Lesion Elevation Prediction from Skin Images Improves Diagnosis

Kumar Abhishek[0000−0002−7341−9617] and Ghassan Hamarneh[0000−0001−5040−7448]

School of Computing Science, Simon Fraser University, Canada
{kabhishe,hamarneh}@sfu.ca

Abstract. While deep learning-based computer-aided diagnosis for skin lesion image analysis is approaching dermatologists' performance levels, there are several works showing that incorporating additional features such as shape priors, texture, color constancy, and illumination further improves the lesion diagnosis performance. In this work, we look at another clinically useful feature, skin lesion elevation, and investigate the feasibility of predicting and leveraging skin lesion elevation labels. Specifically, we use a deep learning model to predict image-level lesion elevation labels from 2D skin lesion images. We test the elevation prediction accuracy on the derm7pt dataset, and use the elevation prediction model to estimate elevation labels for images from five other datasets: ISIC 2016, 2017, and 2018 Challenge datasets, MSK, and DermoFit. We evaluate cross-domain generalization by using these estimated elevation labels as auxiliary inputs to diagnosis models, and show that these improve the classification performance, with AUROC improvements of up to 6.29% and 2.69% for dermoscopic and clinical images, respectively. The code is publicly available at https://github.com/sfu-mial/LesionElevation.

Keywords: skin lesion · lesion elevation · deep learning · diagnosis.

1 Introduction

Skin cancer is highly prevalent globally and the most commonly diagnosed cancer in the USA [5] with over 5 million annual diagnoses [35]. Although it accounts for a small fraction of all skin cancers, melanoma is the deadliest form with an estimated 99,700 diagnoses and 8,290 deaths in 2024 in the USA alone, and timely diagnosis is critical as early detection results in a 99% estimated 5-year survival rate. Deep learning (DL)-based methods have proven to be successful in improving image-based clinical decision support systems with expert-level computer-aided dermatological diagnosis [21,10]. While DL-based methods have demonstrated remarkable performance, there is a considerable body of research showing that incorporating additional features, such as shape priors [33], texture [46], color constancy [36], and illumination [2], can further improve skin lesion image analysis. Another feature that has been shown to enhance lesion diagnosis and clinical management prediction, using both classical machine learning [29] and

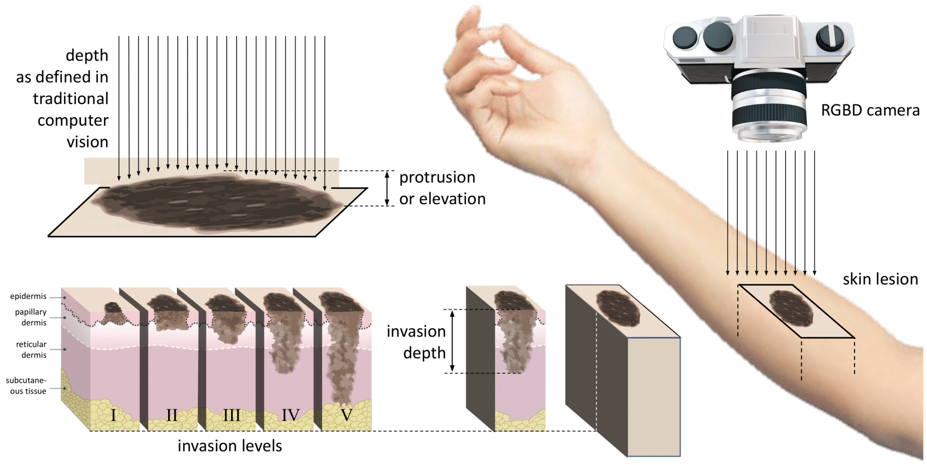
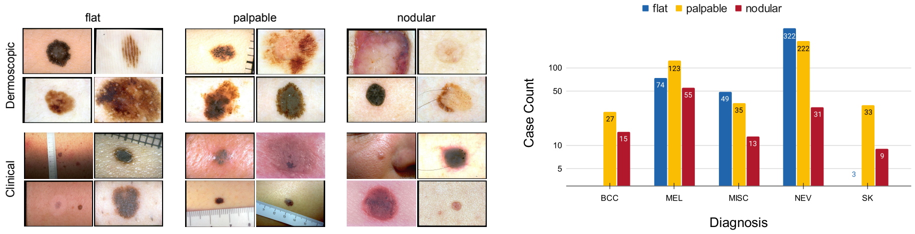


Fig. 1: Visualizing the difference between skin lesion elevation versus depth. Invasion levels inset figure courtesy of Melanoma Institute Australia [31].

