2019 IEEE International Conference on Big Data (Big Data)

Towards Explainable Melanoma Diagnosis: Prediction of Clinical Indicators Using Semi-supervised and Multi-task Learning

Seiya Murabayashi and Hitoshi Iyatomi
Applied Informatics, Graduate School of Science and Engineering, Hosei University, Tokyo, Japan
Email: {seiya.murabayashi.9j@stu., iyatomi@}hosei.ac.jp

***Abstract*—Although image-based melanoma diagnosis has achieved a sufficient level of numerical accuracy, providing objective evidence is essential to enhance the explainability and reliability of this approach. The collection of label information based on quantitative clinical indicators is very expensive, meaning that the amount of labeled data available is limited. In this paper, we propose an effective method for predicting explainable melanoma indicators defined by a 7-point checklist in a situation where only a limited number of labeled data are available. Our proposal effectively utilizes virtual adversarial training as a semi-supervised learning framework with multi-task learning. This approach gives favorable performance for only a very limited number of expensive labeled data. The proposed method improves the final accuracy of melanoma diagnosis calculated based on these predicted indices by 7.5% (making it equivalent to expert dermatologists), based on 9,124 unlabeled images with diagnosis information added to the 226 base labeled training images.**

***Index Terms*—melanoma, computer-aided diagnosis (CAD), deep learning, multi-task learning, 7-point checklist**

I. Introduction

Advanced malignant melanoma is the most aggressive form of skin cancer. According to U.S. statistics, about 96,480 people are newly diagnosed as melanoma patients, and 7,230 people are currently expected to die from melanoma [1]. Correct diagnosis, especially in the early stages, is therefore essential for the reduction of melanoma-related deaths. However, diagnosis of melanoma is often difficult and subjective. According to [1], the accuracy of expert dermatologists in diagnosing melanoma is still estimated to be about 80% with the use of a specially designed magnifying scope called a dermoscope. To overcome these problems, automated analysis procedures for pigmented skin lesions including melanomas have been proposed since the beginning of this century [2–7]. The first three are pre-date deep learning era, and usually require appropriate tumor area segmentation [8], [9], color calibration [10] etc. as preprocessing steps, the remainder are based on deep learning techniques, and can omit these preprocessing stages. Although both traditional and deep learning methods have achieved high levels of identification accuracy that are comparable to those of specialists, these systems only provide diagnosis results without specific evidence.

Hence, there is still room for improvement in terms of reliability, due to their blackbox characteristics. There have been a limited number of studies providing evidence for diagnosis based on the extraction of typical structures [11–14], heat-map representation [15], [16] and quantification of the score defined in the diagnostic guidelines [17]. However, the expressions obtained by the methods in these first two categories are subjective. Providing quantitative diagnosis evidence in accordance with diagnostic guidelines such as a 7-point checklist [18] reduces the problem known as blackboxing in AI-based diagnosis, and greatly improves the reliability of the system. In addition, the collection of labeled training data for this task is very expensive, since dermatologists need to determine each indicator. To make matters worse, the definition of these indicators is also subjective, meaning that variations in the training labels assigned to the items cannot be ignored [17]. It is therefore necessary to ask several experienced dermatologists to do this work, which is both expensive and difficult.

In the field of machine learning, on the other hand, a great deal with research has been done to improve the generalization based on a small amount of training data. Researchers started with constraints from the classic regularizers such as the L_1- and L_2- norms. Furthermore, new and efficient forms of regularization and data augmentation techniques are now being routinely used. Semi-supervised learning is the one of more effective ways to improve the generality of a system. It uses a large amount of unlabeled data for training, in addition to a small amount of labeled data. A systematic review is found in [19]. The methodologies of semi-supervised learning can be categorized into self-training (iterative learning), graph-based models, and vector-based techniques. Following the spread of deep learning techniques, the vector-based techniques have become particularly advanced [20–24]. Most recent vector-based methods use deep networks to obtain a efficient low-dimensional representation of a given task. Metric learning is a technique based on the idea that data with similar properties in a certain context (i.e. data that belong to the same or a similar class in real space) should be close in low-dimensional space [20], [21]. It should be noted that metric learning can be applied to unsupervised or semi-supervised training. In addition, unsupervised or semi-supervised learning techniques based on generative models have also been proposed. Chen et al. [22] improved the robustness of their classifier by

[1] Key statistics for melanoma skin cancer, American Cancer Society, accessed Sep.18, 2019. https://www.cancer.org/cancer/melanoma-skin-cancer/about/key-statistics.html

978-1-7281-0858-2/19/$31.00 © 2019 IEEE
978-1-7281-0858-2/19/$31.00 ©2019 IEEE

generating pseudo-data that were similar to specific data in the metric space, but which were easily misclassified into different classes. This can be considered an effective combination of metric learning and generative modelling. Although this technique performed well, a large amount of labeled training data was required to obtain an accurate probability distribution for the input data space. For skin lesion classification, Kitada et al. [23] used a semi-supervised learning framework to tackle the problem of imbalanced data by learning features from unlabeled lesion images, and observed an increase in discrimination performance.

