Uncertainty-Aware AI for Clinical Decision Support

Rohini Banerjee

CMU-CS-25-102

May 2025

Computer Science Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Artur W. Dubrawski, Chair
László A. Jeni

Submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer Science.

Partial support from U.S. Department of Defense contracts W81XWH-19-C0083 and W81XWH-19-C0101

Keywords: Uncertainty quantification, trustworthy AI, interpretable AI, XAI, medical image segmentation, clinical decision support systems, self-supervised pretraining, deep ensembles, Monte carlo networks, masked image modeling

In loving memory of Amit Bandopadhay.

Abstract

Building interpretable-by-design AI models that intuitively communicate model uncertainty is vital to engendering physician and patient trust. We develop uncertainty-guided deep learning systems for two pertinent healthcare settings. Efficient intravascular access in trauma and critical care is a high-stakes intervention affording minimal tolerance for error. Autonomous needle insertion systems can be useful in austere environments due to the lack of skilled medical personnel. However, inaccuracies in vessel segmentation modeling can result in vessel damage and hemorrhage. The risk can be mitigated via predictive uncertainty estimation to assess model reliability. Thus, we introduce MSU-Net, a novel multistage approach to semantic vessel segmentation in ultrasound images that combines the predictive power of Monte Carlo networks and deep ensembles. We demonstrate significant improvements, 27.7% over the state-of-the-art, while enhancing model reliability through a 20.9% stronger discrimination in epistemic uncertainty between correct and incorrect predictions.

Next, we investigate the robustness of predictive modeling in quantifying the severity of rash manifestations associated with Cutaneous Dermatomyositis (CDM), a rare and currently incurable autoimmune disorder. Given the importance of telemedicine for remote disease monitoring and timely intervention, we address challenges of data scarcity and patient diversity by integrating a novel BERT-style self-supervised learning framework to CNN-based models. Pretrained via masked image modeling on demographically diverse images, our model achieves over a 40% improvement in fine-tuning performance on high-resolution in-clinic hand images from a limited cohort of 23 CDM patients. We achieve 83% accuracy on a held-out patient set, surpassing the clinical benchmark of 70–75% accuracy. To our knowledge, this is the first work to integrate uncertainty estimation into such architectures, enabling robustness under distributional shift in skin tone unseen during fine-tuning.

Our contributions lay the groundwork for developing accurate, statistically rigorous, clinically actionable deep learning models that can be aware of their limitations and communicate this awareness to their users. Future work aims to improve the interpretability of models for equitable clinical decision support.

Acknowledgments

I would like to thank my wonderful advisor, Professor Artur Dubrawski, for guiding me through my research and working with me on a number of fascinating projects including RoboTRAC, the DARPA Triage Challenge (DTC), and DART in collaboration with UPMC. Under his tutelage, I was able to freely explore the exciting field of AI in healthcare and finally find my dream field of work. I also feel incredibly grateful to be a member of the Auton Lab surrounded by such brilliant researchers. At UPMC, I have enjoyed innovating new solutions with Dr. Rohit Aggarwal and Dr. Nantakarn Pongtarakulpanit and I admire their dedication to pioneering new technology for their patients.

To my friends along the years—thank you for consistently lending a shoulder to support me and encouraging me as I explored new beginnings in my research. Thank you for always believing in my potential and celebrating my achievements with such effusive enthusiasm.

Lastly, I would like to thank my parents and brother, who have always been my number one supporters since I started my research journey. I could not have done any of this without their unwavering encouragement and unconditional support. To my father Robindranath, who has always made me feel seen and understood throughout my Master's. To my mother Sangeeta, whose kindness and wisdom carried me through the toughest challenges. To my brother Siddharth, whose unwavering optimism has taught me to smile and always find the silver lining. You have truly guided me to be a better researcher, peer, and person every day throughout my college journey. My success and drive are rooted in the love, guidance, and strength I've received from my family.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Clinical decision support systems (CDSSs) are a network of software systems used primarily at the point-of-care to improve healthcare delivery by enhancing medical decisions with targeted clinical knowledge and demographic patient data. While earlier CDSSs retrieved information from knowledge-based systems, modern technologies leverage the powerful and efficient capabilities of Artificial Intelligence (AI) algorithms. Recent Deep Learning (DL) systems rival the performance of board-certified specialists; they can perform rudimentary diagnostic tasks to detect diabetic retinopathy or arrhythmia on electrocardiograms (Sutton et al., 2020), generate precise vessel segmentations (Banerjee et al., 2025), or perform automated cancerous lesion detection in a fraction of the time of a trained dermatologist (Yu et al., 2017). However, while DL is an attractive solution for solving complex diagnostic tasks, adopting such algorithms in CDSS is challenging for the following reasons;

Clinical data quality. Acquiring high quality, clean data is non-trivial. For visual recognition tasks, noisy image data often arises from inconsistent lighting conditions, a lack of calibration in technology settings, and operator bias. Acquisition artifacts can obscure the desired object in an image, leading to data leakage (Berseth, 2017) or performance degradation in downstream tasks (Wang et al., 2021). Furthermore, the scarcity of high-quality labeled training data is a well-known issue in the medical community, as labeling is time-consuming and requires domain expertise. The quality of labels is critical to model performance; improper labeling can propagate misinformation and result in inaccurate predictions, endangering patient lives and clinical reputations. Consequently, most labeled clinical datasets contain only a few hundred to a few thousand samples, insufficient for training modern, sophisticated architectures. The limited amount of training data provides a weak representation of larger populations, which can diminish the utility of DL in CDSSs. Distribution shifts between training and testing data can significantly affect a model's generalizability, as a model trained on data from one clinic may perform poorly when applied to patients from a different clinic.

Model interpretability. DL architectures are often described as "opaque" or "black-box", alluding to the lack of interpretability of their internal decision-making processes. In healthcare, standalone models can diminish autonomy, requiring clinicians to use model predictions without sufficiently understanding how they are retrieved. Similarly, opacity makes it challenging

to identify preexisting biases, which often go unnoticed and are unknowingly perpetuated. As modern architectures grow in sophistication, the trade-off between interpretability and performance becomes increasingly difficult to balance. In patient-centered care, the transparency and open disclosure of predictive model uncertainty is an ethical and moral imperative (Simpkin and Armstrong, 2019) to address the lack of interpretability in standalone DL models. Uncertainty estimates enable physicians to subjectively abstain from using model predictions heuristically (Kompa et al., 2021), sparsely leveraging a clinician's expertise without causing fatigue. Without a way to communicate its predictive uncertainty, clinicians are likely to exhibit a stronger algorithmic aversion to the CDSS (Dietvorst et al., 2015). Buffering the tendency to abandon an algorithm the first instance of an error through uncertainty estimates is vital for the continued adoption of DL in CDSSs.

Engendering physician trust is two-fold. First, models must be equipped to quantify their uncertainty. Second, and crucially, uncertainty estimates must be reliable. After all, a model that claims high certainty in every prediction is not meaningful. To address these challenges, we design accurate and efficient DL systems for two relevant healthcare settings. Our proposed systems seek to enhance human capabilities by surpassing the performance of novice, semi-trained, and trained clinicians at several notable diagnostic tasks. Thus, our research aims to answer the following questions:

- **RQ1:** What additional considerations do we need to make when designing DL systems for healthcare applications?
- **RQ2:** What methods of uncertainty quantification are most effective in communicating predictive uncertainty?
- **RQ3:** How reliable are the uncertainty estimates of our models?
- **RQ4:** How can we ensure that our systems establish clinical relevance?

1.1 Notable Implications

1.1.1 Saving Lives at Ground Zero

Fluid resuscitation in trauma patients is an urgent high-stakes intervention that demands accurate localization of femoral vessels and affords minimal tolerance for error, as imprecise needle placement can result in complications as severe as catastrophic hemorrhage. In austere environments, autonomous robotic systems can be used to substantially automate vascular access with minimal supervision, yet ensuring precise identification of the optimal needle insertion site without human oversight is an ongoing challenge. To address this, we propose MSU-Net, a novel multistage deep learning framework focused on the first and foundational step of this procedure: high-precision segmentation of femoral vessels. Despite the complexity of this task, our approach achieves a 92.5% Dice Score Coefficient (DSC), along with a statistically significant reduction in false negative rates at the 95% confidence level—key metrics for minimizing the risk of misidentified vessels.

Crucially, our framework advances model trustworthiness, demonstrating a 20.9% improvement in the discrimination of epistemic uncertainty estimates between correct and incorrect predictions. This enhanced calibration is critical for clinician trust in AI-assisted decision-making. Our findings have direct and immediate relevance for the development of safe, autonomous needle-insertion systems. By accurately delineating vascular structures and reducing prediction ambiguity, our method contributes to safer procedural planning by minimizing the risk of puncture in high-risk areas such as the femoral bifurcation. We lay the groundwork for dependable AI-clinician collaboration in time-sensitive trauma care scenarios.

1.1.2 Democratizing Preventative Care via Telemedicine

Delivering timely, high-quality care through telemedicine has the potential to significantly enhance patient outcomes and strengthen clinician-patient relationships. As telehealth continues to redefine healthcare as fast, accessible, and increasingly remote, the integration of AI systems into virtual clinical workflows becomes critical—particularly for the early detection and treatment of rare, debilitating conditions such as Cutaneous Dermatomyositis (CDM). In this work, we propose a novel image-based predictive modeling pipeline that leverages BERT-style self-supervised pretraining to autonomously estimate rash severity in CDM patients. Our approach achieves a 40.6% improvement in accuracy over traditional transfer learning methods and demonstrates strong generalization in out-of-distribution patients. These results exceed clinician standards, suggesting the feasibility of deploying our system in the University of Pittsburgh Medical Center (UPMC) telemedicine platforms to support personalized care.

Chapter 2

Trauma Care in a Rucksack

2.1 Introduction

Trauma, the leading cause of death among young individuals in the U.S. (Wallace and Regunath, 2025), often results in blood loss, which requires rapid fluid resuscitation for vital organ oxygenation. In severe cases, hemorrhage is one of the leading causes of death within the first hour (Verhoeff et al., 2018). Endovascular resuscitation is the most popular approach, with vascular access being the critical first step. Femoral arterial and venous access is the most practical approach (Manning et al., 2021), requiring methodical identification of the femoral vessels to assess the optimal location for needle insertion for cannulation. Compared to other points of access, femoral vessels provide an easily accessible and relatively safe route for rapid intravenous access. However, due to the loss of pulses that can make arterial localization challenging, this procedure requires the expertise of a trained physician.

In austere settings, access to timely medical care and expertise is difficult due to limited access, dangerous conditions, time constraints, and the lack of medical infrastructure. A U.S. Department of Defense analysis of battlefield mortality found that one in four pre-hospital combat deaths and one in two in-hospital combat deaths were potentially preventable (Latif et al., 2023), most of which were attributable to traumatic hemorrhage. During battlefield triage, environments can become hazardous for human intervention due to exposure to chemical or nuclear waste, poor air quality, and the presence of flammable or explosive materials. Autonomous robotic systems can assist in intravenous fluid administration when medical experts are unavailable, providing support in emergencies. These systems can also guide non-experts in accurately performing phlebotomy tasks, empowering them to contribute effectively in dire medical situations.

Ultrasound (US) imaging is widely used for femoral vessel localization for several reasons. First, its affordability, speed, safety and portability make it ideal for interventions, unlike cumbersome CT or MRI imaging systems that use ionizing radiation. Second, US-guided catheterization is commonly used to help distinguish artery from vein in order to prevent inadvertent punctures (Manning et al., 2021). Third, US guidance can help locate the vessel bifurcation; a *safe* needle insertion location is at least 2 cm away from this critical landmark. Despite advances in autonomous needle insertion systems (Chen et al., 2022), a critical challenge persists: inaccurate vessel predictions can have life-threatening consequences. For example, the incorrect prediction

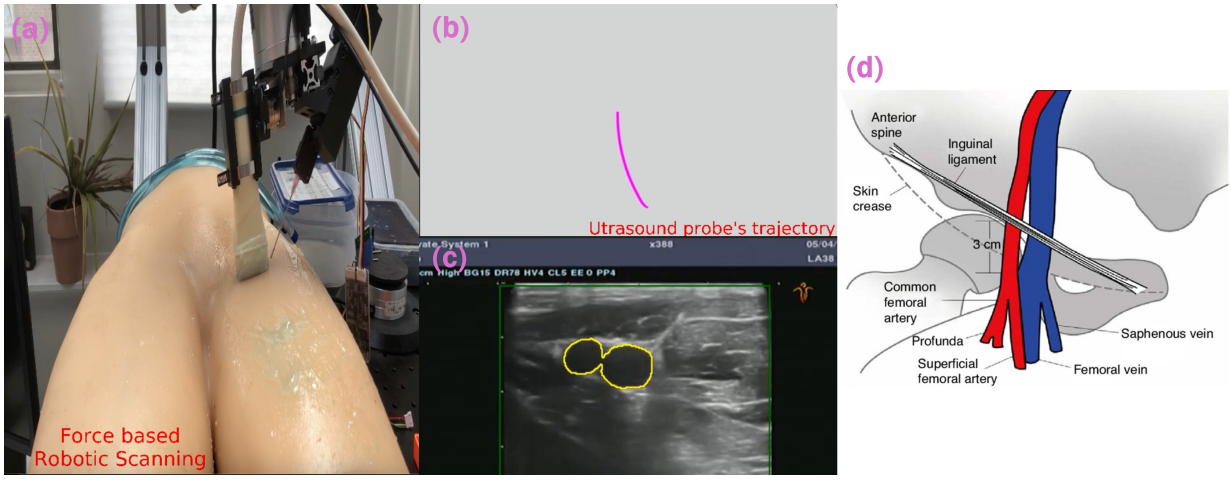


Figure 2.1: **Ultrasound scanning apparatus for vessel localization.** Our apparatus uses (a) 5 MHz linear transducer to scan a phantom model simulating femoral vessels, producing two-dimensional transverse images and recording (b) the US probe trajectory simultaneously. (c) Expert clinicians annotate each image using CVAT and (d) illustrates the ideal femoral arterial puncture site (Chen et al., 2022).

of two adjacent vessels as a single vessel could lead to laceration of the vessel wall and cause hemorrhage upon needle insertion. In severe cases, such hemorrhage can lead to death.

Accurate vessel segmentation is a crucial precursor to vessel localization. First, the US probe scans the patient's femoral region to collect two-dimensional (2D) transverse US images alongside their poses (see Figure 2.1a-c). Next, vessel segmentations are performed for each 2D slice to discriminate vessels from the background. Finally, using both robot poses and our predicted segmentation masks, we can apply interpolation heuristics to localize the bifurcation in order to identify the optimal insertion position.

Uncertainty estimation is essential in vessel segmentation as it provides insight into the reliability of the model's predictions. In cases where the model predicts poor segmentations, uncertainty estimation can help identify regions of the prediction that are less certain or more error-prone. By quantifying uncertainty, the model can highlight areas where further attention or caution is needed, guiding non-experts to avoid making decisions based on unreliable predictions. It follows that uncertainty estimation not only improves model performance, but also ensures patient safety by reducing the risk of harmful, incorrect interventions.

Hence, we introduce MSU-Net, a novel MultiStage Monte Carlo U-Net, performing accurate and efficient semantic vessel segmentations. We equip our models with the capacity to communicate model uncertainty, enabling transparent communication of the model's limitations. Our contributions include **(1)** identifying the first known improvement in uncertainty estimation for ultrasound images and **(2)** demonstrating significant improvements in model performance with MSU-Net. In addition to quantitative achievements, our method produces visual uncertainty maps that are consistent with clinician evaluations of vessels. Notably, we achieve these results with little additional resources required. Thus, our research addresses the following questions:

- ❖ **RQ1:** How can we accurately model complex vessel structures using automated systems?
- ❖ **RQ2:** How can we improve the reliability of uncertainty quantification methods in US image segmentation?
- ❖ **RQ3:** How can we translate quantitative or qualitative uncertainty estimates into intuitive results for non-experts?

2.2 Uncertainty Quantification in Deep Learning

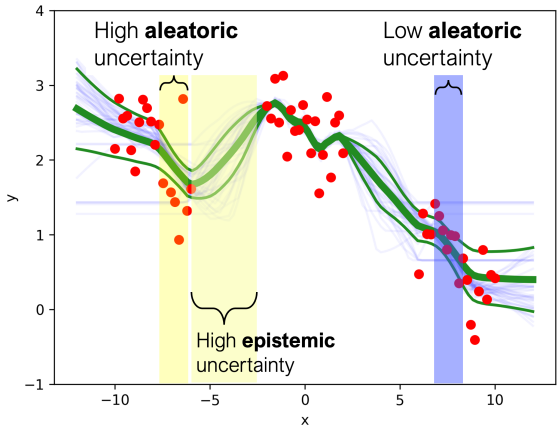


Figure 2.2: **Aleatoric and epistemic uncertainty in deep learning.** We utilize a synthetic regression dataset from (Wilson and Izmailov, 2020) to illustrate aleatoric and epistemic uncertainties. We generate a deep ensemble of 50 independently trained regression neural networks. Each network consists of fully-connected layers with ReLU activations and weights following a Gaussian prior $\omega_j \sim \mathcal{N}(0, \sigma^2)$. A lack of data manifests as high epistemic uncertainty, evidenced by high variability in the predictions of each network.

Uncertainty quantification is vital for assessing model reliability. Traditionally, the frequentist approach relies on a single point estimate of network weights and uses class likelihoods as confidence measures. Consider a segmentation model $f^{\widehat{\omega}}$ with learned weights $\widehat{\omega}$. Given some input $\mathbf{x}^{(i)} \sim \mathcal{D}$ sampled from our training dataset $\mathcal{D}$ and an activation function $\sigma_A : \mathbb{R} \rightarrow [0, 1]$ applied element-wise

$$\widehat{\mathbf{p}}^{(i)} \triangleq \sigma_A \left(f^{\widehat{\omega}} \left(\mathbf{x}^{(i)} \right) \right) \tag{2.1}$$

where $\widehat{\mathbf{p}}^{(i)}$ represents the *confidence* of the model prediction. These probabilities can be further binarized to generate our class predictions. However, these likelihoods often overestimate accuracy (Guo et al., 2017), and the popular metric used to quantify confidence, Expected Calibration Error (ECE), has been criticized for bias and inconsistency (Gruber and Buettner, 2022). This motivates the need for alternative approaches to accurately quantify model uncertainty. Predictive uncertainty decomposes into aleatoric and epistemic components (Ghoshal et al., 2019). Aleatoric uncertainty accounts for inherent noise in observations, while epistemic uncertainty arises from limited training data and model parameter uncertainty. We illustrate these uncertainties through a toy univariate regression dataset in Figure 2.2. Recent advancements in Bayesian inference and Bayesian neural networks have provided robust frameworks to quantify both forms of uncertainty by estimating posterior distributions over model weights.

2.2.1 Exact Bayesian Inference is Intractable

We wish to compute the posterior predictive distribution

$$p(\mathbf{y}|\mathbf{x}, \mathcal{D}) \coloneqq \int_{\omega} p(\mathbf{y}|\omega, \mathbf{x}, \mathcal{D}) p(\omega|\mathcal{D}) d\omega \approx \frac{1}{T} \sum_{t} p(\mathbf{y}|\omega_t, \mathbf{x}, \mathcal{D}) \text{ for } w_t \sim p(\omega|\mathcal{D}) \tag{2.2}$$

but integrating over the state space of ω is computationally intractable, requiring the Monte Carlo Integration (MCI) of T random samples described at the end of Equation 2.2. This equation is also referred to as the Bayesian model average (BMA). MCI estimates the integration over ω by averaging the values of model predictions $p(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}, \omega_t)$ evaluated at random samples. Although in expectation MCI converges to the true predictive distribution, it is infeasible to generate a sufficient number of random samples in practice. Markov Chain Monte Carlo (MCMC) is a popular alternative in which dependent samples are generated by a Markov chain, and the samples converge to the target posterior distribution. Yet, MCI and MCMC are only worth the prohibitive cost for experiments with shallow networks when substantial computational resources are available. Next, we see how variational inference can address these limitations.

2.2.2 Variational Inference

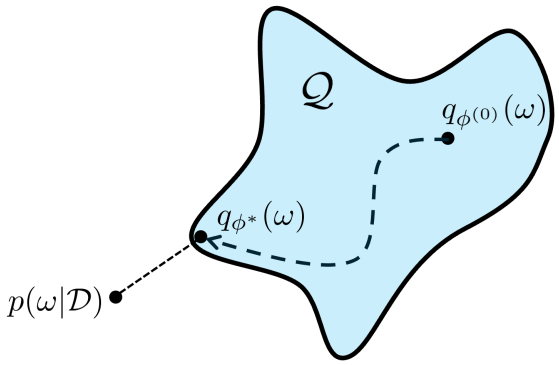


Figure 2.3: **Visualizing variational inference (VI).** We posit a family of approximate distributions $\mathcal{Q}$ and select our initial distribution $q_{\phi^{(0)}}(\omega) \in \mathcal{Q}$. VI iteratively minimizes the KL divergence between the approximating distribution and the conditional likelihood $p(\omega|\mathcal{D})$ to find the optimal $q_{\phi^*}(\omega)$ that lies within $\mathcal{Q}$. In practice, however, we equivalently maximize the ELBO criterion (Equation 2.3).

Consider parameterizing the (often) intractable posterior $p(\omega|\mathcal{D})$ from Equation 2.2 by a family of approximate densities $\mathcal{Q}$. Hence, we can reformulate approximating the posterior as a standard optimization problem by minimizing the divergence between the true posterior and our parameterized version $q_\phi(\omega)$. Kullback-Leibler (KL) divergence is a natural choice for this criterion, as it measures the dissimilarity between two probability distributions p and q. However, in practice, we maximize the evidence lower bound (ELBO) as a proxy for the sake of tractability

$$\text{ELBO}(q_\phi) \coloneqq \mathbb{E}_{q_\phi}[\log p(\omega, \mathcal{D})] - \mathbb{E}_{q_\phi}[\log q_\phi(\omega)] \tag{2.3}$$

which is equivalent to minimizing the KL divergence (see Appendix A.1.2 for derivation). Finally, we optimize for

$$q_{\phi^*}(\omega) \triangleq \underset{q_\phi(\omega) \in \mathcal{Q}}{\text{argmax}} \text{ ELBO}(q_\phi) \tag{2.4}$$

This procedure is known as Variational Inference (VI), and the member $q_{\phi^*} \in \mathcal{Q}$ that achieves the tightest bound serves as our best approximation.

However, there are inherent limitations to this approach. VI is sensitive to the choice of parameterization. If the variational family cannot adequately capture the true posterior, the approximation will be poor, leading to biased or inaccurate results. Furthermore, while VI tends to be faster than MCMC, it does not benefit from the asymptotic guarantee of achieving the target distribution that MCMC is guaranteed (Blei et al., 2017).

2.2.3 Monte Carlo Dropout

(Gal and Ghahramani, 2016) introduced Monte Carlo Dropout (MCD) for approximate Bayesian inference in deep neural networks, using dropout to generate stochastic forward passes that approximate the VI solution. In dropout, for a given layer ℓ in a neural network, the nodes in ℓ are randomly retained with a probability of p and dropped with a probability of $1 - p$. (Damianou and Lawrence, 2013) showed that dropout at every weight layer is mathematically equivalent to an approximation of a probabilistic deep Gaussian process.

Formulating MCD for convolutional neural networks relies on strategic modifications to the standard architecture. First, dropout layers must be situated at each convolutional and fully-connected layer. Here, the node retention rate p_ℓ for each layer ℓ acts as a tunable hyperparameter. Next, training the network with L_2 weight decay and dropout is equivalent to minimizing the ELBO criterion, which is the desired VI objective. In this way, dropout implicitly optimizes a variational approximation to the desired posterior distribution.

Bayesian approximation using MCD has been extensively applied: (Kendall et al., 2016) developed Bayesian SegNet for scene understanding, while (Dechesne et al., 2021) used it in U-Net for high-accuracy image segmentation. Yet, single-model architectures are now supplanted by model ensembles due to difficulties in capturing inherent variability.

2.2.4 Deep Ensembles

An alternative to multiple forward passes of the *same* network through MCD is to train multiple *independent* networks to assemble a deep ensemble. Deep ensemble members, formed by Maximum A Posteriori (MAP) retraining of the same architecture multiple times, converge at different local minima. It follows that deep ensembles are BMA (Wilson and Izmailov, 2020).

(Lakshminarayanan et al., 2017) found that deep ensembles produced more accurate and better calibrated predictive distributions compared to MCD. Figure 2.4 illustrates the ability of deep ensembles to identify multiple modes of the loss landscape over MCD.

Ensemble performance improvements are positively correlated with the diversity of its members. Ensembling multiple models offers various strategies to encourage diversity and generalizability. Techniques such as bagging (Breiman, 1996), stacking (Wolpert, 1992), and boosting (Freund and Schapire, 1999) achieve this diversity. In their work, (Lakshminarayanan et al., 2017) found that selecting different random weight initializations achieved the best performance.

Of course, this technique is considerably more computationally intensive than MCD, as it requires training and storing multiple models for inference. However, training can be easily parallelized with multiple GPUs and (Lakshminarayanan et al., 2017) found that performance gains

plateaued after an ensemble size of $M = 10$ deep networks. In contrast, $M >> 100$ non-DL models are required to achieve comparable performance on the same task. We acknowledge that this technique may not be computationally feasible in all contexts. However, for our purposes, the benefits of deep ensembles are worth the relatively minimal increase in computational costs.