DL methods [27,26,37,3,32], is lesion elevation. However, learning-based methods have yet to incorporate lesion elevation prediction into computerized diagnosis.

The American Cancer Society's ABCDE criteria include elevation (E) as one of the components [43]. Moreover, in the clinical setting, dermatologists often palpate the skin to examine the lesion when making a diagnosis [16]. Case in point, a study showed that palpation alone, without any visual assessment, was sufficient to correctly diagnose 14 of 16 cases [17]. With the rise of teledermatology, partly accelerated by external factors such as COVID-19 [4], one of the major reasons for dermatologists' dissatisfaction with teledermatology is the inability to palpate lesions [18,20]. This is particularly pressing for "store-and-forward" teledermatology, where images are captured and submitted alongside patient history, which has been adopted for its efficiency and low cost, is imperfect since "*even good quality photos are two-dimensional; raised lesions . . . for example, may be difficult to distinguish from flat lesions of a similar colour*" [16]. Clinical and dermoscopic images of lesions do not capture elevation, and while it is recommended to capture tangential views of lesions in teledermatology, measuring elevation is not easy to do with limited camera views, making the examinations "less complete" [8,25]. Therefore, while teledermatology has the potential to improve triage, access to care for underserved communities, and patient convenience [34,15,40,24], it would greatly benefit from being able to leverage lesion elevation information as a proxy for in-person palpation. Solutions to bridge this gap could be either in the form of patient side hardware [28], which is expensive to develop, maintain, and deploy, or purely software-based approaches to estimate lesion elevation from single RGB images, which we focus on.

Before proceeding, it is worth clarifying the difference in terminology vis-à-vis lesion "elevation" and "depth" (Fig. 1), and how these terms differ in their



(a) Sample dermoscopic and clinical images from derm7pt showing the three image-level lesion elevation labels.

(b) Distribution of elevation. Note the absence of a clear diagnosis-to-elevation mapping.

Fig. 2: derm7pt dataset: (a) sample images categorized by elevation labels and (b) distribution of elevation labels across diagnoses.

usage in dermatology compared to traditional computer vision. Lesion elevation refers to the lesion's surface and how it protrudes above the outer skin surface (epidermis). On the other hand, lesion depth or thickness, unlike the definition of depth in computer vision, refers to the depth of invasion of melanoma underneath the skin surface and is used for melanoma staging, measured using scales such as Breslow's depth [9] and Clark level [11].

While lesion elevation has been used along with other clinical metadata (e.g., gender, lesion location, and age) for skin lesion image analysis tasks and has shown to improve performance [26,3], to the best of our knowledge, there is no work that explores either the utility of elevation alone as a metadata, or the feasibility of predicting elevation from 2D RGB skin lesion images. In this work, we pose three research questions: (i) can we predict, with sufficient accuracy, lesion elevation from a single lesion image?; (ii) does lesion elevation alone, without any other metadata, improve lesion diagnosis?; (iii) can we leverage an elevation prediction model to infer elevations on datasets without ground truth elevation, thus potentially improving the diagnosis accuracies thereon? Our results show that the answer is affirmative to all these questions.

2 Method

The dataset: Let $(\mathcal{X}, \mathcal{Y}, \mathcal{E})$ be the dataset of images, diagnosis labels, and elevation labels. Specifically, $\mathcal{X} = \{X_i\}_{i=1}^{N}$ is the set of skin images with corresponding diagnosis labels $\mathcal{Y} = \{Y_i\}_{i=1}^{N}$ and single image-level elevation labels $\mathcal{E} = \{E_i\}_{i=1}^{N}$, where $X_i \in \mathbb{R}^{H \times W \times 3}$, $Y_i \in \{1, 2, \cdots, N_D\}$, and $E_i \in \{1, 2, \cdots, N_E\}$, and N_D and N_E denote the total number of diagnosis and image-level elevation class labels, respectively.

