MELANOMA DETECTION WITH UNCERTAINTY QUANTIFICATION

SangHyuk Kim[1,2] *Edward Gaibor*[1] *Brian Matejek*[2] *Daniel Haehn*[1]

[1] Department of Computer Science, University of Massachusetts Boston, Boston, Massachusetts, USA
[2]Computer Science Laboratory, SRI International, Arlington, Virginia, USA

ABSTRACT

Early detection of melanoma is crucial for improving survival rates. Current detection tools often utilize data-driven machine learning methods but often overlook the full integration of multiple datasets. We combine publicly available datasets to enhance data diversity, allowing numerous experiments to train and evaluate various classifiers. We then calibrate them to minimize misdiagnoses by incorporating uncertainty quantification. Our experiments on benchmark datasets show accuracies of up to 93.2% before and 97.8% after applying uncertainty-based rejection, leading to a reduction in misdiagnoses by over 40.5%. Our code and data are publicly available[1], and a web-based interface[2] for quick melanoma detection of user-supplied images is also provided.

***Index Terms*—** melanoma detection, machine learning, uncertainty quantification, calibration, medical imaging

1. INTRODUCTION

Melanoma is a severe skin cancer responsible for around 55,500 deaths annually [1]. Deep Neural Networks (DNNs) are effective in melanoma detection, but their evaluation lacks standardization, complicating comparisons. For example, DNN-based melanoma classification using segmented features [2] has shown promising results on the ISIC'16 [3] and ISIC'17 [4] datasets. Transfer learning methods [5] have also achieved high accuracy on ISIC'18 [6]. Additionally, Vision Transformers [7] have demonstrated exceptional accuracy on a private dataset. However, many studies operate under inconsistent testing conditions.

Melanoma lesions exhibit significant variability, making detection challenging in real-world settings. While DNNs require extensive datasets for accuracy, only a few studies have integrated multiple datasets within a unified framework. Additionally, although DNN predictions can be accurate, they may be confidently incorrect, underscoring the need for reliable diagnostic methods.

We develop a framework for consistent experimentation with datasets and DNNs to minimize misdiagnoses. It includes four modules: Input, Melanoma Recognition, Uncertainty Analysis, and Integration, as shown in Fig. 1. Our methodology involves importing multiple datasets, uniform preprocessing, and evaluating classifiers in classification, calibration, and uncertainty scores. We filter out uncertain predictions to enhance the classification and calibration performance of the remaining samples. Ultimately, we aim to develop reliable classifiers for clinical use and help researchers identify optimal dataset and classifier combinations.

[1]Our GitHub repository, `https://mpsych.org/melanoma`
[2]Our Web interface, `https://mpsych.github.io/melanoma/`

2. METHODS

2.1. Input

Most public melanoma datasets include images and associated metadata about skin lesions. Image sizes range from 147x147 to 3096x3096 pixels, with class counts from 2 to 9, consistently including melanoma labels. These datasets vary in structure—some use a single folder with a separate CSV for labels and splits, while others organize by folders or categories. We parse these diverse structures to load them into compatible environments and create merged datasets with a common interface.

2.2. Melanoma Recognition

Classification. We use a unified framework for preprocessing, augmentation, training, validation, and inference, enabling direct performance comparisons across experiments. This approach combines datasets with DNN models and applies standard metrics like precision, sensitivity, specificity, F1, accuracy, and AUC-ROC. The softmax classifier categorizes results as either "Melanoma" or "Non-Melanoma".
Calibration. We evaluate our models' reliability using Expected Calibration Error (ECE) [8] and the Brier score [9], analyzing how different data combinations affect these metrics to ensure model confidence aligns with actual outcomes.