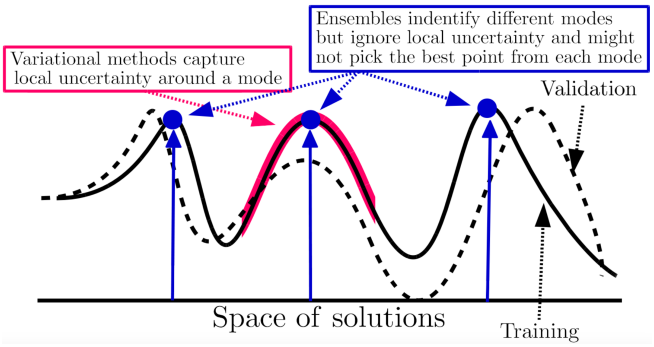


Figure 2.4: **Monte carlo dropout vs. deep ensembles in the loss landscape.** Toy illustration from (Fort et al., 2020). x-axis indicates ω values and y-axis plots the NLL. Deep ensembles formed by MAP estimation are more likely to identify different modes of loss compared to dropout which relies on approximating local variational inference methods.

2.3 Related Work

2.3.1 Image Segmentation

The popular U-Net architecture (Ronneberger et al., 2015) continues to be the gold standard for medical image semantic segmentation. The U-Net is a convolutional neural network consisting of a contracting path to capture context and a symmetric expanding path to enable precise localization. Together, these paths constitute its distinct encoder-decoder structure.

The encoder consists of convolutional layers followed by max-pooling layers to downsample and extract hierarchical features. Each layer learns a different part-based representation of the original input; shallow layers extract low-level features such as edges, textures, and shapes, whereas deeper layers extract full object components. Subsequently, the decoder gradually increases the spatial resolution of feature maps through upsampling and skip connections to reconstruct the final segmentation mask. As a result, the U-Net produces a predicted map of size $(H \times W \times L)$ for an input of size $(H \times W)$ with L distinct semantic classes.

The U-Net architecture features a *copy-and-crop* operation enabling it to segment fine-grained details and learn precise boundaries, making it an ideal choice for our application. First, skip connections between the encoder and decoder components help the model retain important spatial information from the encoder, which is essential for accurate segmentation of fine structures in medical images. The encoder processes the image and progressively reduces its spatial resolution to capture high-level features. The features of each encoder block are then *copied* to the symmetric decoder block, allowing the decoder to use this high-resolution information during upsampling. This enables the decoder to focus on more precise localization of objects. To adjust for different sizes of outputs due to a loss of spatial resolution, a *crop* operation is utilized to align the spatial dimensions of the feature maps of the encoder and decoder.

The U-Net is particularly effective due to its strong performance in limited data environments. This is beneficial for supervised medical imaging tasks that require costly expert annotations. Its

ability to perform with limited data can considerably reduce the time that expert clinicians spend on the laborious task of manually generating annotations.

The U-Net architecture has been successfully integrated in a variety of medical image segmentation tasks. In 2022 alone, almost 3,000 research works cited U-Net as a baseline (Azad et al., 2024). (Zhang et al., 2024) devised VM-UNet, leveraging state-space models to address limitations in long-range modeling capabilities and demonstrating competitive results in skin lesion segmentation on the ISIC 2017 and 2018 archive datasets. (AL Qurri and Almekkawy, 2023) improved U-Net performance on CT scan and US datasets by incorporating attention and spatial normalization mechanisms. (Arun et al., 2023) evaluated 3D Bayesian U-Nets on point-of-care ultrasound (POCUS), motivating further exploration in austere settings.

2.3.2 Multistage Neural Network Ensembles

Although deep ensembles improve uncertainty estimation over single models (Lakshminarayanan et al., 2017), they are limited by naïve aggregation strategies such as simple or weighted averaging or majority vote. (Yang et al., 2002) improves traditional ensembles by using a secondary neural network to adaptively assign weights, leveraging the flexibility and nonlinear modeling of neural networks. (Lai et al., 2006) designed a multistage reliability-based neural network ensemble learning approach to discriminate good from poor creditors. Similarly, (Yin et al., 2022) designed Paw-Net, a two-stage ensemble for semantic segmentation, which achieves higher Intersection-over-Union (IoU) scores by integrating outputs of multiple U-Nets specialized in different classes.

2.4 Multistage Monte Carlo U-Net (MSU-Net)

Our MSU-Net architecture draws on these prior works to leverage the capabilities of both deep ensembles and MCD for effective vessel segmentation. First, we overproduce a set of candidate U-Net models. Next, we select members with the strongest diversity based on our decorrelation maximization algorithm to construct a deep ensemble. Finally, we combine their outputs using a final U-Net fitted with MCD for precise segmentation. This constitutes our three-stage strategy, as shown in Figure 2.5.

2.4.1 Bootstrapping Diverse Networks

Ensemble strength is determined by member diversity. (Lakshminarayanan et al., 2017) suggests using random weight initializations to achieve this diversity. However, to manage limitations due to data scarcity, we leverage encoder weights pretrained on the ImageNet1k dataset for each U-Net. Fine-tuning from pretrained weights is common practice to achieve faster convergence speed and improved performance on the target task (Hidy et al., 2024). As a consequence, diversity through different weight initializations is inapplicable. Instead, we train multiple models, or *candidates*, using bootstrapping, inspired by the "bootstrap aggregating" (bagging) procedure. This is achieved by repeatedly sampling the training dataset with replacement to generate

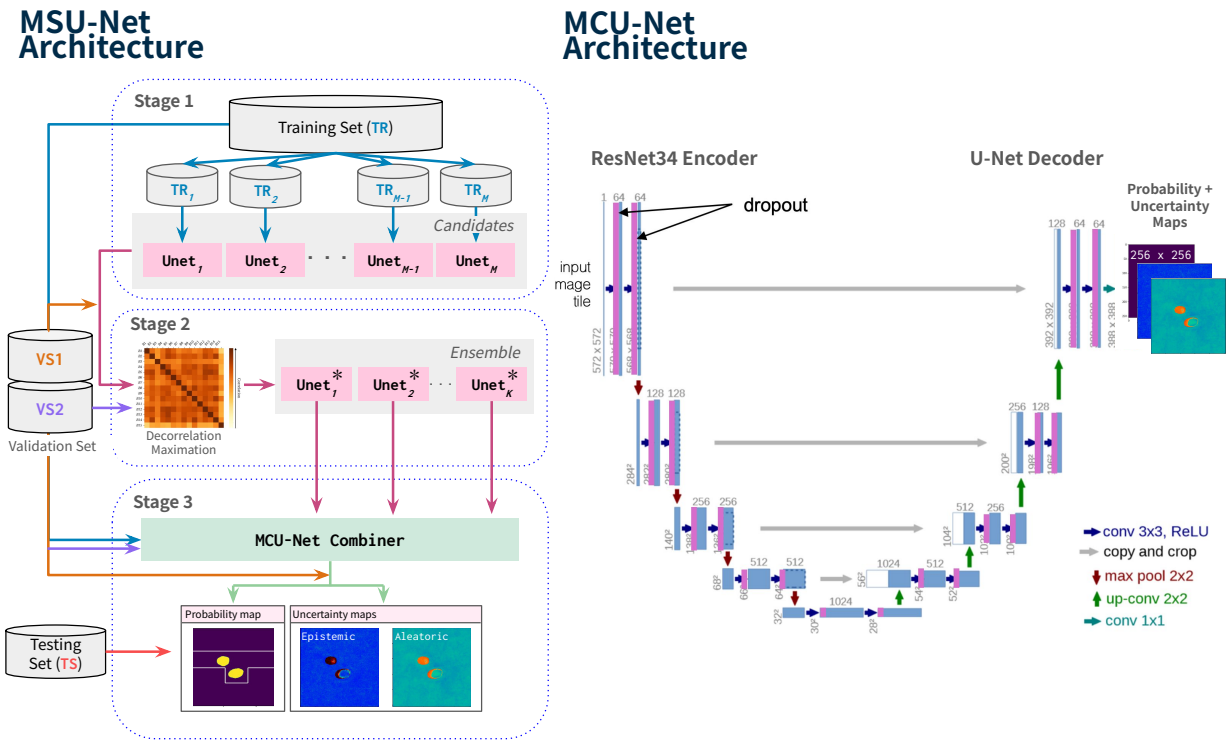


Figure 2.5: **Proposed MultiStage Monte Carlo U-Net (MSU-Net) architecture.** Candidate U-Nets are trained on bootstrap samples from the original training set (TR) and externally validated on VS1. Decorrelated ensemble members are chosen using VS2 to compose our final deep ensemble. A Monte Carlo U-Net (MCU-Net), shown on the right, is trained on ensemble outputs to predict the final segmentation mask. We rigorously validate our results on an independent, held-out testing set (TS).

M training subsets, each as large as the original training dataset (further details provided in Appendix A.3).

We train $M = 15$ bootstrapped models (Lakshminarayanan et al., 2017) and optimize hyperparameters, such as training epochs, using the held-out VS1 validation set. Early stopping is implemented to prevent each candidate model from overfitting. We will refer to the set of M candidates by $\{f^{\widehat{\omega}_1}, f^{\widehat{\omega}_2}, \cdots, f^{\widehat{\omega}_M}\}$.

2.4.2 Decorrelation Maximization

In stage 2, we select $K \leq M$ candidates to form a diverse and efficient ensemble, aiming to minimize correlation and reduce computational costs. Using a modified decorrelation maximization method from (Lai et al., 2006), we compute the Brier score loss matrix on $\{f^{\widehat{\omega}_1}, f^{\widehat{\omega}_2}, \cdots, f^{\widehat{\omega}_M}\}$ using a validation set VS2 within a specified region of interest (ROI) selected to address class imbalance (refer to Appendix A.4.2 for details). We calculate the mean, variance, and covariance of the Brier scores to build the correlation matrix R, which quantifies the strength of correlations within model pairs. The matrix R is represented in block form for each candidate $f^{\widehat{\omega}_i}$ as

$$R \xrightarrow{\text{extract principal submatrix}} \begin{bmatrix} R_{-i} & r_i \\ r_i^T & 1 \end{bmatrix} \tag{2.5}$$

where R_{-i} is the principal submatrix of R resulting from deleting the ith row and ith column. Subsequently, we compute the plural-correlation coefficient

$$\rho_i^2 := r_i^T R_{-i}^{-1} r_i \tag{2.6}$$

by which candidate $f^{\widehat{\omega}_i}$ is kept if $\rho_i^2 \leq \gamma$ for some threshold γ, else discarded, in order to build our final ensemble. Empirical trials testing γ thresholds indicate that $K = 3$ performs almost equally as well as $K = 15$, suggesting that minimal additional training will suffice for our new architecture. Our full procedure is detailed in Algorithm 1.

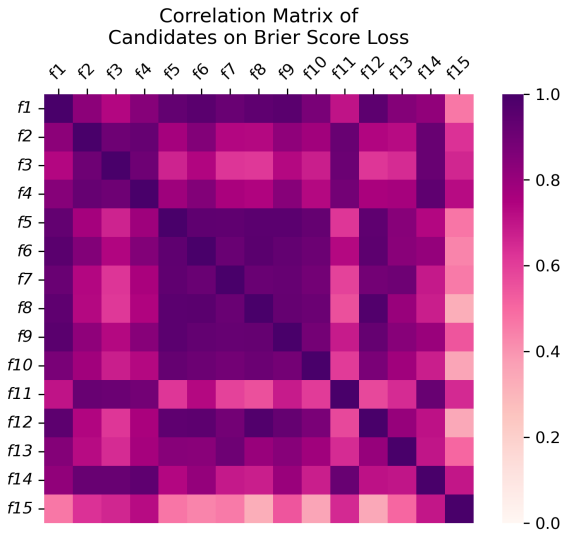


Figure 2.6: **Pair-wise correlation matrix of brier score loss error on VS2.** Each candidate U-Net is evaluated on VS2 and the correlation of their losses is subsequently computed. Darker colors indicate stronger positive correlation, whereas lighter colors indicate weaker correlation. Thus, loss correlation is utilized as a viable proxy for model diversity. We aim to select models that minimize loss correlation.

Algorithm 1 Decorrelation maximization

Input: Candidate models $\{f^{\widehat{\omega}_1}, f^{\widehat{\omega}_2}, \cdots, f^{\widehat{\omega}_M}\}$, correlation matrix R, threshold γ
Output: Chosen ensemble members $\{f_1^*, f_2^*, \cdots, f_K^*\}$
`ensemble` $\leftarrow$ `[]`
for $f^{\widehat{\omega}_i}$ in $\{f^{\widehat{\omega}_1}, f^{\widehat{\omega}_2}, \cdots, f^{\widehat{\omega}_M}\}$ **do**
    Rewrite R in block matrix form for candidate model $f^{\widehat{\omega}_i}$, i.e. $R \rightarrow \begin{bmatrix} R_{-i} & r_i \\ r_i^T & 1 \end{bmatrix}$
    $\rho_i^2 \leftarrow r_i^T R_{-i}^{-1} r_i$
    if $\rho_i^2 \leq \gamma$ **then**
        `ensemble.append`$\left(f^{\widehat{\omega}_i}\right)$
    else
        Discard $f^{\widehat{\omega}_i}$
    end if
end for
return `ensemble`

We illustrate pair-wise correlations of our candidate U-Nets in Figure 2.6. Note that lighter colors denote weaker correlations and vice versa. For example, $f^{\widehat{\omega}_{15}}$ appears to have the weakest correlations with all other candidates. We present a detailed analysis of patterns in loss correlations in Appendix A.4.1.

2.4.3 Monte Carlo U-Net (MCU-Net)

In stage 3, we train a final Monte Carlo U-Net (MCU-Net) combiner taking ensemble member outputs as inputs and outputting segmentation and uncertainty maps.

We introduce stochasticity into our inference process by situating dropout layers after the ReLU activation in each convolutional and fully-connected layer of the U-Net architecture. Empirical tests indicate that drop rates of 0.4 and 0.5 provide the most training stability while enabling enough stochasticity for meaningful results. These dropout layers are turned on during *both* training and inference. We then perform T forward passes, or Monte Carlo samples, of the MCU-Net during inference to compute (1) the average model prediction and (2) the model uncertainty maps. Our empirical results show no significant improvement beyond $T = 30$. The average model prediction is computed by

$$\bar{\mathbf{p}}^{(i)} \triangleq \frac{1}{T} \sum_{t=1}^{T} \sigma_A \left(f^{\widehat{\omega_t}} \left(\mathbf{x}^{(i)} \right) \right) \tag{2.7}$$

for an input image $\mathbf{x}^{(i)}$ and activation function $\sigma_A : \mathbb{R} \rightarrow [0, 1]$ applied element-wise. Notice how our predicted probability map in Equation 2.7 is equivalent to the frequentist notion of confidence defined in Equation 2.1.

Predictive uncertainty. We can generate pixel-wise predictive uncertainty maps during inference. In the binary segmentation problem, each pixel can be treated as an independent binary

classification instance, making $\mathbf{y}$ discrete. Thus, our predictive uncertainty decomposes to

$$
\begin{aligned}
Var_{\Omega}(\mathbf{y}) \triangleq & \underbrace{\frac{1}{T}\sum_{t=1}^{T}\left[\text{diag}\left\{p(\mathbf{y} \mid \mathbf{x}, \widehat{\omega}_t\right\} - p(\mathbf{y} \mid \mathbf{x}, \widehat{\omega}_t)^{\otimes 2}\right]}_{\text{aleatoric uncertainty}} \\
& + \underbrace{\frac{1}{T}\sum_{t=1}^{T}\left[p(\mathbf{y} \mid \mathbf{x}, \widehat{\omega}_t) - p_{\widehat{\omega}}(\mathbf{y} \mid \mathbf{x})\right]^{\otimes 2}}_{\text{epistemic uncertainty}}
\end{aligned}
\tag{2.8}
$$

for some parameterization $\omega \in \Omega$ where $p^{\otimes 2} \triangleq pp^T$. Let $\widehat{\mathbf{p}}_t$ represent the predicted probability maps from the t-th forward pass, thus we substitute $\widehat{\mathbf{p}}_t$ for $p(\mathbf{y} \mid \mathbf{x}, \widehat{\omega}_t)$ in Equation 2.8. Additionally, MCD computes an approximation to the predictive posterior distribution, hence we can substitute in Equation 2.7 for $p_{\widehat{\omega}}(\mathbf{y} \mid \mathbf{x})$ above. Aleatoric uncertainty captures the inherent randomness within our data, while epistemic uncertainty captures uncertainty across the chosen model parameters.

Since our baseline is a single MCU-Net, we compute the segmentation and uncertainty maps of our baseline analogously to MSU-Net.

2.5 Fine-tuning Setup

For image segmentation, we use U-Net with a ResNet34 backbone in Pytorch from the Segmentation Models library (Iakubovskii, 2019) with encoders pretrained on ImageNet1k. We perform fine-tuning experiments using NVIDIA RTX A6000 GPUs.

2.5.1 Dataset

We used a US scanning system to scan the CAE Blue Phantom anthropomorphic gel model simulating human femoral vessels. Equipped with a 5MHz linear transducer, the system can scan up to 5 cm in depth, producing 2D transverse ultrasound images (see setup in Figure 2.1). Expert clinicians annotate these images using the Computer Vision Annotating Tool (CVAT) (Sekachev et al., 2020), which are subsequently cropped and resized to 256×256 pixels. The dataset is divided into training, validation, and testing subsets. The validation set is further randomly divided into two disjoint sets, VS1 and VS2.

2.5.2 Loss Specification

We place a Gaussian prior on the weights of each candidate U-Net in stage 1 by optimizing with L_2 weight decay to mimic MAP estimation. This is necessary to allow members to explore different basins of attraction. We similarly train the final MCU-Net combiner in stage 3 with added dropout to implicitly optimize a variational approximation to the posterior predictive distribution. For binary segmentation, we utilize the binary cross-entropy loss criterion. However, our

dataset exhibits severe class imbalance due to a considerably larger proportion of background (negative) pixels than vessels per frame. On average, less than 4% of all pixels are vessels (see Appendix A.4.2). To alleviate class imbalance, we also include a term for Dice loss, which is robust and beneficial for segmentation tasks with imbalanced labels. We combine both losses to balance between achieving pixel-wise accuracy and boundary alignment

$$\mathcal{L}_{\text{seg}} \triangleq \underbrace{-\frac{1}{N}\sum_{k=1}^{N} y_k \cdot \log \hat{p_k} + (1-y_k)\cdot \log\left(1-\hat{p_k}\right)}_{\mathcal{L}_{\text{BCE}}} + \underbrace{1 - \frac{2\sum_{k=1}^{N} y_k \hat{p_k} + \epsilon}{\sum_{k=1}^{N} y_k^2 + \sum_{k=1}^{N} \hat{p_k}^2 + \epsilon}}_{\mathcal{L}_{\text{DICE}}} + \frac{\lambda}{2}||\omega||_2^2 \tag{2.9}$$

where ϵ is added to maintain numerical stability during training. For smoothed Dice loss, we set $\epsilon = 1.0$. For competitive convergence rates, we opt for the AdamW optimizer initialized with $\beta_1 = 0.9, \beta_2 = 0.999$ exponential decay rates, and a learning rate of $1e^{-}4$. Appendix A.4.4 specifies the exact hyperparameter settings for the experimental results found in Section 2.7.

2.6 Evaluation

2.6.1 Quantifying Uncertainty Quality

In the frequentist paradigm, given $\widehat{p_{ij}} = \sigma_A\left(f^{\widehat{\omega}}(x_{ij})\right)$, the Expected Calibration Error (ECE)

$$\text{ECE} \coloneqq \sum_{m=1}^{M} \frac{|B_m|}{n} \text{ abs} \left(\underbrace{\frac{1}{|B_m|}\sum_{i\in B_m} \mathbb{1}\left[\text{bin}(\widehat{p_{ij}}) = y_{ij}\right]}_{\text{acc}(B_m)} - \underbrace{\frac{1}{|B_m|}\sum_{i\in B_m} \widehat{p_{ij}}}_{\text{conf}(B_m)} \right) \tag{2.10}$$

is commonly used to evaluate model calibration. However, deep networks notoriously overestimate their prediction accuracy, resulting in statistically biased ECE estimates. An apparent source of bias is extreme class imbalance, which is particularly relevant to our problem context. The vast number of "easy" background pixels can artificially inflate the proportion of highly confident predictions, reducing the ECE. We illustrate this deficiency in Appendix A.6.1. Moreover, the ECE value is highly sensitive to the bin count M.

However, Bayesian (and adjacent) approaches do not have well-defined or consistent metrics to evaluate the quality of model uncertainty estimates, evidently due to the "black-box" nature of DL. Model quality depends on accurately reflecting uncertainty: low for correct predictions and high for incorrect ones, improving calibration and user trust. As such, we devise a novel two-fold approach that evaluates both the strength and quality of the model uncertainty estimates. First, we use the Rényi divergence (RD) statistic, a generalization of KL divergence, to measure the dissimilarity between the epistemic uncertainty distributions of correct (p) from incorrect (q) predictions. RD quantifies the compression gain achievable by mixing two codes p and q (van Erven and Harremos, 2014). Second, we examine whether the mean epistemic uncertainty for the set of correct predictions is significantly lower than the mean of incorrect predictions.

Distribution divergence estimation. We use a nonparametric estimator of RD that is conditionally L_2-consistent using only k-nearest-neighbor statistics to reduce computational effort (Poczos and Schneider, 2011). For i.i.d. samples $X_{1:n_0} = (X_1, \cdots, X_{n_0})$ from a distribution with density p and $Y_{1:n_1} = (Y_1, \cdots, Y_{n_1})$ from a distribution with density q, $\rho_k(i)$ denotes the k-th nearest neighbor of observation X_i in $X_{1:n_0}$ and $v_k(i)$ the k-th nearest neighbor of X_i in $Y_{1:n_1}$. With $B_{k,\alpha} = \frac{\Gamma(k)^2}{\Gamma(k-\alpha+1)\Gamma(k+\alpha-1)}$ where $\Gamma(x) = (x-1)!$, we can estimate RD by

$$\widehat{R}_\alpha(p \,||\, q) \triangleq \frac{1}{\alpha - 1} \, \log \left(n_0^{-1} \sum_{i=1}^{n_0} \left(\frac{(n_0 - 1)\rho_k(i)}{n_1 v_k(i)} \right)^{1-\alpha} B_{k,\alpha} \right) \tag{2.11}$$

The hyperparameter α controls how much the divergence measure weighs different parts of the distributions. We select $k = 4$ for k-nearest neighbors and $\alpha = 0.85$ (Poczos and Schneider, 2011) for numerical stability and robustness against heavy tails. We develop a vectorized approach to computing estimated RD, achieving up to 176x faster execution (Appendix A.2).

Distribution mean estimation. In addition to measuring the distributional divergence, we also verify whether the mean epistemic uncertainty for correct predictions is significantly lower than the mean uncertainty for incorrect predictions. To do so, we use kernel density estimation with a Gaussian kernel and the optimal bandwidth via Silverman's rule of thumb over $100,000$ samples. We quantify the difference in mean epistemic uncertainty by

$$\Delta\widehat{\mu} \triangleq (\widehat{\mu}_{\text{incorr}} - \widehat{\mu}_{\text{corr}}) \tag{2.12}$$

2.6.2 Model Performance Metrics

Dice score coefficient (DSC). DSC is a similarity metric commonly used in image segmentation to measure the degree of overlap between two sets A and B. In the segmentation context, we can flatten the binarized predictions $\widehat{\mathbf{p}}$ and ground truth masks $\boldsymbol{y}$ and evaluate the DSC by

$$DSC \coloneqq \frac{2\sum_{k=1}^{N} y_k \hat{p}_k + \epsilon}{\sum_{k=1}^{N} y_k^2 + \sum_{k=1}^{N} \hat{p_k}^2 + \epsilon} \tag{2.13}$$

Jaccard index. Commonly referred to as the Intersection-over-Union (IoU) score, the Jaccard index is another popular metric to evaluate model performance in image segmentation tasks. IoU and DSC are very similar, however IoU penalizes under- and over-segmentation more than DSC. We evaluate IoU for the sake of completeness by

$$IoU \coloneqq \frac{DSC}{2 - DSC} \tag{2.14}$$

Type I and Type II errors. We compute the true negatives (TN), true positives (TP), false negatives (FN) and false positives (FP) by considering individual pixels as independent binary classification instances. We then calculate the type I error rate, or the rate of false positives (FPR), and the type II error rate, or the rate of false negatives (FNR).

2.6.3 Bootstrapping 95% Confidence Intervals

Apart from leveraging bootstrapping to inject diversity into our ensemble members, we also use bootstrapping to construct confidence intervals on our test statistics. Using the bootstrapping procedure described in Section A.3, we construct $B = 1000$ bootstrap samples of the testing set, $\{\mathcal{D}_{\text{TS}}{}'^1, \mathcal{D}_{\text{TS}}{}'^2, \cdots, \mathcal{D}_{\text{TS}}{}'^B\}$ on which we evaluate our metrics defined in Section 2.6.2. Setting $\alpha = 0.05$, we use the percentile method to construct $(1-\alpha)\%$ confidence intervals

$$[\, Q_{\tilde{t}}(\alpha/2),\ Q_{\tilde{t}}(1-\alpha/2)\,] \tag{2.15}$$

where $Q_{\tilde{t}}$ is the quantile function on our test statistic $\tilde{t}$.