A diagnosis prediction model: A diagnosis prediction model f_D, parameterized by Θ_D, is trained to generate disease predictions from images,

$$\widehat{Y}_i = f_D(X_i; \Theta_D). \tag{1}$$

Leveraging ground truth elevation labels for diagnosis prediction: The f_D model architecture can also be modified to take the elevation label as an auxiliary input for diagnosis prediction,

$$\widehat{Y}_i = f_{DE}(X_i \oplus E_i; \Theta_{DE}), \tag{2}$$

where $\oplus$ is a combination operator.
Predicting elevation labels: Additionally, since we have images with corresponding image-level elevation labels, we can also train a DL-model g to predict elevation labels from an input image,

$$\widehat{E}_i = g(X_i; \Phi), \tag{3}$$

where $\widehat{E}_i \in \mathbb{R}^{N_E}$ is an N_E-element probabilistic prediction of the elevation label. We denote elevation class label with the highest predicted probability as $\widehat{E}_i^{\max} = \arg\max_j \widehat{E}_{ij}$. For example, if $N_E = 3$ and $\widehat{E}_i = [0.1, 0.7, 0.2]$, then $\widehat{E}_i^{\max} = 2$.
Leveraging predicted elevation labels for diagnosis models: Finally, given a trained elevation prediction model g, we use this model to infer elevation labels (Eqn. 3) on datasets without ground truth elevation, and use these labels as auxiliary inputs to re-train the diagnosis prediction model,

$$\widehat{Y}_i = f_{D\widehat{E}}(X_i \oplus \widehat{E}_i; \Theta_{D\widehat{E}}). \tag{4}$$

More details about exact model architectures, losses, and metrics for evaluation are discussed in the next section.

3 Results and Discussion

Datasets: Since skin lesion elevation data is expensive and difficult to acquire, there are, to the best of our knowledge, only 2 publicly available datasets with lesion elevation labels: PAD-UFES-20 [38] and derm7pt [26]. PAD-UFES-20 contains 2,298 smartphone camera images of skin lesions with binary labels indicating if a lesion is elevated or not, whereas derm7pt contains 1,011 cases with clinical and dermoscopic images with 3 elevation class labels: "flat", "palpable", and "nodular", and because of the relatively more granular elevations and the presence of two imaging modalities in the latter, we use derm7pt for our experiments. We partition the dataset with elevation label-based stratification into training, validation, and testing sets in the ratio of 70:15:15, accounting for the inherent class imbalance: "flat": 448 cases, "palpable": 440 cases, "nodular": 123 cases. See Fig. 2 (a,b) for sample images from different elevation labels and diagnosis-wise distribution of elevation labels, respectively. We group the diagnosis labels in derm7pt into 5 classes as originally proposed by Kawahara et al. [26]: BCC (basal cell carcinoma), MEL (melanoma), NEV (nevi), SK (seborrheic keratosis), and MISC (miscellaneous). Although some elevation labels appear more or less frequently with certain diagnoses, we note that there is no direct diagnosis-elevation mapping, and elevation labels are distributed across

Table 1: Results (accuracy and area under the ROC curve (for skin lesion elevation prediction from clinical and dermoscopic images of derm7pt. Reported values are the mean ± std. dev. averaged over 3 runs. Numbers in [·] present the 95% CI values. Bold values denote the best values for the metrics.

