What Can We Learn from Inter-Annotator Variability in Skin Lesion Segmentation?

Kumar Abhishek[1 [0000−0002−7341−9617]], Jeremy Kawahara[2 [0000−0002−6406−5300]], and Ghassan Hamarneh[1 [0000−0001−5040−7448]]

[1] School of Computing Science, Simon Fraser University, Canada
[2] AIP Labs, Budapest, Hungary
{kabhishe,hamarneh}@sfu.ca, jeremy@aip.ai

Abstract. Medical image segmentation exhibits intra- and inter-annotator variability due to ambiguous object boundaries, annotator preferences, expertise, and tools, among other factors. Lesions with ambiguous boundaries, e.g., spiculated or infiltrative nodules, or irregular borders per the ABCD rule, are particularly prone to disagreement and are often associated with malignancy. In this work, we curate IMA++, the largest multi-annotator skin lesion segmentation dataset, on which we conduct an in-depth study of variability due to annotator, malignancy, tool, and skill factors. We find a statistically significant (p<0.001) association between inter-annotator agreement (IAA), measured using Dice, and the malignancy of skin lesions. We further show that IAA can be accurately predicted directly from dermoscopic images, achieving a mean absolute error of 0.108. Finally, we leverage this association by utilizing IAA as a "soft" clinical feature within a multi-task learning objective, yielding a 4.2% improvement in balanced accuracy averaged across multiple model architectures and across IMA++ and four public dermoscopic datasets. The code is available at https://github.com/sfu-mial/skin-IAV.

Keywords: dermatology · skin lesion segmentation· inter-rater variability · multi-task learning.

1 Introduction

Medical image segmentation is a foundational task in modern healthcare, enabling precise quantitative analysis, the development of downstream diagnostic or prognostic models, and treatment planning [4]. However, the process of delineating structures in medical images, whether performed manually or semi-automatically, is prone to variability, leading to intra- and inter-annotator differences. The sources of this variability are multifactorial, including, but not limited to, ambiguous boundaries, varying interpretations of imaging characteristics, discrepancies in annotation protocols, and differences in annotator experience or skill levels. In clinical practice, lesions that lack well-defined boundaries and are therefore often difficult to segment, such as spiculated or infiltrative nodules, are often strongly associated with malignancy [16,37], suggesting that poorly-defined boundaries may be associated with the underlying disease severity.

Specific to skin image analysis, skin lesion segmentation (SLS) [7,18,30] can play an important role in computing segmentation-based clinical features (e.g., irregular borders in the ABCD [22] rule), where the presence of certain clinical features can be used to distinguish melanoma from benign lesions and enhance the interpretability of deep learning-based diagnosis methods [31,26]. However, reliably computing clinical features derived from lesion segmentations can be challenging due to annotator segmentation variability. For example, irregular borders or psuedopods, which are clinical features strongly associated with melanoma [21,44], can be difficult to delineate and may contribute to annotator variability. Thus, in this work, we hypothesize that the level of annotator (dis)agreement in SLS may itself be related to malignancy. Despite numerous works on modeling annotation styles [47,2,41], segmentation selection or aggregation [35,29,42], and studying variability in expert segmentations [24,13,34] and non-expert annotations of clinical features [9,33], no prior research has formally investigated if an association exists between the quantitative level of **i**nter-**a**nnotator segmentation **a**greement (IAA) and lesion malignancy.

Addressing this gap, we first formally examine whether a systematic relationship exists between IAA levels and lesion malignancy. Using a newly curated dataset, IMA++, we demonstrate a significant association: malignant lesions exhibit systematically lower levels of IAA compared to benign lesions. Based on this observation, we treat the IAA as a type of clinical feature that quantifies how ambiguous a lesion is to annotate, which may serve as a proxy for existing clinical features (e.g., irregular border, pseudopods). Driven by this association, our next contribution seeks to predict per-image IAA scores directly from the dermoscopic image using deep regression models, avoiding the segmentation step and allowing us to leverage this signal without requiring multiple annotations during inference. Finally, motivated by multi-task learning's ability to enhance individual task performance [28,48], and following works that simultaneously predict diagnosis and associated clinical features, like the 7-point criteria to improve diagnostic accuracy and interpretability [20,26,31], our approach views IAA as a "soft" clinical feature. Unlike traditional multi-task methods that jointly predict the diagnosis with segmentation [38,45] or related clinical features such as ABCD [33] (which can be ambiguous due to inter-annotator differences), we hypothesize that training a model to learn the variability in human interpretation implicitly captures complex morphological characteristics indicative of malignancy, such as border irregularity and asymmetry, which are often difficult to formalize or are influenced by annotator subjectivity.

