find influential training images that lead to validation misclassifications in an existing MIL model and retrain it after removing these images. The final model is chosen based on the best validation AUC. Training labels come from Reader 1, while evaluation uses scores from two readers. For each misclassified validation visit, the image with the highest attention score among the 46 is chosen, and gradients are traced to remove the top k influential training images, $x_{\text{rem}} \in S_{\text{train}}$. After removing influential images, N becomes variable as it decreases with the number of removed images. 'Baseline' refers to the MIL models without data attribution for both ConvNeXT and ViT encoders.

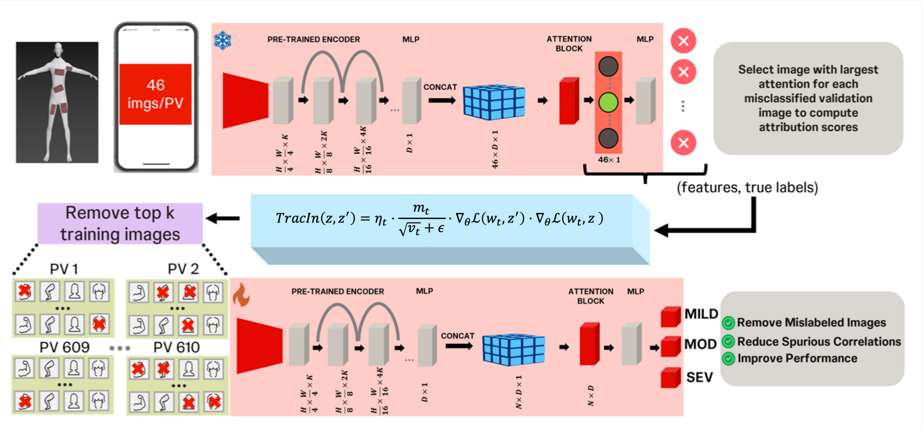
46
imgs/PV
PRE-TRAINED ENCODER
MLP
ATTENTION BLOCK
MLP
$\frac{H}{4} \times \frac{W}{4} \times K$
$\frac{H}{8} \times \frac{W}{8} \times 2K$
$\frac{H}{16} \times \frac{W}{16} \times 4K$
$D \times 1$
CONCAT
$46 \times D \times 1$
46× 1
Select image with largest attention for each misclassified validation image to compute attribution scores
(features, true labels)
$TracIn(z,z') = \eta_t \cdot \frac{m_t}{\sqrt{v_t}+\epsilon} \cdot \nabla_\theta \mathcal{L}(w_t,z') \cdot \nabla_\theta \mathcal{L}(w_t,z)$
Remove top k training images
PV 1
PV 2
PV 609
PV 610
PRE-TRAINED ENCODER
MLP
ATTENTION BLOCK
MLP
CONCAT
$N \times D \times 1$
$N \times D$
MILD
MOD
SEV
Remove Mislabeled Images
Reduce Spurious Correlations
Improve Performance

Fig. 1. The overall MIL framework for multiclass PsO severity classification, including data attribution. PASI classification thresholds are mild (0–5): class 0, moderate (5–10): class 1, and severe (>10): class 2.

2.2 Overall Architecture

Our architecture consists of a pre-trained encoder, an attention block, and a multi-layer perceptron (MLP) for multi-class classification, as shown in Fig.1. Given that our dataset includes body images captured by patients, we found ImageNet [3] pre-trained encoders effective for transfer learning. We evaluated encoder architectures such as ConvNeXT and Vision Transformer (ViT). Here we describe the pre-trained encoder for multiclass PsO classification using ConvNeXT as we obtain best Baseline results using this encoder. The ConvNeXT encoder employs convolutional and downsampling blocks, producing feature dimensions of $R^{H/4 \times W/4 \times K}$, $R^{H/8 \times W/8 \times 2K}$, $R^{H/16 \times W/16 \times 4K}$, and $R^{H/32 \times W/32 \times 8K}$ for an input RGB image of size $H \times W \times C$ where K is the number of channels. The encoder output feature dimension is reshaped to $R^{D \times 1}$ via an MLP where $D = 768$. This is repeated N times, where N corresponds to the images per pv (initially 46/pv). The resulting features are concatenated into a tensor of size $R^{N \times D \times 1}$, input into the attention block, and the output is passed to a final MLP to generate class probabilities for three classes.