Model		Clinical Images		Dermoscopic Images	
Architecture	# Params (M)	Accuracy ↑	AUROC ↑	Accuracy ↑	AUROC ↑
MobileNetV2	2.228	0.8234 ± 0.0515 [0.7954, 0.8514]	0.7474 ± 0.0496 [0.7251, 0.7697]	0.8039 ± 0.0552 [0.7780, 0.8298]	0.7789 ± 0.0536 [0.7549, 0.8029]
MobileNetV3L	4.206	0.7969 ± 0.0561 [0.7712, 0.8226]	0.7326 ± 0.0535 [0.7111, 0.7541]	0.7908 ± 0.0576 [0.7659, 0.8157]	0.7481 ± 0.0552 [0.7260, 0.7702]
EfficientNet-B0	4.011	0.8190 ± 0.0582 [0.7914, 0.8466]	0.7444 ± 0.0570 [0.7222, 0.7666]	0.8257 ± 0.0573 [0.7978, 0.8536]	0.8088 ± 0.0563 [0.7825, 0.8351]
EfficientNet-B1	6.517	0.8013 ± 0.0604 [0.7752, 0.8274]	0.7284 ± 0.0602 [0.7071, 0.7497]	0.8344 ± 0.0604 [0.8056, 0.8632]	0.8033 ± 0.0576 [0.7774, 0.8292]
DenseNet-121	6.957	0.8146 ± 0.0589 [0.7874, 0.8418]	0.7405 ± 0.0568 [0.7186, 0.7624]	0.8301 ± 0.0589 [0.8018, 0.8584]	0.7931 ± 0.0544 [0.7680, 0.8182]
VGG-16	14.724	**0.8543 ± 0.0632** [0.8229, 0.8857]	**0.8220 ± 0.0610** [0.7941, 0.8499]	**0.8475 ± 0.0592** [0.8173, 0.8777]	**0.8152 ± 0.0582** [0.7883, 0.8421]
ResNet-18	11.178	0.8190 ± 0.0582 [0.7914, 0.8466]	0.7321 ± 0.0555 [0.7106, 0.7536]	0.7996 ± 0.0536 [0.7740, 0.8252]	0.7653 ± 0.0530 [0.7422, 0.7884]
ResNet-50	23.514	0.7660 ± 0.0607 [0.7425, 0.7895]	0.6927 ± 0.0576 [0.6732, 0.7122]	0.8083 ± 0.0576 [0.7820, 0.8346]	0.7586 ± 0.0536 [0.7359, 0.7813]

all diagnoses in our dataset (an exception is that BCC and SK have almost no "flat" elevations).

In addition, we also use five other datasets for diagnosis prediction that do not contain ground truth elevation labels: (i) ISIC 2016 [23], (ii) ISIC 2017 [13], (iii) ISIC 2018 [12], (iv) MSK [1], and (v) DermoFit [7], where (i) is a binary classification dataset and all others are multi-class classification datasets. Note that (i)-(iv) are dermoscopic image datasets while (v) contains clinical images. For (i)-(iii), we use the standard dataset partitions, and for (iv), (v), we generate training, validation, and testing partitions in 70:10:20 ratio.

Experiment 1: Can we predict skin lesion elevation labels from images alone? To test the feasibility of predicting skin lesion elevation labels directly from images, we train eight different DL model architectures on the derm7pt dataset. Specifically, we train elevation prediction models g that, given a skin lesion image X_i, predict the elevation label $\widehat{E}_i$ (Eqn. 3). We choose the architectures from a variety of families, covering a large range of model sizes (see parameter counts in Table 1): ResNet-18 and ResNet-50, MobileNetV2 and MobileNetV3L, DenseNet-121, EfficientNetB0 and EfficientNetB1, and VGG-16. For all architectures except VGG-16, we modify the final layer to predict 3 classes ($N_E = 3$ for derm7pt). However, since a large number of parameters in VGG-16 emanate from the fully-connected layers, we modify the architecture by replacing these fully-connected layers with a global average pooling layer [44]. We use ImageNet-pretrained weights for initialization. All models are trained

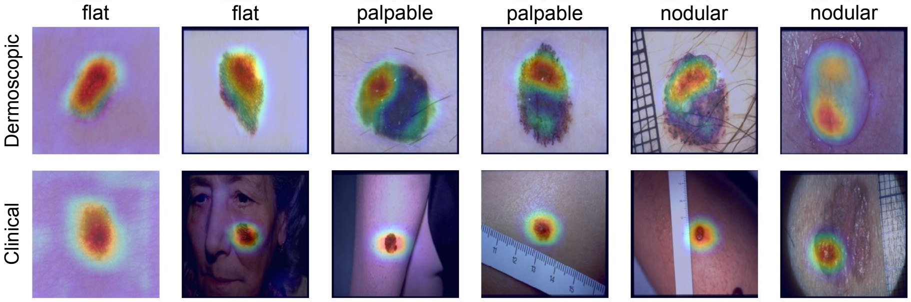


Fig. 3: Visualizing class activation maps for skin lesion elevation label prediction for dermoscopic and clinical images, generated through GradCAM. Notice how the activation areas are focused around the lesion regions, indicating that the prediction model g does not learn to rely on spurious features or "shortcuts".

for 50 epochs with stochastic gradient descent and momentum of 0.9, weight decay of 1e-4, batch size of 32, and a learning rate of 1e-2 which was decayed by a factor of 0.1 every 10 epochs. All images are resized to 224 × 224 and we augment the images with horizontal and vertical flips and rotations in multiple of 90°. To account for the inherent class imbalance, we use the cross-entropy loss with median frequency balancing to assign class weights, i.e. class-wise weights in the loss calculation are weighted by the ratio of the median of class frequencies in the entire training set to each class's frequency [19,6]. The model with the best area under the ROC curve (AUROC) on the validation set was chosen for evaluation. All experiments were repeated 3 times for robust results.