2.3. Uncertainty Analysis

Our third module estimates neural network uncertainties using Shannon's entropy [10]. This method directly assesses the model's confidence in its predictions while considering the entire distribution of predicted probabilities, making it

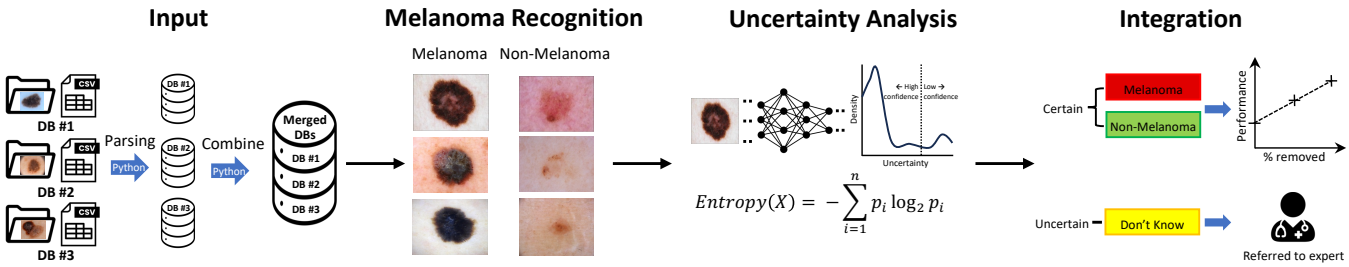

Fig. 1. Our framework includes aggregating diverse datasets, optimizing predictions through 1,296 experiments with 54 combinations of combined datasets and 24 CNNs, filtering uncertain predictions with entropy scores, referring ambiguous cases for human evaluation, and enhancing classification performance.

Fig. 2. We developed a web application that allows users to estimate a melanoma risk score of user-supplied images using web-based inference without upload (edge computing).

sensitive to model indecision. We filter out uncertain predictions to avoid confident false diagnoses from the softmax classifier. The uncertainty scores are defined by:

$$H(X) = -\sum_{i=1}^{n} p_i \log(p_i) \tag{1}$$

Here, p_i is the probability of the i-th class, and $\log(p_i)$ measures the associated uncertainty. By multiplying p_i with $\log(p_i)$ and summing across all classes, we assess the model's uncertainty $H(X)$ within [0, 1]. Lower values indicate higher confidence, while higher values signify greater uncertainty.

2.4. Integration

We classify images into an "Uncertain" class by combining neural network predictions with uncertainty scores and rejecting low-confidence predictions with high entropy. This enhances accuracy by retaining only confident predictions and excluding uncertain cases from evaluation.

2.5. Web Application

We present a web application[2] for melanoma detection that utilizes a MeshNet [11] architecture with volumetric dilated convolutions to reduce parameters for web deployment (see Fig. 2). Users can process images from their devices and detect melanoma within seconds using client-side inference. The model is trained on ten-combined datasets [3, 4, 6, 12, 13, 14, 15, 16, 17, 18].

3. EXPERIMENTAL SETUP

3.1. Datasets and Preprocessing

We utilize publicly available skin disease datasets [3, 4, 6, 12, 13, 14, 15, 16, 17, 18] and evaluate our experiments using provided test sets [4, 6, 15, 18]. When a validation set is not provided, we split the training set into an 80:20 ratio for creating a validation set. These datasets include a "Melanoma" label, and we convert other labels to "Non-Melanoma" for binary classification. To address the high class imbalance leaning towards "Non-Melanoma," we apply oversampling and data augmentation techniques, such as random resized crop, color jitter, zooming, and random flip, to prevent overfitting. We also create combined datasets using up to 10 datasets to train and evaluate various DNNs on classification and calibration, assessing the impact of added data on these metrics.

3.2. Melanoma Recognition

We use various open-source models in `PyTorch`, including ResNet [20], DenseNet [19], VGG [21], and EfficientNet [22] with ImageNet [23] pre-trained weights, resizing images to 224x224 pixels. We freeze the 2D convolutional layers and adjust the final layer for binary classification. Based on our empirical experiments, we train the models for a minimum of 30 epochs, gradually increasing to a maximum of 50 epochs when all ten datasets are combined. The training uses Stochastic Gradient Descent (SGD) [24] with a learning rate of 0.001, momentum of 0.9, a decay factor of 0.1 every 7 epochs, a batch size of 32, and a cross-entropy loss function.

3.3. Rejection by Uncertainty

We perform uncertainty quantification on the predictions from trained models. ECE is defined as:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)| \tag{2}$$

where M is the number of bins, B_m is the set of samples in the m-th bin, $\text{acc}(B_m)$ and $\text{conf}(B_m)$ are the accuracy and

Table 1. Classification performance of our top-performing models in our curated benchmarks.