To summarize, we make the following contributions: **(1)** We curate, to our knowledge, the largest SLS dataset, IMA++, comprising 5111 masks from 15 unique annotators, and present the largest-scale study of intra- and inter-annotator variability in this context. **(2)** We empirically demonstrate, using rigorous statistical methods, that inter-annotator agreement (IAA) is significantly associated with lesion malignancy in IMA++. **(3)** We show that IAA scores can be predicted, with reasonably low error, directly from image content alone, without requiring any segmentations at inference time. **(4)** We demonstrate, through ex-

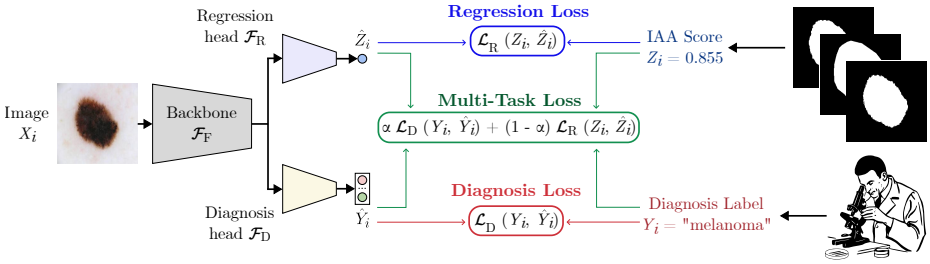


Fig. 1: Regression ($\mathcal{M}_1$), diagnosis-only ($\mathcal{M}_2$), and multi-task ($\mathcal{M}_{\text{MT}}$) models.

tensive evaluation on multiple datasets, that multi-task models jointly predicting diagnosis and IAA outperform diagnosis-only models.

2 Methods

2.1 Agreement distribution shift across disease classes

Let $(\mathcal{X}, \mathcal{Y}, \mathcal{S})$ denote a dataset of N images $\{X_i\}_{i=1}^N$, corresponding N diagnoses $\{Y_i\}_{i=1}^N$, and N sets of multiple segmentation masks $\mathcal{S} = \{\{S_{ik}\}_{k=1}^{K_i}\}_{i=1}^N$, where $K_i \geq 2$ is the number of masks for X_i. Let $\mathcal{Z} = \{Z_i\}_{i=1}^N$ be the set of corresponding inter-annotator agreement (IAA) scores, where $Z_i = g(\{S_{ik}\}) \in \mathbb{R}$ is computed per image based on the multiple segmentations, where $g(\cdot)$ uses either overlap-based (e.g., Dice similarity coefficient) or boundary-based (e.g., Hausdorff distance) metrics.

First, we wish to rigorously evaluate if there exists a systematic difference between the IAA scores for benign and malignant lesions. In particular, we examine the relationship between the probability of sampling a certain value from the IAA distribution of benign versus malignant lesions. To this end, we apply first-order stochastic dominance (FOSD) testing: a distribution $f_A(x)$ is said to first-order stochastically dominate a distribution $f_B(x)$, if $F_A(x) \leq F_B(x) \forall x$, with a strict inequality for some x, where $F_A(x)$ and $F_B(x)$ are the cumulative distribution functions (CDFs) of $f_A(x)$ and $f_b(x)$, respectively; loosely put, $f_A(x)$ is more likely to generate higher values of x than $f_B(x)$. This first-order stochastic dominance is denoted as $F_A \succeq_1 F_B$.

We define Z_{ben} and Z_{mal} as the subsets of IAA scores $\mathcal{Z}$ corresponding to benign and malignant lesions, respectively. We conduct two separate one-sided tests of FOSD [6]: (1) testing whether malignant Dice scores stochastically dominate benign scores, with the hypothesis $H_{\text{mal}\succeq_1\text{ben}} : F_{\text{mal}}(x) \leq F_{\text{ben}}(x)\ \forall x$, and vice versa (2) testing whether benign scores stochastically dominate malignant, with the hypothesis $H_{\text{ben}\succeq_1\text{mal}} : F_{\text{ben}}(x) \leq F_{\text{mal}}(x)\ \forall x$. As a complementary analysis, we also compare the two distributions using a Mann–Whitney U test [27].

2.2 Image-based prediction of inter-annotator agreement

Next, we examine the ability to predict the IAA score for an image based on the image content alone and without access to the corresponding segmentations during inference. Given an image X_i, we predict the target $\hat{Z}_i = \mathcal{M}_1(X_i; \Theta_\mathrm{F}, \Theta_\mathrm{R})$, where $\mathcal{M}_1 = \mathcal{F}_\mathrm{R} \circ \mathcal{F}_\mathrm{F}$, $\circ$ denotes function composition, and $\mathcal{F}_\mathrm{F}$ and $\mathcal{F}_\mathrm{R}$ are the feature-extracting backbone and the regression head, parameterized by Θ_F and Θ_R, respectively (Fig. 1). The regression model $\mathcal{M}_1$ is trained by minimizing a regression loss $\mathcal{L}_\mathrm{R}$:

$$\Theta_\mathrm{F}^*, \Theta_\mathrm{R}^* = \arg\min_{\Theta_\mathrm{F}, \Theta_\mathrm{R}} \sum_{i=1}^{N} \mathcal{L}_\mathrm{R}(Z_i, \hat{Z}_i). \tag{1}$$

