Verification of Medical Image Classifiers under Clinically Aligned Perturbations

Nick DiSanto
Department of Computer Science
Vanderbilt University
Nashville, TN, USA
nicolas.c.disanto@vanderbilt.edu

Ehsan Khodapanah Aghdam
Department of Computer Science
Vanderbilt University
Nashville, TN, USA
ehsan.khodapanah.aghdam@vanderbilt.edu

Abstract

Formal verification allows the possibility to analyze how neural networks respond across ranges of inputs. This provides guarantees that extend beyond finite test evaluation. Robustness is often defined using L_∞-bounded perturbations, which are mathematically convenient because they fit well within existing tools and frameworks. However these pixel-independent perturbations have very little direct applicability for medical imaging tasks. Therefore, in this work, we study formal robustness verification for medical image classifiers under both standard L_∞ perturbations and a domain-aligned brightness-shift perturbation designed to approximate exposure, illumination, and scanner variability. To do so, we develop an end-to-end verification pipeline for MedMNIST classifiers, evaluating multilayer perceptron and convolutional models on the PneumoniaMNIST and DermaMNIST datasets. In order to analyze the robustness of these models' classifications, we use Marabou as a constraint-based verifier and α, β-CROWN as a bound-propagation verifier, and do so across 320 instance-pairs. Our findings show that brightness-shift perturbations expose more severe failure modes than independently bounded L_∞ noise and increase verification difficulty across both tools. Additionally, we find consistency and agreement across decided instances between Marabou and α, β-CROWN, but for Marabou's *unknown* evaluations, α, β-CROWN resolves 68 of the 71 cases. These findings show that perturbation design directly affects both the clinical relevance and tractability of formal robustness evaluation. Additionally, this shows that independently bounded L_∞ specifications do not enforce globally coherent acquisition-driven intensity changes as effectively as brightness shifts do. Code is available at https://github.com/NickDiSanto/Verification-MedicalNN-Perturbations.

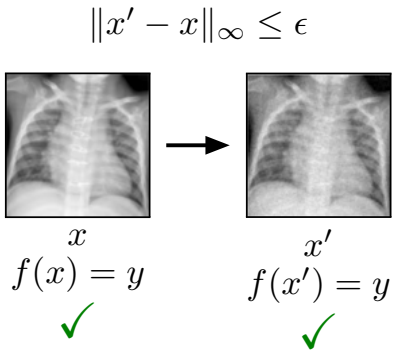
I. Introduction

The importance of robust classifications from deep neural networks in medical imaging tasks is increasing rapidly as these tools become more widely used. This is particularly true in high-risk medical image analysis tasks, such as disease classification from chest X-rays and lesion classification from dermoscopic images. Although models often achieve strong empirical performance on structured and independent experiments, their robustness under structured input perturbation remains an open area of work. This is particularly important because in downstream clinical settings images can vary due to a variety of acquisition conditions, scanner calibration, illumination, and preprocessing pipelines. As a result, a classifier that performs well on a held-out test set may still be sensitive to arbitrary input variations.

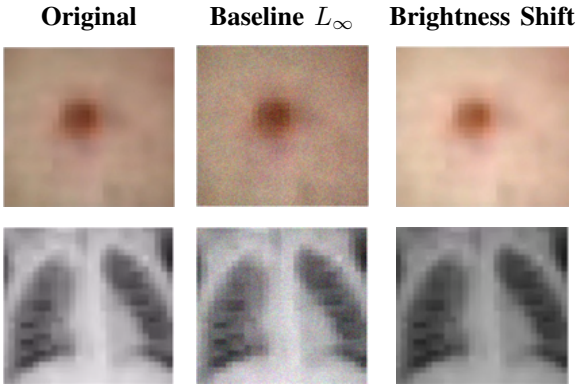
Formal verification seeks to quantify the robustness of these tools by providing a complementary approach to empirical evaluation. Rather than testing a model on a finite set of perturbed inputs, verification seeks to establish that a property holds for all inputs within a specified perturbation set. In neural network classification, this is typically formulated as a local robustness property, which asks whether the model's correct prediction remains invariant within an L_∞-bounded neighborhood. This formulation is widely used in verification benchmarks and is supported by tools such as Marabou [1] and α, β-CROWN [2].

