2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)
April 4-7, 2018, Washington, D.C., USA

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)

Noel C. F. Codella[1†], *David Gutman*[2†], *M. Emre Celebi*[3], *Brian Helba*[4], *Michael A. Marchetti*[5], *Stephen W. Dusza*[5], *Aadi Kalloo*[5], *Konstantinos Liopyris*[5], *Nabin Mishra*[6], *Harald Kittler*[7], *Allan Halpern*[5‡]

[1] IBM T. J. Watson Research Center, Yorktown Heights, NY, USA
[2] Emory University, Atlanta, GA, USA
[3] University of Central Arkansas, Conway, AR, USA
[4] Kitware, Clifton Park, NY, USA
[5] Memorial Sloan-Kettering Cancer Center, New York, NY, USA
[6] Missouri University of Science and Technology, Rolla, MO USA
[7] Medical University of Vienna, Vienna, Austria

ABSTRACT

This article describes the design, implementation, and results of the latest installment of the dermoscopic image analysis benchmark challenge. The goal is to support research and development of algorithms for automated diagnosis of melanoma, the most lethal skin cancer. The challenge was divided into 3 tasks: lesion segmentation, feature detection, and disease classification. Participation involved 593 registrations, 81 pre-submissions, 46 finalized submissions (including a 4-page manuscript), and approximately 50 attendees, making this the largest standardized and comparative study in this field to date. While the official challenge duration and ranking of participants has concluded, the dataset snapshots remain available for further research and development.

Index Terms— Dermatology, dermoscopy, melanoma, skin cancer, challenge, deep learning, dataset

1. INTRODUCTION

The most prevalent form of cancer in the United States is skin cancer, with 5 million cases occurring annually [1, 2, 3]. Melanoma, the most dangerous type, leads to over 9,000 deaths a year [2, 3]. Even though most melanomas are first discovered by patients [4], the diagnostic accuracy of unaided expert visual inspection is only about 60% [5].

Dermoscopy is a recent technique of visual inspection that both magnifies the skin and eliminates surface reflection. Research has shown that with proper training, diagnostic accuracy with dermoscopy is 75%-84% [5, 6, 7]. In an attempt to improve the scalability of dermoscopic expertise, procedural algorithms, such as "3-point checklist," "ABCD rule," "Menzies method," and "7-point checklist," were developed [6, 8]. However, many clinicians forgo these methods in favor of relying on personal experience, as well as the "ugly duckling" sign (outliers on patient) [9].

Recent reports have called attention to a growing shortage of dermatologists per capita [10]. This has increased interest in techniques for automated assessment of dermoscopic images [11, 12, 13, 14]. However, most studies have used isolated silos of data for analysis that are not available to the broader research community. While an earlier effort to create a public archive of images was made [14, 15], the dataset was too small (200 images) to fully represent scope of the task.

The International Skin Imaging Collaboration (ISIC) has begun to aggregate a large-scale publicly accessible dataset of dermoscopy images. Currently, the dataset houses more than 20,000 images from leading clinical centers internationally, acquired from a variety of devices used at each center. The ISIC dataset was the foundation for the first public benchmark challenge on dermoscopic image analysis in 2016 [16, 17]. The goal of the challenge was to provide a fixed dataset snapshot to support development of automated melanoma diagnosis algorithms across 3 tasks of lesion analysis: segmentation, dermoscopic feature detection, and classification.

In 2017, ISIC hosted the second instance of this challenge, featuring an expanded dataset. In the following sections, the datasets, tasks, metrics, participation, and the results of this challenge are described.

[†] The first two authors contributed equally to this work.
[‡] Corresponding author.

978-1-5386-3636-7/18/$31.00

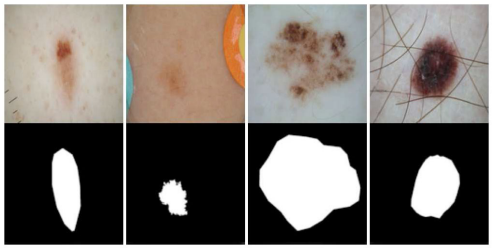

Fig. 1. Images from "Part 1: Lesion Segmentation." *Top:* Original images. *Bottom:* Segmentation masks.