2.3 Integrating inter-annotator agreement and diagnosis prediction

Given an image X_i, a typical image-based diagnosis model $\mathcal{M}_2$ predicts $\hat{Y}_i = \mathcal{M}_2(X_i; \Theta_\mathrm{F}, \Theta_\mathrm{D})$, where $\mathcal{M}_2 = \mathcal{F}_\mathrm{D} \circ \mathcal{F}_\mathrm{F}$, and $\mathcal{F}_\mathrm{D}$ is the diagnosis head parameterized by Θ_D. The diagnosis model is optimized by minimizing a diagnosis (classification) loss $\mathcal{L}_\mathrm{D}$:

$$\Theta_\mathrm{F}^*, \Theta_\mathrm{D}^* = \arg\min_{\Theta_\mathrm{F}, \Theta_\mathrm{D}} \sum_{i=1}^{N} \mathcal{L}_\mathrm{D}(Y_i, \hat{Y}_i). \tag{2}$$

Finally, inspired by previous works on multi-task learning in medical imaging [48] and skin images in particular [46,20,38,45,1,33], we investigate whether simultaneous prediction of IAA and diagnosis improves the accuracy of the latter. To this end, we train $\mathcal{M}_\mathrm{MT}$ to simultaneously predict $\hat{Y}_i$ and $\hat{Z}_i$ such that $(\hat{Y}_i, \hat{Z}_i) = \mathcal{M}_\mathrm{MT}(X_i; \Theta_\mathrm{F}, \Theta_\mathrm{D}, \Theta_\mathrm{R})$, where $\mathcal{M}_\mathrm{MT} = (\mathcal{F}_\mathrm{R} \circ \mathcal{F}_\mathrm{F}, \mathcal{F}_\mathrm{D} \circ \mathcal{F}_\mathrm{F})$ is a multi-task prediction model with prediction heads for diagnosis (classification) and IAA score (regression) that share the same backbone (Fig. 1). $\mathcal{M}_\mathrm{MT}$ is trained by minimizing a (weighted) sum of the two tasks' objectives:

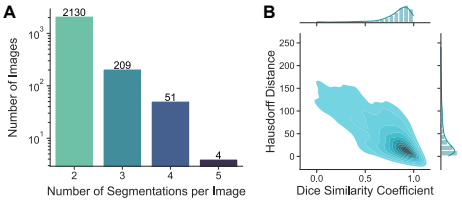
$$\Theta_\mathrm{F}^*, \Theta_\mathrm{D}^*, \Theta_\mathrm{R}^* = \arg\min_{\Theta_\mathrm{F}, \Theta_\mathrm{D}, \Theta_\mathrm{R}} \sum_{i=1}^{N} \Big[\alpha \cdot \mathcal{L}_\mathrm{D}(Y_i, \hat{Y}_i) + (1 - \alpha) \cdot \mathcal{L}_\mathrm{R}(Z_i, \hat{Z}_i) \Big], \tag{3}$$

where α is a loss-weighting hyperparameter. Note that $\alpha = 0$ and $\alpha = 1$ are equivalent to regression-only ($\mathcal{M}_1$) and diagnosis-only ($\mathcal{M}_2$) models, respectively. More details about exact model architectures, losses (Eqns. 1, 2, 3), datasets, training, and evaluation are discussed in the next section.

3 Results and Discussion

3.1 Datasets and analysis

A new curated dataset: Prior work on multi-annotator skin lesion segmentation has produced either **(a)** large datasets without annotator-level information



	Malignancy		Annotator	
	Benign	Malignant	Same	Different
Dice	0.791 0.225	0.753 0.227	0.900 0.131	0.772 0.221
p-value	4.77E-06		1.85E-35	
Cohen's *d*	0.798		2.714	
	Tool		Skill	
	Same	Different	Same	Different
Dice	0.862 0.157	0.747 0.231	0.806 0.167	0.710 0.258
p-value	2.45E-69		1.17E-05	
Cohen's *d*	2.447		1.816	

Fig. 2: IMA++ dataset statistics: (A) number of segmentations per image, (B) pairwise agreement metrics (Dice and Hausdoff distance), (C) intra- and inter-factor agreement (mean$_{\text{std. dev.}}$ of Dice) with p-value and Cohen's d.

(e.g., Ribeiro et al. [34,35], Mirikharaji et al. [29]: 2223 images, 4647 total segmentations) or, **(b)** small datasets with annotator or style metadata (e.g., Zepf et al. [47]: 100 images, 300 total segmentations, 3 styles; Abhishek et al. [2]: 454 images, 1058 total segmentations, 10 unique annotators). In this work, we curate and publicly release a new dataset from the ISIC Archive, called **ISIC MultiAnnot++** (**IMA++** hereafter). It contains 2394 dermoscopic images segmented by 15 unique annotators, where 2130 images have 2 masks, 209 images have 3 masks, 51 images have 4 masks, and 4 images have 5 masks, resulting in a total of 5111 segmentation masks (Fig. 2). To the best of our knowledge, IMA++ is the largest public multi-annotator skin lesion segmentation dataset in terms of both mask and annotator counts.