2.3 Removing Influential Images

We compute the impact of individual training samples on validation predictions by integrating gradient-based data attribution [20] as outlined in Algorithm 1, which estimates influence scores by computing gradient similarity between training and validation examples. For each misclassified validation image, we rank training samples based on their cumulative influence and remove the k most influential ones, refining S_{train}. In the MIL model, we handle this by passing

Algorithm 1 : Tracing Gradients in GRASP-PsONet for TracIn Score with Adam

```
Input: Model $F$, validation point $z' \in S_{\text{val}}$, training point $z \in S_{\text{train}}$, training checkpoints $\{w_t\}_{t=1}^{T}$, mini-batches $B_t$ for $t = 1, 2, \ldots, T$, batch size $b$, Adam moments $m_t, v_t$, learning rate $\eta_t$, loss function $\mathcal{L}(w, z)$.
Output: TracIn attribution score $\text{TracIn}(z', z)$.
Initialize: $\text{TracIn}(z', z) \leftarrow 0$
for $t = 1$ to $T$ do
    if $z \in B_t$ then
        Compute gradients: $\nabla_\theta \mathcal{L}(w_t, z')$ (w.r.t. $z'$) and $\nabla_\theta \mathcal{L}(w_t, z)$ (w.r.t. $z$)
        Update the TracIn score:
            $\text{TracIn}(z', z) \leftarrow \text{TracIn}(z', z) + \eta_t \cdot \frac{m_t}{\sqrt{v_t} + \epsilon} \cdot \nabla_\theta \mathcal{L}(w_t, z') \cdot \nabla_\theta \mathcal{L}(w_t, z)$
    end if
end for
Normalize by batch size: $\text{TracIn}(z', z) \leftarrow \frac{1}{b}\text{TracIn}(z', z)$
return $\text{TracIn}(z', z)$
```

a binary vector indicating image presence, which is multiplied with the attention block to mask absent images. The influence of a training sample z on a validation example z' is quantified by the TracIn score, $\text{TracIn}(z', z)$, which approximates the loss reduction in z' when z is utilized during training. Specifically, $\mathcal{L}(w_{t+1}, z') = \mathcal{L}(w_t, z') + \nabla \mathcal{L}(w_t, z') \cdot (w_{t+1} - w_t) + \mathcal{O}(\|w_{t+1} - w_t\|^2)$. Ignoring higher-order terms yields $\mathcal{L}(w_{t+1}, z') = \mathcal{L}(w_t, z') + \nabla \mathcal{L}(w_t, z') \cdot (w_{t+1} - w_t)$. For Adam optimizers, $w_{t+1} - w_t = -\eta_t \frac{m_t}{\sqrt{v_t + \epsilon}}$, where m_t and v_t are the first and second moments, η_t is the learning rate, and ϵ ensures numerical stability. Substituting, we get $\mathcal{L}(w_{t+1}, z') - \mathcal{L}(w_t, z') = \text{TracIn}(z', z) = \eta_t \frac{m_t}{\sqrt{v_t + \epsilon}} \nabla \mathcal{L}(w_t, z')$. For mini-batches B_t with batch size $b \geq 1$, the final computation of $\text{TracIn}(z', z)$ is shown in Algorithm 1. This method identifies training images that drive misclassifications due to spurious correlations like lighting or labeling errors. Since our labels are at the patient visit level, with each visit containing 46 images, we track loss evolution on the most attended image per visit for misclassified validation patients to find influential training examples. We are able to maintain data versatility as we get a score for each image and selectively drop images rather than removing all 46 images of a particular patient.