However, standard L_∞ perturbations are only a partial proxy for variability in empirical medical image analysis. Independently bounded perturbations may be inherently arbitrary, which means they do not encode the global intensity changes caused by varying exposure or scanner calibration. Therefore, in this study we compare L_∞ robustness with a globally coupled brightness-shift perturbation, formalized as a low-dimensional and structured subset of the corresponding L_∞ ball. Our contributions are:

- We develop an end-to-end formal verification pipeline for MedMNIST classifiers, supporting multiple model architectures, perturbation families, and verification tools.
- We formalize a clinically aligned brightness-shift perturbation and compare it with standard independently bounded L_∞ specifications.
- We show that brightness-shift perturbations reveal different robustness failures and increase verification difficulty, with effects varying across dataset modality and model architecture.
- We compare Marabou and α, β-CROWN, showing that they agree on all decided instances but differ substantially in decisiveness under challenging settings.



(a) Local robustness: a perturbation x' inside the bounded set $\mathcal{B}(x)$ must preserve the predicted label, $f(x') = y$.



(b) Concrete examples of the two perturbation families on DermaMNIST (top) and PneumoniaMNIST (bottom) samples. The baseline column applies an L_∞ noise; the brightness-shift column applies a global intensity offset.

Fig. 1: Local robustness and the two perturbation families studied in this paper. (a) the formal property $\hat{y}(x') = y$ for every $x' \in \mathcal{B}(x)$; (b) representative original / L_∞ / brightness-shift inputs on the two medical-imaging modalities.

II. Related Work

A. Neural Network Verification

Modern feed-forward neural network verifiers often fall broadly into two families: constraint-based approaches and bound-propagation approaches. Constraint-based approaches extend Simplex-style solvers to handle piecewise-linear activations such as ReLU. Reluplex [3] introduced this paradigm for the ACAS Xu collision-avoidance system, and Marabou [1] generalized it into a scalable and sound framework with multiple input front-ends.

Complementary to these methods, bound-propagation approaches compute layer-wise upper and lower bounds over the input region. AI^2 [4] formulated this process using abstract interpretation, while CROWN [5] introduced linear bound propagation that underpins many modern verifiers. DeepPoly [6] improves precision through a polyhedral abstraction, and α, β-CROWN [2] combines bound propagation with optimizable parameters and branch-and-bound search for improved scalability. The VNN-COMP competition [7] provides standardized benchmarks and defines the VNNLIB specification format used in this work.

B. Medical Imaging Benchmarks

MedMNIST v2 [8] provides a collection of lightweight biomedical classification tasks at 28×28 resolution, including the PneumoniaMNIST and DermaMNIST datasets used in this work. Its small input dimensionality (and model requirements) make it one of the few medical imaging benchmarks suitable for systematic neural network verification experiments.

Formal verification has been successfully applied to small neural networks in safety-critical domains, most notably ACAS Xu [3]. Additionally, prior work on dermoscopic lesion classification has also shown that CNN predictions can be brittle under both artificial transformations and natural variation between repeated lesion photographs [9]. However, to our knowledge, verification of medical imaging models under clinically motivated perturbation families has not been studied at a larger scale.

Unlike prior work, which primarily evaluates robustness under norm-bounded perturbations, we focus on perturbation models grounded in imaging physics and acquisition variability. The proposed brightness-shift perturbation is motivated by exposure variation in radiography [10] and illumination variability in dermoscopy [11], enabling a more domain-aligned evaluation of robustness behavior.