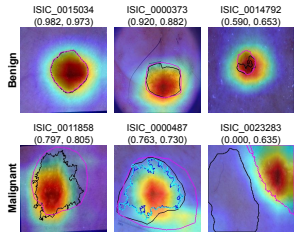
Each mask contains information about the tool used: (T1) manual polygon tracing by a human expert, (T2) semi-automated flood-fill with expert-defined parameters, or (T3) a fully-automated segmentation reviewed and accepted by a human expert; and the skill level of the manual reviewer: (S1) expert or (S2) novice. We partition the images in IMA++ into training, validation, and testing splits in the ratio of 70:15:15, stratified by malignancy, number of segmentations per image, and Dice score range: low (< 0.5), medium, and high (> 0.8).

Calculation of IAA scores: All images and binary segmentations are resized to 256×256. For each image X_i, we compute the Dice and Hausdorff distance between all $\binom{K_i}{2}$ unique pairs of segmentation masks (Fig. 2B shows the full distribution of pairwise scores). Although previous IAA studies have used Cohen's kappa [35] and Fleiss' kappa [2], these metrics measure categorical agreement and fail to capture spatial overlap between annotations; thus, we adopt the Dice metric, which is standard in IAA studies in medical imaging [36,3,17]. For each image, we average the pairwise Dice scores to obtain a single IAA score. Although most lesions tend to exhibit high agreement between annotators ([344, 818] out of 2394 images have Dice above [0.95, 0.90]), a notable subset shows poor agreement (23 images have 0 Dice), highlighting the wide range of inter-annotator agreement, and in line with the previous study by Ribeiro et al. [34].

Is malignancy associated with IAA? Fig. 2C reveals a notable difference in IAA scores between benign and malignant lesions: benign lesions tend to exhibit higher Dice scores (0.791 ± 0.215 vs. 0.753 ± 0.227). However, a comparison of

Model	Params (M)	MACs (G)	MAE			MSE			p-value
			Benign	Malignant	Overall	Benign	Malignant	Overall	
VGG-16	14.72	15.36	$0.118_{0.166}$	$0.134_{0.188}$	$0.121_{0.171}$	$0.027_{0.059}$	$0.035_{0.077}$	$0.029_{0.064}$	1.54E-05
ResNet-18†	11.31	1.81	$0.103_{0.158}$	$0.127_{0.178}$	$0.108_{0.162}$	$0.025_{0.065}$	$0.032_{0.068}$	$0.026_{0.064}$	1.15E-08
ResNet-50	24.03	4.09	$0.124_{0.175}$	$0.143_{0.193}$	$0.128_{0.180}$	$0.031_{0.074}$	$0.038_{0.081}$	$0.032_{0.076}$	1.41E-37
MobileNetV2†	2.55	0.30	$0.103_{0.157}$	$0.129_{0.182}$	$0.109_{0.163}$	$0.025_{0.067}$	$0.033_{0.070}$	$0.026_{0.068}$	3.15E-15
MobileNetV3L	3.22	0.21	$0.106_{0.156}$	$0.131_{0.183}$	$0.111_{0.162}$	$0.024_{0.063}$	$0.033_{0.070}$	$0.026_{0.065}$	6.53E-09
DenseNet-121	7.22	2.83	$0.131_{0.182}$	$0.141_{0.191}$	$0.133_{0.184}$	$0.033_{0.074}$	$0.037_{0.082}$	$0.034_{0.076}$	5.21E-32
EfficientNet-B0	4.34	0.38	$0.110_{0.164}$	$0.138_{0.191}$	$0.116_{0.170}$	$0.027_{0.068}$	$0.036_{0.076}$	$0.029_{0.070}$	1.01E-15
EfficientNet-B1†	6.84	0.57	$0.107_{0.165}$	$0.121_{0.177}$	$0.110_{0.167}$	$0.027_{0.074}$	$0.032_{0.088}$	$0.028_{0.077}$	1.63E-07
ConvNeXt-T	28.02	4.47	$0.130_{0.195}$	$0.155_{0.207}$	$0.135_{0.199}$	$0.039_{0.103}$	$0.048_{0.100}$	$0.041_{0.102}$	2.62E-28
Swin-T	27.72	4.50	$0.131_{0.188}$	$0.152_{0.203}$	$0.135_{0.192}$	$0.035_{0.089}$	$0.043_{0.092}$	$0.037_{0.090}$	3.70E-37
SwinV2-T	27.78	5.96	$0.127_{0.195}$	$0.155_{0.207}$	$0.133_{0.198}$	$0.039_{0.105}$	$0.048_{0.102}$	$0.041_{0.104}$	6.02E-19
ViT-B/16	86.00	16.86	$0.122_{0.179}$	$0.149_{0.206}$	$0.128_{0.186}$	$0.032_{0.082}$	$0.044_{0.102}$	$0.035_{0.086}$	3.35E-22
ViT-B/32	87.65	4.37	$0.129_{0.181}$	$0.149_{0.202}$	$0.133_{0.186}$	$0.033_{0.080}$	$0.041_{0.091}$	$0.035_{0.083}$	5.94E-44