Test set	Network	Train sets	Before rejection / After rejection						Threshold	Epoch	Param #
			Precision (↑)	Specificity (↑)	Sensitivity (↑)	F-1 (↑)	Accuracy (↑)	AUC-ROC (↑)			
ISIC2017 [4]	DenseNet201 [19]	[A–E,G,I,J]	**83.8%** / 85.3%	91.5% / 91.8%	86.3% / 92.2%	**86.0%** / 88.0%	**90.5%** / 91.9%	94.5% / 95.7%	0.1	46	18.1M
	DenseNet201 [19]	[A–F,H]	82.3% / **92.8%**	90.1% / **96.7%**	**87.2%** / **94.5%**	84.8% / **94.1%**	89.5% / **96.3%**	**95.4%** / **96.3%**	0.2	44	18.1M
	ResNet152 [20]	[A-D,I,J]	82.0% / 87.0%	**91.7%** / 93.9%	77.8% / 89.6%	83.2% / 89.1%	89.0% / 93.1%	93.9% / 94.4%	0.2	42	58.1M
ISIC2018 [6]	ResNet152 [20]	[D]	75.0% / 81.1%	93.1% / 96.4%	**65.5%** / 70.9%	**77.0%** / 82.3%	89.9% / 94.2%	91.0% / 92.7%	0.15	30	58.1M
	ResNet152 [20]	[A,C–E]	**75.2%** / 86.5%	93.4% / 97.9%	63.7% / **71.7%**	76.7% / **86.5%**	**90.0%** / 95.8%	91.1% / 94.5%	0.2	38	58.1M
	ResNet152 [20]	[A–D,I,J]	74.6% / **88.5%**	**94.3%** / **98.5%**	61.4% / 67.7%	75.9% / 85.5%	89.7% / **96.0%**	**91.7%** / **94.6%**	0.2	42	58.1M
7-point criteria [15]	ResNet152 [20]	[A–E,G]	**78.9%** / **88.8%**	**90.1%** / **97.1%**	65.3% / 65.8%	**78.2%** / **84.3%**	**83.8%** / **89.6%**	85.0% / **86.6%**	0.2	42	58.1M
	ResNet152 [20]	[A,C–E]	77.5% / 82.3%	89.1% / 92.6%	64.4% / **67.8%**	77.0% / 81.1%	82.8% / 86.5%	84.6% / 85.4%	0.1	38	58.1M
	ResNet152 [20]	[A–E]	77.4% / 84.6%	88.1% / 94.3%	**67.3%** / 67.0%	77.5% / 82.4%	82.8% / 87.5%	**85.1%** / 86.3%	0.11	40	58.1M
Kaggle [18]	DenseNet201 [19]	[A–E,G,I,J]	91.2% / 94.1%	85.6% / 88.7%	97.3% / 99.3%	90.9% / 93.7%	90.9% / 93.8%	98.1% / 98.5%	0.08	46	18.1M
	DenseNet201 [19]	[A–D,I,J]	**93.2%** / **97.7%**	**89.2%** / **96.1%**	98.0% / 98.0%	**93.2%** / **97.8%**	**93.2%** / **97.8%**	98.1% / 99.1%	0.19	42	18.1M
	ResNet101 [20]	[A–G,I,J]	92.5% / 97.6%	87.0% / 95.7%	**98.7%** / **99.6%**	92.3% / 97.6%	92.3% / 97.6%	**98.5%** / **99.3%**	0.18	48	42.5M

Notes: A: ISIC'16 B: ISIC'17 C: ISIC'18 D: ISIC'19 E: ISIC'20 F: 7-point criteria G: PH2 H: PAD_UFES_20 I: MEDNODE J: Kaggle

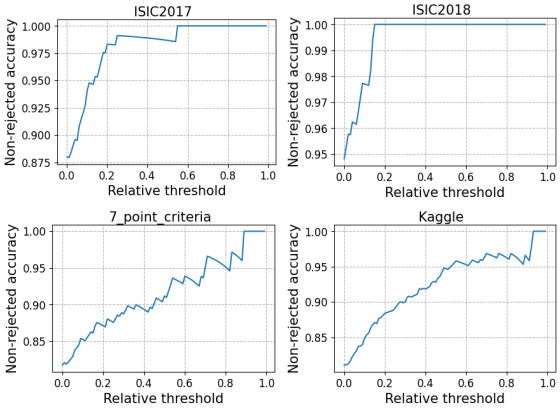

Fig. 3. Accuracy improves with varying rejection thresholds on validation sets, commonly in the range of 0-0.2.