III. Problem Formulation

A. Local Robustness

We now formalize the robustness property studied in this work and the verification setting used throughout the experiments. Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a classifier and let x be a correctly classified input with label $y = \arg\max f(x)$. Given a perturbation set $\mathcal{S}(x)$, local robustness requires that the predicted class remains invariant for all admissible inputs:

$$\forall x' \in \mathcal{S}(x), \quad \arg\max f(x') = y.$$

Verification is typically performed by checking the negation of this property:

$$\exists x' \in \mathcal{S}(x) \text{ such that } \arg\max f(x') \neq y.$$

Each verification instance yields one of three outcomes:

- **Robust:** no violating input exists within $\mathcal{S}(x)$.
- **Counterexample:** a violating input x' is found.
- **Unknown:** the verifier cannot determine the result within the available resources.

B. VNNLIB Specification

We encode each verification problem using the VNNLIB format [12], which specifies both the input domain and the desired output property. Input constraints define a bounded region around x, which is typically expressed as lower and upper bounds on each input dimension. Output constraints encode a classification violation. For a reference class y, a violation occurs if there exists some class $j \neq y$ such that $f_j(x') \geq f_y(x')$.

Verification seeks to determine whether any input within the constrained region satisfies the violation condition and if no such input exists, the instance is certified robust. Otherwise, a counterexample is produced.

C. Verification Tools

We evaluate two complementary neural network verification approaches:

1) **Marabou [1]** is an SMT-based verifier that encodes neural networks as systems of piecewise-linear constraints. It performs a search over feasible assignments to determine whether a violating input exists. Marabou completes when it terminates, but may return *unknown* on complex instances due to its combinatorial search.
2) **α, β-CROWN [2]** is a bound-propagation-based verifier that computes output bounds over input regions using linear relaxations and optimization. Its method of avoiding explicit enumeration of activation patterns allows it to be potentially more scalable and decisive in practice than Marabou, although its effectiveness depends on the tightness of the computed relaxations and branch-and-bound search.

Fig. 2: Overview of the verification pipeline. A trained model, perturbation set, and formal specification are combined to produce per-instance verification outcomes using multiple tools.

IV. Formalizing Clinical Perturbations

We formalize two perturbation families: a standard independently bounded L_∞ baseline and a globally coupled brightness-shift perturbation motivated by acquisition-driven intensity variation.

A. L_∞ Perturbations (Baseline)

The baseline perturbation set is defined as an L_∞-bounded neighborhood:

$$\mathcal{S}_\infty(x;\epsilon) = \{x' \mid \|x' - x\|_\infty \leq \epsilon\}.$$

Equivalently, each input dimension satisfies

$$x_i - \epsilon \leq x'_i \leq x_i + \epsilon,$$

with values clipped to the valid image range. This formulation is standard in adversarial robustness and directly supported by existing verification tools. Because it permits independent pixel-wise variation, we use it as a baseline for our experiments and comparison to our global empirical perturbation.

B. Brightness-Shift Perturbation

To model clinically relevant acquisition variability, we define brightness shift as a globally coupled perturbation:

$$x' = \text{clip}(x + \beta, 0, 1), \qquad \beta \in [-b, b].$$

Equivalently, before clipping, each pixel satisfies

$$x'_i = x_i + \beta \quad \text{for all pixels } i,$$

where the same scalar brightness variable β is shared across the entire image.

This transformation induces a globally consistent intensity change, approximating exposure and scanner-calibration variability in chest X-rays as well as illumination variability in dermoscopy. Brightness shift is a structured subset of the corresponding L_∞ perturbation space: instead of allowing each pixel to vary independently, it constrains all pixels to move coherently through the shared scalar β. Although this constraint can in principle be expressed by coupling input variables within an L_∞-style specification, most standard verification pipelines assume independent input bounds, making such coupling nontrivial to enforce in practice.

This formulation allows robustness to be evaluated under coherent acquisition-aligned intensity variation rather than only independent pixel-wise perturbations.