3 Experiments

3.1 Dataset and Settings

The dataset consisted of 344 screened patients (220 female, 124 male) who each had 1–4 unique visits as part of the study protocol (baseline, weeks 2, 4, and 8), resulting in 844 total visits. Each visit comprised 46 images, leading to a dataset of 38,824 total images. Data were split into training (70%; 610 visits, 247 patients), validation (10%; 64 visits, 28 patients), and test (20%; 170 visits, 69 patients) sets. Skin tones were categorized using Fitzpatrick types: I ($N = 60$), II ($N = 163$), III ($N = 85$), IV ($N = 22$), V ($N = 12$), and VI ($N = 2$). PASI scores were assigned by one of seven dermatologists from a contracted research organization (Reader 1), with inter-rater variability assessed by an independent eighth rater (Reader 2).

3.2 Implementation Details

Our models were trained using PyTorch [19]. Images were resized to 224×224 and normalized with ImageNet statistics: mean $(0.485, 0.456, 0.406)$ and std $(0.229, 0.224, 0.225)$. Training used the Adam optimizer [10] with $\alpha = 10^{-6}$, weight decay 10^{-4}, batch size 4, and 100 epochs on four NVIDIA A100 GPUs. Weighted sampling addressed patient-level imbalanced PASI distributions.

3.3 Downstream Task and Evaluation Metrics

We implement a gradient tracking data attribution algorithm using pretrained PsO classification checkpoints for the Baseline model and misclassified validation examples. We identify the top-k influential training points causing misclassifications by computing influence scores using labels from Reader 1 and retrain our MIL model after removing these k points, comparing its performance to the Baseline. We control the number of removed images to avoid negative effects on training. In our analysis, 9 out of 64 validation patient visits (each with 46 images) are misclassified. For each misclassified visit, we select the image with the highest attention score to identify the top-100, 200, 300, 400, and 500 influential training samples, representing 3-16% of the training dataset (900–4,500 images out of 28,060). We observe an overlap in flagged training images across misclassified cases, leading to fewer removals than the calculated maximum. For example, targeting the top-500 images for each misclassified validation image could remove up to 4,500 images, but due to overlap, we remove only 3,734 unique images. Thus, the actual removal ranges from 2.8% to 13.3% of the training set. This controlled removal keeps the dataset large enough for effective training while reducing harmful samples. We evaluate this on a multiclass PsO severity classification task, finding that removing influential points and retraining improves AUC-ROC and Cohen's Kappa across two readers.

3.4 Detecting Annotation Inconsistencies

Currently, we have 342 patient visits scored by two readers comprising the entire test set (170 patient visits) and 172 visits from the train set. To assess our algorithm's ability to detect annotation inconsistencies, we conduct an experiment by reconstructing our dataset to include the test data that were independently scored by two readers making it have 780 patient visits. This is possible as we do not want to evaluate further, rather just analyze if our method is able to

Table 1. Performance comparison of ConvNeXT and ViT on the test set using Cohen's Kappa and AUC with different numbers of images removed from the training set.

Encoder	k-value	% Training Images Removed	Reader 1		Reader 2	
			Cohen's Kappa	AUC	Cohen's Kappa	AUC
ConvNeXT	Baseline	0.0% (0/28,060)	0.53	0.85	0.56	0.86
	Top 100 removed	2.8% (794/28,060)	0.52	0.88	0.45	0.86
	Top 200 removed	5.6% (1,576/28,060)	0.54	0.88	0.45	0.87
	Top 300 removed	8.2% (2,292/28,060)	**0.62**	**0.90**	**0.54**	**0.89**
	Top 400 removed	10.8% (3,019/28,060)	0.49	0.85	0.46	0.85
	Top 500 removed	13.3% (3,734/28060)	0.42	0.83	0.38	0.80
ViT	Baseline	0.0% (0/28,060)	0.37	0.72	0.39	0.72
	Top 100 removed	2.8% (794/28,060)	0.41	0.75	0.39	0.75
	Top 300 removed	8.2% (2,292/28,060)	0.44	0.81	0.43	0.78
	Top 500 removed	13.3% (3,734/28,060)	**0.61**	**0.89**	**0.52**	**0.84**