Virtual adversarial training (VAT) [24] introduces well-designed constraints to build smooth discrimination boundaries, based on the hypothesis that a robust classification model should have a smooth boundary and that it can be applied to semi-supervised learning. VAT has achieved excellent results in multiple tasks for a very small number of available labeled datasets. We expect these features of VAT to be very effective for our problem.

As mentioned above, the availability of labeled training data for the diagnostic indicators that we aim to predict is very limited. However, training data with diagnostic labels (e.g. melanoma or nevus tumors) has recently become widely available, following a competition related to automated melanoma diagnosis technology [2]. Multi-task learning [25] is a strategy that solves multiple tasks that are related to each other, and can improve the prediction performance by learning common feature representations. Since a diagnosis is made by summing up the score of diagnostic indicators in clinical practice, the simultaneous training of diagnostic indicators and the use of a large amount of diagnosis information for the related task in a multi-task learning framework is expected to improve the performance.

In this study, we propose an effective method for predicting explainable melanoma indicators, as defined by the 7-point checklist, using a combination of VAT (semi-supervised learning) and multi-task learning. Our method aims to achieve excellent predictive performance by using a very limited amount of data with original labels, as well as a large amount of data with “different” labels. The purpose of this study is not to pursue the level of accuracy of automatic diagnosis often seen elsewhere, but to achieve a quantitative presentation of the basis for diagnosis, in order to improve the explainability and reliability of the results of blackbox systems.

II. Preparation

A. 7-point checklist

First, we introduce a 7-point checklist [18], that is a well-known diagnostic method for melanoma. This checklist requires the identification of seven dermoscopic structures, as shown in Table I. The score for a skin lesion is determined as the weighted sum of the structures present in it. Using this checklist, the total score (TS) is calculated as

$$\mathrm{TS} = (\#.\mathrm{major} \times 2) + (\#.\mathrm{minor}), \tag{1}$$

TABLE I: Dermoscopic structures and associated scores defined in the 7-point checklist

Major criteria	Score
S1. Atypical pigment network	$\times$ 2
S2. Blue-whitish veil	$\times$ 2
S3. Atypical vascular pattern	$\times$ 2
Minor criteria	**Score**
S4. Irregular streaks	$\times$ 1
S5. Irregular pigmentation	$\times$ 1
S6. Irregular dots / globules	$\times$ 1
S7. Regression structures	$\times$ 1

where #. major and #. minor are the number of major and minor dermoscopic structures, present in the image, respectively, and the TS is therefore in the range zero to 10. If TS $>= 3$, then the lesion is considered to be malignant. According to [1], the average sensitivity and specificity shown by 40 expert dermatologists based on this criterion were 75.0% and 76.2%, respectively.

B. Datasets and definition of gold standard

We used two datasets in this study. Dataset A is an indicator-labeled dataset consisting of 226 dermoscopy images with gold standard scores, as defined in the 7-point checklist, and diagnosis information labeling them as melanoma or nevus. These images are selected from [3], and contain 104 melanomas and 122 nevi. Dataset B is a diagnosis-labeled dataset containing 9,124 different dermoscopy images, consisting of 1,237 melanomas and 7,887 nevi drawn from the ISIC2018 dataset [26], and external data also taken from [3]. This diagnosis-labeled dataset does not contain scores from the 7-point checklist. Confirmed diagnostic information is available for all cases in both datasets (benign pigmented skin lesions or malignant melanoma), and we leverage this diagnosis information effectively. Note that obtaining the final diagnosis information is much easier than obtaining these indicators. As mentioned earlier, the recognition of dermoscopic structures is highly subjective, even for expert dermatologists, and as a result no gold standard has been established. In this study, the dermoscopic structures defined by the 7-point checklist were determined by four experienced dermatologists, and the average was used, so that the range for each indicator was $[0, 1]$. Accordingly, each of the labeled training dermoscopy images had an eight-dimensional binary vector (corresponding to each item in the 7-point checklist and a diagnosis).