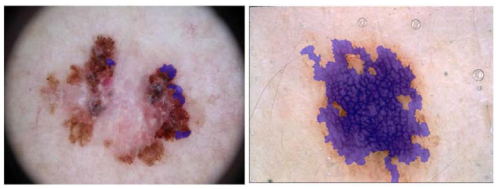

Fig. 2. Images from "Part 2: Dermoscopic Feature Classification". Ground truth labels highlighted in purple. *Left:* Streaks. *Right:* Pigment Network.

2. DATASET DESCRIPTIONS & TASKS

The 2017 challenge consisted of 3 tasks: lesion segmentation, dermoscopic feature detection, and disease classification. For each, data consisted of images and corresponding ground truth annotations, split into training (n=2000), validation (n=150), and holdout test (n=600) datasets. Predictions could be submitted on validation and test datasets. The validation submissions provided instantaneous feedback in the form of performance evaluations, as well as ranking in comparison to other participants. Test submissions only provided feedback after the submission deadline. The training, validation, and test datasets continue to be available for download from the following address: http://challenge2017.isic-archive.com/

Part 1: Lesion Segmentation Task: Participants were asked to submit automated predictions of lesion segmentations from dermoscopic images in the form of binary masks. Lesion segmentation training data included the original image, paired with the expert manual tracing of the lesion boundaries also in the form of a binary mask, where pixel values of 255 were considered inside the area of the lesion, and pixel values of 0 were outside (Fig. 1).

Part 2: Dermoscopic Feature Classification Task: Participants were asked to automatically detect the following four clinically defined dermoscopic features: "network," "negative network," "streaks," and "milia-like cysts," [18, 19]. Pattern detection involved both localization and classification (Fig. 2). To reduce the variability and dimensionality of spatial feature annotations, the lesion images were subdivided into superpixels using the SLIC algorithm [20]. Lesion dermoscopic feature data included the original lesion image and a corresponding set of superpixel masks, paired with superpixel-wise expert annotations for the presence or absence of the dermoscopic features. Validation and test sets included images and superpixels without annotation.

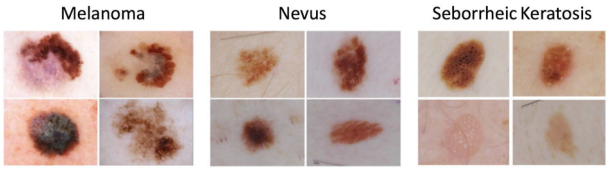

Fig. 3. Example images from "Part 3: Disease Classification." Ground truth labels written above.

Part 3: Disease Classification Task: Participants were asked to classify images as belonging to one of 3 categories (Fig. 3), including "melanoma" (374 training, 30 validation, 117 test), "seborrheic keratosis" (254, 42, and 90), and "benign nevi" (1372, 78, 393), with classification scores normalized between 0.0 to 1.0 for each category (and 0.5 as binary decision threshold). Lesion classification data included the original image paired with the gold standard diagnosis, as well as approximate age (5 year intervals) and gender when available.

3. EVALUATION METRICS

Details of evaluation metrics have been previously described [16, 17]. For classification decisions, any confidence above 0.5 was considered positive for a category. For segmentation tasks, pixel values above 128 were considered positive, and pixel values below were considered negative.

For evaluation of classification decisions, the area under curve (AUC) measurement from the receiver operating characteristic (ROC) curve was computed [16].

Additionally, for classification of melanoma, specificity was measured on the operating curve where sensitivity was equal to 82%, 89%, and 95%, corresponding to dermatologist classification and management performance levels, and theoretically desired sensitivity levels, respectively [17].

Segmentation submissions were compared using the Jaccard Index, Dice coefficient, and pixel-wise accuracy [16]. Participant ranking used Jaccard.

4. RESULTS

The 2017 challenge saw 593 registrations, 81 pre-submissions, and 46 finalized submissions (including a 4 page arXiv paper with each). The associated workshop at ISBI 2017 saw approximately 50 attendees. To date, this has been the largest standardized and comparative study in this field, accounting for the size of the dataset, the number of algorithms evaluated, and the number of participants. In the following, the results for each challenge part are investigated.