Unfortunately, while the naïve bootstrap is a powerful inferential tool, it cannot be employed to construct 95% confidence intervals on our RD estimator. Nearest neighbor estimators are sensitive to perturbations in the underlying distribution, therefore their limited variance cannot be consistently estimated by a naïve Efron-type bootstrap (Abadie and Imbens, 2008). Since this behavior may result in a non-negligible positive bias in bootstrap estimates, we instead apply a direct M-out-of-N (MooN) type bootstrap (Walsh and Jentsch, 2023) shown in Algorithm 2.

Algorithm 2 M-out-of-N (MooN) bootstrapping

Input: Distributions p, q, degree of undersampling $\gamma \in (0, 1]$
Output: Bootstrap estimates of nonparametric Rényi divergence statistic
$n_0, n_1 \leftarrow |p|, |q|$
$\alpha_n \leftarrow \frac{n_0}{n_1}$
$N^* \leftarrow \left\lfloor (n_0 + n_1)^\gamma + \frac{1}{2} \right\rfloor$
$n_0^* \leftarrow \left\lfloor \frac{\alpha_n}{1+\alpha_n} N^* + \frac{1}{2} \right\rfloor$
$n_1^* \leftarrow N^* - n_0^*$
`samples` $\leftarrow$ `[]`
for i in range 1000 **do**
    `boot_p` $\leftarrow$ `resample(`p`, n_samples=`n_0^*`)` with replacement
    `boot_q` $\leftarrow$ `resample(`q`, n_samples=`n_1^*`)` with replacement
    `samples` $\stackrel{+}{=}$ `[`$\widehat{R}_{\alpha=0.85}$`(boot_p, boot_q)]`
end for
return `samples`

For MooN-type bootstrap, $\gamma = 0.8$ is selected to maintain the largest proportion of original data while achieving the closest coverage probability of 0.950 for 95% confidence intervals.

Finally, we conduct permutation tests to assess the statistical deterministicity of our model quality results. For total $B = 1000$ iterations, we combine all correct and incorrect predictions into a single pool, randomly assign labels as if there were no significant differences between the two groups, and calculate the approximate RD value. We can then compute the empirical p-value from all B iterations to assess the likelihood of achieving our results.

2.7 Results

2.7.1 Improving Clinical Relevance via Model Precision

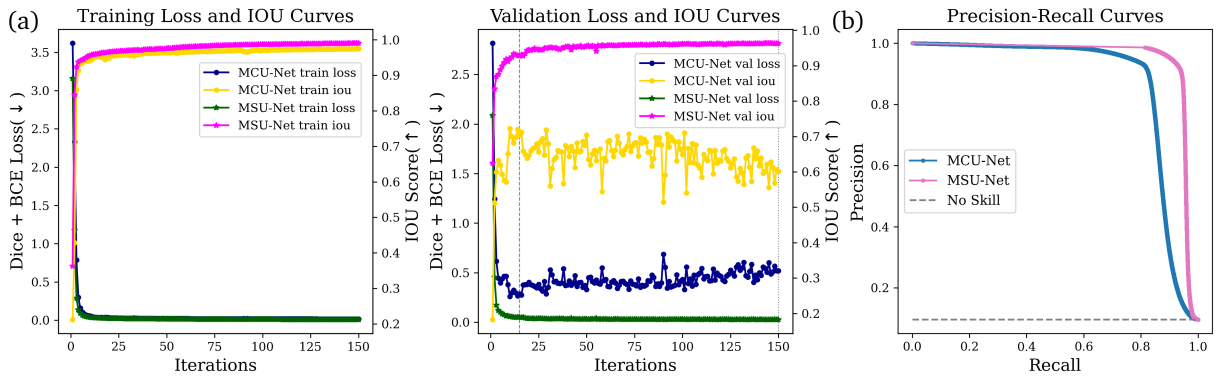


Figure 2.7: **Training and validation curves for MSU-Net vs. MCU-Net.** (a) Training (left) and validation (right) curves for MSU-Net and MCU-Net. Combined loss and DSC values computed on VS1 are plotted at the end of every epoch. Early stopping is delineated by gray lines. (b) Precision-recall curves indicate improved MSU-Net performance.

Figure 2.7a shows training behavior for both models. MSU-Net shows a more stable convergence and consistently outperforms MCU-Net during validation performed after each epoch. We additionally utilize precision-recall curves, seen in Fig. 2.7b, which are resilient to unbalanced classes since they only focus on positive class predictions. Performance results are displayed in Table 2.1. MSU-Net achieves a 27.7% better mean DSC and 18.1% better mean IoU. We achieve significant improvements in sensitivity and false negative rate scores at alpha level 0.05, while other metrics remain similar. We refer to Appendix A.5.1 for confusion matrices for both architectures.

	DSC(↑)	IoU(↑)	Specificity(↑)	Sensitivity(↑)	FPR(↓)	FNR(↓)
MCU-NET	0.648	0.679	**0.998**	0.673	**0.002**	0.327
MSU-NET	**0.925**	**0.860**	0.996	**0.890**	0.004	**0.110**

Table 2.1: **Model performance on test dataset.** Arrow indicates direction of better performance.

2.7.2 Improving Reliability of Epistemic Uncertainties

Our model quality results are shown in Table 2.2. Permutation tests for MCU-Net and MSU-Net, p-val ≤ 0.003 for both, indicate that the observed separation between correct and incorrect predictions is statistically significant. At the 95% confidence level, we see no overlap between their confidence intervals, revealing that the ability of MSU-Net to distinguish correct from incorrect predictions is significantly better than that of MCU-Net. We verify this in Figure 2.8.

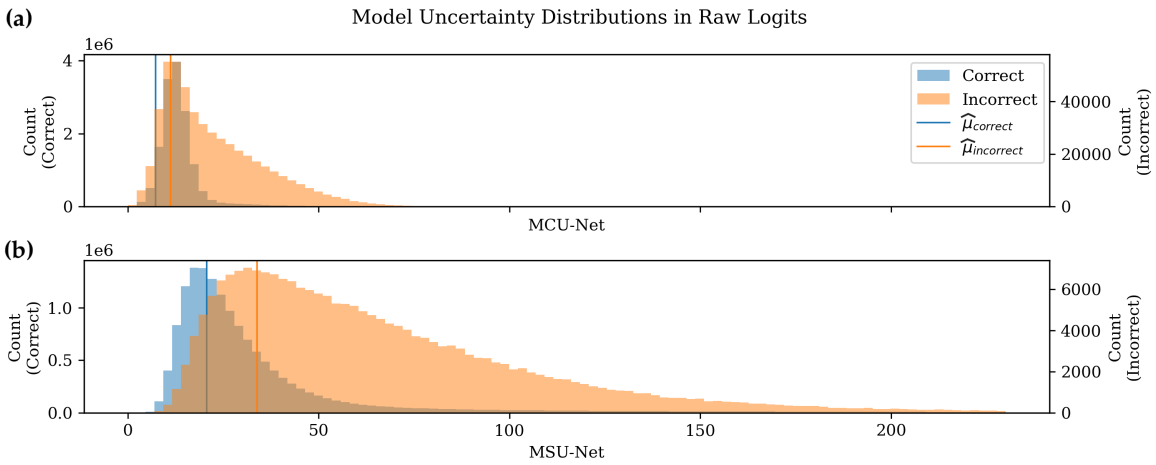


Figure 2.8: **Epistemic uncertainty distributions for correct and incorrect segmentations** for (a) MCU-Net and (b) MSU-Net. Our approach yields a markedly better differentiation of incorrect (orange) from correct (blue) predictions.

	$\widehat{\mu}_{corr}$	$\widehat{\mu}_{incorr}$	$\Delta\widehat{\mu}(\uparrow)$	$\widehat{R}_\alpha(corr \,\|\|\, incorr)(\uparrow)$	95% CI on $\widehat{R_\alpha}$	p-val($\downarrow$)
MCU-NET	7.230	11.229	3.999	0.429	[0.426, 0.453]	0.003
MSU-NET	20.783	33.876	**13.093**	**0.638**	**[0.603, 0.667]**	0.003

Table 2.2: **Model quality on test dataset.** Arrows indicates direction of better performance.

Qualitative uncertainty maps visually validate our findings and capture local variations in model performance. MCU-Net exhibits indiscriminately low epistemic uncertainty in large patches where the model fails to segment a vessel and is highly sensitive to noise in the background class as shown by high aleatoric uncertainty outside of the vessels in Fig. 2.9. In contrast, MSU-Net provides more interpretable uncertainty values, demonstrating increased epistemic uncertainty for semantically challenging pixels at the bottom of vessels and decreased epistemic uncertainty for clearer vessel tops. Furthermore, MSU-Net excels by capturing the intrinsic variability in vessel shapes. This is evident from the increased aleatoric uncertainty around vessel walls, reflecting the diverse vessel structures inherent to each subject.

We provide additional examples of our vessel segmentations along with the corresponding epistemic uncertainty maps in Appendix A.6.2.

2.7.3 Effects of Ablating Model Stages on Segmentation Capabilities

We methodically test the individual effects of each stage through ablation studies. We validate the following meaningful configurations of stages 1, 2, and 3 using the DSC and IoU metrics on the VS1 validation set:

- **MCU-Net**: Our baseline (single network)

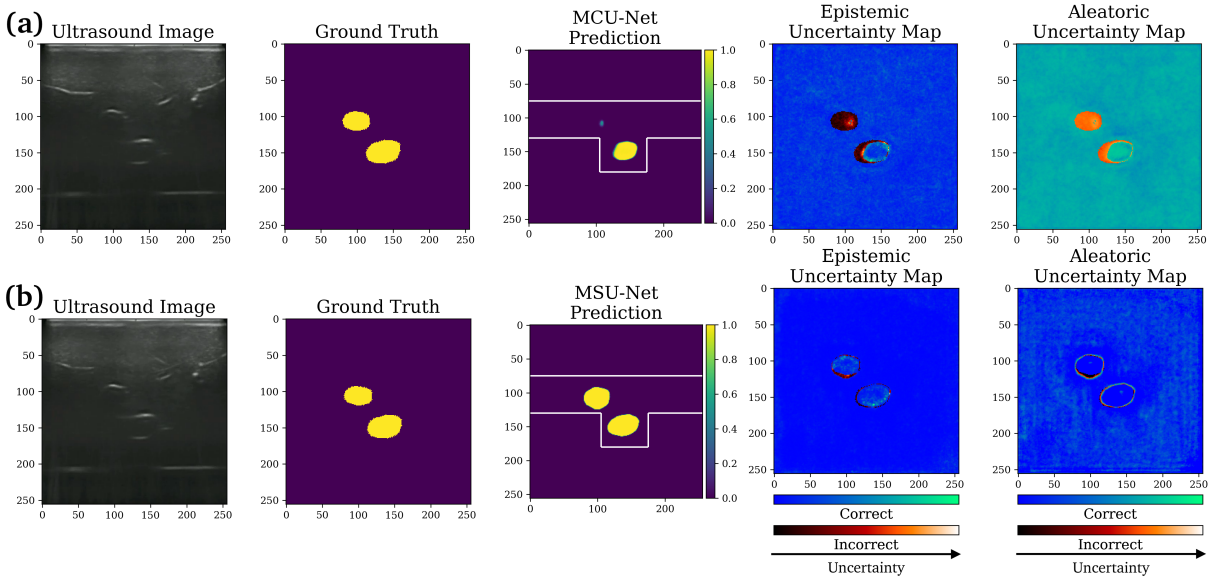


Figure 2.9: **Qualitative epistemic and aleatoric uncertainty maps** for (a) MCU-Net and (b) MSU-Net. Darker colors indicate lower uncertainty, while lighter colors indicate higher uncertainty. Evaluations are confined to white-outlined region of interest to address class imbalance.

- **Deep Ensemble**: A deep ensemble of $M = 15$ bagged networks
- **Decorrelated Deep Ensemble**: A deep ensemble of $K = 3$ bagged networks optimally selected through decorrelation maximization
- **MSU-Net**: Our proposed framework with a decorrelated ensemble of bagged networks and a final MCU-Net combiner

Model	**Stage 1**	**Stage 2**	**Stage 3**	**DSC**(↑)	**ΔDSC**	**IoU**(↑)	**ΔIoU**
MCU-Net	✗	✗	✓	0.649	−0.272	0.534	−0.310
Deep Ensemble	✓	✗	✗	0.782	−0.139	0.689	−0.155
Decorrelated Deep Ensemble	✓	✓	✗	0.821	−0.100	0.742	−0.102
MSU-Net	✓	✓	✓	**0.921**	**+0.0**	**0.844**	**+0.0**

Table 2.3: **Results from ablation studies on MSU-Net model stages.** While MSU-Net outperforms all other configurations, we find that our decorrelation maximization algorithm considerably improves the performance of naïve deep ensembles. Stage 2 is not evaluated independently, as decorrelation is only meaningful when used in conjunction with ensembles.

The addition of each stage improves both DSC and IoU scores on the VS1 validation set. Deep ensembles improve segmentation performance over a single Monte carlo network. Interestingly, our decorrelation procedure that *reduces* the size of the ensemble to $K = 3$ networks exhibits a 3.9% increase in DSC and 3.5% increase in IoU over the original deep ensemble of

$M = 15$ networks. These results suggest that training more models is not always the optimal strategy. Rather, training a robust, diverse ensemble can provide stronger predictive power. There is benefit in motivating networks to explore different local minima of the loss landscape.

MSU-Net generally outperforms MCU-Net. Although it performs marginally worse at specificity and false positive rates (FPR), the precision-recall curves from Fig. 2.7b indicate that MSU-Net achieves an average precision score of 0.936, which is 18% higher than MCU-Net's score of 0.755, compared to the baseline score of 0.09 for a "no-skill" classifier. As such, at the same level of recall, MSU-Net correctly classifies a higher proportion of pixels that are actually vessels than MCU-Net. Crucially, MSU-Net achieves a considerably lower false negative rate (FNR) than MCU-Net. In this context, not recognizing a real vessel can have more severe consequences for a critically injured person than mistakenly identifying a vessel that is not actually there. MSU-Net improves credibility not only through higher quality results, but also through more accurate results while avoiding potentially disastrous deficiencies of predictive modeling in the context of medical image segmentation.

2.8 Discussion and Future Work

Our proposed multistage learning ensemble framework, MSU-Net, significantly improves uncertainty quantification and accuracy in femoral vessel segmentation in ultrasound images, outperforming traditional Monte Carlo U-Nets. The integration of bagging and decorrelation techniques ensures that the ensemble models are diverse and robust. We empirically show that deep ensembles strongly benefit from diversity and provide an intuitive decorrelation maximization algorithm to produce well-calibrated, high performing deep ensembles. Our results indicate a 27.7% improvement in the mean DSC, with better sensitivity and lower false negative rates, increasing transparency and trustworthiness. These advancements are achieved despite minimal additional training, making it a valuable tool for guiding autonomous systems and assisting non-experts in high-stakes medical environments. MSU-Net's differentiation between correct and incorrect predictions, measured by Rényi divergence and observed in qualitative uncertainty maps, highlights its ability to identify and address segmentation errors. Our qualitative maps enable clinicians to interpret model uncertainty results without requiring expertise in DL. Future research will focus on refining ensemble selection and validating our findings on live animal and human data, extending beyond the current phantom data and binary segmentation context.

2.9 Acknowledgments

We thank Nico Zevallos for gathering experiment data. This work was partially supported by the U.S. Department of Defense contracts W81XWH-19-C0083 and W81XWH-19-C0101.

This work has been peer-reviewed and published in the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI). This chapter closely follows (Banerjee et al., 2025).

Chapter 3

Advancing AI Trust in Personalized Medicine

3.1 Introduction

3.1.1 Dermatomyositis

Dermatomyositis (DM) is a type of Idiopathic Inflammatory Myopathies (IIMs, myositis), characterized by progressive muscle weakness and inflammatory rashes and other organ involvement (DeWane et al., 2020). Rashes are often pruritic, photosensitive, and considerably impact patients' quality of life (QOL) (Goreshi et al., 2011; Kleitsch et al., 2023). DM patients experience a poorer QOL across all subscales of the Short Form 36, particularly in the areas of vitality and mental health, compared to the general population. (Hundley et al., 2006) evaluated 71 patients with DM or DM sine myositis against two QOL measures, Skindex-16 and Dermatology Life Quality Index, identifying a strong correlation between DM symptoms and higher (worse) QOL scores, especially in women. In addition, their scores are markedly lower than those of other chronic diseases such as psoriasis and atopic dermatitis. Treatment options for DM are currently limited, often accompanied by significant side effects and inadequate efficacy, highlighting the critical need for novel therapies through well-designed randomized controlled trials. However, improvements in DM activity can lead to a statistically significant reduction in associated pruritus (Robinson et al., 2015), the main contributor to the worsening of psychological symptoms in affected patients and their ability to function in daily tasks. Evidently, there is an incentive for early detection of DM to mitigate harmful symptoms that generally intensify over time.

3.1.2 DART Study

Cutaneous dermatomyositis (CDM) is the skin manifestation of dermatomyositis, where the rash is the most prominent feature. Although most commonly evaluated on the hand, rashes may also occur on the face, neck, upper chest, and back, and may be accompanied by swelling in these affected areas. Traditional in-clinic diagnosis requires clinical evaluations of rashes, which are expert-dependent, subjective, semi-quantitative, and have variable inter-rater reliability. Our investigation of the level of agreement between two rheumatologists (MDs) for CDM severity

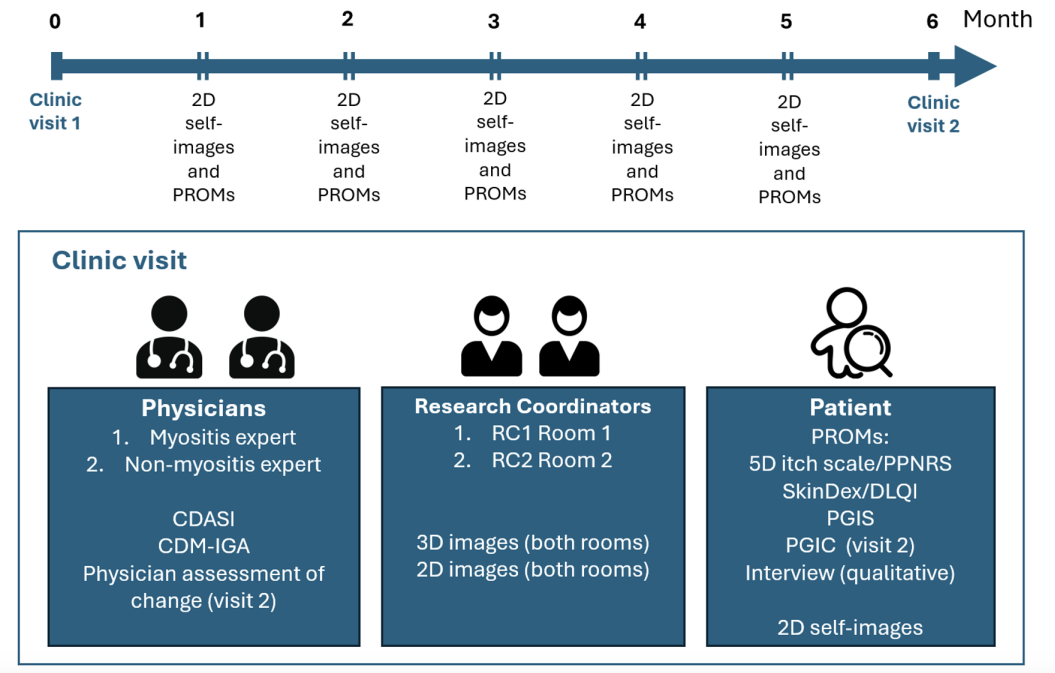


Figure 3.1: **Dermatomyositis Assessment of Rash via Telemedicine (DART) study design.** Each participant completes two clinic visits, with the second occurring approximately six months after the first visit. Demographics, disease characteristics, and classification of disease are collected in clinic visit 1. All outcome evaluations, including physician assessment, imaging procedures and PROMs are conducted during both clinic visit 1 and 2, except for physician and patient global impression of change conducted only in visit 2 (Aggarwal and Pongtarakulpanit, 2025).

scoring found the weighted Cohen's kappa to be 0.561 ± 0.0615 at the 95% confidence level. (Landis and Koch, 1977) identifies a Cohen's kappa value of 0.41–0.60 as moderate agreement. These limitations may be amplified in larger multicenter clinical trials, especially in centers lacking expertise in the field.

Rapid advancements in image-based machine learning technologies present an attractive solution for enhancing the reliability and reproducibility of rheumatological assessment. Therefore, clinicians from the Division of Rheumatology and Clinical Immunology at the University of Pittsburgh Medical Center (UPMC) are exploring the feasibility of ML applications in telemedicine to assess CDM skin rashes compared to traditional in-clinic evaluations. CDM patients have been prospectively enrolled in an observational study called DART (Dermatomyositis Assessment of Rash via Telemedicine), with eligiblity determined by the 2017 EULAR/ACR criteria. DART aims to demonstrate automated, objective, quantitative, and reproducible image-based ML outcome measures of CDM disease activity using **(1)** in-clinic 3D imaging, **(2)** in-clinic 2D smartphone imaging, and **(3)** patient 2D self-imaging using mobile applications (Aggarwal and Pongtarakulpanit, 2025).

In this study, two independent MDs (MD_1 and MD_2) of varying expertise in myositis assess participants' rash manifestations during both in-clinic and telemedicine visits, spaced approximately 2-4 weeks apart. A second in-clinic visit is conducted approximately 6 months from the first visit. The Cutaneous Dermatomyositis Disease Area and Severity Index (CDASI) (Tiao et al., 2017) is used to score the severity of hand rashes during these evaluations by rating disease severity and activity. In addition, patient-reported outcome measures (PROMs) are acquired during in-clinic visits. Figure 3.1 illustrates the DART study design.

However, several challenges persist in applying ML towards the task of automated CDM rash severity scoring. First, the prevalence of DM, and specifically CDM, is quite rare; A retrospective study in the U.S., including subjects of all ages based on hospital discharge diagnoses from 1963 to 1982, reported an annual incidence of dermatomyositis and polymyositis of 5.5 cases per million inhabitants (Oddis et al., 1990). The rarity of CDM, the specialized expertise required to assess rash severity, and the time-consuming labeling procedure contribute to data scarcity. Second, inconsistencies during image acquisition can negatively affect model training. The 2D and 3D images are collected through different imaging technologies, lacking the standardization required for robust model performance (Rodríguez-Rodríguez et al., 2024).

To address these challenges, we contribute a novel DL framework that leverages powerful pretraining capabilities to overcome data scarcity, generating relevant and reliable severity predictions. Our fully-automated system predicts scores comparable to the standard expert-rated CDASI, exceeding accuracy thresholds set by current DM experts. We aim to answer the following research questions:

- **RQ1:** Which modeling paradigms can overcome data scarcity issues?
- **RQ2:** How do we overcome a lack of diversity in our limited dataset?
- **RQ3:** How reliable are the uncertainty estimates of our model under dataset shift?

Figure 3.2: **Examples of image modalities in DART study for Patient 001DM1031.** We explore the following three modalities from left to right: in-clinic 3D imaging, in-clinic 2D smartphone imaging, and telemedicine (at-home) 2D smartphone imagining.

3.2 Data Collection

3.2.1 Image Modalities

In-clinic 3D imaging. The VECTRA H1 handheld imaging system provides high-resolution, clinical quality 3D imaging with minimal staff training required (Scientific, 2024). We leverage its proprietary RBX technology to separate unique color signatures of Red and Brown skin components for unequalled visualization of skin conditions. A designated research coordinator acquires 3D images from the VECTRA H1 camera, which are further annotated and exported as 2D images by attending physicians.

In-clinic 2D smartphone imaging. In-clinic 2D images are acquired in a separate room from VECTRA H1 3D images. Hand placement within the images appears consistent among the in-clinic smartphone images. However, images differ in background setting; acquisition artifacts, such as marker labels, are visible in some images but not all.

Telemedicine 2D smartphone imaging. Patients upload telemedicine images using the SkinIO smartphone application (SkinIO, 2024). We observe that these images lack consistency, exhibiting noticeable variations in background, lighting, and angle. As such, their quality is considerably lower than that of the in-clinic data.

Clinicians evaluate scores using the *CDASIver02* form, a partially-validated DM-specific instrument designed to capture the extent of cutaneous disease (Anyanwu et al., 2015). CDASI scores for hand images range from 0 to 14, with higher scores indicating greater severity. Images of the right and left hands are evaluated simultaneously and given one CDASI score to maintain consistency, since rashes tend to distribute symmetrically across hands. Examples of our collected datasets can be found in Appendix B.1 Figure B.1.

3.2.2 Dataset Limitations

In the initial study, a total of 27 CDM patients underwent evaluation, of which 26 were Caucasian-American and 1 African-American. Approximately 82.6% of the patients were female. The patients had a mean age of 48.6 ± 17.4 years and a median disease duration of 38 months.

Upon further examination, 12 of the 88 initial images appeared to have discrepancies of 2 points or more on the CDASI scale between the two rheumatologists. After consulting with the clinicians, we agreed to remove these from our dataset to prevent the high uncertainty and disagreement from affecting our model calibration. This process resulted in a final set of 76 images from 23 patients. Furthermore, MD_1 scores were designated as the reference standard for CDASI scoring, given the increased frequency of patient interactions and greater familiarity with MD_1. Figure 3.3 displays our data analysis on the final set of 76 images.