V. Experimental Setup

A. Datasets

We evaluate on two MedMNIST v2 datasets [8]:

- **PneumoniaMNIST:** a binary classification task on chest X-ray images.
- **DermaMNIST:** a 7-class classification task on dermoscopic images.

Both datasets consist of 28×28 images and are designed to support lightweight benchmarking. Their small input dimensionality makes them suitable for neural network verification while retaining clinically relevant imaging modalities.

B. Models

We consider two verification-friendly architectures:

- **MLP:** a fully connected network with architecture $784 \to 128 \to 128 \to C$, where C denotes the number of output classes.
- **CNN3:** a small convolutional network with two convolutional layers followed by a fully connected classifier.

These choices provide some variation in our evaluation, so there can be a comprehensive comparison of deep learning model capacity, verification tractability, perturbation settings, and datasets.

C. Training Details and Model Performance

All models were trained using standard empirical risk minimization. We use the Adam optimizer with learning rate 10^{-3}, cross-entropy loss, and a batch size of 64. Inputs are resized to 28×28 and normalized to $[0, 1]$. No adversarial training, robust training, or verification-aware training methods are applied. Lightweight data augmentation is used during training to improve generalization. For PneumoniaMNIST, we apply brightness and contrast jitter with magnitude 0.04 and additive Gaussian noise with standard deviation 0.005. For DermaMNIST, we use brightness and contrast jitter with magnitude 0.08 and Gaussian noise with standard deviation 0.01. These augmentations are applied only during training; verification is performed on clean test images.

We report the clean test accuracy of the exact checkpoints used in the verification experiments. PneumoniaMNIST achieves 83.33% accuracy with the MLP and 83.49% with CNN3. DermaMNIST achieves 68.83% with the MLP and 71.27% with CNN3.

D. Verification Protocol

We evaluate robustness across the following dimensions:

- **Datasets:** PneumoniaMNIST, DermaMNIST
- **Models:** MLP, CNN3
- **Verification tools:** Marabou, α, β-CROWN
- **Perturbation families:** L_∞ noise and brightness shift

For each configuration, we select 20 correctly classified test images and verify local robustness under the specified perturbation set. Each verification instance yields one of three outcomes: robust, counterexample, or unknown. The full evaluation consists of 32 experiment cells, corresponding to all combinations of dataset, model, perturbation family, and tool, and produces a total of 640 verification instances.

VI. Results

We evaluate robustness and verification behavior across the full experimental matrix. Across all configurations, we obtain 416 robust outcomes, 145 counterexamples, and 79 unknown outcomes, corresponding to an overall decided rate of 87.7%. Unknown outcomes are concentrated most strongly in the DermaMNIST-CNN3 brightness-shift settings.

A. Effect of Perturbation Type and Dataset

Figure 3 summarizes robustness behavior across datasets, models, and perturbation types. Compared with L_∞ noise, brightness-shift perturbations produce more counterexamples and unknown outcomes. One example is the combination of PneumoniaMNIST with CNN3, which yields 23 counterexamples under brightness shift but only 1 under L_∞ perturbations. Robustness in this case remains relatively stable and modest for PneumoniaMNIST with the MLP, but it becomes much more problematic for DermaMNIST and CNN3. In the hardest DermaMNIST-CNN3 settings, brightness shift significantly influences the outcomes leading to results largely dominated by counterexamples and unknowns.

B. Verification Tool Behavior

Figure 4 compares Marabou and α, β-CROWN on identical verification instances. Both tools are consistent on every decided case and agree on every robust and counterexample decision. However, significant differences arise in cases where Marabou returns unknown. In a majority of these cases, α, β-CROWN resolves the instance, usually by identifying a counterexample that Marabou was unable to find in its 120-second timeframe. Thus, the observed cross-tool difference is primarily one of decisiveness and efficiency, α, β-CROWN resolving many cases that were found to be intractable for Marabou.