Part 1: Lesion Segmentation Task: 21 sets of prediction scores on the final test set were submitted for the segmenta-

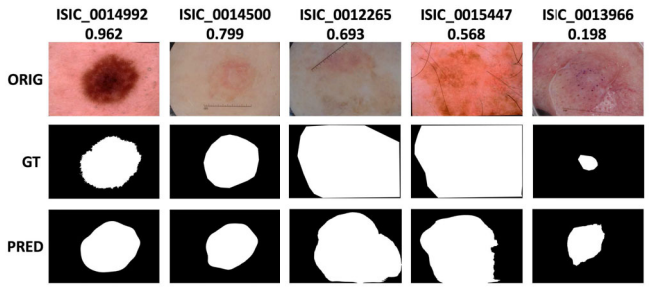

Fig. 4. Part 1 example segmentations from top ranked participant submission. *Top Row:* Original images. *Middle Row:* Ground truth segmentations. *Bottom Row:* Participant predictions. ISIC identifiers and Jaccard Index values are listed at each column head.

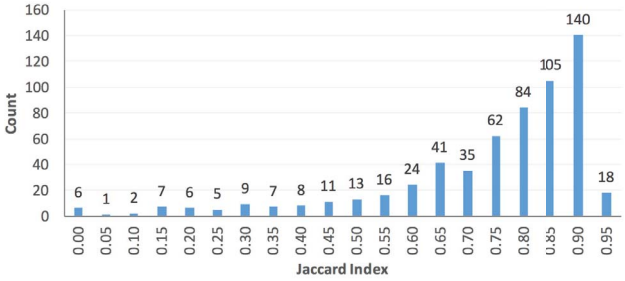

Fig. 5. Histogram of Jaccard Index values for individual images from top segmentation task participant submission.

tion task, and 39 were submitted to the validation set. The top ranked participant achieved an average Jaccard Index of 0.765, accuracy of 93.4%, and Dice coefficient of 0.849, using a variation of a fully convolutional network ensemble (a deep learning approach) [21]. Example segmentations are shown in Fig. 4, and a histogram of individual image Jaccard Index measurements is shown in Fig. 5. Subjectively assessing the quality of the segmentations, one can observe that segmentations of Jaccard Index 0.8 or above tend to appear visually "correct." This observation is consistent with prior reports that measured an inter-observer agreement of 0.786 on a subset of 100 images from the ISIC 2016 Challenge [13]. When Jaccard falls to 0.7 or below, the "correctness" of the segmentation can be debated. 156 out of 600 images (26%) fell at or below a Jaccard of 0.7. 91 images (15.2%) fell at or below Jaccard of 0.6. This suggests a failure rate of 15% to 26%, which is higher than the pixel-wise failure rate of 6.6%.

Part 2: Dermoscopic Feature Classification Task: For the second year in a row, dermoscopic feature classification has received far less participation than other tasks. Only 3 submissions [22, 23] on the test set were received from 2 parties. Whether this is due to the technical framing of the task (how well it maps to existing frameworks), or the perceived importance of the task, is a matter of current investigation.

Regardless, performance levels of those submissions that

Method / Rank	AVG	Network	Neg. Network	Streaks	Milia-Like Cyst
[22] / 1	**0.895**	**0.945**	**0.869**	**0.960**	0.807
[23] / 2	0.833	0.835	0.762	0.896	**0.838**
[23] / 3	0.832	0.828	0.762	0.900	0.837

Table 1. Part 2: Dermoscopic Feature Classification AUC Measurements. AVG = Average across all categories.