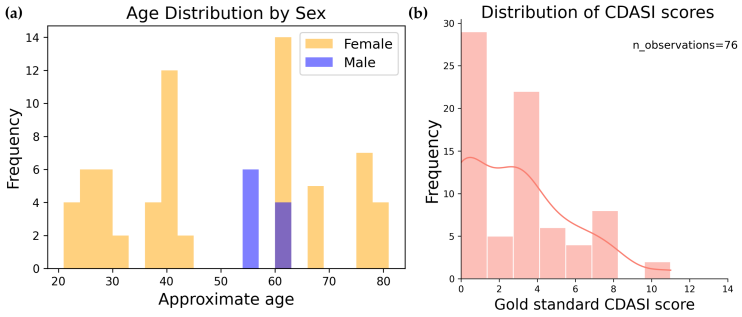


Figure 3.3: **Exploratory data analysis of demographics for post-processed DART data.** Distribution of (a) age and sex demographics and (b) ground truth labels after removing four patients with large MD rating discrepancies during both in-clinic visits.

Evidently, this process further imbalances our dataset. The only African-American patient in our original dataset was removed due to high variability in CDASI scores between MDs. In addition, the label distribution is unimodal with a peak at CDASI of 0, with a strong right skew due to a single outlying patient with a CDASI of 11 during their first visit. The overwhelming majority of "normal" (CDASI of 0) patients in our dataset is likely to have a non-negligible impact on our model predictions.

One technique to alleviate data scarcity is to upsample the original dataset with data augmentations. We choose the following augmentations: horizontal flip, vertical flip, 90° counterclockwise (CCW) rotation, 180° CCW rotation, and 270° CCW rotation, adding 5 new augmented images per original image in our dataset. This process increased our data set by 6x, resulting in a total of 456 images available for analysis.

3.3 Ordinal Regression Experiments

We first explore rash severity scoring as an ordinal regression problem to leverage the natural ordering of CDASI scores $0 \prec 1 \prec \cdots \prec 14$, where low scores represent mild severity, and high scores indicate more severe cases of CDM. However, "distances" between categories may not necessarily be equal. In particular, differences between intermediate scores have smaller distances than differences between higher CDASI scores. Standard approaches for modeling ordinal data involve fitting parallel separating hyperplanes that optimize a certain loss function.

This setup offers efficient learning via strong inductive biases (Lu et al., 2022). For our setting, linear threshold models are most appropriate (Wang et al., 2023), which simultaneously learn an output mapping and thresholds to partition the output in order to make ordinal predictions. The cumulative formulation of the model is particularly convenient for parameter interpretation and data simulation (Gambarota and Altoè, 2024), although adjacent categories (Fullerton and Xu, 2018) and continuation odds ratio (Cole and Ananth, 2001) models are heavily cited in the literature as well.

3.3.1 Regression Formulation

Our formulation is derived from the assumption of an underlying latent regression model with a continuous response. Let Y^* represent the underlying latent variable and consider Y_k to be the observed ordinal variable with $K = 15$ coarse, categorical levels. Analogously to the standard regression formulation, Y^* is a function of the linear predictor η_i, thus $Y^* = \eta_i + \epsilon_i$ where ϵ_i is the irreducible error of the model. We must define an appropriate *link* function to map probabilities to the linear predictor, $F(p) = \eta$. Probabilities can be subsequently extracted using the inverse function $p = F^{-1}(\eta)$. Consequently, the log-odds ratio can be modeled by

$$\log\left(\frac{\Pr(Y \leq k)}{1 - \Pr(Y \leq k)}\right) \coloneqq \alpha_k - \boldsymbol{\beta}^T \mathbf{x}_i, \;\; k = \{1, 2, \cdots, K-1\} \tag{3.1}$$

with $\eta_i = \boldsymbol{\beta}^T \mathbf{x}_i$, and the logit link $F = \log\left(\frac{u_k}{1-u_k}\right)$. Equivalently, we can directly model the *cumulative* probabilities by taking the inverse of the link function

$$\Pr(Y \leq k) \coloneqq F^{-1}(\alpha_k - \boldsymbol{\beta}^T \mathbf{x}_i) \;\; k = \{1, 2, \cdots, K-1\} \tag{3.2}$$

The cumulative *probit* uses the (inverse) standard normal distribution link and is another popular choice. Here, $F = \Phi^{-1}(p) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{p} e^{-z^2/2} dz$. However, probit and logit links usually produce similar statistical results, and we observe negligible differences in our analysis when switching between the two.

Requiring the coefficients of each predictor to be identical across categories results in a crucial simplification of the cumulative logit model to a proportional odds cumulative logit model. The proportional odds assumption (POA) significantly reduces model complexity and convergence speed, only requiring us to model $(K-1)$ unique intercepts and $K + p + 1$ coefficients for p total predictors. POA is a frequently made assumption in healthcare studies (Afroja et al., 2020; Ayyaz et al., 2021; Lall et al., 2002).

3.3.2 Utilizing Clinically-Relevant Handcrafted Features

Earlier investigations reveal crucial correlations between visual image-based features and expert scores: rash area and rash redness had a 0.722 and 0.721 Spearman's Rho correlation value with respect to MD evaluations, respectively. This indicates a substantial degree of association between rash area and redness and CDASI scores. Furthermore, rash redness and area showed a strong positive correlation with patient self-assessment scores, with values of 0.648 and 0.636, respectively. Clinicians additionally hypothesized correlations between hand textures and CDASI

scores during their assessment. To this end, we construct the following handcrafted redness and textural features for each image

- ❖ The ratio between the rash area and hand area
- ❖ The average redness of the rash, measured as the average a*b* value of the rash-only pixels in the L*a*b* color space[1]
- ❖ The ratio between the average redness of the rash and the average redness of the hand, computed as specified above
- ❖ Texture features extracted from gray level co-occurrence matrices (GLCMs) (Haralick et al., 1973), which are histograms of co-occurring grayscale values at varying offsets over an image

Two key characteristics of DM motivate the inclusion of additional demographic features: (1) a higher prevalence in females and (2) a bimodal distribution of incidence rates, with peaks in childhood and between the ages of 40 and 60 years in adulthood (Mainetti et al., 2017). Thus, we collect additional demographic information from patients including their age, sex, height, weight, race, and the presence/absence of antibodies related to CDM manifestation as additional predictors in our model.

3.3.3 Automating Hand Rash Localization

The computation of handcrafted features requires two steps, first determining the hand and general rash region within the image, and then further identifying precise rash clusters. By restricting our analysis to a more fine-grained region of interest and only extracting features from this relevant region, we can successfully reduce the influence of irrelevant background pixels, leading to more accurate feature representations. Evidently, calculating the above features requires adequate segmentations of the hand and rash regions.

As such, we first use the Python `rembg` package (Qin et al., 2020) to remove the background and isolate the hand in each image. However, identifying the rash region within the hand requires a more sophisticated supervised learning approach. For this task, we perform semantic segmentation on our new images to classify pixels as either rash or non-rash. Clinicians used the Pixlr[2] image editing software to manually label coarse rash regions, which we use as ground truth masks for image segmentation.

Coarse rash recognition. The DeepLabV3+ model (Chen et al., 2018) builds on the standard encoder-decoder architecture using atrous convolutions in its encoder to capture multiscale contexts and to expand the receptive field without reducing spatial resolution. Atrous convolutions particularly aid in improved segmentation accuracy at object boundaries, enabling precise rash ROI segmentations. We choose the Xception architecture with pretrained weights as our

[1]The L*a*b* color space, also referred to as the CIELAB color space, is a perceptually uniform color space defined by the International Commission on Illumination in 1976. It decomposes colors based on its luminance (L*), red-green (a*), and blue-yellow (b*) components. This is crucial since our fine-grained K-means clustering algorithm relies on Euclidean distance between color values to identify rash regions.

[2]`https://pixlr.com`

encoder for feature extraction. Initial segmentation results reveal a systematic inability to predict empty masks for "normal" patients without hand rashes. Hence, we add an auxiliary classification head to our encoder composed of a global average pooling layer, dropout with $p = 0.2$, and a fully-connected layer to output binary classification predictions corresponding to the presence (1) or absence (0) of rashes. We additionally modify our loss function to simultaneously train our encoder for classification while also learning precise segmentations from the decoder output

$$\mathcal{L}_{\text{multi-task loss}} \triangleq \underbrace{\mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{CE}}}_{\text{segmentation loss}} + \underbrace{\mathcal{L}_{\text{BCE}}}_{\text{classification loss}} \tag{3.3}$$

We refer back to Equation 2.9 for definitions of our loss terms. As in Section 2.5.2, we use a combined cross-entropy and Dice loss for segmentation to address the under-representation of target pixels in our images. Notably, if our encoder predicts an absence of rashes for an image, the corresponding segmentation mask from the decoder is appropriately zeroed out to reflect our classification prediction. In this way, positive segmentations for "normal" patients incur a heavy loss penalty. We achieve an average DSC improvement of 25.4% using multi-task loss, validated on held-out images during group k-fold cross-validation (see Section 3.3.4), increasing from 45.0% to 65.4%.

Fine-grained rash localization. Manually annotating ground truth masks for our segmentation training is costly, and thus we reconcile with coarser masks for our training procedure above. After segmentation, we further employ the unsupervised K-means clustering algorithm (Lloyd, 1982) on our predicted segmentation masks to extract a more accurate delineation of the rash clusters. We posit that capturing fine-grained clusters induces more precise extraction of handcrafted features. Starting from a random initialization of cluster centroids, the K-means algorithm iteratively partitions a set of data points into k clusters that minimize inertia. We utilize the Euclidean distance between the mean a*b* color values of each centroid to improve our model fit. Finally, we select the cluster with the highest a*b* centroid value, as it is most likely to correspond to the rash region. We refer to Appendix B.2.2 for examples of K-means rash localization maps. Our handcrafted features are computed from the final rash regions after coarse segmentation and K-means localization. We detail our complete handcrafted feature extraction pipeline in Appendix B.2.1.

3.3.4 Regression Training

Normal k-fold cross validation allows for images from the *same* patient to leak from training to validation sets. This results in biased and optimistic evaluations of model capabilities. Instead, patients are randomly partitioned into k groups by ID to perform group k-fold cross-validation. For each fold, we fit a cumulative logistic regression model on a training set of 19 patients using the `clm` function in R's `ordinal` package, and validate our fit on a test set of 4 patients. A modified Newton algorithm is utilized to find the maximum likelihood estimates of the model parameters. Given our large cohort of predictors, we perform (bidirectional) stepwise regression on the Akaike Information Criterion (AIC) to determine a parsimonious proportional odds model with the largest predictive power.

However, our predictors suffer from almost perfect multicollinearity, which poses significant challenges for training stability and convergence. To improve model stability across runs, we opt to first run Principal Component Analysis (PCA) to project our original features onto the directions of highest variance, thereby removing feature correlations. PCA projections align with clinical hypotheses: the first two principal components (PCs) suggest strong positive correlations between the rash redness ratio, rash area ratio, contrast, and dissimilarity, and between the demographic features of age, height, and weight. From Figure 3.4, we can observe that PC1 represents the redness and texture of the hand rash, while PC2 captures key demographics of a patient. We outline our regression procedure in Algorithm 3.

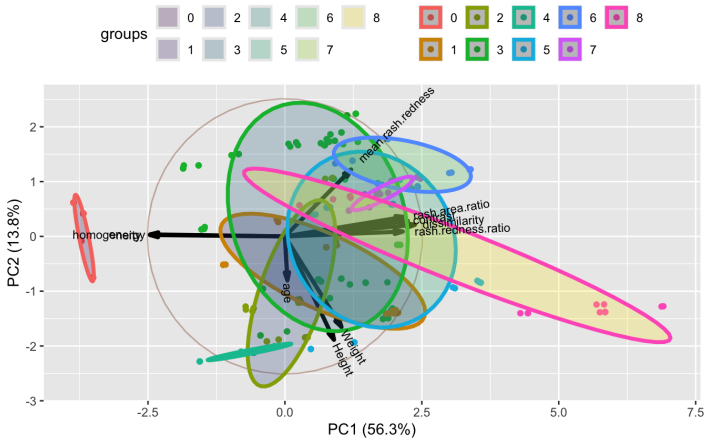


Figure 3.4: **PCA biplot projecting training samples with handcrafted features onto the first two principal components.** Strong positive correlations observed among rash redness and rash area features and also between demographic features of age, height, and weight.

Algorithm 3 PCA-based CLM with Stepwise Regression

Input: Images $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)}) : i = 1, 2, \cdots, N\}$, corresponding PIDs, number of grouped folds K
Output: T model evaluation metrics (e.g., off-by-j-accuracy, Spearman ranked correlation)
Split images into K folds based on PIDS
for k in range K **do**
    1. Partition data into training set $\mathcal{D}^{-k}_{\text{train}}$ and testing set $\mathcal{D}^{k}_{\text{test}}$
    2. Perform PCA on $\mathcal{D}^{-k}_{\text{train}}$ to achieve $\mathbf{\Phi}^{-k}_{\text{train}} = \{(\phi(\mathbf{x}^{(i)}), y^{(i)}) : i = 1, 2, \cdots, N\}$,
    where ϕ projects each sample to the first ℓ PCs explaining at most 90% of the
    original variance
    3. Apply the same transformation on the test set to achieve $\mathbf{\Phi}^{k}_{\text{test}}$
    4. $\texttt{init.model}_k \leftarrow \texttt{clm}(\mathbf{\Phi}^{-k}_{\text{train}})$
    5. $\texttt{stepwise.model}_k \leftarrow \texttt{stepAIC(init.model}_k\texttt{, direction = "both")}$
    6. Evaluate $\texttt{stepwise.model}_k$ performance on $\mathbf{\Phi}^{k}_{\text{test}}$ using selected metrics
    7. Store evaluation metrics for fold k, $\{M_{1k}, M_{2k}, \cdots, M_{Tk}\}$
end for
Compute $\{\frac{1}{K}\sum_{k=1}^{K} M_{tk}\}$ for all M_{tk} for $t = 1, 2, \cdots T,\ k = 1, 2, \cdots, K$
return $\{M_{1*}, M_{2*}, \cdots, M_{T*}\}$

3.3.5 Regression Predictions Correlate with Clinician Assessments

Our ordinal regression predictors differ between PCA and non-PCA settings. Non-PCA regression has a wider selection of predictors for stepwise regression, whereas PCA is limited to only continuous predictors. The full list of available predictors in the non-PCA setting is detailed in Appendix B.2.3. During non-PCA regression, our `clm` consistently identifies the first three principal components to be statistically significant via F-tests during stepwise regression. We plot our group k-fold cross-validation results for PCA regression in Figure 3.5.

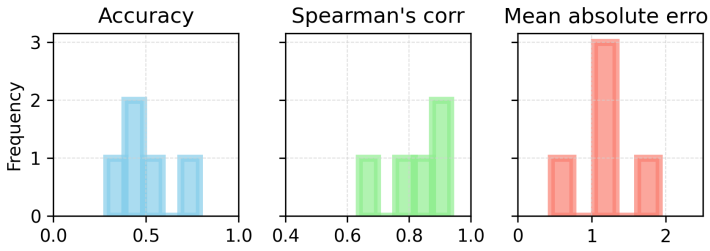


Figure 3.5: **k-fold cross validation results for PCA-transformed ordinal regression.** We evaluate our `clm` model on a held-out PCA-transformed test set using accuracy, correlation, and MAE.

Our results show moderate variability in accuracy and Spearman's rho values across the five folds. Interestingly, accuracy appears to be negatively associated with Spearman's rho values; Higher correlation values tend to be associated with lower accuracy scores. However, this is most likely due to a lack of statistical power caused by a limited sample size. The worst-performing model occurs in Fold 3, with a 28.6% accuracy and MAE of 1.904. The test patients have CDASI scores of $0, 2, 7$ and 8, so we attribute the low performance on this fold to its test set having a higher proportion of severe patients ($\approx 50\%$) compared to other folds. We provide additional regression results in Table B.1 (Appendix B.2.4). The ability of our ordinal regression model to generalize to different levels of rash severity given our featurization seems limited. Our features are highly sensitive to perturbations in hand images within the same class. For example, the K-means algorithm performs poorly on old, mature skin due to roughened skin texture. Hence, a younger patient with the same CDASI as an older patient is likely to have considerably different feature values.

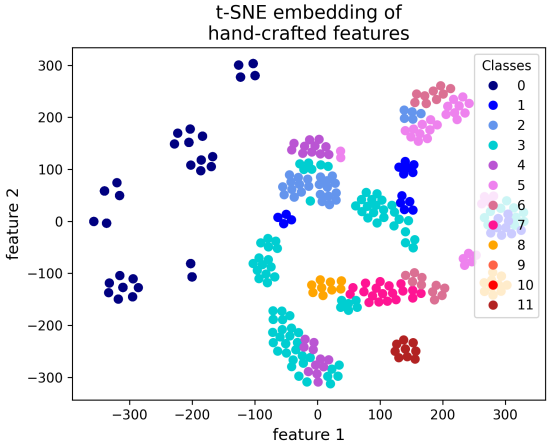


Figure 3.6: **t-SNE plot captures CDASI class separability in two dimensions.** Each color corresponds to a unique CDASI score. Extreme score values (i.e., CDASI$= \{0, 11\}$) exhibit distinct, compact clusters in two dimensions and the largest inter-cluster distance than any other pair of scores.

The t-SNE plot in Figure 3.6 shows a high separability of the CDASI score classes in a

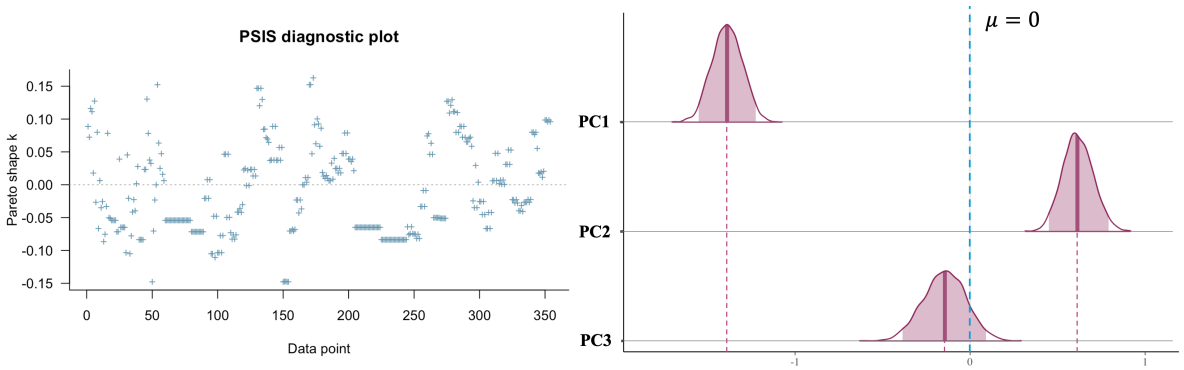


Figure 3.7: **Density plots computed from posterior draws with all chains merged for PCA-transformed Bayesian ordinal regression.** (a) PSIS diagnostic plot and (b) 95% credible posterior interval estimates. The diagnostic plot verifies the reliability of our interval estimates.

reduced-dimensional space. Despite our model's low statistical power, our PCA-transformed handcrafted features provide strong discrimination between CDASI score classes, which can be captured despite dimensionality reduction. Extreme scores (i.e., very low or very high CDASI) generally form distinct, compact clusters, while intermediate scores are grouped in adjacent clusters. This is consistent with clinical rash assessments, with extreme cases being easier to diagnose than minimal to moderate cases. However, the "normal" class appears to have greater intra-cluster variability than any other class. Although we improve our rash segmentations with a multi-task loss function, our model still predicts several false positive masks, which contributes to the high variability in feature representations of "normal" patients.

3.3.6 Bayesian Ordinal Regression

To assess the credibility of our model estimates, we fit a Bayesian ordinal regression model over our PCA-transformed predictors using the `stan_polr` package in R. We specify a jointly uniform Dirichlet prior on the probability of falling in each of the 15 CDASI score categories and a Beta prior on the coefficient of determination such that $R^2 \sim \text{Beta}(p/2, 2p)$, where p is the total number of predictors. Our Beta prior enables regularization on our coefficient values. We then estimate posterior distributions over the selected PCs in our ordinal regression model by drawing Markov Chain Monte Carlo (MCMC) samples. We run 4 chains for $1,000$ warm-up iterations followed by $1,000$ additional iterations. (Vehtari et al., 2024) determines that the diagnostic threshold for Pareto k depends on the sample size S. If $k < \min(1 - 1/\log_{10}(S), 0.7)$, we can confirm that the PSIS estimate and the corresponding Monte Carlo standard error estimates are reliable. From Figure 3.7a, we validate that Pareto $k < 0.61$, therefore we can meaningfully interpret the quantile-based posterior interval estimates in Figure 3.7b.

Demographics play a vital role in determining rash severity. Our results present strong posterior evidence that decreasing PC1 values and increasing PC2 values are significantly associated with an increase in rash severity at the 95% credible level. Therefore, we verify that rash redness and textural information (PC1) and demographic information (PC2) are likely to have a

significant effect on rash severity. Although it is intuitive that the visual characteristics of rashes influence the predicted CDASI score, the impact of demographic information on the score is surprising. However, considering the higher prevalence of DM among women and the elderly, it is evident that patient demographics play a crucial role in understanding rash manifestation.

3.3.7 Limitations

Our handcrafted feature extraction pipeline benefits from being clinically-motivated and intuitive. Parameter estimates are highly interpretable and have a strong correlation with feature importance. Moreover, our model is compact enough to enable Bayesian methods for predictive uncertainty estimation without incurring excessive computational costs. Our regression pipeline enables clinicians to understand how their hypothesized features quantitatively affect CDASI scores. However, this contextualization has several compelling drawbacks. Most prominent is our strict inductive biases. Our proportional odds model makes the limiting assumption that the log-odds of having a certain CDASI score versus all lower scores are a linear function of the input features. Furthermore, POA assumes that the difference between adjacent scores is roughly equivalent across all CDASI scores. This is inconsistent with clinical evaluations suggesting rash redness affects higher scores more than lower scores near zero. Furthermore, our selected handcrafted features may provide at best a limited featurization of the original image. Since our pipeline is not trained end-to-end, it is likely that our selected features, although clinically motivated, may not be adequately tuned for our regression setting. This motivates the need to explore models with less restrictive inductive biases to enable enough flexibility to capture potential non-linear relationships in the image data. Naturally, convolutional neural networks are an attractive alternative, providing automated hierarchical feature extraction *without* requiring explicit domain expertise.

3.4 Related Work

Data scarcity is a pervasive challenge in the development of large-scale machine learning models. More data exposes a model to a wider variety of potential inputs, enabling improved pattern recognition and generalizability, reduced tendency to overfit, and robustness against noise and outliers. Conversely, a lack of data significantly hinders the application of models, often resulting in ineffectual learned feature representations (Alzubaidi et al., 2021), poor out-of-distribution detection (Li et al., 2021), or limited representation of a larger population (Habehh and Gohel, 2021). Transfer learning (Torrey and Shavlik, 2010) is the most prominent technique for overcoming these challenges without requiring additional labeled data. Transfer learning leverages knowledge gained through one task or dataset towards improving model performance on other related tasks or datasets. Several strategies within transfer learning can help reduce computational costs, data scarcity, and generalizability (IBM, 2025), including feature extraction (Belinkov, 2022), self-supervised pretraining (Zoph et al., 2020), and unsupervised pretraining (Ge et al., 2023). In the following sections, we review the strengths and limitations of these methodologies.

3.4.1 Feature Extraction

We can reimagine a CNN model as a feature extractor f by removing its classification head. f generates intermediate representations $\mathbf{y}' \in \mathbb{R}^{(B\times M)}$ of our input $\mathbf{x} \in \mathbb{R}^{(B\times H\times W)}$ in some lower-dimensional latent space. After pretraining its featurization capabilities through supervised or unsupervised learning, f can be utilized as an "off-the-shelf" model for domain-specific tasks by training a shallow classification head h with the appropriate number of classes for the specialized task. Thus, our classification prediction becomes $\mathbf{y} = (h \circ f)(\mathbf{x})$. In particular, this approach freezes the convolutional layers in the model, only using f to *generate* generic features from the image data while training h to *adjust* the parameters for the downstream classification task. The feature extraction framework assumes several benefits: It allows the domains, tasks, and distributions used in training and testing to be different (Pan and Yang, 2010), requires minimal training of a shallow network which reduces computational costs (IBM, 2025), and learns particularly useful feature representations in the medical imaging domain (Agarwal et al., 2021; Kim et al., 2022; Weiss et al., 2016). However, this method is limited by the quality and applicability of common latent features extracted by f. Since f is often trained on generic and broad source datasets, such as COCO or ImageNet, its output features are not guaranteed to be effective for all types of target tasks. When the visual characteristics of the target task are not accurately represented by the source tasks, it is recommended to resort to costly fine-tuning from scratch in order to achieve satisfactory performance (Kim et al., 2022).

3.4.2 Self-Supervised Pretraining

Models trained through supervised learning (SL) often struggle to generalize across diverse target tasks, as they tend to learn feature representations that are highly correlated with the original source tasks (Wolf et al., 2023). Self-supervised learning (SSL) addresses these issues by encouraging the model to efficiently learn useful and generic feature representations from large, unlabeled datasets to prime the fine-tuning of a downstream task. Notably, this technique eliminates the need for costly manual image annotation. Self-supervised pretraining follows a two-stage process: First, the DL model is pretrained with SSL to capture general high-level features of the images. Next, the model is fine-tuned on a small labeled dataset using SL to adjust the learned features to the target domain. SSL has been extensively applied in the medical image domain, demonstrating superior performance over training from scratch (Chen et al., 2021; Dufumier et al., 2021; Ghesu et al., 2022; Tang et al., 2022). Two approaches dominate the literature in self-supervised pretraining: Contrastive learning and masked image modeling.