Table 1 lists the quantitative results for elevation prediction on both clinical and dermoscopic images in the derm7pt dataset. We report mean and the std. dev. of the overall accuracy of classification as well as the AUROC across 3 repeated runs, as well as the 95% confidence intervals (CIs). We observe that while all architectures are able to predict elevation labels reasonably accurately, the VGG-16 model performs the best across both imaging modalities. To ascertain that this performance is not due to the model learning spurious features or "shortcuts" in the images to make the predictions, we generate the class activation maps (CAMs) for the VGG-16 model using GradCAM [42]. Sample CAMs for both modalities and all elevation labels are shown in Fig. 3. We observe that the CAMs are almost completely contained within and around the lesion regions, suggesting that the elevation predictions are indeed based on lesion features. Since VGG-16 most accurately predicts elevation for both modalities, we use this model architecture for all subsequent experiments.

Experiment 2: Do ground truth elevation labels help improve lesion diagnosis? For this experiment, we train lesion diagnosis models f_{DE} (Eqn. 2) that leverage ground truth elevation labels as auxiliary inputs and compare their diagnosis performance to "vanilla" diagnosis models f_D trained without any elevation labels (Eqn. 1). To combine the elevation labels as inputs along

Table 2: Leveraging inferred elevation labels (Eqn. 4), either "discrete" ($f_{D\widehat{E}^{\max}}$) or "probabilistic" ($f_{D\widehat{E}}$) improves diagnosis performance over no elevation labels (f_D). Reported metrics are mean ± std. dev. over 3 repeated runs. We also report statistical significance tests (McNemar's mid-p test) and effect sizes (Cohen's d).