flag images with annotation inconsistencies. Following prior work [20], we utilize self-influence scores, which quantify a training sample's influence on its own loss. As computing self-influence scores is computationally heavy, out of these 342 doubly rated patient visits, we pick 100 patient visits at random but maintain the ratio of same label:different label to approximately 84:16 which is the same as that of the doubly rated dataset. For each of these 100 patient visits, we compute a 46 × 46 self-influence matrix and assign the maximum diagonal value as the self-influence score for that visit. Higher self-influence scores are expected to indicate potential mislabeling, enabling systematic identification of annotation discrepancies.

4 Results

In this section, we present the quantitative results for the PsO severity classification task, and highlight the efficacy of our approach in addressing poor annotations. We report the values of micro-average AUC and linearly weighted Cohen's Kappa after removing images from the training data as summarized in 1. The most favorable outcomes were obtained by removing the top 300 images from the training set per validation misclassification for multi-class severity classification using ConvNeXT encoders. Validation AUROC can be used to decide how many images to remove. Removing the top 300 most "harmful" samples per misclassified patient visit yielded the highest validation AUROC of 89.2% with ConvNeXT, versus 81.4% (top 100), 82.6% (200), 78.7% (400), and 76.6% (500). This trend aligns with test performance in Table 1. We show a few additional experiments on the ViT backbone where removing top 500 images performs the best. Fig 2 displays the confusion matrices for two independent raters on a held-out test set, comparing baseline performance with GRASP-PsONet after removing the top 300 images from training per validation misclassified patient using a ConvNeXT based encoder. This approach significantly improves performance, achieving AUC-ROC scores of 88.8% and 90.2% for the two raters. Performance analysis on subgroups, such as skin tone, show that on the test set our framework achieved AUCs of 90.5% for FST I–III (6624 images) and 88.0% for FST IV–VI (1196 images), compared to baseline AUCs of 84.7% and 87.6%. Type V showed the lowest performance due to limited data, while Types I, IV, and VI exceeded 90%.

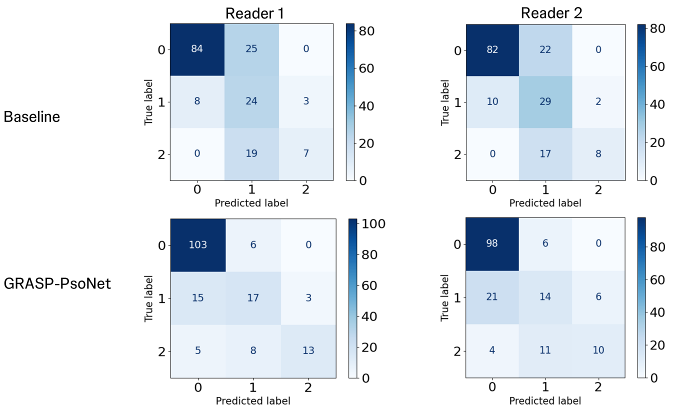

Fig. 2. Confusion matrices for the Baseline and our method on the test set after removing 300 training images per validation misclassification using a ConvNeXT encoder.