(a)



(b)

Fig. 3: Predicting inter-annotator agreement (Dice) on the proposed IMA++ dataset. (a) Quantitative results ($\text{mean}_{std.dev.}$of MAE and MSE, and p-values for 13 model architectures ($\mathcal{M}_1$). † denotes top 3 models by overall MAE. (b) GradCAM++ saliency heatmaps from $\mathcal{M}_1$ (ResNet-18). Each image X_i shows corresponding overlaid segmentations $\{S_{ik}\}_{k=1}^{K_i}$ and (ground truth IAA Z_i, predicted IAA $\hat{Z}_i$) below the ISIC image ID.

means alone can be misleading if the underlying distributions differ in shape or variance. To address this, we compared the full distributions of IAA scores for benign and malignant lesions using the Mann-Whitney U test, which confirmed that agreement is significantly higher for benign lesions ($p < 0.01$), suggesting greater annotator consensus in those cases.

Our FOSD tests (Sec. 2.1), conducted using PySDTest [23], reinforce this conclusion. Using 1,000 bootstrap resampling iterations at significance level $\alpha = 0.001$, we rejected the hypothesis that malignant lesions stochastically dominate benign ones ($H_{\text{mal}\succeq_1\text{ben}}$, $p < 0.001$), while the reverse hypothesis ($H_{\text{ben}\succeq_1\text{mal}}$) was not rejected ($p = 0.923$). Together, these results support that the distribution of inter-annotator agreement for benign lesions first-order stochastically dominates that for malignant lesions, indicating higher segmentation consensus for benign cases. This is likely due to benign lesions often exhibiting more well-defined, homogeneous boundaries, making them easier to segment consistently. In contrast, malignant lesions tend to be more heterogeneous in appearance and morphology, which likely contributes to higher annotation variability.

Impact of other annotation factors on IAA: In addition to malignancy, Fig. 2C summarizes intra- and inter-factor-dependent IAA scores, along with corresponding Mann–Whitney U test p-values and Cohen's d [12] effect sizes. As expected, and consistent with findings in other medical imaging modalities [14,17], intra-annotator agreement is significantly higher than inter-annotator agreement. We also observe that segmentations performed using the same annotation tool tend to show higher agreement. Similarly, annotators with the same skill level exhibit greater consistency, particularly in the case of malignant lesions. To our knowledge, this represents the largest study of annotator variability in skin lesion segmentation to date in terms of dataset size, substantially exceeding the scale of prior work [24,13,32,40,34,35].

Table 1: Comparing the diagnostic performance of $\mathcal{M}_2$ to $\mathcal{M}_{\text{MT}}$ on IMA++ for different values of α (Eqn. 3). $\alpha = 0.9$ performs the best across all architectures.

		ResNet-18		MobileNetV2		EfficientNet-B1	
		Bal. Acc.	AUROC	Bal. Acc.	AUROC	Bal. Acc.	AUROC
Diagnosis Only ($\mathcal{M}_2$)		$0.746_{0.008}$	$0.835_{0.003}$	$0.757_{0.009}$	$0.843_{0.004}$	$0.746_{0.009}$	$0.827_{0.001}$
Multi-Task Learning ($\mathcal{M}_{\text{MT}}$)	$\alpha = 0.1$	$0.711_{0.009}$	$0.785_{0.001}$	$0.748_{0.003}$	$0.859_{0.001}$	$0.744_{0.016}$	$0.826_{0.018}$
	$\alpha = 0.2$	$0.723_{0.009}$	$0.822_{0.002}$	$0.740_{0.007}$	$0.857_{0.035}$	$0.750_{0.000}$	$0.853_{0.002}$
	$\alpha = 0.5$	$0.750_{0.004}$	$0.852_{0.006}$	$0.785_{0.006}$	$0.869_{0.006}$	$0.738_{0.010}$	$0.869_{0.003}$
	$\alpha = 0.8$	$0.757_{0.004}$	$0.852_{0.001}$	$0.797_{0.011}$	$0.879_{0.002}$	$0.767_{0.007}$	$0.873_{0.001}$
	$\alpha = 0.9$	$0.765_{0.002}$	$0.869_{0.002}$	$0.805_{0.004}$	$0.882_{0.001}$	$0.772_{0.009}$	$0.878_{0.003}$

Other datasets: In addition to IMA++, we also conduct experiments on 4 other dermoscopic image datasets: PH^2 [5], derm7pt [20], ISIC 2018 [10,39], ISIC 2019 [11,39,19]. We use the standardized partitions for ISIC 2018, 2019, and split PH^2 and derm7pt into train:valid:test in 70:15:15 ratio stratified by diagnosis.