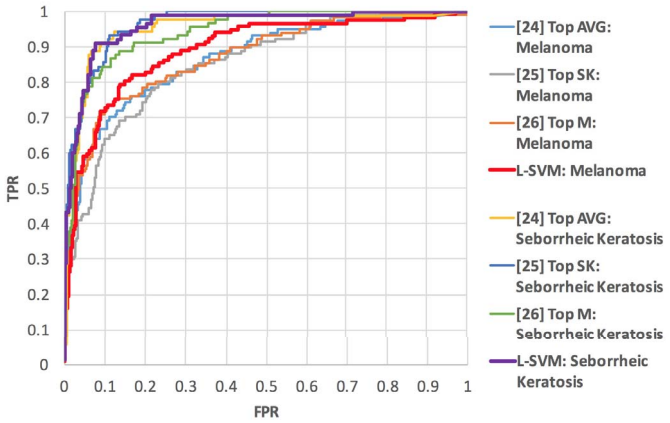

Fig. 6. ROC curves for top 3 submissions to "Part 3: Disease Classification", as well as linear SVM fusion.

were received demonstrated that localization of dermoscopic features is a tractable task for computer vision approaches. Top performance levels are shown in Table 1. AUC was above 0.75 ubiquitously, with an average close to 0.9.

Part 3: Disease Classification Task: The disease classification task received 23 final test set submissions, and 39 validation set submissions. Performance characteristics of the average (AVG) classification winner [24], seborrheic keratosis (SK) classification winner [25], and melanoma (M) classification winner [26], respectively, are shown in Table 2, as well as 3 fusion strategies [17]: score averaging (AVGSC), linear SVM (L-SVM), and non-linear SVM (NL-SVM) using a histogram intersection kernel. Fusion strategies utilize all submissions on the final test set, and are carried out via 3-fold cross-validation. SVM input feature vectors included all disease category predictions. Both SVM methods used probabilistic SVM score normalization, producing an output confidence between 0.0 and 1.0 (with 0.5 as binary threshold), correlating with the probability of disease on a balanced dataset [13]. ROC curves for the 3 submissions and the best fusion strategy (Linear SVM) are shown in Fig. 6.

The 5 major trends observed involve the following: 1) All top submissions implemented various ensembles of deep learning networks. All used additional data sources to train, either from ISIC [24, 26], in-house annotations [25], or external sources [26]. 2) Classification of seborrheic keratosis appears to be an easier task in this dataset, compared to

Method	AVG-AUC	M-AUC	SK-AUC	M-SP82	M-SP89	M-SP95	M-SENS	M-SPEC	SK-SENS	SK-SPEC
[24] Top AVG	0.911	0.868	0.953	0.729	0.588	0.366	0.735	0.851	0.978	0.773
[25] Top SK	0.910	0.856	0.965	0.727	0.555	0.404	0.103	0.998	0.178	0.998
[26] Top M	0.908	0.874	0.943	0.747	0.590	0.395	0.547	0.950	0.356	0.990
AVGSC	0.913	0.872	0.954	0.778	0.605	0.435	0.214	0.988	0.600	0.975
L-SVM	0.926	0.892	0.960	0.834	0.692	0.571	0.718	0.901	0.878	0.931
NL-SVM	0.904	0.853	0.955	0.801	0.449	0.168	0.675	0.909	0.889	0.928

Table 2. Part 3: Disease classification evaluation metrics for top 3 participants, followed by average score, linear SVM, and non-linear SVM 3-fold cross-validation fusion methods. *Key:* M = Melanoma. SK = Seborrheic Keratosis. AVG = Average between two classes. AUC = Area Under Curve. SP82/89/95 = Specificity mearured at 82/89/95% sensitivity. L-SVM = Linear SVM. NL-SVM = Non-Linear SVM. AVGSC = Average Score.

melanoma classification. This may reflect aspects of the disease, or bias in the dataset. The best performance came from the team that added additional weakly labelled pattern annotations to their training data [25]. 3) The top average performer was not the best in any single classification category. 4) The most complex fusion approach (NL-SVM) led to a decrease in performance, whereas simpler methods led to overall improvements in performance, consistent with previous findings [17]. This challenge is the second benchmark to demonstrate that a collaborative among all participants outperforms any single method alone. 5) Not all thresholds balanced sensitivity and specificity. Probabilistic score normalization in fusions is effective at balancing sensitivity and specificity [13, 17].