Contrastive learning is based on the principle that similar images should map to similar embeddings. Hence, it learns features that maximize mutual information between representations at different spatial locations (Henaff, 2020) or views (Tian et al., 2020) of an image. In this framework, a chosen data sample is referred to as the *anchor*, with samples from the same distribution as the anchor being classified as *positive* samples, while those from a different distribution are considered *negative* samples. The goal of contrastive learning is to minimize (or maximize) the distance between the anchor and positive (negative) samples. SimCLR (Chen et al., 2020) seeks to maximize agreement between different augmented versions of the anchor, Nearest-Neighbor Contrastive Learning (NNCLR) (Dwibedi et al., 2021) diversifies the positive sample space by

sampling the nearest neighbor from the latent space of the original image, and ORE (Joseph et al., 2021) incrementally expands the feature space based on semantic similarity by adding "unknown" classes to learn new ones. However, contrastive learning often requires additional augmentations of the existing dataset, which multiplies its effective size. For example, SimCLR trains with twice the number of images, requiring additional computational resources that may not be readily available. In addition, newer methods, such as masked autoencoders, have been found to outperform state-of-the-art contrastive techniques (He et al., 2022) without the prohibitive cost.

Pretraining paradigms for natural language processing (NLP) have been extensively studied since the advent of transformer-based architectures. That is, NLP models benefit from BERT-style pre-training (Devlin et al., 2019), where the model randomly masks some of the original input tokens and then attempts to predict the vocabulary id of the masked tokens based only on its context. In their seminal paper, (Bao et al., 2022) extends this technique to vision transformers (ViTs), forming the *masked image modeling* (MIM) framework. MIM involves partitioning an image into *masked* and *visible* patches, using the well-known reconstruction loss (see Section 3.5.3) to predict masked patches from visible patches similar to BERT. This technique has recently gained traction and several extensions have followed (He et al., 2022; Xie et al., 2022). Most notably, (Tian et al., 2023) introduced Sparse masKed modeling (SparK), leveraging sparse convolutions for MIM during BERT-style pretraining for CNN-based architectures. SparK uses an encoder with a lightweight decoder for pretraining, gathering all unmasked image patches into a sparse image before applying sparse convolutions to encode it. In this way, convolutions are adapted to handle irregular masked inputs. (Tian et al., 2023) validate their findings on the ImageNet1k classification task, and (Wolf et al., 2023) showed improvements using SparK on two different downstream medical imaging tasks with CT scans: identifying brain hemorrhages (145 images) and multi-class classification of 11 different body organs ($13,952$ images). Although model performance on various downstream tasks has been widely studied, to the best of our knowledge, there has been no investigation into the quality of uncertainty estimates using the SparK framework.

3.5 Self-Supervised Pretraining is Key

3.5.1 Pretraining Setup

As in (Tian et al., 2023), we use a modified UNet architecture for pretraining, consisting of a ResNet-style model for the encoder P_θ and a lightweight decoder Q_ω. The ResNet encoder produces feature maps at 4 different resolutions: $\frac{H}{4}\times\frac{W}{4}$, $\frac{H}{8}\times\frac{W}{8}$, $\frac{H}{16}\times\frac{W}{16}$, and $\frac{H}{32}\times\frac{W}{32}$. The decoder contains three successive blocks to upsample sparse feature maps back to the original input space. Prior to image reconstruction, the decoder performs *densifying* by filling in the empty positions in the sparse feature maps. Thus, given an input $\mathbf{x}$, we generate our reconstruction by $\mathbf{x}' = Q_\omega(P_\theta(\mathbf{x}))$.

Figure 3.8: **Pretrained datasets ranked in relevance to fine-tuning dataset.** (Ordered from left to right) ImageNet1k dataset has the least semantic relevance, while UPMC dataset has highest relevance to the DART hand dataset used for the downstream task of rash severity classification.

3.5.2 Pretraining Datasets

ImageNet1k. The full ImageNet dataset is currently the most comprehensive and diverse visual recognition database in the image world (Deng et al., 2009). It boasts a total of 14.2 million cleanly annotated images spread over $21,841$ categories (i.e., *house finch*, *table lamp*, *volcano*, *promontory*). In computer vision applications, typically only $1,000$ of the high-level categories are used, termed *ImageNet1k*. This subset consists of approximately 1.28 million images. Training on ImageNet1k enables the model to learn robust and generalizable features to improve the convergence of training on most downstream tasks.

MIT Fitzpatrick17k. The Fitzpatrick17k dataset contains $16,577$ clinical images with skin condition labels and skin type labels based on the Fitzpatrick scoring scale (Groh et al., 2021, 2022), a photo-typing scale to measure the response of skin types to UV radiation. This six-point scale is generally used to classify human skin color. To evaluate the diagnostic accuracy of this dataset, a board-certified dermatologist manually verified 3% of the dataset. It contains images in 22 diverse categories of skin conditions, however, to retain as much semantic similarity to our target dataset, we only use $10,886$ of the images labeled inflammatory.

Google SCIN. SCIN contains $10,379$ images of skin, nail, or hair conditions, directly contributed by individuals experiencing them (Ward et al., 2024). It consists largely of common allergic, inflammatory, and infectious conditions, most of which show early-stage concerns. Relevant categories include eruptions (56.4%) and contact dermatitis (10.5%). Additionally, the dataset provides an approximately balanced distribution of images for each Fitzpatrick skin type.

UPMC Rheumatology. We incorporate additional data collected by our rheumatologists for pretraining. While we focus our fine-tuning analysis on the in-clinic 3D images captured by the VECTRA H1 camera, we reserve the in-clinic and telemedicine 2D smartphone images (*UPMC-*

Smartphone) for pretraining. We also utilize images depicting DM manifestations across the body (*UPMC-Body*). Although these images do not contain hand rashes, they still provide valuable semantic information about rash appearance, texture, and size, which can enhance our fine-tuning process. After removing images with identifiable information, we are left with 411 images. Including in-clinic and telemedicine images, we have a total of 599 images.

Pretraining Dataset	n images
ImageNet1k	$1,281,167$
Fitzpatrick17k	$10,886$
SCIN	$10,379$
UPMC	
UPMC-Body	411
UPMC-Smartphone	188

Table 3.1: **Number of images in each pretraining dataset.**

Examples of each of the pretrained datasets are shown in Figure 3.8 and the sizes of each are provided in Table 3.1. The diversity in images of different skin tones is highly advantageous for our research, as our processed DART hand dataset currently contains images from only Caucasian-American patients. Due to limited dataset availability, our selected datasets have varying degrees of semantic similarity to our DART hand dataset. We collaborate with UPMC clinicians to assess the relevance of each pretraining dataset to our fine-tuning DART dataset. Clinicians identify ImageNet1k to have the least semantic similarity to our dataset, whereas the UPMC dataset has the most. All images are standardized using ImageNet1k statistics to ensure consistency of the input data distribution.

3.5.3 Loss Specification

We upsample the decoder output until we achieve the original resolution of the input $\mathbf{x} \in \mathbb{R}^{H \times W}$. As per (He et al., 2022), we compute per-patch normalized pixels as targets for L_2-loss, calculating errors only on the masked positions. Let φ represent the function that extracts masked positions from an image. Given target $\mathbf{t}$, a trainable encoder P_θ, and decoder Q_ω, we minimize

$$\mathcal{L}_{\text{recon}} \triangleq \frac{1}{N} \sum_{n=1}^{N} \left(\varphi(\mathbf{t}^{(n)}) - \varphi \underbrace{\left[Q_\omega(P_\theta(\mathbf{x}^{(n)})) \right]}_{\mathbf{x}^{(n)\prime}} \right)^2 \tag{3.4}$$

3.5.4 Results for Pretraining via Hierarchical Masked Image Modeling

Without ImageNet1k, our aggregated pretraining dataset ($21,864$ images) is insufficient to develop a robust inductive bias for effective fine-tuning. As such, we perform a two-stage incremental strategy: First, we initialize pretrained weights on ImageNet1k from (Tian et al.,

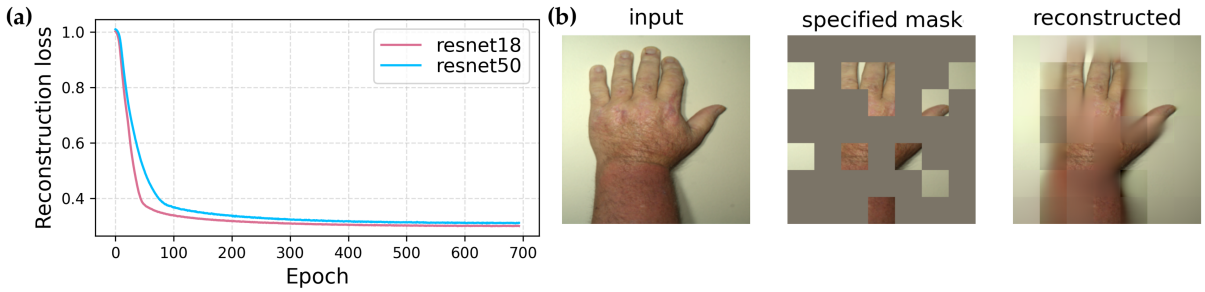


Figure 3.9: **Pretraining ResNet-18 vs. ResNet-50 architectures using hierarchical masked image modeling.** (a) Pretraining curves for two SOTA convolutional neural networks, ResNet-18 and ResNet-50. (b) Example masked reconstruction from SparK encoder-decoder model after pretraining.

2023). Next, we perform additional pretraining for 700 epochs on our domain-specific aggregated datasets from Section 3.5.2. Our training curves are shown in Figure 3.9a. We observe that ResNet-18 exhibits faster pretraining convergence and consistently achieves a lower reconstruction loss at the end of each epoch over ResNet-50. After 700 pretraining epochs, ResNet-18 achieves a loss of 0.300, whereas ResNet-50 achieves only 0.311. In Figure 3.9b we show the results of masked reconstruction after pretraining. We set the mask ratio to 0.6, indicating that 60% of the image patches should be hidden randomly. Our model is able to recover hidden patches during reconstruction (right), despite masking (middle) a majority of patches in our original input (left). ResNet-50 is more susceptible to memorization due to its increased complexity, with approximately 13.87 million more parameters than ResNet-18. Hence, we use the ResNet-18 encoder for the downstream severity scoring task.

Optimizing pretraining epochs. To identify the optimal pretraining epochs, we perform linear probing after the following epochs: $[75, 150, 225, 300, 375, 450, 525, 600, 675, 700]$ by fine-tuning only the classification head on top of the ResNet-18 encoder for a total of 75 epochs each (see Figure 3.10). We achieve the best validation accuracy and the second-best validation MAE after only pretraining for 75 epochs. Model performance on the downstream CDASI classification task appears to degrade after 300 epochs. Thus, self-supervised pretraining for 75-300 epochs with our aggregated dataset is ideal. Evidently, our model learns relevant domain-specific features with minimal pretraining required. For the analyses in the next section, we choose to initialize our fine-tuning procedure with weights obtained after 75 pretraining epochs.

3.6 Fine-tuning for Automated Severity Scoring

3.6.1 Fine-tuning Dataset

UPMC DART. Since we use in-clinic 3D images converted to 2D to fine-tune our encoder (see leftmost pair in Figure 3.2), we must preprocess our images as described in Section 3.2.2. To

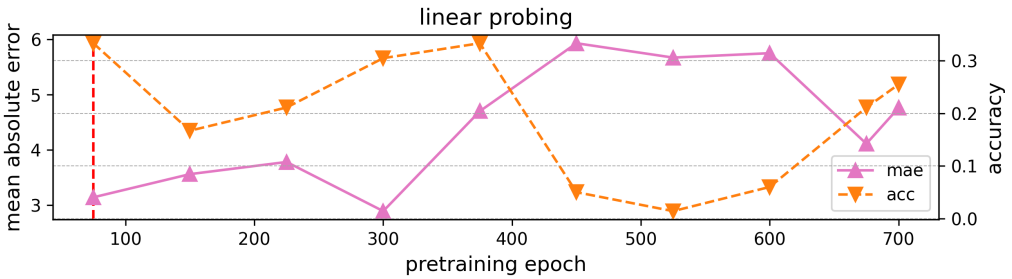


Figure 3.10: **Hyperparameter optimization for pretraining epoch.** Validation accuracy is maximized after only 75 epochs.

enhance our current augmented dataset, we additionally utilize AutoAugment[3] from the `timm` (PyTorch Image Models) library. AutoAugment searches for the optimal combination of transformation policies for a specific dataset, aiming to enhance the generalization ability our model by creating more diverse training samples. We also enable Mixup (Zhang et al., 2017) during training to prevent memorization and sensitivity to adversarial examples. Mixup is a data-agnostic data augmentation routine that constructs virtual training examples $\tilde{\mathbf{x}}$ as linear combinations of existing samples $(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}), (\mathbf{x}^{(j)}, \mathbf{y}^{(j)})$ by

$$\begin{aligned} \tilde{\mathbf{x}} &= \lambda \mathbf{x}^{(i)} + (1-\lambda)\mathbf{x}^{(j)} \\ \tilde{\mathbf{y}} &= \lambda \mathbf{y}^{(i)} + (1-\lambda)\mathbf{y}^{(j)} \end{aligned} \tag{3.5}$$

where $\lambda \sim \text{Beta}(0.1, 0.1)$. However, linear interpolations generated by Mixup tend to be locally ambiguous and unnatural. To balance this effect, we also utilize CutMix (Yun et al., 2019), which replaces an image region with a patch from another training image to generate more natural images. From existing samples, we generate new samples in CutMix by

$$\begin{aligned} \tilde{\mathbf{x}} &= \mathbf{M} \odot \mathbf{x}^{(i)} + (\mathbf{1} - \mathbf{M}) \odot \mathbf{x}^{(j)} \\ \tilde{\mathbf{y}} &= \lambda \mathbf{y}^{(i)} + (1-\lambda)\mathbf{y}^{(j)} \end{aligned} \tag{3.6}$$

where $\mathbf{M} \in \{0,1\}^{W \times H}$ denotes a binary mask indicating where to drop out and fill in from two images and $\lambda \sim \text{Unif}(0,1)$ represents the combination ratio. Similarly to pretraining, we standardize the DART dataset using ImageNet1k statistics to ensure data distribution consistency.

Due to the varying number of samples for each CDASI score, we employ pseudo-randomized group k-fold cross-validation to prevent patient data leakage between training and testing datasets. Specifically, we randomly select groups to ensure that CDASI 11, which contains only a single patient, is not included in the test set. Without this precaution, we would risk evaluating the model on a class it has never encountered during training, thereby distorting its performance.

After pretraining, we discard the decoder and solely fine-tune the encoder on our DART dataset.

[3] `https://timm.fast.ai/AutoAugment`

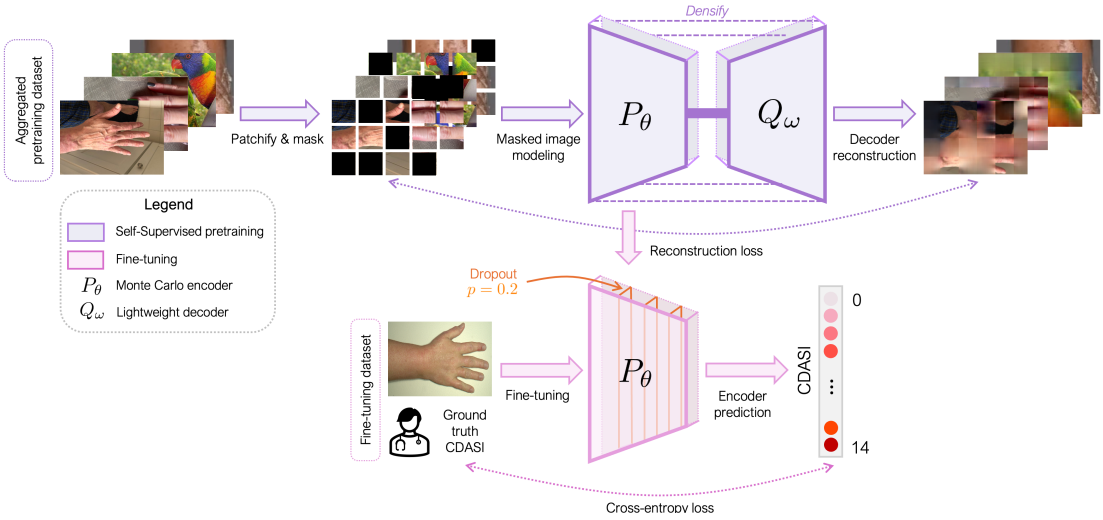


Figure 3.11: **Proposed Monte Carlo (MC) SparKNet pretraining and fine-tuning framework.** Encoder-decoder model is pretrained using BERT-style self-supervised learning on domain-relevant images, then encoder is fine-tuned on DART hand rash dataset. Dropout layers are situated in encoder to perform MC dropout.

3.6.2 Loss Specification

We optimize our encoder P_θ using cross-entropy loss

$$\mathcal{L}_{\text{classification}} \coloneqq -\sum_{n=1}^{N}\sum_{c=1}^{C} \mathbf{y}_c^{(n)} \log \widehat{\mathbf{y}}_c^{(n)} \tag{3.7}$$

and utilize the Layer-wise Adaptive Moments optimizer for Batch training (LAMB) (You et al., 2020) to prevent very large or very small updates during optimization. We initialize the hyperparameters $\beta_1 = 0.9, \beta_2 = 0.999$ and use the square root learning rate scaling rule and linear epoch warm-up scheduling to automatically adjust the learning rate.

3.7 Experiments

3.7.1 Unsupervised Pretraining Improves Fine-tuning Capabilities

Transfer learning strategies. To assess the effects of SparKNet on our fine-tuning dataset, we evaluate three different transfer learning strategies:

(1) **Feature extraction (Fe)**: All ResNet layers are frozen except for the classification head

(2) **Partial fine-tuning (PFt)**: Earlier convolutional blocks are frozen to preserve feature extraction capabilities, and the last few blocks are trainable. For example, "PFt-3-4" indicates that the last two blocks are unfrozen during fine-tuning

(3) **Full fine-tuning (FFt)**: All ResNet layers are unfrozen and trainable

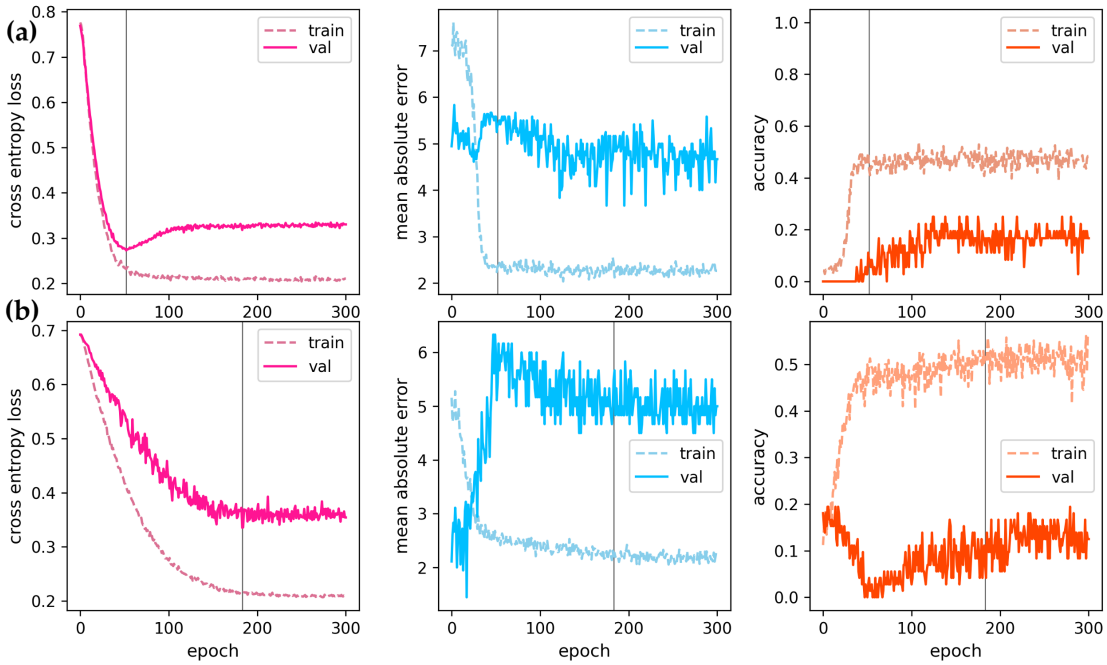


Figure 3.12: **Training and validation curves for full fine-tuning (FFt) paradigm.** We show curves for (a) self-supervised (SSL) pretraining and (b) supervised (SL) pretraining. Gray line delineates early stopping epoch identified by loss on the validation set. Validation MAE appears to decrease as our encoder is trained for longer for both supervised and self-supervised pretraining. While our validation accuracy does not exceed 32% after 120 training epochs for SSL, we cannot achieve beyond 20.5% accuracy for SL during the entire fine-tuning process.

All experiments are conducted on the ResNet-18 architecture, the smallest model in the ResNet family, chosen to avoid overfitting given the limited size of the training dataset. We fine-tune our encoder for 300 total epochs for each strategy and perform early stopping using the held-out validation set to prevent the model from overfitting to the training data. The mean absolute error (MAE), Root Mean Squared Error (RMSE) and model accuracy are evaluated at the end of each training epoch. Although we are treating this as a classification problem, we aim to take advantage of the natural ordering of the rash severity scores when evaluating our model's performance. The MAE and RMSE enable us to assess the relative distance between our predictions and the target score, unlike the accuracy. Figure 3.12 illustrates the training and validation curves for the full fine-tuning technique. Fine-tuning after self-supervised pretraining (SparKNet) appears to be more stable than fine-tuning after supervised pretraining, consistently achieving higher validation accuracy than the baseline model after 120 epochs. Additionally, SparKNet demonstrates faster convergence, beginning to overfit after at most 60 epochs, compared to 185 epochs for the baseline model. Across all transfer learning strategies, we observe that our models typically begin to overfit after approximately 75 epochs. Furthermore, SparKNet does not achieve accuracy greater than 32% on the validation set beyond 120 training epochs. Our inspection of the validation sets indicates that they frequently contain patients with CDASI scores absent in the training dataset, leading to an underestimation of the model's true performance. We show the results for all transfer learning strategies on the test set averaged across five different runs in Table 3.2.

Model	**Pretraining (Pt)**		**Finetuning (Ft)**		
	Supervised?	**Pt Dataset**	**TL Strategy**	**MAE**(↓)	**Accuracy**(↑)
OrdReg (ours)	None	-	FFt	1.160 ± 0.518	0.484 ± 0.187
ResNet-18	✓	ImageNet1k	Fe	1.775 ± 0.117	0.442 ± 0.022
			FFt	1.850 ± 0.101	0.383 ± 0.034
SparKNet-18 (ours)	✗	ImageNet1k, SCIN, Fitz17k, UPMC	Fe	2.344 ± 0.080	0.169 ± 0.014
			PFt-4	1.867 ± 0.293	0.469 ± 0.075
			PFt-3-4	1.167 ± 0.237	$\underline{0.775 \pm 0.041}$
			PFt-2-3-4	$\underline{1.042 \pm 0.195}$	0.683 ± 0.059
			FFt	$\mathbf{1.008 \pm 0.187}$	$\mathbf{0.789 \pm 0.039}$

Table 3.2: **ResNet-18 and SparKNet-18 model performance results across five runs for different fine-tuning strategies.** SparKNet-18 with full fine-tuning achieves the highest performance (highlighted), with partial fine-tuning with the last two blocks unfrozen achieving the second highest accuracy (underlined). We evaluate using MAE(↓) and accuracy(↑) metrics.

SparKNet outperforms costly supervised pretraining strategies. Compared to the baseline FFt model using supervised pretraining, our SparkNet FFt model using self-supervised pretrain-

ing achieves a 40.6% improvement in accuracy and a 0.842 reduction in average MAE when evaluated on the test sets. Additionally, this improvement is accompanied by faster convergence: The supervised pretraining model takes 185 epochs to train, while fine-tuning our self-supervised model requires only 60 epochs. As a result, we achieve a speed-up in both the pretraining and fine-tuning stages. Our self-supervised approach bypasses the costly and time-consuming labeling process during pretraining, while learning richer and domain-specific representations. This allows the model to reach strong performance in fewer epochs compared to the supervised pretraining baseline. Despite the improvements in the FFt approach, we observe poorer performance in SparKNet when all layers are frozen and only the classification head is fine-tuned. Our baseline achieves 44.2% accuracy averaged across the test sets, whereas SparKNet only achieves about 16.9% accuracy. Supervised learning relies on explicit labels, which likely helps our baseline model learn more discriminative features for classification tasks. In contrast, self-supervised learning may be less structured, potentially leading to less effective feature extraction.

Freezing layers tends to restrict fine-tuning performance. From Table 3.2, we observe that a decrease in the number of trainable parameters in our model is associated with a decrease in model performance. FFt consistently outperforms Fe during each run, with FFt achieving a 78.9% accuracy on the held-out test set and Fe only achieving 16.9%. Furthermore, FFt achieves the lowest MAE of 1.008 compared to 2.344 for Fe. PFt is the only strategy that does not consistently improve by unfreezing additional layers. In fact, although unfreezing the last two blocks performs almost as well as FFt, we observe a degradation in performance when unfreezing the last three blocks. It is possible that pretraining the last two blocks balances learning high-level features without overfitting the task-specific details learned in the final block. This may help the model retain useful transferable features while avoiding over-specialization.