All models were trained on an Ubuntu 20.04 workstation with AMD Ryzen 9 5950X, 32 GB RAM, NVIDIA RTX 3090 with Python 3.10.18 and PyTorch 2.7.1. All reported metrics are $\text{mean}_{\textit{std. dev.}}$ over 3 runs with different seeds. All trained models and code are available at https://github.com/sfu-mial/skin-IAV.

3.2 Image-based prediction of inter-annotator agreement

To directly predict IAA scores from images (Sec. 2.2), we evaluate 13 architectures spanning CNNs and Transformers, covering a wide range of capacities in terms of parameters and multiply-accumulate operations (MACs). Each model uses the backbone as a feature extractor with a regression head: `Linear(256)` → `BatchNorm1D` → `ReLU` → `Dropout(0.5)` → `Linear(1)`. All models were trained for 50 epochs using SGD (momentum = 0.9, weight decay = 1e-4, batch size = 32, learning rate = 1e-2 decayed ×0.1 every 10 epochs). We use Smooth-L_1 loss [15] as $\mathcal{L}_{\text{R}}$, selecting the model with the lowest validation MAE. Results are reported in terms of MAE, MSE, and Mann–Whitney U test p-values (Fig. 3a).

All models achieve good predictive performance (MAE $\in [0.10, 0.135]$), suggesting that IAA scores can be inferred from image content alone. Grad-CAM++ [8] visualizations (Fig. 3b) for the best model (ResNet-18) confirm saliency focused on the lesions and their boundaries. Notably, the third malignant example shows the model correctly localizing the lesion and predicting a plausible IAA (0.635), despite the "true" IAA being 0.0, highlighting label noise rather than prediction error. For all subsequent analyses, we use the top 3 performing architectures: ResNet-18, MobileNetV2, and EfficientNet-B1.

Table 2: Evaluating generalization performance on four other dermoscopic image datasets with 3 model architectures ($\mathcal{M}_{\text{MT}}$ with α set to 0.9 based on Table 1).

		ResNet-18		MobileNetV2		EfficientNet-B1	
		Bal. Acc.	AUROC	Bal. Acc.	AUROC	Bal. Acc.	AUROC
PH^2	Diag. Only ($\mathcal{M}_2$)	$0.938_{0.000}$	$0.988_{0.000}$	$0.943_{0.033}$	$0.988_{0.007}$	$0.870_{0.009}$	$0.979_{0.002}$
	Multi-Task ($\mathcal{M}_{\text{MT}}$)	$0.979_{0.009}$	$0.992_{0.000}$	$0.979_{0.009}$	$0.999_{0.002}$	$0.964_{0.009}$	$0.984_{0.004}$
derm7pt	Diag. Only ($\mathcal{M}_2$)	$0.734_{0.009}$	$0.836_{0.009}$	$0.654_{0.007}$	$0.800_{0.003}$	$0.756_{0.037}$	$0.862_{0.015}$
	Multi-Task ($\mathcal{M}_{\text{MT}}$)	$0.748_{0.005}$	$0.846_{0.001}$	$0.792_{0.012}$	$0.887_{0.002}$	$0.774_{0.011}$	$0.861_{0.003}$
ISIC 2018	Diag. Only ($\mathcal{M}_2$)	$0.744_{0.005}$	$0.893_{0.002}$	$0.727_{0.007}$	$0.872_{0.000}$	$0.713_{0.066}$	$0.868_{0.002}$
	Multi-Task ($\mathcal{M}_{\text{MT}}$)	$0.752_{0.003}$	$0.898_{0.001}$	$0.745_{0.007}$	$0.903_{0.003}$	$0.753_{0.012}$	$0.885_{0.047}$
ISIC 2019	Diag. Only ($\mathcal{M}_2$)	$0.670_{0.004}$	$0.853_{0.002}$	$0.623_{0.004}$	$0.849_{0.001}$	$0.657_{0.009}$	$0.869_{0.003}$
	Multi-Task ($\mathcal{M}_{\text{MT}}$)	$0.698_{0.008}$	$0.881_{0.001}$	$0.716_{0.023}$	$0.890_{0.002}$	$0.667_{0.006}$	$0.873_{0.001}$