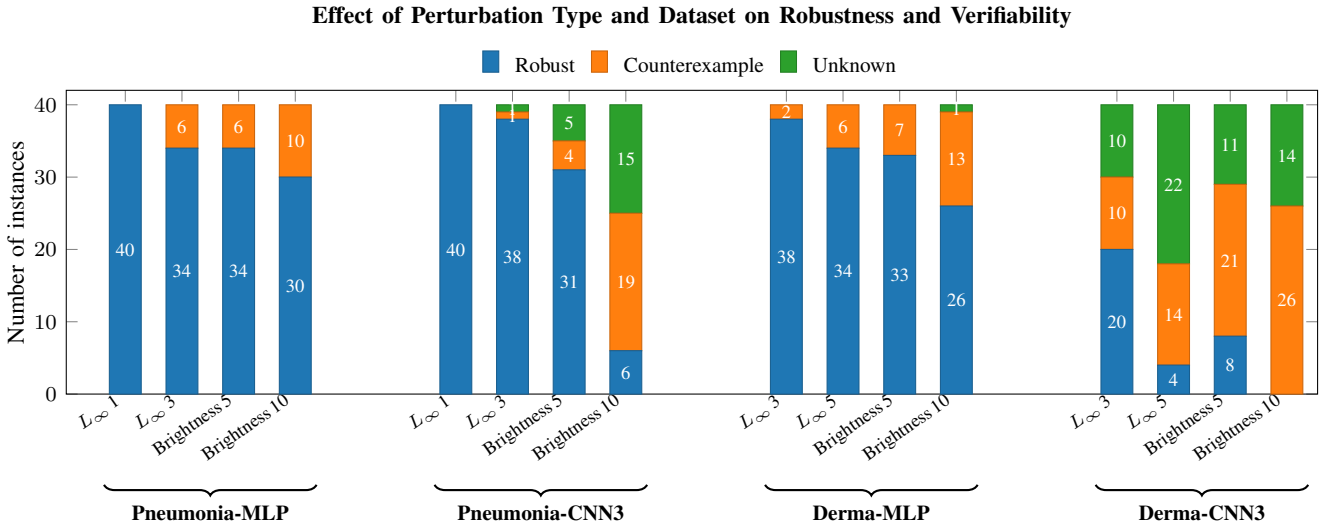


Fig. 3: Per-cell verification outcomes across the evaluation matrix. Each bar aggregates 40 verifier runs (20 instances $\times$ 2 tools), with outcomes grouped as robust, counterexample (CE), and unknown. Bars are organized by dataset and model, with increasing perturbation magnitude shown within each perturbation family. Perturbation magnitudes are reported in 8-bit intensity units (i.e., values correspond to $k/255$ in normalized $[0, 1]$ space).

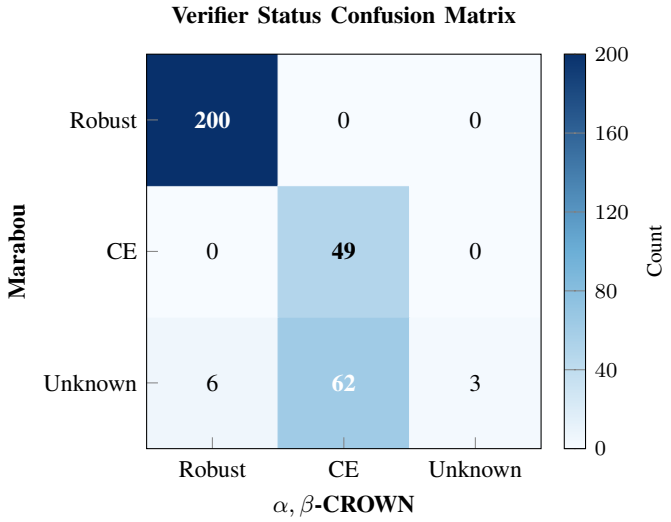


Fig. 4: Cross-tool outcome comparison. Rows correspond to Marabou outcomes and columns to α, β-CROWN outcomes. Diagonal entries indicate agreement; off-diagonal entries capture differences in tool behavior.