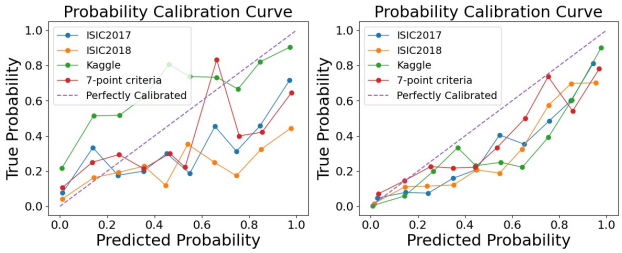

Fig. 4. Calibration curves of ResNet50 as a demonstration. **Left**: Single dataset (ISIC'20). **Right**: Combined datasets (ISIC'16, ISIC'17, ISIC'18, MEDNODE, Kaggle).

the average confidence in the m-th bin, and n is the total number of samples. We use 10 bins for a finer granularity. The Brier Score is defined as:

$$\text{Brier Score} = \frac{1}{n}\sum_{i=1}^{n}(\hat{p}(x_i) - y_i)^2 \tag{3}$$

where $\hat{p}(x_i)$ is the predicted probability for sample x_i, y_i is the true label, and n is the total number of samples. Both metrics range within [0, 1], with 0 indicating optimal calibration. We find that rejecting up to 20% of low-confidence predictions in validation sets improves model accuracy (Fig. 3). We calculate ECE and Brier scores while varying the threshold to minimize their average for each model. After this dynamic rejection, we re-evaluate the remaining samples and compare them to the originals on classification and calibration metrics.

4. RESULTS

Performance Before Rejection. Classification performance is ranked by precision, following the ISIC challenge convention [3]. DenseNet201 and ResNet152 achieve top precisions of 83.8%, 75.2%, 78.9%, and 93.2% across ISIC'17 [4], ISIC'18 [6], 7-point criteria [15], and Kaggle datasets [18] (see Table 1). Table 2 shows that combining datasets enhances performance, with EfficientNetB1 achieving a top AUC-ROC of 90.7% in the SIIM-ISIC challenge [25] with seven-combined datasets. Additionally, models trained on combined datasets align more closely with the ideal calibration curve $y = x$ (see Fig. 4). The time complexity for the models is $O(i \cdot P \cdot H \cdot W)$ where i: number of images, P: number of parameters, and $H \cdot W$: image size. For example, our best model in ISIC'17, which has 600 test samples, has a time complexity of $O(600 \cdot 18.1M \cdot 224 \cdot 224)$ for inference.

Table 2. Leaderboard scores for our top-performing models on the SIIM-ISIC challenge [25].

Network	Trainsets	Private Score (AUC-ROC)
EfficientNetB1 [22]	[A–E,H,I]	**90.7%**
EfficientNetB2 [22]	[A,E]	90.5%
ResNet152 [20]	[A,C–E]	90.3%
ResNet152 [20]	[A–E,G]	90.1%
EfficientNetB1 [22]	[A,C–E]	90.0%
ResNet101 [20]	[A–E,G,F]	90.0%

Notes: A: ISIC'16 B: ISIC'17 C: ISIC'18 D: ISIC'19 E: ISIC'20 F: 7-point criteria G: PH2 H: PAD_UFES_20 I: MEDNODE J: Kaggle

Calibration scores in Table 3 show that models using three or more datasets perform better. Notably, shallower models like VGG and EfficientNet often have better calibration than deeper networks such as ResNet and DenseNet, supporting Guo et al. (2017) [8]. The EfficientNet family exhibits strong

Table 3. Calibration performance of select models based on their classification performance in Table 1.