SparKNet exceeds clinical performance expectations. Our rheumatologists require a model accuracy of at least 70%-75% to achieve clinical relevance and consider adoption in telehealth services. Both our partial fine-tuning and full fine-tuning strategies exceed clinical expectations, achieving an average accuracy of 77.5% and 78.9%, respectively. In addition, we observe a considerable 30.5% improvement in accuracy over our ordinal regression model, demonstrating superior discrimination of visual rash characteristics at different CDASI score values.

We recognize that clinicians do not make a clear distinction between pairs of successive CDASI scores, motivating a coarser scale of rash severity scores. Therefore, we also analyze model performance after reducing the granularity of the classes to help interpret rash severity in broader terms. We evaluate off-by-k accuracy for $k = \{0, 1, 2\}$, which considers predictions to be correct as long as they are within k classes from the target, providing a more lenient evaluation of the classifier's performance (Table 3.3). Off-by-0 accuracy is considered exact accuracy. As expected, both SparKNet and the baseline ResNet show improved accuracy scores as we simplify the CDASI scoring scale. Our results indicate a higher performance for SparkNet across all scales, suggesting stronger generalization capabilities than the baseline ResNet.

Model	Exact Accuracy(↑)	Off-by-1 Accuracy(↑)	Off-by-2 Accuracy(↑)
RESNET-18	0.383 ± 0.034	0.454 ± 0.040	0.458 ± 0.042
SPARKNET-18	$\mathbf{0.789 \pm 0.039}$	$\mathbf{0.796 \pm 0.047}$	$\mathbf{0.819 \pm 0.042}$

Table 3.3: **Comparison of baseline ResNet-18 and SparKNet-18 models across different levels of CDASI score granularity.** SparKNet-18 (ours) consistently achieves higher accuracy (↑) for each scale.

3.7.2 Uncertainty Quantification with SparKNet

Why SparKNet? The Monte Carlo Dropout (MCD) technique for uncertainty quantification relies on situating dropout layers after each convolutional and fully-connected layer (see Section 2.2.3). Thus, MCD cannot be implemented in a meaningful or effective manner in transformer-based models, such as Vision Transformers (ViTs), which lack convolutional backbones. When applied, dropout in transformer ViTs is often structured and token-specific, which fails to induce meaningful stochasticity in the final output distribution required for the MCD technique.

SparKNet addresses these challenges by leveraging the benefits of BERT-style pretraining, commonly used in transformer-based models, *while maintaining the capability to perform deep uncertainty quantification*. To implement MCD, we situate dropout layers after each sparse convolutional layer and the classification head in our SparKNet model. We set the drop rate to $p = 0.2$ for each dropout layer and train with weight decay to enforce L_2 regularization on our model weights. During inference, $T = 50$ Monte Carlo samples are generated. We refer to our model with MCD as MC-SparKNet.

SparKNet ensembles. In addition to MCD, we also train a deep ensemble of SparKNet models to assess the relative reliability of uncertainty estimates. However, unlike our approach in (Banerjee et al., 2025), we cannot use weight initialization or bootstrapping techniques to introduce diversity into our ensemble, as our current method hinges on utilizing self-supervised pretrained weights. Instead, we opt to slightly perturb the pretrained weights of each ensemble member with Gaussian noise to push each member to explore different basins of attraction during fine-tuning. Evidently, overly distorting the weights can lead an ensemble member to converge at a suboptimal minimum. To prevent this, we use the standard deviation (SD) of the parameters as the SD hyperparameter of Gaussian noise. Our noise injection procedure is detailed in Algorithm 4. We empirically find that scaling our noise term $\eta_{\ell,j}$ by $\gamma_\ell = 0.05$ at each layer ℓ produces the most reliable ensemble predictions. As in MCD, we train each ensemble member with weight decay to enable MAP estimation for model weights. We refer to our Deep Ensemble as DE-SparKNet.

MC-SparKNet provides more reliable uncertainty estimates. We evaluate both model performance and uncertainty estimation quality for both uncertainty quantification techniques: DE-SparKNet and MC-SparKNet (Table 3.4). Both models are fine-tuned using the full-finetuning (FFt) transfer learning strategy. MC-SparKNet outperforms DE-SparKNet with a 20.3% im-

Algorithm 4 Gaussian Noise Injection on Pretrained Weights

Input: Trainable parameters $\omega_{\ell,j}$ for each tensor element j in layer $\ell \in \{1, 2, \cdots, L\}$.
Output: Perturbed weights $\tilde{\omega}_{\ell,j}$
for layer ℓ in range L **do**
    $\sigma_\ell \leftarrow \mathtt{sd}(\omega_{\ell,*})$
    for element j in range J **do**
        $\tilde{\omega}_{\ell,j} \leftarrow \omega_{\ell,j} + \gamma_\ell \cdot \eta_{\ell,j}, \quad \eta_{\ell,j} \sim \mathcal{N}(0, \sigma_\ell^2)$ ▷ **additive Gaussian noise**
    end for
end for
return New weight initialization $\tilde{\omega}_{\ell,j}$ for fine-tuning

provement in accuracy, alongside notable reductions in MAE and RMSE.

To assess uncertainty quality, we compute the estimated Rényi divergence across fifteen different model runs. MC-SparKNet exhibits a 38.2% higher divergence than DE-SparKNet, indicating a stronger ability to distinguish between correct and incorrect predictions. In contrast, DE-SparKNet suffers from a 16.1% reduction in accuracy relative to the single FFt SparKNet, suggesting a considerable drop in performance when ensembling. This performance degradation is unexpected, as ensembles typically benefit from averaging the outputs of multiple well-performing models. However, in our case, the ensemble constituents likely converge to disparate and sub-optimal local minima. This divergence reduces the ensemble's predictive coherence and contributes to its poorer discrimination between correct and incorrect predictions (Figure 3.13b).

Compared to deep ensembles, MC Dropout maintains a more consistent inductive bias across forward passes, since all predictions are drawn from the same weight space with randomized dropout masks. This structural consistency may lead to more reliable epistemic uncertainty estimates, especially when fine-tuning on small datasets. In contrast, ensemble members trained independently can drift toward different subspaces of the loss landscape, introducing more variance but not necessarily capturing meaningful uncertainty. MC Dropout may offer superior uncertainty calibration in low-data or transfer learning settings.

	Model Performance			**Model Quality**
Model	Accuracy(↑)	MAE(↓)	RMSE(↓)	$\widehat{R}_\alpha(incorr \mid\mid corr)(\uparrow)$
SparKNet †	$\underline{0.789 \pm 0.039}$	$\mathbf{1.008 \pm 0.187}$	$\mathbf{2.369 \pm 0.573}$	–
DE-SparKNet	0.628 ± 0.038	1.471 ± 0.205	2.557 ± 1.147	0.307 ± 0.140
MC-SparKNet	$\mathbf{0.831 \pm 0.042}$	$\underline{1.017 \pm 0.255}$	$\underline{2.470 \pm 1.237}$	$\mathbf{0.689 \pm 0.083}$

† represents our single FFt SparKNet model from Table 3.2.

Table 3.4: **Model performance and quality evaluation across single SparKNet, DE-SparKNet, and MC-SparKNet architectures.** MC-SparKNet exhibits best accuracy (bold) and second best (underline) MAE and RMSE scores evaluated on held-out test patients. Compared to DE-SparKNet, it provides stronger discrimination between incorrect and correct classifications.

Figure 3.13: **Epistemic uncertainty distributions for correct and incorrect classifications** for (a) MC-SparKNet and (b) DE-SparKNet. Although both techniques exhibit moderate overlap, MC-SparKNet provides slightly stronger discrimination between correct and incorrect predictions.

3.7.3 Out-of-Distribution Detection

Out-of-distribution (OOD) detection refers to a model's ability to detect anomalous data that is inconsistent with the training data distribution. Models that have strong generalizability tend to have robust OOD detection capabilities. In this section, we examine the OOD detection capabilities of SparKNet versus the baseline ResNet against two different forms of OOD samples:

(1) Same imaging conditions, but different visual characteristics (African-American patient)

(2) Same visual characteristics, but different imaging conditions (UPMC-Smartphone dataset)

SparKNet improves out-of-distribution detection performance. Our results in Table 3.5 evaluate model performance on several semantically challenging out-of-distribution (OOD) cases: two OOD-lighting patients (P1005 and P1035), the full set of OOD-lighting patients, and the only available African-American patient (OOD-AA). The notation DART→OOD-* refers to fine-tuning on the DART hand dataset and evaluating on the specified OOD dataset.

Across all OOD-lighting patients, SparKNet achieves a substantial reduction in prediction error, reducing MAE by 0.43 and RMSE by 0.65 compared to our ResNet baseline. Notably, even in the more challenging setting of the OOD-AA patient, SparKNet reduces MAE by 0.28 and RMSE by 0.51. These improvements highlight SparKNet's enhanced ability to generalize beyond the distribution trained on during fine-tuning.

Crucially, this generalizability can be attributed in part to pretraining on datasets with diverse skin tones, such as SCIN and Fitzpatrick17k. Although the downstream DART dataset contains only Caucasian-American patients, pretraining on demographically diverse images enables SparKNet to better handle variations in skin tone and lighting at inference time. This result underscores the importance of incorporating diversity during pretraining—not only for inclusivity, but also for improving the robustness of underrepresented groups in downstream tasks.

Model	Metric	DART→OOD-Lighting			DART→OOD-AA
		P1005	P1035	All	P1009
ResNet-18	MAE(↓)	2.567 ±0.25	3.795 ±0.21	2.494 ±0.14	2.750 ±0.53
	RMSE(↓)	2.944 ±1.52	3.844 ±1.18	3.120 ±1.13	3.011 ±1.61
SparKNet-18	MAE(↓)	**2.405** ±0.55	**2.750** ±0.40	**2.062** ±0.14	**2.475** ±0.24
	RMSE(↓)	**2.632** ±1.30	**2.915** ±1.55	**2.466** ±0.77	**2.505** ±1.08

Table 3.5: **Preliminary out-of-distribution (OOD) detection evaluation for baseline ResNet-18 and SparKNet-18.** We compute MAE(↓) and RMSE(↓) values for both models for two individual OOD-lighting patients (PIDs P1005 and P1035), all OOD-lighting patients at once, and the single available African-American (OOD-AA) patient. SparKNet-18 consistently exhibits lower error across all OOD patients.

SparKNet uncertainties are effective out-of-distribution detectors. We observe two interesting patterns in Figure 3.14: First, OOD samples categorized by camera lighting (green) generally exhibit similar aleatoric uncertainty but considerably higher epistemic uncertainty compared to in-distribution (IND) samples from the same patients. In contrast, OOD samples based on skin tone (orange) tend to show more pronounced aleatoric uncertainty with similar levels of epistemic uncertainty. These observations intuitively align with the definitions of aleatoric and epistemic uncertainty. The distribution of OOD-lighting samples is not present in the training set, leading to high epistemic uncertainty during inference. However, images of African-American patients are captured with the same VECTRA H1 camera as the training images, meaning they primarily exhibit higher variability compared to the training set. Thus, while the epistemic uncertainties are similar to those of IND samples, their aleatoric uncertainties are more pronounced.

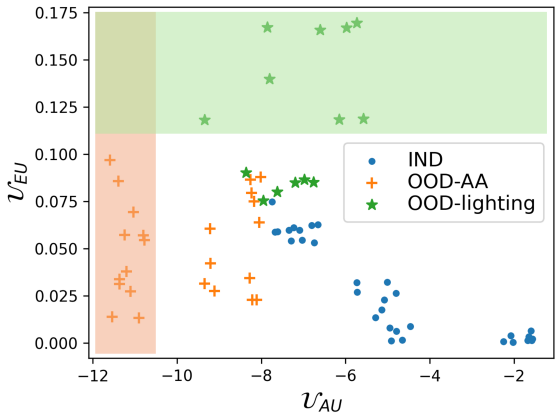


Figure 3.14: **MC-SparKNet uncertainty estimates for in-distribution (IND) and out-of-distribution (OOD) samples.** IND samples (blue) tend to have lower aleatoric and epistemic uncertainties. OOD samples, either due to skin tone (OOD-AA) or lighting (OOD-lighting) differences, tend to have more extreme uncertainty values, indicating well-calibrated uncertainties to the input distribution.

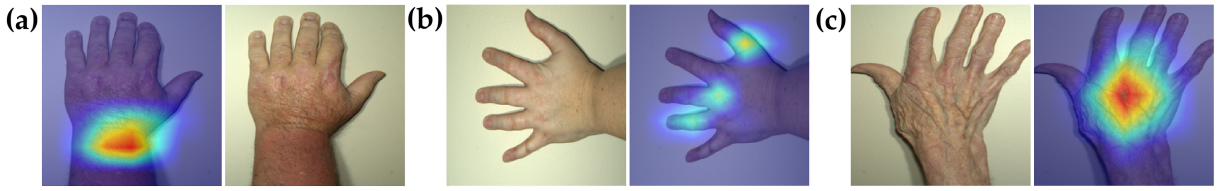


Figure 3.15: **Grad-CAM visualizations highlight salient regions that contribute to SparKNet's predictions.** (a) and (b) demonstrate a correlation between localized rash areas and the regions that SparKNet focuses on for predictions. (c) For mature skin, the focus region exhibits greater variability, likely due to added complexity from more textured skin.

3.7.4 Assessing Model Explainability

The previous sections stressed the implications of evaluating model *interpretability*. However, we now shift our focus to evaluating model *explainability*. Interpretability centers on simplifying the mechanisms or structures to make a model's decision-making process more easily understood by humans, whereas explainability focuses on providing clear, human-understandable rationale for why a model made a particular decision or prediction.

Gradient-Weighted Class Activation Mapping (Grad-CAM) (Gildenblat and contributors, 2021) is a widely-used post-hoc explainable AI (XAI) technique, leveraging the gradients of a particular class to produce intuitive saliency maps that highlight the key regions of an image relevant to that class. This technique is valuable for visualizing which areas of the image are influential in the model's decision. By incorporating Grad-CAM, we can provide clinicians with enhanced transparency and insight into our model's reasoning, which is essential for building trust and facilitating its integration into clinical practice.

We illustrate examples of Grad-CAM saliency maps on our DART hand images in Figure 3.15. First, we observe strong correlations between rash redness and area with the regions highlighted by Grad-CAM, indicating that SparKNet relies on the precise rash regions when making its classification prediction. Additionally, regions of higher focus appear on the knuckles and joints of the hand. This provides evidence that SparKNet's DL features are correlated with the original handcrafted features we explored earlier. Second, mature skin appears to be more challenging for SparKNet than younger skin, as the highlighted region in Figure 3.15c is less precise and has greater variability than (a) or (b).

Thus, our results indicate a strong potential for exploring the relationship between DL and handcrafted clinical features as a part of future work.

3.8 Future Work

Limitations. We provide the foundation for the development of uncertainty-guided AI systems for automated cutaneous dermatomyositis rash severity prediction. Our evaluations are limited to the small number of existing patients in the DART study and we intend to assess the robustness of our proposed methodologies on a larger cohort of patients when available. In the future, we

aim to broaden our scope to include a wider range of clinics, enhancing the generalizability of our models across diverse data distributions.

Future work. First, we want to further investigate the effects of self-supervised learning on our target task of CDASI severity scoring to better understand (a) the effect of semantic relevance of pretraining dataset on our downstream performance and (b) whether the minimum number of target samples to achieve reasonable performance is the same between pretraining with supervised learning or with self-supervised learning.

Second, we plan to evaluate the performance of our SparKNet framework directly against that of UPMC clinicians using the same test datasets. This comparison aims to provide an objective assessment of our model framework relative to expert human judgment, strengthening insights into the clinical applicability and reliability of our system.

3.9 Acknowledgments

We thank Dr. Rohit Aggarwal and Dr. Nantakarn Pongtarakulpanit for gathering experiment data and Prakruthi Pradeep for her preliminary analysis of the DART dataset.

Chapter 4

Conclusion

Our research seeks to identify contexts that make machine learning particularly challenging in healthcare settings. At every stage, we make sure to ask the right questions: *Can our models account for data scarcity? How can we convey the uncertainty of our predictions to non-experts in a meaningful way?* We particularly remain cognizant of this last challenge, as we expect that non-experts should be able to interact with machine learning systems without a literacy barrier. Building human-AI trust in an intuitive way is essential for AI adoption in high-stakes clinical settings.

Designing for healthcare settings introduces additional constraints. Two primary limitations make integrating deep learning systems into healthcare challenging: **(1)** a lack of quality data and **(2)** a lack of model transparency. It is imperative to innovate new techniques to improve model robustness without requiring expensive labeled data. Self-supervised learning (SSL) and unsupervised learning (UL) techniques can learn rich, transferable representations from unlabeled data, which are often more abundant than labeled data in medical settings. These techniques prime models for efficient and effective downstream task performance.

Uncertainty quantification in deep learning. Monte Carlo dropout (MCD) and deep ensembles represent two leading Bayesian approximation methodologies to assess parameter uncertainty in deep neural networks. Deep ensembles involve training multiple networks independently, with the expectation that, given sufficient diversity, these networks will converge to distinct basins of attraction within the loss landscape. However, if some ensemble members become sub-optimal, the ensemble's aggregate performance may deteriorate along with the quality of uncertainty estimates. In such scenarios, MCD can offer a more robust alternative. These uncertainty quantification techniques not only provide valuable diagnostic insights into model confidence, but also facilitate more informed decision-making in safety-critical applications, model calibration, and out-of-distribution detection.

Engineering interpretable-by-design AI is challenging. Although both are desirable, model performance and the quality of uncertainty estimates assess distinct aspects of a model's behavior. As such, strong predictive accuracy does not imply high-quality uncertainty estimates, and

improvements in one do not *necessarily* translate to improvements in the other. Although our proposed frameworks demonstrate clinical relevance, additional considerations are required to ensure the reliability of their uncertainties. We saw how this challenge was particularly pronounced in the context of predictive modeling of dermatomyositis. Similarly, using XAI tools like Grad-CAM enhances model transparency for end users. These visualizations are crucial for giving clinicians insight into the model's reasoning.

Evaluating AI performance against clinical assessment induces reliability. Directly comparing AI models to clinician performance is a parallel step toward establishing the reliability and trustworthiness of automated systems in healthcare. By benchmarking MSU-Net and SparKNet against practicing UPMC clinicians, we aim to quantify the alignment between human and AI judgment. This approach not only provides an objective measure of the model's capabilities but also highlights scenarios where AI may offer complementary strengths or require further refinement. When combined with uncertainty quantification, these insights can guide clinicians to interpret AI judgments with appropriate caution, especially in high-uncertainty scenarios, thereby reducing the risk of overreliance and promoting more informed, collaborative decision-making.

Ultimately, our goal is to design reliable, interpretable-by-design AI systems that perform well in real-world applications, not just in theory. Communicating expectations with clinicians is the first foundational step toward achieving clinical relevance with our algorithms. After all,

Acknowledging what you don't know is as vital, if not more so,
as acknowledging what you do know.

Appendix A

Trauma Care in a Rucksack

A.1 Bayesian Variational Inference

A.1.1 Bayesian learning

Consider learnable weights ω and our dataset $\mathcal{D}$. Given some prior over the weights $p(\omega)$, we can iteratively compute the posterior $p(\omega|\mathcal{D})$ using the likelihood $p(\mathcal{D} \mid \omega)$ by

$$p(\omega \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \omega)p(\omega)}{p(\mathcal{D})} = \frac{p(\mathcal{D} \mid \omega)p(\omega)}{\int_{\omega'} p(\mathcal{D} \mid \omega')p(\omega')d\omega'} \tag{A.1}$$

This computation is often referred to as *exact inference* and requires the specification of both the prior and likelihood.

A.1.2 Minimizing KL divergence is equivalent to maximizing ELBO

Consider the KL divergence of $q_\phi(\omega)$ from $p(\omega|D)$, where q_ϕ is a member of the family of distributions $\mathcal{Q}$. We show that minimizing this expression is equivalent to maximizing for ELBO. First, using Equation A.1 from above

$$\begin{aligned}
\mathrm{KL}(q_\phi(\omega) \mid\mid p(\omega \mid \mathcal{D})) &= \int_\omega q_\phi(\omega) \log \frac{q_\phi(\omega)}{p(\omega \mid \mathcal{D})} \\
&= \int_\omega q_\phi(\omega) \log q_\phi(\omega) - \int_\omega q_\phi(\omega) \log p(\omega \mid \mathcal{D}) \\
&= \mathbb{E}_{q_\phi}\left[\log q_\phi(\omega)\right] - \int_\omega q_\phi(\omega) \log \frac{p(\omega, \mathcal{D})}{p(\mathcal{D})} \\
&= \mathbb{E}_{q_\phi}\left[\log q_\phi(\omega)\right] - \int_\omega q_\phi(\omega) \log p(\omega, \mathcal{D}) + \log p(\mathcal{D}) \\
&= \underbrace{\mathbb{E}_{q_\phi}\left[\log q_\phi(\omega)\right] - \mathbb{E}_{q_\phi}\left[\log p(\omega, \mathcal{D})\right]}_{-\mathrm{ELBO}} + \log p(\mathcal{D})
\end{aligned}$$

Although $\log p(\mathcal{D})$ is intractable, we observe that it has no relation to the parameters of q_ϕ. When optimizing with respect to ϕ, this term can be treated as a constant and effectively ignored, hence

$$\underset{\phi}{\text{argmin}}\ \text{KL}(q_\phi(\omega)\ ||\ p(\omega \mid \mathcal{D})) \equiv \underset{\phi}{\text{argmax}}\ \text{ELBO}(q_\phi)$$

A.2 Rényi Divergence Estimation

A.2.1 Vectorized implementation

Our contribution (Table A.1) reduces the total execution time to compute the Rényi divergence estimator (Poczos and Schneider, 2011) from more than 100 minutes to just under 1.5 minutes with negligible loss of accuracy on large datasets using a single NVIDIA RTX A6000 GPU. Thus, when used in the boostrapping procedure, we reduce the computation time of our bootstrapped confidence interval **from** > 70 ***days*** **to just** 22 **hours**.

Sample size N	**Init. exec. time**(s)	**Vec. exec. time**(s)	**Speedup**
2000	0.326	**0.00297**	122.02x
20000	3.657	**0.0208**	176.15x
200000	36.314	**0.2709**	134.02x
15100000	6017.9	**80.97**	74.32x

Table A.1: **Improved execution times for vectorized computation of non-parametric Rényi divergence estimator.** Our vectorized approach achieves between 75 to 175 times the speedup over the baseline implementation.

A.3 Bootstrapping

A.3.1 Naïve Efron-type bootstrap

The naïve Efron-type bootstrap requires sampling our original dataset $\mathcal{D}$ of size n with replacement until we generate n new samples to form our bootstrapped dataset $\mathcal{D}'$:

Algorithm 5 Naïve Efron-type bootstrap

Input: $\mathcal{D} = \{\mathbf{x}^{(i)}, \mathbf{y}^{(i)} \mid \forall i\}$ and $n = |\mathcal{D}|$
Output: Bootstrapped dataset $\mathcal{D}'$
$\mathcal{D}' \leftarrow \{\}$
for $i = 1, 2, \cdots, n$ **do**
  Uniformly sample $(\mathbf{x}'^{(i)}, \mathbf{y}'^{(i)}) \sim \mathcal{D}$
  Add $(\mathbf{x}'^{(i)}, \mathbf{y}'^{(i)})$ to $\mathcal{D}'$
  Replace $(\mathbf{x}'^{(i)}, \mathbf{y}'^{(i)})$ in $\mathcal{D}$
end for
return $\mathcal{D}'$

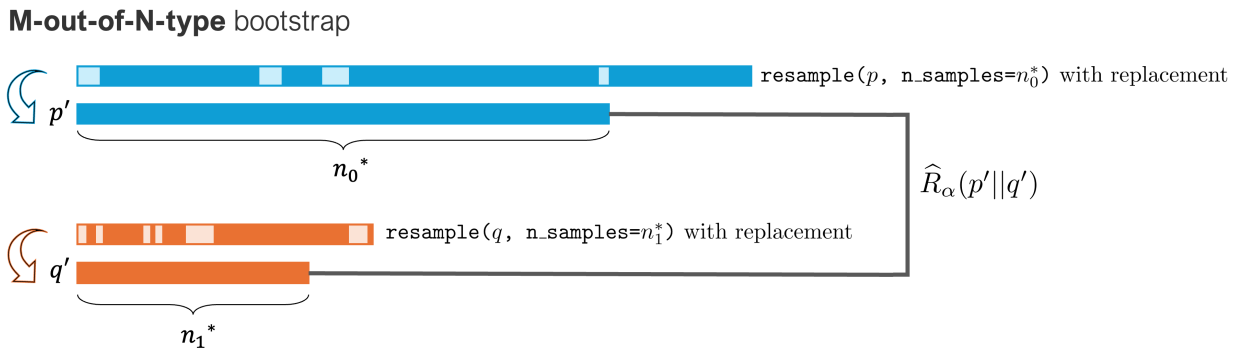


Figure A.1: **Visualizing the M-out-of-N (MooN) bootstrap.** Instead of sampling from the original sets we instead subsample the sets at the same rate to form p', q' and compute the Rényi divergence estimator on the new sets.