3.3 Integrating inter-annotator agreement and diagnosis prediction

Finally, we leverage this link between malignancy and inter-annotator agreement and investigate whether jointly learning to predict IAA improves diagnostic performance (Sec. 2.3) by comparing diagnosis-only models ($\mathcal{M}_2$) with multi-task models ($\mathcal{M}_{\text{MT}}$). As before, we use Smooth-L_1 loss for $\mathcal{L}_{\text{R}}$ and focal loss [25] for $\mathcal{L}_{\text{D}}$. The multi-task architecture shares a common backbone and employs two heads: a regression head (as in Sec. 3.2) and a classification head (`Linear(256)` $\rightarrow$ `BatchNorm1D` $\rightarrow$ `ReLU` $\rightarrow$ `Dropout(0.5)` $\rightarrow$ `Linear(`n_{classes}`)`). To study the impact of loss weighting, we vary α in Eqn. 3, assigning lower ($\alpha \in \{0.1, 0.2\}$), equal ($\alpha = 0.5$), and higher ($\alpha \in \{0.8, 0.9\}$) emphasis on the diagnosis loss $\mathcal{L}_{\text{D}}$. All models are trained under the same setup as Sec. 3.2, except we select the model with the highest balanced accuracy on the validation set. We report balanced accuracy and AUROC in Table 1. Across all architectures, we find that $\alpha = 0.9$ yields the best diagnostic performance. Moreover, multi-task models ($\mathcal{M}_{\text{MT}}$) with equal or greater emphasis on $\mathcal{L}_{\text{D}}$ ($\alpha \geq 0.5$) consistently outperform diagnosis-only models ($\mathcal{M}_2$), confirming our hypothesis that inter-annotator agreement prediction serves as a beneficial auxiliary task for diagnosis.

To assess generalizability, we fine-tune the $\mathcal{M}_2$ and $\mathcal{M}_{\text{MT}}$ models (trained on IMA++ with $\alpha = 0.9$ for $\mathcal{M}_{\text{MT}}$) on external datasets: PH^2, derm7pt, ISIC 2018, and ISIC 2019. Since these datasets lack multiple annotations and thus have no IAA labels, we freeze the regression head of $\mathcal{M}_{\text{MT}}$ before fine-tuning. Fine-tuning is conducted for 15 epochs using SGD (momentum = 0.9, weight decay = 1e-4, batch size = 32, learning rate = 1e-3 with $\times 0.1$ decay every 3 epochs). Results in Table 2 show that $\mathcal{M}_{\text{MT}}$ outperforms $\mathcal{M}_2$ across all datasets and architectures, suggesting that the performance gains from learning to predict IAA on IMA++ may be transferable to new datasets.

4 Conclusion

We studied the problem of inter-annotator agreement (IAA) for skin lesion segmentation, and demonstrated, through agreement metrics and statistical tests,

a clear relationship between IAA and malignancy. We showed that IAA can be predicted from image content alone. Across five dermoscopic datasets, we further showed that incorporating IAA prediction as an auxiliary task in a multi-task diagnosis model improves performance over diagnosis-only models. To support this study, we curated IMA++, the largest publicly available multi-annotator skin lesion segmentation dataset, in terms of both the number of segmentations and unique annotators. To our knowledge, this is the most extensive IAA study in skin image analysis. Future work would assess how to test for null of non-dominance against dominance [43], evaluate other boundary-based metrics such as Hausdorff distance and boundary-F1 score, explore groupwise IAA measures instead of the pairwise measures: Dice and Hausdorff distance, and examine how inter-annotator variability impacts the ABCD score.

Acknowledgments. The authors are grateful for the computational resources provided by NVIDIA Corporation and Digital Research Alliance of Canada (formerly Compute Canada). Partial funding for this project was provided by the Natural Sciences and Engineering Research Council of Canada (NSERC RGPIN/06752-2020).

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Abhishek, K., et al.: Predicting the clinical management of skin lesions using deep learning. Sci Rep (2021)
2. Abhishek, K., et al.: Segmentation style discovery: Application to skin lesion images. In: MICCAI ISIC (2024)
3. Ahamed, S., et al.: Comprehensive evaluation and insights into the use of deep neural networks to detect and quantify lymphoma lesions in PET/CT images. arXiv preprint arXiv:2311.09614 (2023)
4. Asgari Taghanaki, S., et al.: Deep semantic segmentation of natural and medical images: A review. Artificial Intelligence Review (2021)
5. Ballerini, L., et al.: A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions. In: Color Medical Image Analysis (2013)
6. Barrett, G.F., et al.: Consistent tests for stochastic dominance. Econometrica (2003)
7. Celebi, M.E., et al.: A state-of-the-art survey on lesion border detection in dermoscopy images. Dermoscopy Image Analysis (2015)
8. Chattopadhay, A., et al.: Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In: WACV (2018)
9. Cheplygina, V., et al.: Crowd disagreement about medical images is informative. In: MICCAI LABELS (2018)
10. Codella, N., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). arXiv preprint arXiv:1902.03368 (2019)
11. Codella, N.C., et al.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In: ISBI (2018)