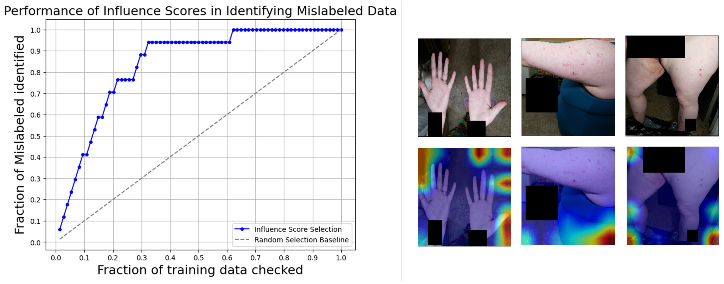

Fig. 3. Left: Proportion of correctly identified mislabeled samples using a rank-ordered list of self-influence scores. **Right:** Example of Grad-CAM maps on training images with inter-rater discrepancies where the model misclassified the label.

A key aspect of our study involves leveraging self-influence scores to correctly identify inter-rater discrepancies within the training dataset. Fig 3 presents a performance curve illustrating the effectiveness of this approach. The x-axis represents the fraction of the training dataset inspected, while the y-axis indicates the cumulative proportion of mislabeled samples identified. The solid blue line demonstrates that by reviewing only the top 30% of ranked samples, we successfully identify over 90% of the mislabeled cases. This result highlights the advantage of self-influence scores in prioritizing label verification efforts, offering an automated strategy for dataset quality checks. We also show heatmaps on some flagged training images. Interestingly, for high-influence images, GRAD-CAM reveals that the model is focusing on incorrect/irrelevant regions, suggesting that these images are problematic and affect training due to spurious correlations.

5 Conclusion and Future Work

In this work, we introduced a novel framework for PsO severity classification that leverages score-based influence functions to refine training data. By tracing gradients from the optimization process, we identify and remove the most influential training images using misclassified images from validation data to improve performance and generalizability. We also demonstrate the capacity of this method to specifically flag images which have discrepancies in annotations. In the future, instead of removing such problematic images, we can send them for re-scoring or quality check, thereby alleviating the need for complete dataset checking and re-scoring.

Acknowledgments. This work was sponsored by Johnson and Johnson Innovative Medicine.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bayasi, N., Hamarneh, G., Garbi, R.: Culprit-prune-net: Efficient continual sequential multi-domain learning with application to skin lesion classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 165–175. Springer (2021)
2. Berth-Jones, J., Grotzinger, K., Rainville, C., Pham, B., Huang, J., Daly, S., Herdman, M., Firth, P., Hotchkiss, K.: A study examining inter-and intrarater reliability of three scales for measuring severity of psoriasis: Psoriasis area and severity index, physician's global assessment and lattice system physician's global assessment. British Journal of Dermatology **155**(4), 707–713 (2006)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
4. Hara, S., Nitanda, A., Maehara, T.: Data cleansing for models trained with sgd. Advances in Neural Information Processing Systems **32** (2019)
5. He, Y., Chen, M., Yang, Z., Lu, Y.: Data-centric diet: Effective multi-center dataset pruning for medical image segmentation. arXiv preprint arXiv:2308.01189 (2023)
6. Holste, G., Jiang, Z., Jaiswal, A., Hanna, M., Minkowitz, S., Legasto, A.C., Escalon, J.G., Steinberger, S., Bittman, M., Shen, T.C., et al.: How does pruning impact long-tailed multi-label medical image classifiers? In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 663–673. Springer (2023)