The success of the bootstrap, particularly in injecting diversity across different runs of Algorithm 5, results from the observation that $\mathcal{D}'$ contains only $\approx 63.2\%$ samples from the original dataset $\mathcal{D}$. Consider the following informal proof: Since we are uniformly sampling from $\mathcal{D}$, in any iteration i a sample t has exactly $\frac{1}{n}$ chance of being selected and thus $1 - \frac{1}{n}$ chance of not being selected. Furthermore, since we are sampling with replacement, subsequent draws are independent, and we can calculate the probability of *never* selecting sample t as $(1 - \frac{1}{n})^n$. Now consider increasing our dataset size n. As n tends to infinity, we observe that our probability of never selecting sample t converges to

$$\lim_{n\to\infty} \left(1 - \frac{1}{n}\right)^n = \frac{1}{e} \approx 0.368$$

It follows that the chance of seeing samples from the original dataset is approximately 63.2%.

A.3.2 M-out-of-N (MooN) bootstrap

An alternative to the non-parametric bootstrap is the M-out-of-N (MooN) bootstrap, particularly in cases involving a non-smooth estimator, such as extreme order statistics. Instead of sampling from the original set $\mathcal{D}$ of size n, we instead subsample $\mathcal{D}$ forming $\mathcal{D}'$. Theoretically, our subsample size $m = |\mathcal{D}'|$ is determined by the estimator's convergence rate. However, in practice we select a value for m that maintains the appropriate $(1 - \alpha)$ coverage probability for a $(1 - \alpha) * 100\%$ confidence interval. Our implementation for the Rényi divergence estimator is detailed in Algorithm 2 and visualized in Figure A.1.

A.4 Training Details

A.4.1 Model error as a proxy for diversity

In Section 2.4.2 we discuss the procedure of decorrelation maximization to prune our set of candidates. In particular, we leverage model error as a proxy for model diversity. (Lai et al., 2006) describes using binarized errors

$$\text{error}(x_{ij}, y_{ij}, f^{\widehat{\omega}_m}) = \mathbb{1}\left[\text{bin}\left[\sigma_A\left(f^{\widehat{\omega}_m}(x_{ij})\right)\right] \neq y_{ij}\right] \tag{A.2}$$

where $\text{bin}\left[\sigma_A\left(f^{\widehat{\omega}_m}(x_{ij})\right)\right] = 1$ when $\sigma_A\left(f^{\widehat{\omega}_m}(x_{ij})\right) \geq 0.5$ and 0 otherwise. However, Brier score loss provides a more fine-grained representation of the model prediction error as it uses prediction probabilities

$$BS(x_{ij}, y_{ij}, f^{\widehat{\omega}_m}) = \left(\sigma_A\left(f^{\widehat{\omega}_m}(x_{ij})\right) - y_{ij}\right)^2 \tag{A.3}$$

and we choose to evaluate correlations with Brier score loss instead. We note that each pair (x_{ij}, y_{ij}) corresponds to a single pixel and its ground truth label, and σ_A refers to the sigmoid activation function. As an example, consider three models predicting the following probabilities of five pixels: $f_1 \rightarrow \{0.9, 0.8, 0.9, 0.95, 0.97\}$, $f_2 \rightarrow \{0.5, 0.98, 0.6, 0.52, 0.75\}$ and $f_3 \rightarrow \{0.89, 0.81, 0.93, 0.95, 0.97\}$. f_1 and f_3 evidently appear to be highly correlated compared to f_2 and f_3, yet all three will exhibit the same error profile when using binarized errors. Fortunately, Brier score loss can account for this kind of behavior.

We then use Equation A.3 to generate our loss correlation matrix in Figure 2.6. We refer to Figure A.2 to better illustrate the patterns in Brier scores for different pairs of models with varying strengths in correlations.

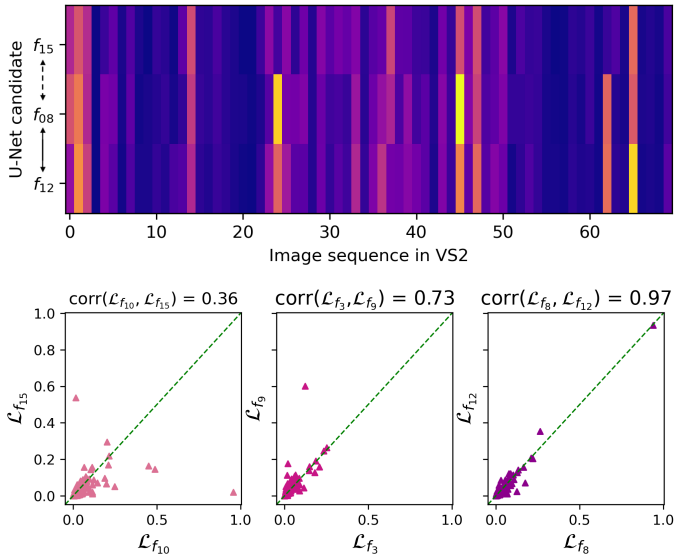


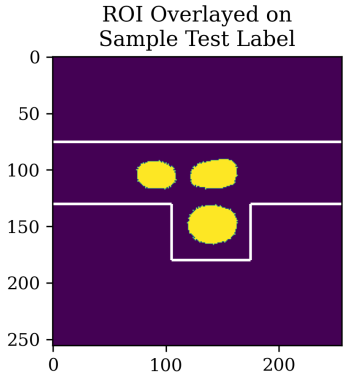
Figure A.2: **Visualizing model error patterns as a proxy for ensemble diversity.** (Top) Brier score loss patterns of three different U-Net candidates across a subset of images in VS2. Visually, $f^{\widehat{\omega}_8}$ appears to have a stronger correlation with $f^{\widehat{\omega}_{12}}$ than $f^{\widehat{\omega}_{15}}$. (Bottom) From left to right: Brier score loss plotted for pairs of models with little, medium, and strong positive correlation.

From Figure A.2 it is evident that $f^{\widehat{\omega}_8}$ and $f^{\widehat{\omega}_{12}}$ produce errors of similar magnitudes for the same images, whereas the errors between $f^{\widehat{\omega}_8}$ and $f^{\widehat{\omega}_{15}}$ appear fairly distinct. From these results, our decorrelation maximization algorithm will keep $f^{\widehat{\omega}_{15}}$ in the final ensemble over $f^{\widehat{\omega}_8}$ or $f^{\widehat{\omega}_{12}}$.

A.4.2 ROI for class imbalance

Our dataset in particular suffers from severe class imbalance, since an overwhelming proportion of total pixels constitute the background and not the vessel. The table in Figure A.3 categorizes the average proportion of pixels labeled vessels at different stages of the US scan, denoted as

"pre-bifurc", "mid-bifurc", and "post-bifurc". Evidently, at most about 3.8% of pixels on average during any particular stage are labeled as vessels. In addition to adding a Dice loss term during optimization to alleviate this issue, we additionally constrain our evaluations to a predefined region of interest (ROI) illustrated on the left in Figure A.3. These boundaries are chosen such that the vessel pixels never exceed this ROI during the entire US scan.



	Average % of pixels labeled as vessels		
	Pre-bifurc	Mid-bifurc	Post-bifurc
Training Set (TR)	0.84	1.75	1.62
Validation Set (VS)	3.50	3.42	2.28
Testing Set (TS)	3.45	3.78	2.46

Figure A.3: **Predefined region of interest (ROI) to account for severe class imbalance.** (Left) Illustration of the ROI on an example ground truth post-bifurcation segmentation mask. (Right) Average percentage of pixels labeled as vessels for different stages of the US scan. For example, "pre-bifurc" stands for the entire sequence of frames pre-bifurcation.

A.4.3 Dataset distribution

We divide frames from the US scan into different datasets for segmentation training as shown in Table A.2.

Dataset	n images
Training Set (TR)	1392
Validation Set (VS)	
VS1	453
VS2	454
Testing Set (TS)	856

Table A.2: Number of images in each dataset subset.

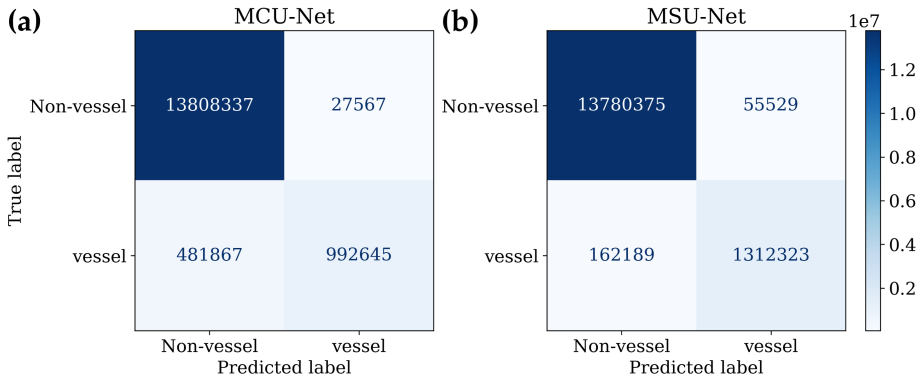


Figure A.4: **Confusion matrices for MCU-Net vs. MSU-Net.** MSU-Net predicts a larger number of true and false positives compared to MCU-Net. However, MSU-Net reduces the number of false negatives by almost 3x from MCU-Net.

A.4.4 Selected hyperparameters

Table A.3 details the optimal hyperparameters used in our experiments.

Task	**Hyperparameter**	**Value**	
		MCU-Net	**MSU-Net**
Model architecture	Ensemble size	1	3
	Ensemble technique	-	Bagging
	Correlation metric	-	Plural-correlation coefficient
Drop rates	`Conv1` Layer	0.5	0.4
	`Conv2` Layer	0.5	0.5
Model training	Early stopping epoch	15	150
AdamW optimizer	lr	$1e^{-4}$	$1e^{-3}$
	β_1	0.9	0.9
	β_2	0.999	0.999

Table A.3: **Selected hyperparameters for each model architecture.** MCU-Net shows signs of overfitting considerably earlier than MSU-Net. Drop rates are identified through empirical tests to find the optimal balance between training stability and performance.

A.5 Model Performance

A.5.1 Confusion matrices

We show the confusion matrices for MCU-Net vs. MSU-Net in Figure A.4.

A.6 Predictive Uncertainty

A.6.1 Expected calibration error

Our results in Figure A.5 indicate that the baseline MCU-Net is as well-calibrated as MSU-Net, as both models have approximately 1% ECE. However, our evaluations in Section 2.7 demonstrate that the ECE fails to capture the potentially catastrophic deficiencies of MCU-Net.

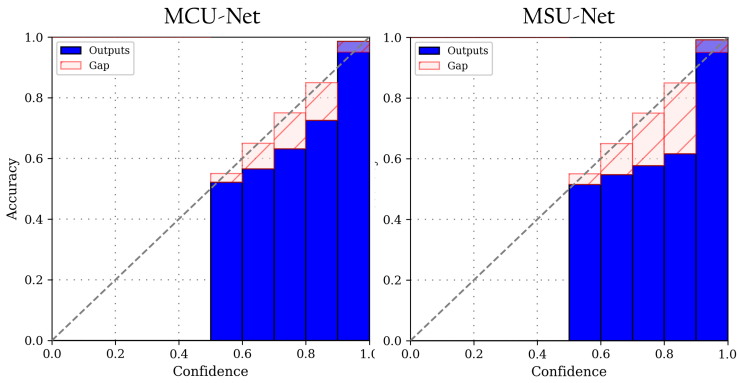


Figure A.5: **Reliability diagrams for MCU-Net vs. MSU-Net.** Confidence values on the x-axis are binned and plotted against accuracy on the y-axis for MCU-Net (left) and MSU-Net (right). Values closer to the diagonal indicate better calibration.

A.6.2 Additional vessel segmentation examples

We show several examples of predictions and their corresponding model (epistemic) uncertainty maps in Figure A.6. MSU-Net achieves high precision in predicting vessel boundaries and is able to effectively capture bifurcation behavior.

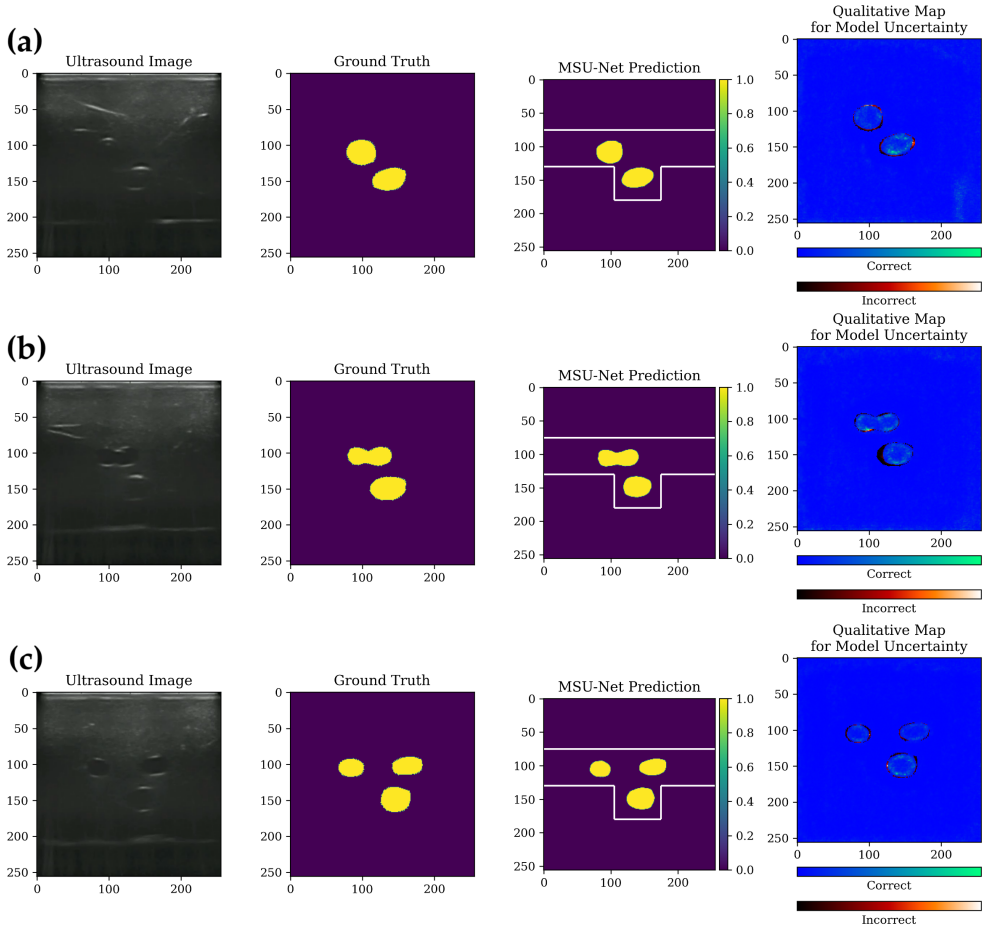


Figure A.6: **Example MSU-Net vessel segmentations with corresponding qualitative epistemic uncertainty maps.** Segmentations are provided during (a) pre-bifurc, (b) mid-bifurc, and (c) post-bifurc stages. Qualitative maps appropriately indicate lower uncertainty for *correct* predictions within vessels and higher uncertainty for *incorrect* predictions near outer vessel walls.

Appendix B

Advancing AI Trust in Personalized Medicine

B.1 Dermatomyositis Rash Manifestation

Figure B.1 illustrates the datasets we collected over the course of this project. As part of the DART study, we collected in-clinic and telehealth images using two different imaging modalities. In-clinic VECTRA H1 images are used as our fine-tuning dataset, while in-clinic smartphone, telehealth, and body images are used during SSL pretraining.

B.2 Ordinal Regression with Handcrafted Clinical Features

B.2.1 Pipeline

See Figure B.2 for the full ordinal regression pipeline. We first segment the hand and coarse rash region from the image. Then, we apply the K-means algorithm to the predicted segmentation masks, enabling detailed rash localization. We compute our clinically-motivated, handcrafted features, which are then transformed through PCA to remove multicollinearity issues. Our engineered features are finally used as predictors in ordinal regression.

B.2.2 K-means improves rash region localization

We show examples of the K-means rash localization algorithm in Figure B.3. First, we attempt to find a large number of clusters in the a*b* channels of the CIELab color space through K-means. However, if this fails to meaningfully extract the localized rash region, we reduce the number of clusters and re-run the K-means algorithm until the predictions are coherent.

B.2.3 Ordinal regression predictors

Redness features. `mean.rash.redness, rash.area.ratio, rash.redness.ratio`

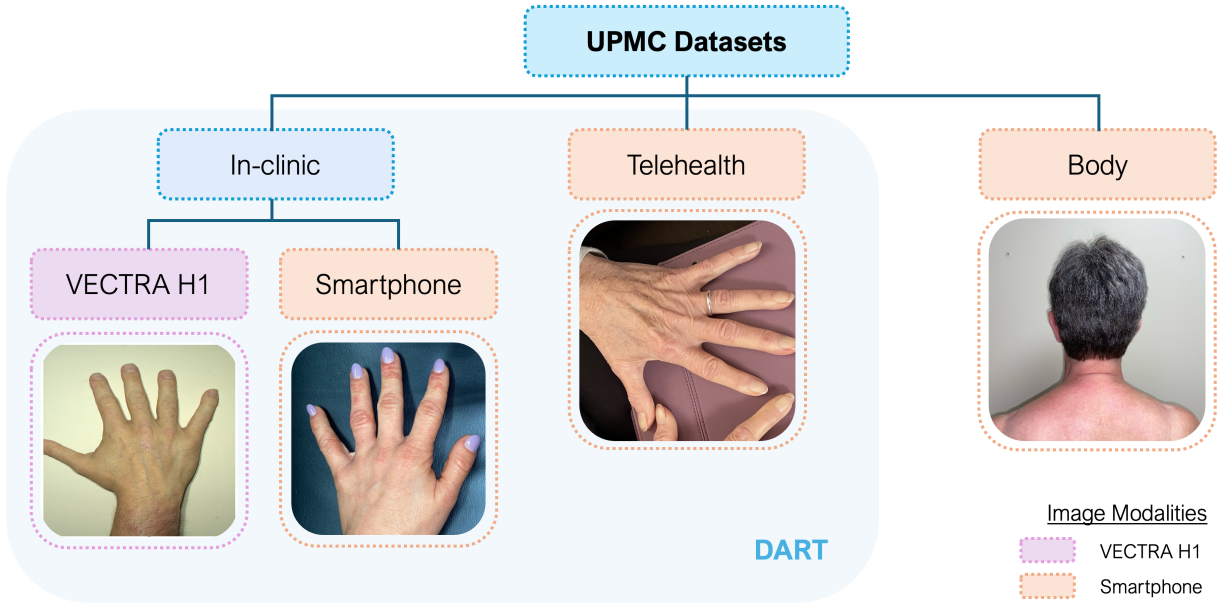


Figure B.1: **UPMC datasets utilized in study.** Images from the VECTRA H1 camera were converted from 3D to 2D. All other images were taken on smartphone devices.

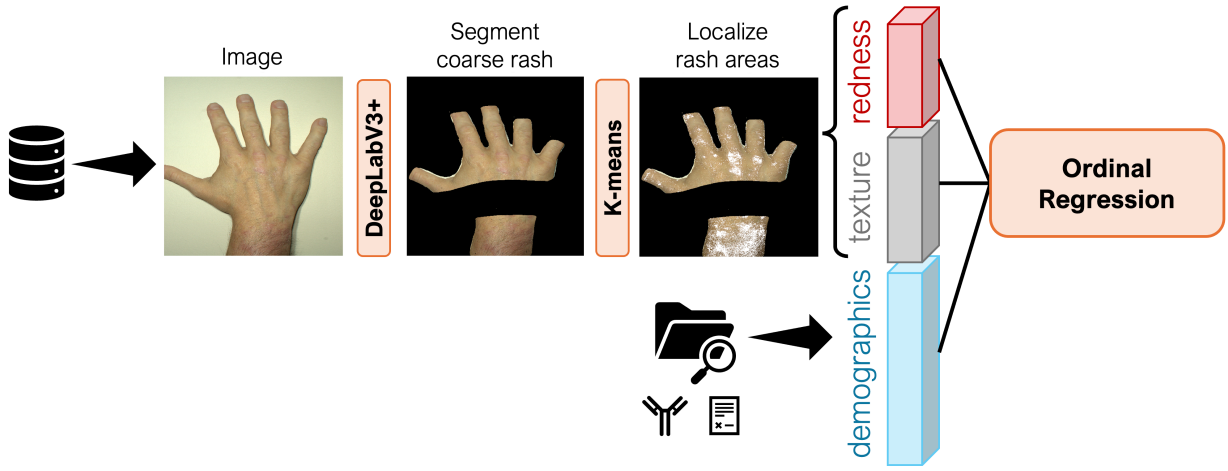


Figure B.2: **Full ordinal regression pipeline using clinically motivated, handcrafted features.** We segment the coarse rash regions from input images, followed by K-means to identify fine-grained rash areas. We construct our handcrafted features from the fine-grained rash areas, which are then concatenated with demographics features, such as antibody expression and age (see Appendix B.2.3), and passed into our regression algorithm.

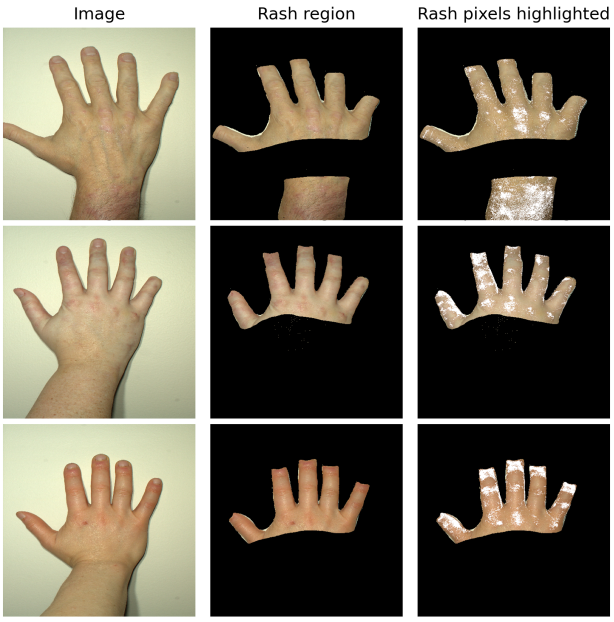


Figure B.3: **Example K-means fine-grained rash region localization procedure in handcrafted regression pipeline.** Example images (left) displayed alongside segmented coarse masks (middle) and predicted K-means clusters overlaid in white (right). K-means can identify fine-grained regions of rash manifestation using the relative pixel redness, even when the rash is not directly on the hand. The top patient exhibits severe rash manifestations on the wrist, which is still captured by K-means.

Texture features. `contrast, dissimilarity, homogeneity, energy, correlation`

Demographic features. `Ab, Jo, PL7, PL12, EJ, PMScl, Ku, U1, U2, Ro TIF, MJ, MDA, Sex, Age, Weight, Height`

`Ab-MDA` are myositis-specific autoantibodies associated with phenotypical features (Marasandra Ramesh et al., 2022). They are often used for sub-classification of DM patients. Several of the original features are removed since they only contain one level, thus acting like a constant, and do not contribute any predictive power to the model.

B.2.4 Additional ordinal regression results

Model	Metrics			
	MAE(↓)	Exact Acc(↑)	Off-by-1 Acc(↑)	Off-by-2 Acc(↑)
POM-REG	1.722 ± 0.840	0.416 ± 0.194	0.547 ± 0.148	0.781 ± 0.145
PCA-REG	$\mathbf{1.160 \pm 0.518}$	$\mathbf{0.484 \pm 0.187}$	$\mathbf{0.624 \pm 0.118}$	$\mathbf{0.852 \pm 0.114}$

Table B.1: **Evaluating ordinal regression strategies for predicting CDASI score for rash severity.** POM-REG performs ordinal regression directly on the original predictors, hence suffers convergence issues due to high multicollinearity. PCA-REG addresses these challenges by performing principal component analysis (PCA) dimensionality reduction on the predictors prior to model fit. We observe that PCA-REG consistently outperforms POM-REG across four different performance metrics. ORDREG in Table 3.2 refers to PCA-REG.