C. Virtual adversarial training

VAT is an effective training method based on the hypothesis that a robust classification model should have a smooth boundary. Local distribution smoothness (LDS) was introduced into VAT as a specially designed regularization metric, and is a negative measure of the local smoothness of the conditional label distribution around the input against local perturbation.

[2] ISIC 2018: Skin Lesion Analysis Towards Melanoma Detection https://challenge2018.isic-archive.com/

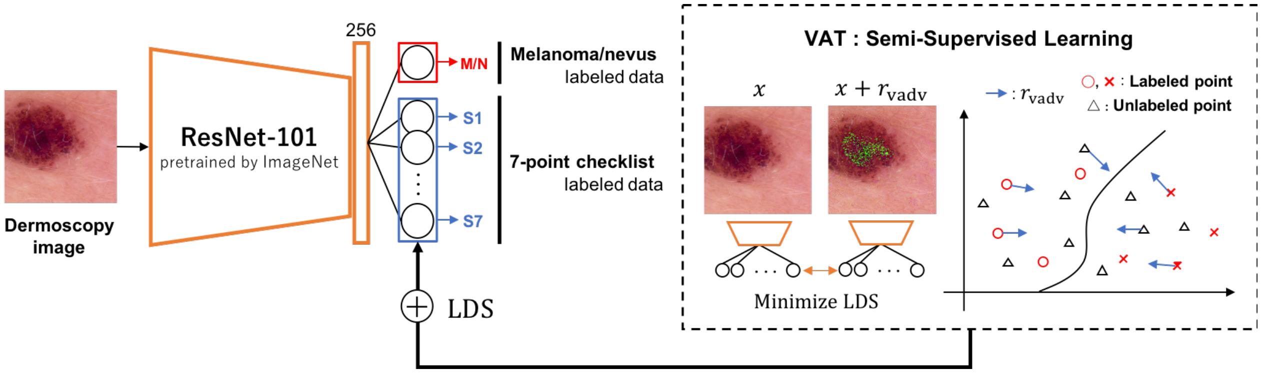


Fig. 1: Overview of proposed model: prediction of clinical indicators using semi-supervised and multi-task learning

The definition of LDS for a labeled or unlabeled input x is as follows:

$$\text{LDS}(x, \theta) := D[p(y|x, \theta), p(y|x + r_{\text{vadv}}, \theta)] \tag{2}$$

$$r_{\text{vadv}} := \arg\max_{r;||r||_2 \leq \epsilon} D[p(y|x, \theta), p(y|x + r, \theta)], \tag{3}$$

where y is the predicted label, θ are the current model parameters, r_{vadv} is the virtual adversarial perturbation that makes the largest impact for the output around the input x, $p(\cdot)$ is the posterior probability, i.e. the output probability distribution of the model, and $D[p, p']$ is a non-negative function that measures the differences between two probabilistic distributions p and p', such as KL-divergence or cross-entropy, respectively. Unlike adversarial training [27], unlabeled data can also be used, since VAT uses LDS to measure the effects of perturbation around the input based on the current model θ. Hence, VAT is capable of semi-supervised training. In this effective algorithm, calculation of r_{vadv} involves only two forwards and one back-propagation steps. The objective loss function of the entire VAT is defined as a linear combination of the conventional classification loss, the above LDS (in eq.(2)) and the additional conditional entropy of the current model output $p(y|x, \theta)$. The reader is referred to the original article [24] for details.

III. Prediction of Explainable Melanoma Indicators

A. Prediction model

Our objective in this study is to accurately predict explainable melanoma indicators as defined by a 7-point checklist, based on a limited set of labeled data. In order to achieve this, our proposed model uses VAT and multi-task learning. Fig. 1 shows an overview of our proposed model.

VAT is expected to generate a robust predictor by utilizing a large amount of unlabeled data. Here, we focus on the fact that information on diagnostic results (melanoma or nevus tumor) can be used even with unlabeled data. We leverage this different type of label information effectively using multi-task learning. Although the amount of labelled data (226) is much smaller than that of the unlabeled data (9,124), diagnosis information is available for both datasets. In this context, a combination of VAT and multi-task learning is expected to give a synergistic effect.

Our proposed model is basically consists of ResNet-101 convolutional neural networks pre-trained with the ImageNet dataset [28]. Using multi-task learning, our network simultaneously predicts the score for seven indicators and the possibility of melanoma (i.e. a total eight dimensions in the output) to make up for the lack of an available labeled dataset. We assign a sigmoid activation function at the output layer to ensure that the network yields the associated output.