5. DISCUSSION & CONCLUSION

The International Skin Imaging Collaboration (ISIC) archive was used to host the second public challenge on Skin Lesion Analysis Toward Melanoma Detection at the International Symposium on Biomedical Imaging (ISBI) 2017. The challenge was divided into 3 tasks: segmentation, feature detection (4 classes), and disease classification (3 classes). 2000 images were available for training, 150 for validation, and 600 for testing. The challenge involved 593 registrations, 81 pre-submissions, and 46 finalized submissions, making it the largest standardized and comparative study in this field.

Analysis of segmentation results suggest that the average Jaccard Index may not accurately reflect the number of images where automated segmentation falls outside inter-observer variability. Future challenges may adjust the evaluation metric based on this observation. For example, a binary error may be more appropriate (segmentation failure or success), computed by either using multiple segmentations per image to determine a segmentation difference tolerance threshold, or by choosing a fixed threshold as an estimator based on prior studies.

Poor participation was noted in dermoscopic feature detection; however, submitted systems achieved reasonable performance. Future challenges may adjust the technical format of the task to more closely align with existing image detection benchmarks to better facilitate ease of participation. For example, the output can be formatted as a segmentation or bounding-box detection task.

Analysis of the classification task demonstrates that ensembles of deep learning approaches and additional data led to the highest performance. In addition, collaborative fusions of all participant systems outperformed any single system alone. With the exception of [25], submitted methods generate little human interpretable evidence of disease diagnosis. Future work or challenges may give more focus to this need for proper integration into clinical workflows.

Limitations of this study included dataset bias (not all diseases, ages, devices, or ethnicities were represented equally across categories), and incomplete dermoscopic feature annotations. Reliance on single evaluation metrics rather than combinations may also be a limitation [27]. Future challenges will attempt to address these issues in conjunction with the community.

6. REFERENCES

[1] Rogers HW, Weinstock MA, Feldman SR, Coldiron BM.: "Incidence estimate of nonmelanoma skin cancer (keratinocyte carcinomas) in the US population, 2012" JAMA Dermatol vol. 151, no. 10, pp. 1081-1086. 2015.

[2] "Cancer Facts & Figures 2017". American Cancer Society, 2017. Available: https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2017.html

[3] Siegel, R.L., Miller, K.D., and Jemal, A.: "Cancer statistics, 2017," CA: A Cancer Journal for Clinicians, vol. 67, no. 1, pp. 7-30. 2017.

[4] Brady, M.S., Oliveria, S.A., Christos, P.J., Berwick, M., Coit, D.G., Katz, J., Halpern, A.C.: "Patterns of detection in patients with cutaneous melanoma." Cancer. vol. 89, no. 2, pp. 342-7. 2000.

[5] Kittler, H., Pehamberger, H., Wolff, K., Binder, M.: "Diagnostic accuracy of dermoscopy". The Lancet Oncology. vol. 3, no. 3, pp. 159-165. 2002.

[6] Carli, P., et al.: "Pattern analysis, not simplified algorithms, is the most reliable method for teaching dermoscopy for melanoma diagnosis to residents in dermatology". Br J Dermatol. vol. 148, no. 5, pp. 981-4. 2003.

[7] Vestergaard, M.E., Macaskill, P., Holt, P.E., et al.: "Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting." Br J Dermatol. vol. 159, pp. 669-676. 2008.

[8] Argenziano, G. et al.: "Dermoscopy of pigmented skin lesions: Results of a consensus meeting via the Internet" J. American Academy of Dermatology. vol. 48, no. 5, 2003.

[9] Gachon, J., et. al.:"First Prospective Study of the Recognition Process of Melanoma in Dermatological Practice". Arch Dermatol. vol. 141, no. 4, pp. 434-438, 2005.

[10] Kimball, A.B., Resneck, J.S. Jr.: "The US dermatology workforce: a specialty remains in shortage." J Am Acad Dermatol. vol. 59, no. 5, pp. 741-5. 2008.