Dataset	**Experiment**	**Metrics**						**Statistical Tests**	
		Bal. Acc. ↑	Accuracy ↑	Precision ↑	Recall ↑	F1-score ↑	AUROC ↑	p-value	Cohen's d
DermoFit [7]	f_D	0.8145 ± 0.0170	0.9331 ± 0.0051	0.8121 ± 0.0194	0.8145 ± 0.0170	0.8103 ± 0.0149	0.8856 ± 0.0092	-	-
	$f_{D\widehat{E}}$	0.8586 ± 0.0003	0.9480 ± 0.0003	0.8449 ± 0.0004	0.8586 ± 0.0003	0.8511 ± 0.0002	0.9125 ± 0.0001	9.87e-03	4.1348
	$f_{D\widehat{E}^{\max}}$	0.8541 ± 0.0009	0.9497 ± 0.0010	0.8466 ± 0.0019	0.8541 ± 0.0009	0.8500 ± 0.0015	0.9108 ± 0.0007	7.08e-03	3.8626
MSK [1]	f_D	0.6004 ± 0.0010	0.8446 ± 0.0159	0.6156 ± 0.0302	0.6004 ± 0.0010	0.5843 ± 0.0091	0.7374 ± 0.0017	-	-
	$f_{D\widehat{E}}$	0.6514 ± 0.0019	0.8833 ± 0.0018	0.7228 ± 0.0037	0.6514 ± 0.0019	0.6726 ± 0.0013	0.7747 ± 0.0014	4.04e-12	23.9526
	$f_{D\widehat{E}^{\max}}$	0.6352 ± 0.0047	0.8878 ± 0.0011	0.7169 ± 0.0210	0.6352 ± 0.0047	0.6638 ± 0.0021	0.7632 ± 0.0038	4.10e-11	8.7647
ISIC 2016 [23]	f_D	0.7008 ± 0.0307	0.8100 ± 0.0474	0.7208 ± 0.0524	0.7008 ± 0.0307	0.6998 ± 0.0338	0.7008 ± 0.3070	-	-
	$f_{D\widehat{E}}$	0.7344 ± 0.0124	0.8545 ± 0.0131	0.7615 ± 0.0022	0.7344 ± 0.0124	0.7467 ± 0.0059	0.7344 ± 0.0124	7.36e-02	0.1547
	$f_{D\widehat{E}^{\max}}$	0.7574 ± 0.0183	0.8391 ± 0.0165	0.7513 ± 0.0213	0.7574 ± 0.0183	0.7515 ± 0.0045	0.7574 ± 0.0183	8.75e-02	0.2603
ISIC 2017 [13]	f_D	0.6926 ± 0.0207	0.8296 ± 0.0072	0.7303 ± 0.0160	0.6926 ± 0.0207	0.7060 ± 0.0133	0.6926 ± 0.0207	-	-
	$f_{D\widehat{E}}$	0.7417 ± 0.0030	0.8500 ± 0.0060	0.7634 ± 0.0118	0.7417 ± 0.0030	0.7513 ± 0.0036	0.7417 ± 0.0030	3.06e-02	3.3198
	$f_{D\widehat{E}^{\max}}$	0.7555 ± 0.0040	0.8583 ± 0.0044	0.7776 ± 0.0095	0.7555 ± 0.0040	0.7644 ± 0.0018	0.7555 ± 0.0040	9.80e-03	4.2192
ISIC 2018 [12]	f_D	0.7949 ± 0.0303	0.9450 ± 0.0055	0.7601 ± 0.0426	0.7949 ± 0.0303	0.7690 ± 0.0413	0.8808 ± 0.0190	-	-
	$f_{D\widehat{E}}$	0.8481 ± 0.0016	0.9668 ± 0.0021	0.8314 ± 0.0064	0.8481 ± 0.0016	0.8390 ± 0.0043	0.9132 ± 0.0014	7.54e-03	2.4051
	$f_{D\widehat{E}^{\max}}$	0.8524 ± 0.0024	0.9641 ± 0.0032	0.8250 ± 0.0102	0.8524 ± 0.0024	0.8376 ± 0.0046	0.9143 ± 0.0012	3.52e-03	2.4885

with the lesion image (i.e., the $\oplus$ operator in Eqn. 2), we concatenate the one-hot encoded elevation labels for each image to the output of VGG-16's global-average pooling layer, which is then passed to the final classification layer, thus adding only a minimal number of parameters ($N_E \times N_D$, i.e., the number of elevation labels $\times$ the number of diagnosis classes). The training details (optimizer, loss, number of epochs, learning rate) for both f_{DE} and f_D remain the same.

We observe that for clinical images, leveraging ground truth elevation labels for diagnosis prediction (f_{DE}) improves the performance [overall accuracy, AUROC]: [0.8569, 0.6820] compared to diagnosis without elevation (f_D): [0.8464, 0.6331]. A similar improvement is noted for dermoscopic images: the performance with elevation labels: [0.9216, 0.8703] is an improvement over that of a "vanilla" diagnosis model: [0.9137, 0.8431]. This improvement in AUROC of 4.89% and 2.72% for clinical and dermoscopic images, respectively, is consistent with findings from previous works [26,3] that showed that using elevation labels along with other metadata is beneficial for lesion diagnosis prediction.

Experiment 3: Can inferred elevation labels improve lesion diagnosis? Having established that it is possible, with a reasonable accuracy, to predict elevation labels from lesion images, and that elevation labels improve lesion diagnosis, the natural next question is if we can infer lesion elevation on datasets that do not contain elevation labels, and if diagnosis prediction models trained with these inferred elevation labels also improve diagnosis accuracy. Therefore, given a trained elevation prediction model g, we infer elevation labels for all images in 5 skin lesion datasets that do not have elevation labels: ISIC 2016, ISIC 2017, ISIC 2018, MSK, and DermoFit. We note that there is a considerable domain shift between these skin lesion datasets [45], and therefore we use modality specific elevation prediction models for inferring elevation labels, i.e., the elevation prediction model g trained on derm7pt's dermoscopic images is used for the first 4 datasets, and g trained on derm7pt's clinical images is used for DermoFit. Next, for each dataset, we train three prediction models: (i) diagnosis prediction without any elevation labels (f_D), (ii) diagnosis prediction with probabilistic "soft" inferred elevation labels ($f_{D\widehat{E}}$), and (iii) diagnosis prediction with "discrete" inferred elevation labels ($f_{D\widehat{E}^{\max}}$). Model training details remain the same as Experiment 1, except the models are trained for longer (20 epochs for ISIC 2018 and 50 epochs for the other datasets), since these datasets are larger than derm7pt. We report several classification metrics: balanced accuracy, overall accuracy, precision, recall, F1-score, and AUROC, and train each model thrice for robustness. In addition to these metrics, we also perform statistical analysis: McNemar's mid-p test [30,22] and effect size (Cohen's d [14]) for comparing $\{f_{D\widehat{E}}, f_{D\widehat{E}^{\max}}\}$ AUROC predictions to those from f_D.