Testset	Network	Trainsets	Before rejection		After rejection	
			Brier score (↓)	ECE (↓)	Brier score (↓)	ECE (↓)
ISIC2017 [4]	DenseNet201 [19]	A–D,H,I	**0.0832**	**0.0158**	**0.0478**	0.0173
	VGG19 [21]	[A–E,H,I]	0.1353	0.0165	0.1140	0.0178
	ResNet152 [20]	[D]	0.1353	0.0165	0.1043	**0.0113**
ISIC2018 [6]	VGG19 [19]	A–C,I,J	**0.0886**	**0.0047**	0.0563	**0.0044**
	EfficientNetB1 [22]	[A,C–E]	0.1015	0.0246	0.0789	0.0103
	VGG16 [21]	[A–C,I,J]	0.0887	0.0131	**0.0529**	0.0140
7-point criteria [15]	EfficientNetB1 [22]	[A–E,H,I]	0.1455	**0.0288**	0.1218	0.0329
	ResNet152 [20]	[D]	**0.1396**	0.0396	**0.1119**	**0.0321**
	VGG19 [21]	[A–E,G–I]	0.1428	0.0455	0.1274	0.0458
Kaggle [18]	VGG16 [21]	[A,E,G]	0.1577	**0.0187**	0.1357	0.0221
	EfficientNetB6 [22]	[A–C,E,G]	0.1212	0.0219	0.0919	**0.0091**
	ResNet152 [20]	[A–D,I,J]	**0.0543**	0.0238	**0.0200**	0.0122

Notes: A: ISIC'16 B: ISIC'17 C: ISIC'18 D: ISIC'19 E: ISIC'20 F: 7-point criteria G: PH2 H: PAD_UFES_20 I: MEDNODE J: Kaggle

stability and accuracy in melanoma detection, as shown in Table 2 and 3. Its depthwise convolution with one filter per input channel decreases complexity, enhancing generalization and calibration.

Performance After Rejection. ResNet152 improves precision by at least 10% points (p), representing percentage point differences, as shown in Table 1 (after rejection). Other models also demonstrate overall gains. The highest precisions after rejection are 92.8% (9%p ↑), 88.5% (13.3%p ↑), 88.5% (9.6%p ↑), and 94.1% (0.4%p ↑) for benchmarks from ISIC'17, ISIC'18, 7-point criteria, and Kaggle, respectively. ResNet152 is the only model that consistently improves in ECE, indicating that deeper networks benefit more from rejection techniques.

Ablation Studies. We conduct additional studies to evaluate our framework's effectiveness. Our findings show a significant reduction in false positives and negatives. Using EfficientNetB1—demonstrated as reliable in Tables 2 and 3—we successfully prevent 353 misdiagnoses (see Fig. 5). This includes 52% false positives (non-melanoma diagnosed as melanoma) and 48% false negatives (melanoma diagnosed as non-melanoma). We reduce 66.5% of false negatives on average across test sets. Notably, 81% of false negatives in the Kaggle [18] test set are reduced.

5. CONCLUSIONS

We propose a melanoma framework integrating unified training and testing with uncertainty-based rejection to enhance classification and calibration while reducing misdiagnoses. Future plans include collecting real-time diagnosis data from dermatologists and using Generative Adversarial Networks (GAN) for iterative retraining. We also aim to extend the framework to other medical image data, like mammograms for breast cancer detection.

6. ACKNOWLEDGMENTS

This material is based, in part, upon work supported by the Defense Logistics Agency (DLA) and the Advanced Research Projects Agency for Health (ARPA-H) under Contract Number SP4701-23-C-0073. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of DLA or ARPA-H. This research was funded in part by the Massachusetts Life Sciences Center through grant Bits-to-Bytes 34428.

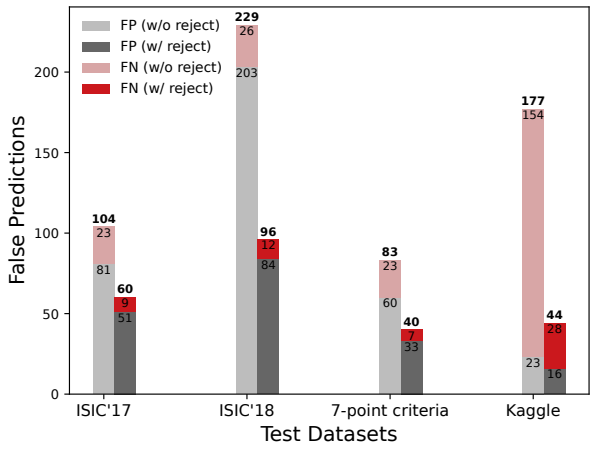

Fig. 5. A plot comparing false diagnoses before and after applying uncertainty-based rejection across benchmarks.