VII. Discussion

A. Perturbation Design and Verification Behavior

The results have interesting implications on how perturbation design affects robustness and verifiability. Independently bounded L_∞ perturbations are useful as a standard baseline, but they are limited as a direct model of structured exposure differences in medical imaging. Brightness shift perturbations instead force all pixels to vary through a shared scalar parameter, which allows for a more domain-aligned failure mode. The increase in counterexamples under brightness shift suggests that the models judged robust under standard L_∞ benchmarks may remain sensitive to clinically plausible appearance changes, aligning with the initial hypothesis about empirical grounding. Structured perturbations also increase the incidence of unknown outcomes, indicating that realistic perturbation models can make verification harder despite being motivated by a simpler acquisition-level intensity change.

B. Verification Tools, Models, and Dataset Complexity

The increased difficulty under structured perturbations is reflected in the behavior of the verification tools, as Marabou and α, β-CROWN show significant differences in their ability to resolve challenging cases in the time allotment. In particular, α, β-CROWN frequently produces definitive outcomes where Marabou returns unknown under brightness-shift perturbations. This pattern reflects the practical tradeoff between the high-level approaches of each verification tool. Marabou provides strong guarantees when it terminates, whereas α, β-CROWN can resolve many cases that remain intractable for Marabou.

These differing results become increasingly evident when compared across model architecture and datasets. Convolutional models introduce greater representational complexity than fully connected networks, while DermaMNIST is more challenging than PneumoniaMNIST because of its multi-class structure and greater visual variability. The hardest regime is therefore DermaMNIST with CNN3 under brightness-shift perturbations, where low robustness and reduced solver decisiveness coincide.

C. Analysis of Undecided Instances

A small number of instances remained entirely undecided, primarily for CNN3 on DermaMNIST under brightness-shift perturbations. Qualitatively, these cases seem to appear near decision boundaries where small global intensity changes can alter the model prediction. This combination increases branching for exact solvers and weakens relaxation bounds for bound-propagation methods, so undecided outcomes are concentrated in inputs that are prediction-sensitive and verification-hard.

D. Limitations and Future Work

This study has three main limitations. First, the models were intentionally selected to be small in order to make high-volume experiments and verification tractable, but this admittedly limits its direct applicability to modern deep models. Second, brightness shift captures only one class of acquisition variability. Recent medical-imaging robustness benchmarks such as ROOD-MRI [13] emphasize that clinically deployed models should also be evaluated under realistic distribution shifts and acquisition artifacts. Future work should therefore extend this verification framework to contrast changes, blur, geometric transformations, and structured artifacts such as lesion markers. Third, each configuration for our experiments uses 20 correctly

classified test instances, which was sufficient to reveal consistent trends but likely not enough to support strong statistical claims about a broader population. Future work should scale the evaluation, incorporate richer perturbation families, and analyze counterexamples with domain-expert input.

VIII. Conclusion

We presented a verification pipeline for medical image classifiers under both standard L_∞ and clinically aligned brightness-shift perturbations. The results show that an empirically-grounded intensity change inspired by acquisition variation leads to more consistent counterexamples and unresolved verification instances than independently bounded L_∞ specifications. Marabou and α, β-CROWN agree on all decided cases but differ substantially in decisiveness, illustrating the practical tradeoff in their approaches when applied to medical imaging classification robustness analysis. Overall, the study shows that perturbation design directly determines both the clinical relevance and tractability of formal robustness evaluation and should be a primary consideration for empirical evaluation of models in high-risk environments.