Bibliography

Alberto Abadie and Guido W. Imbens. On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6):1537–1557, 2008. doi: https://doi.org/10.3982/ECTA6474. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA6474. 2.6.3

Sohani Afroja, Md Rasel Kabir, and Md Akhtarul Islam. Analysis of determinants of severity levels of childhood anemia in Bangladesh using a proportional odds model. *Clinical Epidemiology and Global Health*, 8(1):175–180, 2020. 3.3.1

Deevyankar Agarwal, Gonçalo Marques, Isabel de la Torre-Díez, Manuel A Franco Martin, Begoña García Zapiraín, and Francisco Martín Rodríguez. Transfer Learning for Alzheimer's Disease through Neuroimaging Biomarkers: A Systematic Review. *Sensors*, 21(21):7259, 2021. 3.4.1

Rohit Aggarwal and Nantakarn Pongtarakulpanit. Clinical Observational (CO) Studies in Musculoskeletal, Rheumatic, and Skin Diseases, 2025. 3.1, 3.1.2

Ahmed AL Qurri and Mohamed Almekkawy. Improved UNet with Attention for Medical Image Segmentation. *Sensors*, 23(20), 2023. 2.3.1

Laith Alzubaidi, Jinglan Zhang, Amjad J. Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, J. Santamaría, Mohammed A. Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 8(1):53, 2021. 3.4

C O Anyanwu, D F Fiorentino, L Chung, C Dzuong, Y Wang, J Okawa, K Carr, K J Propert, and V P Werth. Validation of the Cutaneous Dermatomyositis Disease Area and Severity Index: characterizing disease severity and assessing responsiveness to clinical change. *Br J Dermatol*, 173(4):969–974, Oct 2015. 3.2.1

Nishanth Thumbavanam Arun, Leonard Weiss, Andrew Schoenling, Marek Radomski, Frank Guyette, Napoleon Roux, Brittany Daley, Michael J Morris, Howie Choset, and John Galeotti. Temporal Monte Carlo Dropout for Robust Uncertainty Quantification: Application to Point-of-Care Ultrasound-guided Nerve Blocks. In *Medical Imaging with Deep Learning, short paper track*, 2023. URL https://openreview.net/forum?id=WvReNPBoB9F. 2.3.1

Rida Ayyaz, Muhammad Asad Meraj, and Fakhar Mustafa. Exploring the Fundamental Risk Factors of Child Malnutrition: An Application of Proportional Odds Model (POM). *Pakistan Journal of Public Health*, 11(2):113–119, 2021. 3.3.1

Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen, Ehsan Adeli, and Dorit

Merhof. Medical Image Segmentation Review: The Success of U-Net. *IEEE Trans Pattern Anal Mach Intell*, 46(12):10076–10095, Dec 2024. ISSN 1939-3539 (Electronic); 0098-5589 (Linking). doi: 10.1109/TPAMI.2024.3435571. 2.3.1

Rohini Banerjee, Cecilia G. Morales, and Artur Dubrawski. Enhanced Uncertainty Estimation in Ultrasound Image Segmentation with MSU-Net. In Alberto Gomez, Bishesh Khanal, Andrew King, and Ana Namburete, editors, *Simplifying Medical Ultrasound. ASMUS 2024*, volume 15186 of *Lecture Notes in Computer Science*, pages 143–153, Cham, 2025. Springer Nature Switzerland. 1, 2.9, 3.7.2

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT Pre-Training of Image Transformers, 2022. URL https://arxiv.org/abs/2106.08254. 3.4.2

Yonatan Belinkov. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1):207–219, March 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/. 3.4

Matt Berseth. ISIC 2017 - Skin Lesion Analysis Towards Melanoma Detection, 2017. URL https://arxiv.org/abs/1703.00523. 1

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877, April 2017. ISSN 1537-274X. doi: 10.1080/01621459.2017.1285773. URL http://dx.doi.org/10.1080/01621459.2017.1285773. 2.2.2

Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996. doi: 10.1007/BF00058655. URL https://doi.org/10.1007/BF00058655. 2.2.4

Edward Chen, FNU Abhimanyu, Ananya Bal, Nishanth Thumbavanam Arun, Andrew Schoenling, Joo Yoon, Kelvin Kwofie, Lenny Weiss, Marek Radomski, Frank Guyette, Howie Choset, and John Galeotti. Multi-Class Bayesian Segmentation of Robotically Acquired Ultrasound Enabling 3D Site Selection along Femoral Vessels for Planning Safer Needle Insertion, 2022. URL https://openreview.net/forum?id=4FasnP6jfel. 2.1, 2.1

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, 2018. URL https://arxiv.org/abs/1802.02611. 3.3.3

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020. 3.4.2

Xiaocong Chen, Lina Yao, Tao Zhou, Jinming Dong, and Yu Zhang. Momentum Contrastive Learning for Few-Shot COVID-19 Diagnosis from Chest CT Images. *Pattern Recognition*, 113:107826, 2021. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2021.107826. URL https://www.sciencedirect.com/science/article/pii/S0031320321000133. 3.4.2

S R Cole and C V Ananth. Regression models for unconstrained, partially or fully constrained continuation odds ratios. *Int J Epidemiol*, 30(6):1379–1382, Dec 2001. ISSN 0300-5771 (Print); 0300-5771 (Linking). doi: 10.1093/ije/30.6.1379. 3.3

Andreas Damianou and Neil D. Lawrence. Deep Gaussian Processes. In Carlos M. Carvalho and Pradeep Ravikumar, editors, *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pages 207–215, Scottsdale, Arizona, USA, 29 Apr–01 May 2013. PMLR. URL https://proceedings.mlr.press/v31/damianou13a.html. 2.2.3

Clément Dechesne, Pierre Lassalle, and Sébastien Lefèvre. Bayesian U-Net: Estimating Uncertainty in Semantic Segmentation of Earth Observation Images. *Remote Sensing*, 13(19), 2021. ISSN 2072-4292. doi: 10.3390/rs13193836. URL https://www.mdpi.com/2072-4292/13/19/3836. 2.2.3

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848. 3.5.2

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1423. URL https://doi.org/10.18653/v1/n19-1423. 3.4.2

Madeline E DeWane, Reid Waldman, and Jun Lu. Dermatomyositis: Clinical features and pathogenesis. *J Am Acad Dermatol*, 82(2):267–281, Feb 2020. ISSN 1097-6787 (Electronic); 0190-9622 (Linking). doi: 10.1016/j.jaad.2019.06.1309. 3.1.1

Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *Journal of Experimental Psychology: General*, 144(1):114–126, 2015. doi: https://doi.org/10.1037/xge0000033. 1

Benoit Dufumier, Pietro Gori, Julie Victor, Antoine Grigis, Michele Wessa, Paolo Brambilla, Pauline Favre, Mircea Polosan, Colm McDonald, Camille Marie Piguet, Mary Phillips, Lisa Eyler, and Edouard Duchesnay. Contrastive Learning with Continuous Proxy Meta-data for 3D MRI Classification. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 58–68, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87196-3. 3.4.2

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a Little Help from My Friends: Nearest-Neighbor Contrastive Learning of Visual Representations. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9568–9577, 2021. doi: 10.1109/ICCV48922.2021.00945. 3.4.2

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep Ensembles: A Loss Landscape Perspective, 2020. URL https://arxiv.org/abs/1912.02757. 2.4

Yoav Freund and Robert E. Schapire. A short introduction to boosting. 1999. URL https://api.semanticscholar.org/CorpusID:9621074. 2.2.4

Andrew S. Fullerton and Jun Xu. Constrained and Unconstrained Partial Adjacent Category

Logit Models for Ordinal Response Variables. *Sociological Methods & Research*, 47(2): 169–206, 2018. doi: 10.1177/0049124115613781. URL https://doi.org/10.1177/0049124115613781. 3.3

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/gal16.html. 2.2.3

Filippo Gambarota and Gianmarco Altoè. Ordinal regression models made easy: A tutorial on parameter interpretation, data simulation and power analysis. *International Journal of Psychology*, 59(6):1263–1292, 2024. doi: https://doi.org/10.1002/ijop.13243. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ijop.13243. 3.3

Jiawei Ge, Shange Tang, Jianqing Fan, and Chi Jin. On the Provable Advantage of Unsupervised Pretraining, 2023. URL https://arxiv.org/abs/2303.01566. 3.4

Florin C Ghesu, Bogdan Georgescu, Awais Mansoor, Youngjin Yoo, Dominik Neumann, Pragneshkumar Patel, Reddappagari Suryanarayana Vishwanath, James M Balter, Yue Cao, Sasa Grbic, and Dorin Comaniciu. Contrastive self-supervised learning from 100 million medical images with optional supervision. *J Med Imaging (Bellingham)*, 9(6):064503, Nov 2022. ISSN 2329-4302 (Print); 2329-4310 (Electronic); 2329-4302 (Linking). doi: 10.1117/1.JMI.9.6.064503. 3.4.2

Biraja Ghoshal, Allan Tucker, Bal Sanghera, and Wai Lup Wong. Estimating Uncertainty in Deep Learning for Reporting Confidence to Clinicians when Segmenting Nuclei Image Data. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 318–324, 2019. doi: 10.1109/CBMS.2019.00072. 2.2

Jacob Gildenblat and contributors. PyTorch library for CAM methods. https://github.com/jacobgil/pytorch-grad-cam, 2021. 3.7.4

Renato Goreshi, Monika Chock, Kristen Foering, Rui Feng, Joyce Okawa, Matt Rose, David Fiorentino, and Victoria Werth. Quality of life in dermatomyositis. *J Am Acad Dermatol*, 65(6):1107–1116, Dec 2011. ISSN 1097-6787 (Electronic); 0190-9622 (Print); 0190-9622 (Linking). doi: 10.1016/j.jaad.2010.10.016. 3.1.1

Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1820–1828, 2021. doi: 10.1109/CVPRW53098.2021.00201. 3.5.2

Matthew Groh, Caleb Harris, Roxana Daneshjou, Omar Badri, and Arash Koochek. Towards Transparency in Dermatology Image Datasets with Skin Tone Annotations by Experts, Crowds, and an Algorithm. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2), November 2022. doi: 10.1145/3555634. URL https://doi.org/10.1145/3555634. 3.5.2

Sebastian G. Gruber and Florian Buettner. Better Uncertainty Calibration via Proper Scores for

Classification and Beyond. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088. 2.2

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/guo17a.html. 2.2

Hafsa Habehh and Suril Gohel. Machine Learning in Healthcare. *Curr Genomics*, 22(4):291–300, Dec 2021. ISSN 1389-2029 (Print); 1875-5488 (Electronic); 1389-2029 (Linking). doi: 10.2174/1389202922666210705124359. 3.4

Robert M. Haralick, K. Shanmugam, and Its'Hak Dinstein. Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3(6):610–621, 1973. doi: 10.1109/TSMC.1973.4309314. 3.3.2

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, 2022. doi: 10.1109/CVPR52688.2022.01553. 3.4.2, 3.5.3

Olivier Henaff. Data-Efficient Image Recognition with Contrastive Predictive Coding. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4182–4192. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/henaff20a.html. 3.4.2

Gábor Hidy, Bence Bakos, and András Lukács. Enhancing pretraining efficiency for medical image segmentation via transferability metrics, 2024. URL https://arxiv.org/abs/2410.18677. 2.4.1

Jennifer L Hundley, Christie L Carroll, Wei Lang, Beverly Snively, Gil Yosipovitch, Steven R Feldman, and Joseph L Jorizzo. Cutaneous symptoms of dermatomyositis significantly impact patients' quality of life. *J Am Acad Dermatol*, 54(2):217–220, Feb 2006. ISSN 1097-6787 (Electronic); 0190-9622 (Linking). doi: 10.1016/j.jaad.2004.12.015. 3.1.1

Pavel Iakubovskii. Segmentation Models Pytorch. https://github.com/qubvel/segmentation_models.pytorch, 2019. 2.5

IBM. Transfer Learning, 2025. URL https://www.ibm.com/think/topics/transfer-learning. Accessed: 2025-04-06. 3.4, 3.4.1

K J Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards Open World Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021)*, 2021. 3.4.2

Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding, 2016. URL https://arxiv.org/abs/1511.02680. 2.2.3

Hee E. Kim, Alejandro Cosa-Linan, Nandhini Santhanam, Mahboubeh Jannesari, Mate E. Maros, and Thomas Ganslandt. Transfer learning for medical image classification: a literature review. *BMC Medical Imaging*, 22(1):69, 2022. doi: 10.1186/s12880-022-00793-7. URL https://doi.org/10.1186/s12880-022-00793-7. 3.4.1

Julianne Kleitsch, Jeffrey D. Weiner, Rachita Pandya, Josef S. Concha, Darosa Lim, and Victoria P. Werth. The physical and emotional impact of cutaneous dermatomyositis: a qualitative study. *Archives of Dermatological Research*, 315(8):2431–2435, 2023. doi: 10.1007/s00403-023-02625-2. URL https://doi.org/10.1007/s00403-023-02625-2. 3.1.1

Benjamin Kompa, Jasper Snoek, and Andrew L. Beam. Second Opinion Needed: Communicating Uncertainty in Medical Artificial Intelligence. *npj Digital Medicine*, 4(1):4, 2021. doi: 10.1038/s41746-020-00367-3. URL https://doi.org/10.1038/s41746-020-00367-3. 1

Kin Keung Lai, Lean Yu, Shouyang Wang, and Ligang Zhou. Credit Risk Analysis Using a Reliability-Based Neural Network Ensemble Model. In Stefanos Kollias, Andreas Stafylopatis, Włodzisław Duch, and Erkki Oja, editors, *Artificial Neural Networks – ICANN 2006*, pages 682–690, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-38873-9. 2.3.2, 2.4.2, A.4.1

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf. 2.2.4, 2.3.2, 2.4.1

R Lall, M J Campbell, S J Walters, and K Morgan. A review of ordinal regression models applied on health-related quality of life assessments. *Stat Methods Med Res*, 11(1):49–67, Feb 2002. ISSN 0962-2802 (Print); 0962-2802 (Linking). doi: 10.1191/0962280202sm271ra. 3.3.1

J. Richard Landis and Gary G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174, 1977. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2529310. 3.1.2

Rana K Latif, Sean P Clifford, Jeffery A Baker, Rainer Lenhardt, Mohammad Z Haq, Jiapeng Huang, Ian Farah, and Jerrad R Businger. Traumatic hemorrhage and chain of survival. *Scand J Trauma Resusc Emerg Med*, 31(1):25, May 2023. ISSN 1757-7241 (Electronic); 1757-7241 (Linking). doi: 10.1186/s13049-023-01088-8. 2.1

Wen Li, Samaneh Kazemifar, Ti Bai, Dan Nguyen, Yaochung Weng, Yafen Li, Jun Xia, Jing Xiong, Yaoqin Xie, Amir Owrangi, and Steve Jiang. Synthesizing CT images from MR images with deep learning: model generalization for different datasets through transfer learning. *Biomedical Physics & Engineering Express*, 7(2):025020, feb 2021. doi: 10.1088/2057-1976/abe3a7. URL https://dx.doi.org/10.1088/2057-1976/abe3a7. 3.4

S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982. doi: 10.1109/TIT.1982.1056489. 3.3.3

Fred Lu, Francis Ferraro, and Edward Raff. Continuously Generalized Ordinal Regression for Linear and Deep Models. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, pages 28–36. SIAM, 2022. 3.3

Carlo Mainetti, Benedetta Terziroli Beretta-Piccoli, and Carlo Selmi. Cutaneous Manifestations of Dermatomyositis: a Comprehensive Review. *Clin Rev Allergy Immunol*, 53(3): 337–356, Dec 2017. ISSN 1559-0267 (Electronic); 1080-0549 (Linking). doi: 10.1007/s12016-017-8652-1. 3.3.2

James E. Manning, Ernest E. Moore, Jonathan J. Morrison, Regan F. Lyon, Joseph J. Dubose, and James D. Ross. Femoral vascular access for endovascular resuscitation. *Journal of Trauma and Acute Care Surgery*, 91(4):E104–E113, October 2021. ISSN 2163-0755. doi: 10.1097/TA.0000000000003339. Publisher Copyright: © 2021 Wolters Kluwer Health, Inc. All rights reserved. 2.1

Harshita Marasandra Ramesh, Sai Sreeya Gude, Shravya Venugopal, Nikhil Chowdary Peddi, Sai Sravya Gude, and Sravya Vuppalapati. The Role of Myositis-Specific Autoantibodies in the Dermatomyositis Spectrum. *Cureus*, 14(3):e22978, Mar 2022. ISSN 2168-8184 (Print); 2168-8184 (Electronic); 2168-8184 (Linking). doi: 10.7759/cureus.22978. B.2.3

C V Oddis, C G Conte, V D Steen, and T A Jr Medsger. Incidence of polymyositis-dermatomyositis: a 20-year study of hospital diagnosed cases in Allegheny County, PA 1963-1982. *J Rheumatol*, 17(10):1329–1334, Oct 1990. ISSN 0315-162X (Print); 0315-162X (Linking). 3.1.2

Sinno Jialin Pan and Qiang Yang. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. doi: 10.1109/TKDE.2009.191. 3.4.1

Barnabas Poczos and Jeff Schneider. On the Estimation of α-Divergences. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 609–617, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL `https://proceedings.mlr.press/v15/poczos11a.html`. 2.6.1, 2.6.1, A.2.1

Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R. Zaiane, and Martin Jagersand. U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognition*, 106:107404, October 2020. ISSN 0031-3203. doi: 10.1016/j.patcog.2020.107404. URL `http://dx.doi.org/10.1016/j.patcog.2020.107404`. 3.3.3

E S Robinson, R Feng, J Okawa, and V P Werth. Improvement in the cutaneous disease activity of patients with dermatomyositis is associated with a better quality of life. *Br J Dermatol*, 172(1): 169–174, Jan 2015. ISSN 1365-2133 (Electronic); 0007-0963 (Print); 0007-0963 (Linking). doi: 10.1111/bjd.13167. 3.1.1

JoséA Rodríguez-Rodríguez, Ezequiel López-Rubio, Juan A Ángel-Ruiz, and Miguel A Molina-Cabello. The Impact of Noise and Brightness on Object Detection Methods. *Sensors (Basel)*, 24(3), Jan 2024. ISSN 1424-8220 (Electronic); 1424-8220 (Linking). doi: 10.3390/s24030821. 3.1.2

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation, 2015. 2.3.1

Canfield Scientific, 2024. URL https://www.canfieldsci.com/imaging-systems/vectra-h1-3d-imaging-system/. Vectra H1 3D imaging system. 3.2.1

Boris Sekachev, Nikita Manovich, Maxim Zhiltsov, Andrey Zhavoronkov, Dmitry Kalinin, Ben Hoff, TOsmanov, Dmitry Kruchinin, Artyom Zankevich, DmitriySidnev, Maksim Markelov, Johannes222, Mathis Chenuet, a andre, telenachos, Aleksandr Melnikov, Jijoong Kim, Liron Ilouz, Nikita Glazov, Priya4607, Rush Tehrani, Seungwon Jeong, Vladimir Skubriev, Sebastian Yonekura, vugia truong, zliang7, lizhming, and Tritin Truong. opencv/cvat: v1.1.0, August 2020. URL https://doi.org/10.5281/zenodo.4009388. 2.5.1

Arabella L Simpkin and Katrina A Armstrong. Communicating Uncertainty: a Narrative Review and Framework for Future Research. *J Gen Intern Med*, 34(11):2586–2591, Nov 2019. ISSN 1525-1497 (Electronic); 0884-8734 (Print); 0884-8734 (Linking). doi: 10.1007/s11606-019-04860-8. 1

SkinIO. SkinIO: AI Skin Care, 2024. URL https://www.skinio.com/. 3.2.1

Reed T. Sutton, David Pincock, Daniel C. Baumgart, Daniel C. Sadowski, Richard N. Fedorak, and Karen I. Kroeker. An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digital Medicine*, 3(1):17, 2020. doi: 10.1038/s41746-020-0221-y. URL https://doi.org/10.1038/s41746-020-0221-y. 1

Yucheng Tang, Dong Yang, Wenqi Li, Holger R. Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20698–20708, 2022. doi: 10.1109/CVPR52688.2022.02007. 3.4.2

Keyu Tian, Yi Jiang, Qishuai Diao, Chen Lin, Liwei Wang, and Zehuan Yuan. Designing bert for convolutional networks: Sparse and hierarchical masked modeling, 2023. URL https://arxiv.org/abs/2301.03580. 3.4.2, 3.5.1, 3.5.4

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive Multiview Coding. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 776–794, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58621-8. 3.4.2

J Tiao, R Feng, S Bird, J K Choi, J Dunham, M George, T C Gonzalez-Rivera, J L Kaufman, N Khan, J J Luo, R Micheletti, A S Payne, R Price, C Quinn, A I Rubin, A G Sreih, P Thomas, J Okawa, and V P Werth. The reliability of the Cutaneous Dermatomyositis Disease Area and Severity Index (CDASI) among dermatologists, rheumatologists and neurologists. *Br J Dermatol*, 176(2):423–430, Feb 2017. ISSN 1365-2133 (Electronic); 0007-0963 (Print); 0007-0963 (Linking). doi: 10.1111/bjd.15140. 3.1.2

Lisa Torrey and Jude Shavlik. Transfer Learning. In Emilio Soria Olivas, editor, *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, pages 242–264. IGI Global, 2010. URL https://www.igi-global.com/chapter/transfer-learning/36988. 3.4

Tim van Erven and Peter Harremos. Rényi Divergence and Kullback-Leibler Divergence.

IEEE Transactions on Information Theory, 60(7):3797–3820, 2014. doi: 10.1109/TIT.2014.2320500. 2.6.1

Aki Vehtari, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry. Pareto Smoothed Importance Sampling, 2024. URL http://jmlr.org/papers/v25/19-556.html. 3.3.6

Kevin Verhoeff, Rachelle Saybel, Pamela Mathura, Bonnie Tsang, Vanessa Fawcett, and Sandy Widder. Ensuring adequate vascular access in patients with major trauma: a quality improvement initiative. *BMJ Open Qual*, 7(1):e000090, 2018. ISSN 2399-6641 (Electronic); 2399-6641 (Linking). doi: 10.1136/bmjoq-2017-000090. 2.1

Heather A Wallace and Hariharan Regunath. Fluid Resuscitation. Jan 2025. 2.1

Christopher Walsh and Carsten Jentsch. Nearest neighbor matching: M-out-of-N bootstrapping without bias correction vs. the naive bootstrap. *Econometrics and Statistics*, 2023. ISSN 2452-3062. doi: https://doi.org/10.1016/j.ecosta.2023.04.005. URL https://www.sciencedirect.com/science/article/pii/S245230622300031X. 2.6.3

Xiaohong Wang, Xudong Jiang, Henghui Ding, Yuqian Zhao, and Jun Liu. Knowledge-aware Deep Framework for Collaborative Skin Lesion Segmentation and Melanoma Recognition. *Pattern Recogn.*, 120(C), December 2021. ISSN 0031-3203. doi: 10.1016/j.patcog.2021.108075. URL https://doi.org/10.1016/j.patcog.2021.108075. 1

Xingyu Wang, Yanzhi Song, and Zhouwang Yang. A Natural Threshold Model for Ordinal Regression. *Neural Processing Letters*, 55(4):4933–4949, 2023. doi: 10.1007/s11063-022-11073-4. URL https://doi.org/10.1007/s11063-022-11073-4. 3.3

Abbi Ward, Jimmy Li, Julie Wang, Sriram Lakshminarasimhan, Ashley Carrick, Bilson Campana, Jay Hartford, Pradeep K. Sreenivasaiah, Tiya Tiyasirisokchai, Sunny Virmani, Renee Wong, Yossi Matias, Greg S. Corrado, Dale R. Webster, Margaret Ann Smith, Dawn Siegel, Steven Lin, Justin Ko, Alan Karthikesalingam, Christopher Semturs, and Pooja Rao. Creating an Empirical Dermatology Dataset Through Crowdsourcing With Web Search Advertisements. *JAMA Network Open*, 7(11):e2446615, November 2024. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2024.46615. URL http://dx.doi.org/10.1001/jamanetworkopen.2024.46615. 3.5.2

Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big Data*, 3(1):9, 2016. doi: 10.1186/s40537-016-0043-6. URL https://doi.org/10.1186/s40537-016-0043-6. 3.4.1

Andrew G Wilson and Pavel Izmailov. Bayesian Deep Learning and a Probabilistic Perspective of Generalization. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4697–4708. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/322f62469c5e3c7dc3e58f5a4d1ea399-Paper.pdf. 2.2, 2.2.4

Daniel Wolf, Tristan Payer, Catharina Silvia Lisson, Christoph Gerhard Lisson, Meinrad Beer, Michael Götz, and Timo Ropinski. Self-Supervised Pre-Training with Contrastive and Masked

Autoencoder Methods for Dealing with Small Datasets in Deep Learning for Medical Imaging. *Scientific Reports*, 13(1):20260, 2023. doi: 10.1038/s41598-023-46433-0. URL https://doi.org/10.1038/s41598-023-46433-0. 3.4.2

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992. ISSN 0893-6080. doi: https://doi.org/10.1016/S0893-6080(05)80023-1. URL https://www.sciencedirect.com/science/article/pii/S0893608005800231. 2.2.4

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. SimMIM: A Simple Framework for Masked Image Modeling, 2022. URL https://arxiv.org/abs/2111.09886. 3.4.2

Shuang Yang, Antony Browne, and Phil Picton. Multistage Neural Network Ensembles. volume 2364, pages 91–97, 06 2002. ISBN 978-3-540-43818-2. doi: 10.1007/3-540-45428-4_9. 2.3.2

Binqian Yin, Qinhong Hu, Yingying Zhu, Chen Zhao, and Keren Zhou. Paw-Net: Stacking ensemble deep learning for segmenting scanning electron microscopy images of fine-grained shale samples. *Computers & Geosciences*, 168:105218, 2022. ISSN 0098-3004. doi: https://doi.org/10.1016/j.cageo.2022.105218. URL https://www.sciencedirect.com/science/article/pii/S0098300422001674. 2.3.2

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large Batch Optimization for Deep Learning: Training BERT in 76 minutes, 2020. URL https://arxiv.org/abs/1904.00962. 3.6.2

Lequan Yu, Hao Chen, Qi Dou, Jing Qin, and Pheng-Ann Heng. Automated Melanoma Recognition in Dermoscopy Images via Very Deep Residual Networks. *IEEE Transactions on Medical Imaging*, 36(4):994–1004, 2017. doi: 10.1109/TMI.2016.2642839. 1

Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6022–6031, 2019. doi: 10.1109/ICCV.2019.00612. 3.6.1

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization, 2017. 3.6.1

Mingya Zhang, Yue Yu, Sun Jin, Limei Gu, Tingsheng Ling, and Xianping Tao. VM-UNET-V2: Rethinking Vision Mamba UNet for Medical Image Segmentation, 2024. 2.3.1

Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking Pre-training and Self-training. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3833–3845. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/27e9661e033a73a6ad8cefcde965c54d-Paper.pdf. 3.4