7. Hossain, M.I., Zamzmi, G., Mouton, P.R., Salekin, M.S., Sun, Y., Goldgof, D.: Explainable ai for medical data: current methods, limitations, and future directions. ACM Computing Surveys **57**(6), 1–46 (2025)
8. Huang, K., Wu, X., Li, Y., Lv, C., Yan, Y., Wu, Z., Zhang, M., Huang, W., Jiang, Z., Hu, K., et al.: Artificial intelligence–based psoriasis severity assessment: Real-world study and application. Journal of Medical Internet Research **25**, e44932 (2023)
9. Kamran, S.A., Lucas, M.V., Lutnick, B., Parmar, C., Pal, B., Shah, A.P., Apfel, D., Fakharzadeh, S., Miller, L., Yip, S., et al.: Pso-net: Development of an automated psoriasis assessment system using attention-based interpretable deep neural networks. arXiv preprint arXiv:2501.18782 (2025)
10. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
11. Koh, P.W., Liang, P.: Understanding black-box predictions via influence functions. In: International conference on machine learning. pp. 1885–1894. PMLR (2017)
12. Li, Y., Wu, Z., Zhao, S., Wu, X., Kuang, Y., Yan, Y., Ge, S., Wang, K., Fan, W., Chen, X., et al.: Psenet: Psoriasis severity evaluation network. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 800–807 (2020)
13. Liu, Y., Jain, A., Eng, C., Way, D.H., Lee, K., Bui, P., Kanada, K., de Oliveira Marinho, G., Gallegos, J., Gabriele, S., et al.: A deep learning system for differential diagnosis of skin diseases. Nature medicine **26**(6), 900–908 (2020)
14. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Advances in neural information processing systems **30** (2017)
15. Owen, A.B.: Sobol'indices and shapley value. SIAM/ASA Journal on Uncertainty Quantification **2**(1), 245–251 (2014)
16. Owen, A.B., Prieur, C.: On shapley value for measuring importance of dependent inputs. SIAM/ASA Journal on Uncertainty Quantification **5**(1), 986–1002 (2017)
17. Pal, A., Chaturvedi, A., Garain, U., Chandra, A., Chatterjee, R.: Severity grading of psoriatic plaques using deep cnn based multi-task learning. In: 2016 23rd International Conference on Pattern Recognition (ICPR). pp. 1478–1483. IEEE (2016)
18. Pal, A., Chaturvedi, A., Garain, U., Chandra, A., Chatterjee, R., Senapati, S.: Severity assessment of psoriatic plaques using deep cnn based ordinal classification. In: International Workshop on Computer-Assisted and Robotic Endoscopy. pp. 252–259. Springer (2018)
19. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
20. Pruthi, G., Liu, F., Kale, S., Sundararajan, M.: Estimating training data influence by tracing gradient descent. Advances in Neural Information Processing Systems **33**, 19920–19930 (2020)
21. Resneck Jr, J., Kimball, A.B.: The dermatology workforce shortage. Journal of the American Academy of Dermatology **50**(1), 50–54 (2004)
22. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)
23. Schaap, M.J., Cardozo, N.J., Patel, A., de Jong, E.M., van Ginneken, B., Seyger, M.M.: Image-based automated psoriasis area severity index scoring by convolutional neural networks. Journal of the European Academy of Dermatology and Venereology **36**(1), 68–75 (2022)

24. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International conference on machine learning. pp. 3319–3328. PMLR (2017)
25. Xing, Y., Zhong, S., Aronson, S.L., Rausa, F.M., Webster, D.E., Crouthamel, M.H., Wang, L.: Deep learning-based psoriasis assessment: Harnessing clinical trial imaging for accurate psoriasis area severity index prediction. Digital Biomarkers **8**(1), 13–21 (2024)
26. Yao, D., Guo, J., Wu, H., Lu, C.: A study examining interrater reliability of the psoriasis area and severity index (pasi) after training with a new handbook for a multicenter clinical trial. In: JOURNAL OF THE AMERICAN ACADEMY OF DERMATOLOGY. vol. 79, pp. AB25–AB25. MOSBY-ELSEVIER 360 PARK AVENUE SOUTH, NEW YORK, NY 10010-1710 USA (2018)
27. Yeh, C.K., Kim, J., Yen, I.E.H., Ravikumar, P.K.: Representer point selection for explaining deep neural networks. Advances in neural information processing systems **31** (2018)
28. Zhou, Y., Sheng, Y., Gao, J., Zhang, X.: Dermatology in china. In: Journal of Investigative Dermatology Symposium Proceedings. vol. 17, pp. 12–14. Elsevier (2015)