Author Contributions

Both authors contributed to the experimental design, analysis, report, and presentations. Nick DiSanto led the verification methodology, tool integration, experiment execution, aggregation pipeline, visualizations, and final manuscript drafting. Ehsan Khodapanah Aghdam contributed to model training and ONNX export, VNNLIB property generation, figure preparation, result analysis, and initial report drafting.

Disclosure of Generative AI Use

The authors used Claude and ChatGPT for assistance with their research. All technical claims, results, and experimental decisions are the authors'. LLMs assisted simply with structuring and editing, not with generating or interpreting results.

References

[1] G. Katz, D. A. Huang, D. Ibeling, K. Julian, C. Lazarus, R. Lim, P. Shah, S. Thakoor, H. Wu, A. Zeljić, D. L. Dill, M. J. Kochenderfer, and C. Barrett, "The Marabou framework for verification and analysis of deep neural networks," in *Computer Aided Verification (CAV)*, pp. 443–452, 2019.

[2] S. Wang, H. Zhang, K. Xu, X. Lin, S. Jana, C.-J. Hsieh, and J. Z. Kolter, "β-CROWN: Efficient bound propagation with per-neuron split constraints for complete and incomplete neural network robustness verification," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[3] G. Katz, C. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer, "Reluplex: An efficient SMT solver for verifying deep neural networks," in *Computer Aided Verification (CAV)*, pp. 97–117, 2017.

[4] T. Gehr, M. Mirman, D. Drachsler-Cohen, P. Tsankov, S. Chaudhuri, and M. Vechev, "AI2: Safety and robustness certification of neural networks with abstract interpretation," in *IEEE Symposium on Security and Privacy*, pp. 3–18, 2018.

[5] H. Zhang, T.-W. Weng, P.-Y. Chen, C.-J. Hsieh, and L. Daniel, "Efficient neural network robustness certification with general activation functions," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[6] G. Singh, T. Gehr, M. Püschel, and M. Vechev, "An abstract domain for certifying neural networks," *Proceedings of the ACM on Programming Languages*, vol. 3, no. POPL, pp. 1–30, 2019.

[7] C. Brix, S. Bak, C. Liu, and T. T. Johnson, "The fourth international verification of neural networks competition (VNN-COMP 2023): Summary and results," *arXiv preprint arXiv:2312.16760*, 2023.

[8] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni, "MedMNIST v2 – a large-scale lightweight benchmark for 2D and 3D biomedical image classification," *Scientific Data*, vol. 10, no. 1, p. 41, 2023.

[9] R. C. Maron, S. Haggenmüller, C. von Kalle, J. S. Utikal, F. Meier, F. F. Gellrich, A. Hauschild, L. E. French, M. Schlaak, K. Ghoreschi, *et al.*, "Robustness of convolutional neural networks in recognition of pigmented skin lesions," *European journal of cancer*, vol. 145, pp. 81–91, 2021.

[10] P. Pfeiffer *et al.*, "Effects of different exposure values on diagnostic accuracy of digital images," *Quintessence International*, 2000. PMID: 11203933.

[11] K. L. Hanlon, G. Wei, L. Correa-Selm, and J. M. Grichnik, "Dermoscopy and skin imaging light sources: a comparison and review of spectral power distribution and color consistency," *Journal of Biomedical Optics*, vol. 27, no. 8, p. 080902, 2022.

[12] S. Demarchi, D. Guidotti, L. Pulina, and A. Tacchella, "Supporting standardization of neural networks verification with vnn-lib and coconet," in *Proceedings of the 6th Workshop on Formal Methods for ML-Enabled Autonomous Systems (FoMLAS 2023)*, 2023.

[13] L. Boone, M. Biparva, P. M. Forooshani, J. Ramirez, M. Masellis, R. Bartha, S. Symons, S. Strother, S. E. Black, C. Heyn, *et al.*, "Rood-mri: Benchmarking the robustness of deep learning segmentation models to out-of-distribution and corrupted data in mri," *NeuroImage*, vol. 278, p. 120289, 2023.