Quantitative results in Table 2 show that leveraging estimated lesion elevation labels consistently improves diagnosis performance across all datasets: up to 6.29% and 2.69% improvements in AUROC for dermoscopic and clinical images, respectively. Moreover, for all datasets except ISIC 2016, this improvement is statistically significant at $p < 0.05$. Similarly, Cohen's d estimates indicate "huge" effect sizes for these four datasets and "small" effect size for ISIC 2016 [41]. While both "soft" and "discrete" estimates of the elevation label appear to improve diagnosis performance, interestingly, there does not appear to be a consistent pattern of one of them outperforming the other. This is especially surprising since the "soft" labels would convey the uncertainty associated with the elevation prediction, and intuitively, they would be more informative. Nevertheless, we shelve this observation for a future investigation.

4 Conclusion

In this work, we showed that it is possible to predict image-level lesion elevation labels directly from 2D RGB skin lesion images with sufficient accuracy, and that these estimated elevation labels do indeed help improve lesion diagnosis on other datasets, improving AUROC by up to 6.29% and 2.69% on dermoscopic and clinical images, respectively. The ability to predict lesion elevation from 2D images, in addition to improving computer-aided diagnosis, offers the potential to improve teledermatology consults by offering practitioners access to useful estimates of clinical information otherwise unavailable in virtual consultations. Our experiments with off-the-shelf monocular depth prediction models [39] from natural computer vision failed to generate any usable depth maps (see Fig. SM1 in the Supplementary Material), and we postulate that this may be because of the difference in scale of depth that these models are trained on (several orders of magnitude larger than skin lesion elevation) as well as the scene anisotropy of the images they are trained on (natural images generally have a depth anisotropy where the lower parts of the image are closer to the camera plane, which is typically not true for skin lesion images). Therefore, in future work, we would like to explore the feasibility, accuracy, and utility of reconstructing dense elevation maps from single RGB images, specific to skin lesions. Another future direction would be improving the elevation prediction accuracy, which may help reach the upper bound of performance improvement achieved when using ground truth elevation labels. Finally, we would also like to explore using multiple datasets for training elevation labels' prediction models to alleviate any potential biases emanating from using a single dataset (derm7pt).

Acknowledgments. The authors would like to thank Jeremy Kawahara, Manolis Savva, and Aditi Jain for helpful discussions on this work, and 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. International Skin Imaging Collaboration (ISIC): Melanoma Project - ISIC Archive. https://www.isic-archive.com/ (2016), [Online. Accessed 01-03-2024]
2. Abhishek, K., et al.: Illumination-based transformations improve skin lesion segmentation in dermoscopic images. In: CVPRW (2020)
3. Abhishek, K., et al.: Predicting the clinical management of skin lesions using deep learning. Sci Rep (2021)
4. AlAbdulkareem, A.: Palpation in dermatology, will covid-19 be the last straw? Dermatol Ther (2021)
5. American Cancer Society: Cancer Facts & Figures 2024 (2024), https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/2024-cancer-facts-figures.html [Accessed: 01-03-2024]
6. Badrinarayanan, V., et al.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell (2017)
7. 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)
8. Bashshur, R.L., et al.: The empirical foundations of teledermatology: A review of the research evidence. Telemed J E Health (2015)
9. Breslow, A.: Thickness, cross-sectional areas and depth of invasion in the prognosis of cutaneous melanoma. Ann Surg (1970)
10. Brinker, T.J., et al.: Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. Eur J Cancer (2019)
11. Clark Jr, W.H., et al.: The histogenesis and biologic behavior of primary human malignant melanomas of the skin. Cancer Res (1969)
12. Codella, N., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC). arXiv:1902.03368 (2019)
13. 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)