7. COMPLIANCE WITH ETHICAL STANDARDS

All datasets used in this study were anonymized and publicly available, for which no ethical approval was required.

8. REFERENCES

[1] Dirk Schadendorf, Alexander CJ Van Akkooi, Carola Berking, Klaus G Griewank, Ralf Gutzmer, Axel Hauschild, Andreas Stang, Alexander Roesch, and Selma Ugurel, "Melanoma," *The Lancet*, vol. 392, no. 10151, pp. 971–984, 2018.

[2] Kalyanakumar Jayapriya and Israel Jeena Jacob, "Hybrid fully convolutional networks-based skin lesion segmentation and melanoma detection using deep feature," *International Journal of Imaging Systems and Technology*, vol. 30, no. 2, pp. 348–357, 2020.

[3] David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern, "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 preprint arXiv:1605.01397*, 2016.

[4] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, 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 *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE, 2018, pp. 168–172.

[5] Douglas de A Rodrigues, Roberto F Ivo, Suresh Chandra Satapathy, Shuihua Wang, Jude Hemanth, and Pedro P Reboucas Filho, "A new approach for classification skin lesion based on transfer learning, deep learning, and iot system," *Pattern Recognition Letters*, vol. 136, pp. 8–15, 2020.

[6] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, 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.

[7] Farnaz Haghshenas, Adam Krzyżak, and Stanislaw Osowski, "Comparative study of deep learning models in melanoma detection," in *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*. Springer, 2024, pp. 121–131.

[8] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger, "On calibration of modern neural networks," in *International conference on machine learning*. PMLR, 2017, pp. 1321–1330.

[9] Glenn W Brier, "Verification of forecasts expressed in terms of probability," *Monthly weather review*, vol. 78, no. 1, pp. 1–3, 1950.

[10] Lukasz Rudnicki, "Shannon entropy as a measure of uncertainty in positions and momenta," *Journal of Russian Laser Research*, vol. 32, pp. 393–399, 2011.

[11] Alex Fedorov, Jeremy Johnson, Eswar Damaraju, Alexei Ozerin, Vince Calhoun, and Sergey Plis, "End-to-end learning of brain tissue segmentation from imperfect labeling," in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 3785–3792.

[12] Marc Combalia, Noel CF Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C Halpern, Susana Puig, et al., "Bcn20000: Dermoscopic lesions in the wild," *arXiv preprint arXiv:1908.02288*, 2019.

[13] Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, et al., "A patient-centric dataset of images and metadata for identifying melanomas using clinical context," *Scientific data*, vol. 8, no. 1, pp. 34, 2021.

[14] Teresa Mendonça, Pedro M Ferreira, Jorge S Marques, André RS Marcal, and Jorge Rozeira, "Ph 2-a dermoscopic image database for research and benchmarking," in *2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, 2013, pp. 5437–5440.

[15] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh, "Seven-point checklist and skin lesion classification using multitask multimodal neural nets," *IEEE journal of biomedical and health informatics*, vol. 23, no. 2, pp. 538–546, 2018.

[16] Andre GC Pacheco, Gustavo R Lima, Amanda S Salomao, Breno Krohling, Igor P Biral, Gabriel G de Angelo, Fábio CR Alves Jr, José GM Esgario, Alana C Simora, Pedro BC Castro, et al., "Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones," *Data in brief*, vol. 32, pp. 106221, 2020.

[17] Ioannis Giotis, Nynke Molders, Sander Land, Michael Biehl, Marcel F Jonkman, and Nicolai Petkov, "Mednode: A computer-assisted melanoma diagnosis system using non-dermoscopic images," *Expert systems with applications*, vol. 42, no. 19, pp. 6578–6585, 2015.

[18] Claudio Fanconi, "Skin cancer: Malignant vs. benign," `https://www.kaggle.com/datasets/fanconic/skin-cancer-malignant-vs-benign`.

[19] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[21] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[22] Mingxing Tan and Quoc Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[24] Léon Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*. Springer, 2010, pp. 177–186.

[25] Anna Zawacki, "Siim-isic melanoma classification," `https://kaggle.com/competitions/siim-isic-melanoma-classification`, 2020.