[11] Mishra, N.K., Celebi, M.E.: "An Overview of Melanoma Detection in Dermoscopy Images Using Image Processing and Machine Learning" arxiv.org: 1601.07843. Available: http://arxiv.org/abs/1601.07843

[12] Ali, A.A., Deserno, T.M.: "A Systematic Review of Automated Melanoma Detection in Dermatoscopic Images and its Ground Truth Data" Proc. of SPIE Vol. 8318 83181I-1

[13] Codella NCF, Nguyen B, Pankanti S, Gutman D, Helba B, Halpern A, Smith JR. "Deep learning ensembles for melanoma recognition in dermoscopy images" IBM Journal of Research and Development, vol. 61, no. 4/5, 2017. Available: https://arxiv.org/pdf/1610.04662.pdf

[14] Barata, C., Ruela, M., et al.: "Two Systems for the Detection of Melanomas in Dermoscopy Images using Texture and Color Features". IEEE Systems Journal, vol. 8, no. 3, pp. 965-979, 2014.

[15] Mendonca, T., Ferreira, P.M., Marques, J.S., Marcal, A.R., Rozeira, J.: "PH2 - a dermoscopic image database for research and benchmarking". Conf Proc IEEE Eng Med Biol Soc. pp. 5437-40, 2013.

[16] Gutman D, Codella N, Celebi E, Helba B, Marchetti M, Mishra N, Halpern A. "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)". eprint arXiv:1605.01397 [cs.CV]. 2016. Available: https://arxiv.org/abs/1605.01397

[17] Marchetti M, et al. "Results of the 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging challenge: Comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images". Journal of the American Academy of Dermatology, 2017. In Press.

[18] Braun, R.P., Rabinovitz, H.S., Oliviero, M., Kopf, A.W., Saurat, J.H.:"Dermoscopy of pigmented skin lesions.". J Am Acad Dermatol. vol. 52, no. 1, pp. 109-21. 2005.

[19] Rezze, G.G., Soares de S, B.C., Neves, R.I.: "Dermoscopy: the pattern analysis". An Bras Dermatol., vol. 3, pp. 261-8. 2006.

[20] Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Susstrun,S.: "SLIC Superpixels", EPFL Technical Report 149300, June 2010.

[21] Yuan Y, Chao M, Lo YC. "Automatic skin lesion segmentation with fully convolutional-deconvolutional networks". International Skin Imaging Collaboration (ISIC) 2017 Challenge at the International Symposium on Biomedical Imaging (ISBI). Available: https://arxiv.org/pdf/1703.05165.pdf

[22] Kawahara J, Hamarneh G. "Fully Convolutional Networks to Detect Clinical Dermoscopic Features". International Skin Imaging Collaboration (ISIC) 2017 Challenge at the International Symposium on Biomedical Imaging (ISBI). Available: https://arxiv.org/abs/1703.04559

[23] Li Y, Shen L. "Skin Lesion Analysis Towards Melanoma Detection Using Deep Learning Network". International Skin Imaging Collaboration (ISIC) 2017 Challenge at the International Symposium on Biomedical Imaging (ISBI). Available: https://arxiv.org/abs/1703.00577

[24] Matsunaga K, Hamada A, Minagawa A, Koga H. "Image Classification of Melanoma, Nevus and Seborrheic Keratosis by Deep Neural Network Ensemble". International Skin Imaging Collaboration (ISIC) 2017 Challenge at the International Symposium on Biomedical Imaging (ISBI). Available: https://arxiv.org/abs/1703.03108

[25] Daz IG. "Incorporating the Knowledge of Dermatologists to Convolutional Neural Networks for the Diagnosis of Skin Lesions". International Skin Imaging Collaboration (ISIC) 2017 Challenge at the International Symposium on Biomedical Imaging (ISBI). Available: https://arxiv.org/abs/1703.01976

[26] Menegola A, Tavares J, Fornaciali M, Li LT, Avila S, Valle E. "RECOD Titans at ISIC Challenge 2017". International Skin Imaging Collaboration (ISIC) 2017 Challenge at the International Symposium on Biomedical Imaging (ISBI). Available: https://arxiv.org/pdf/1703.04819.pdf

[27] Fishbaugh, J, et al. "Data-Driven Rank Aggregation with Application to Grand Challenges." International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 754-762. 2017.