14. Cohen, J.: Statistical power analysis for the behavioral sciences. routledge (2013)
15. Coustasse, A., et al.: Use of teledermatology to improve dermatological access in rural areas. Telemed J E Health (2019)
16. Cox, N.H.: Palpation of the skin – an important issue. J R Soc Med (2006)
17. Cox, N.H.: A literally blinded trial of palpation in dermatologic diagnosis. J Am Acad Dermatol (2007)
18. Eedy, D., et al.: Teledermatology: A review. Br J Dermatol (2001)
19. Eigen, D., et al.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: ICCV (2015)
20. English, J., et al.: Has teledermatology in the uk finally failed? Br J Dermatol (2007)
21. Esteva, A., et al.: Dermatologist-level classification of skin cancer with deep neural networks. Nature (2017)
22. Fagerland, M.W., et al.: The mcnemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional. BMC Medical Research Methodology (2013)
23. Gutman, D., et al.: Skin lesion analysis toward melanoma detection: A challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC). arXiv:1605.01397 (2016)
24. Hwang, J.K., et al.: Review of teledermatology: lessons learned from the covid-19 pandemic. Am J Clin Dermatol (2024)
25. Jahnke, M., et al.: Pediatric teledermatology (Oct 2022), https://www.hmpgloballearningnetwork.com/site/thederm/cover-story/pediatric-teledermatology [Accessed: 01-03-2024]
26. Kawahara, J., et al.: Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE J Biomed Health Inform (Mar 2019)
27. Kharazmi, P., et al.: A feature fusion system for basal cell carcinoma detection through data-driven feature learning and patient profile. Skin Res Technol (2018)
28. Kim, K.: Roughness based perceptual analysis towards digital skin imaging system with haptic feedback. Skin Res Technol (2016)
29. Li, X., et al.: Depth data improves skin lesion segmentation. In: MICCAI (2009)
30. McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika (1947)
31. Melanoma Institute Australia: Melanoma diagnosis, https://melanoma.org.au/for-patients/melanoma-diagnosis/ [Accessed: 01-03-2024]
32. Mendes, C.F.S.d.F., et al.: Deep and handcrafted features from clinical images combined with patient information for skin cancer diagnosis. Chaos, Solitons & Fractals (2022)
33. Mirikharaji, Z., et al.: Star shape prior in fully convolutional networks for skin lesion segmentation. In: MICCAI (2018)
34. Morenz, A.M., et al.: Evaluation of barriers to telehealth programs and dermatological care for american indian individuals in rural communities. JAMA Dermatol (2019)
35. Nagarajan, P., , et al.: Keratinocyte carcinomas: Current concepts and future research priorities. Clin Cancer Res (2019)
36. hua Ng, J., et al.: The effect of color constancy algorithms on semantic segmentation of skin lesions. In: SPIE Medical Imaging (2019)
37. Pacheco, A.G., et al.: The impact of patient clinical information on automated skin cancer detection. Comput Biol Med (2020)
38. Pacheco, A.G., et al.: PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. Data Brief (2020)

39. Ranftl, R., et al.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Trans Pattern Anal Mach Intell (2022)
40. Santosa, A., et al.: Teledermatology in an emergency department: benefits and gaps. BMC Emerg Med (2023)
41. Sawilowsky, S.S.: New effect size rules of thumb. J Mod Appl Stat Methods (2009)
42. Selvaraju, R.R., et al.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: ICCV (2017)
43. Strayer, S.M., et al.: Diagnosing skin malignancy: Assessment of predictive clinical criteria and risk factors. J Fam Pract (2003), https://www.mdedge.com/familymedicine/article/60122/diagnosing-skin-malignancy-assessment-predictive-clinical-criteria-and
44. Yan, Y., et al.: Melanoma recognition via visual attention. In: IPMI (2019)
45. Yoon, C., et al.: Generalizable feature learning in the presence of data bias and domain class imbalance with application to skin lesion classification. In: MICCAI (2019)
46. Zhang, L., et al.: Automatic skin lesion segmentation by coupling deep fully convolutional networks and shallow network with textons. J Med Imaging (2019)