12. Cohen, J.: Statistical power analysis for the behavioral sciences. Routledge (2013)
13. Fortina, A.B., et al.: Where's the naevus? inter-operator variability in the localization of melanocytic lesion border. Skin Res Technol (2012)
14. Fu, M.C., et al.: Interrater and intrarater agreements of magnetic resonance imaging findings in the lumbar spine: significant variability across degenerative conditions. Spine J (2014)
15. Girshick, R.: Fast R-CNN. In: ICCV (2015)
16. Griff, S.K., et al.: Chapter 16 - breast cancer. In: Bragg, D.G., et al. (eds.) Oncologic Imaging (2002), https://doi.org/10.1016/B0-72-167494-1/50019-0
17. Gut, D., et al.: Use of superpixels for improvement of inter-rater and intra-rater reliability during annotation of medical images. Med Image Anal (2024)
18. Hasan, M.K., et al.: A survey, review, and future trends of skin lesion segmentation and classification. Comput Biol Med (2023)
19. Hernández-Pérez, C., et al.: BCN20000: Dermoscopic lesions in the wild. Sci Data (2024)
20. Kawahara, J., et al.: Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE J Biomed Health Inform (2018)
21. Kaya, S., et al.: Abrupt skin lesion border cutoff measurement for malignancy detection in dermoscopy images. BMC Bioinformatics (2016)
22. Korotkov, K., et al.: Computerized analysis of pigmented skin lesions: A review. Artif Intell Med (2012)
23. Lee, K., et al.: PySDTest: a Python/Stata package for stochastic dominance tests. arXiv preprint arXiv:2307.10694 (2023)
24. Li, X., et al.: Estimating the ground truth from multiple individual segmentations with application to skin lesion segmentation. In: MIUA (2010)
25. Lin, T.Y., et al.: Focal loss for dense object detection. In: ICCV (2017)
26. Lucieri, A., et al.: ExAID: A multimodal explanation framework for computer-aided diagnosis of skin lesions. Comput Methods Programs Biomed (2022)
27. Mann, H.B., et al.: On a test of whether one of two random variables is stochastically larger than the other. The Annals of Mathematical Statistics (1947)
28. Maurer, A., et al.: The benefit of multitask representation learning. Journal of Machine Learning Research (2016)
29. Mirikharaji, Z., et al.: D-LEMA: Deep learning ensembles from multiple annotations-application to skin lesion segmentation. In: CVPR ISIC (2021)
30. Mirikharaji, Z., et al.: A survey on deep learning for skin lesion segmentation. Med Image Anal (2023)
31. Patrício, C., et al.: Coherent concept-based explanations in medical image and its application to skin lesion diagnosis. In: CVPR SAIAD (2023)
32. Peruch, F., et al.: Simpler, faster, more accurate melanocytic lesion segmentation through MEDS. IEEE Trans Biomed Eng (2013)
33. Raumanns, R., et al.: ENHANCE (enriching health data by annotations of crowd and experts): A case study for skin lesion classification. Machine Learning for Biomedical Imaging (2021)
34. Ribeiro, V., et al.: Handling inter-annotator agreement for automated skin lesion segmentation. arXiv preprint arXiv:1906.02415 (2019)
35. Ribeiro, V., et al.: Less is more: Sample selection and label conditioning improve skin lesion segmentation. In: CVPR ISIC (2020)
36. Sampat, M.P., et al.: Measuring intra-and inter-observer agreement in identifying and localizing structures in medical images. In: ICIP (2006)
37. Sohns, C., et al.: Value of the BI-RADS classification in MR-mammography for diagnosis of benign and malignant breast tumors. Eur Radiol (2011)

38. Song, L., et al.: An end-to-end multi-task deep learning framework for skin lesion analysis. IEEE J Biomed Health Inform (2020)
39. Tschandl, P., et al.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci Data (2018)
40. Tschandl, P., et al.: Domain-specific classification-pretrained fully convolutional network encoders for skin lesion segmentation. Comput Biol Med (2019)
41. Wang, J., et al.: Contour-aware multi-expert model for ambiguous medical image segmentation. IEEE Trans Med Imaging (2025)
42. Wang, S., et al.: MSE-Nets: Multi-annotated semi-supervised ensemble networks for improving segmentation of medical image with ambiguous boundaries. arXiv preprint arXiv:2311.10380 (2023)
43. Whang, Y.J.: Econometric analysis of stochastic dominance: Concepts, methods, tools, and applications (2019)
44. Williams, N.M., et al.: Assessment of diagnostic accuracy of dermoscopic structures and patterns used in melanoma detection: A systematic review and meta-analysis. JAMA Dermatol (2021)
45. Xie, Y., et al.: A mutual bootstrapping model for automated skin lesion segmentation and classification. IEEE Trans Med Imaging (2020)
46. Yang, X., et al.: A novel multi-task deep learning model for skin lesion segmentation and classification. arXiv preprint arXiv:1703.01025 (2017)
47. Zepf, K., et al.: That label's got style: Handling label style bias for uncertain image segmentation. arXiv preprint arXiv:2303.15850 (2023)
48. Zhao, Y., et al.: Multi-task deep learning for medical image computing and analysis: A review. Comput Biol Med (2023)