B. Training and experiments

We evaluated our model using 10-fold cross-validation, and compared the prediction performance of each indicator in the 7-point checklist. The final diagnostic performance was compared under the following training conditions:

(1) Labeled data alone (baseline)
(2) Baseline + VAT
(3) Baseline + VAT + multi-task learning (MTL)

In scenario (1), we trained the network only with the indicator-labeled dataset A. The size of the training dataset was 90% of that of the labeled dataset in each fold. Note that the number of output dimensions of the network is seven, the same as the number of indicators defined the 7-point checklist. In senario (2), we first train the network in the same way as in (1) and then train it with both the labeled data and the unlabeled dataset (9,124 data points) using the VAT framework. In (3), the training schema is the same as in (2), but we also use malignancy information and train the network with the multi-task learning. Thus, the number of outputs is eight. Here, the larger the score for each feature defined in the 7-point checklist (i.e. larger the total score, TS, in eq.(1)) the higher the degree of malignancy. Thus, for efficiency, the calculation of the adversarial direction r_{vadv} of VAT at each learning step is determined based on the average of all seven (or eight) items.

We also calculated diagnostic results based on the predicted 7-point checklist. (i.e. calculate the total score, TS, from eq.(1) and check if TS $>= 3$). In addition, we built two classifiers using the same ResNet-101 networks that determined only whether the input image was a melanoma or nevus, for ref-

TABLE II: Prediction error for the 7-point checklist

ID	MAE			SD
	Baseline	+VAT	+VAT +MTL	Dermatologists
S1	0.342	0.324	0.329	0.250
S2	0.200	0.141	0.154	0.209
S3	0.106	0.090	0.071	0.072
S4	0.244	0.327	0.369	0.154
S5	0.265	0.174	0.191	0.297
S6	0.314	0.299	0.322	0.250
S7	0.255	0.206	0.213	0.148
Avg.	**0.242**	**0.222**	**0.236**	**0.197**

TABLE III: Diagnosis performance based on predicted indicators of the 7-point checklist

	Diagnosis performance		
	Sensitivity [%]	Specificity [%]	AUC
1) Baseline	59.6	82.8	0.712
2) +VAT	63.6	92.3	0.780
3) +VAT +MTL	72.7	84.6	**0.787**
Dermatologists (gold standard)	76.0	80.3	**0.781**

erence. These were trained using the indicator-labeled dataset A and the diagnosis-labeled dataset B.

We also calculated the standard deviation of the determinations made by dermatologists for each indicator and the diagnostic results based on the predicted 7-point checklist.

IV. Result

Table II summarizes the prediction performance for each item defined in the 7-point checklist. Here, we add the standard deviation of the gold standard results (as assigned by four expert dermatologists) to the table as a reference. The range of each indicator was $[0, 1]$, and prediction performance was evaluated using the mean absolute error (MAE). Table III compares the diagnosis performance based on indicators predicted under conditions (1) to (3) with the gold standard results. Note that the diagnosis is calculated based on the definitions of the 7-point checklist (see Section II-A).

Table IV shows the diagnosis performances of the two classifiers that only discriminated between melanoma or nevus, for comparison purposes. These results were not based on the 7-point checklist.

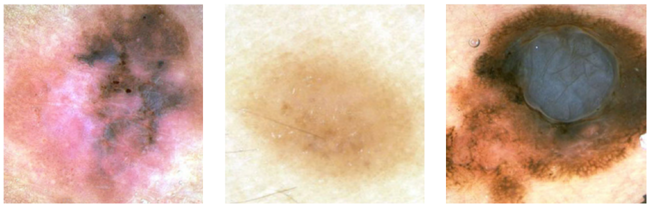
With the introduction of VAT, the prediction error of each indicator in the checklist was reduced by about 2% on average, and the accuracy of melanoma diagnosis calculated based on these indicators improved by about 6.8% in area under the ROC curve (AUC). With the additional introduction of multi-task learning to learn tumor malignancy simultaneously, the improvement in the predictive ability for each item was limited to 0.6% compared to the baseline, while the AUC significantly improved by 7.5% (AUC = 0.787) compared to the baseline (AUC = 0.712). This score is equal to or better than the results based on indicators provided by four experts as the gold standard (SE = 76.0%, SP = 80.3%, AUC = 0.781) or results from the literature (SE = 75.0%, SP = 76.2%) [1]. However, the diagnosis performances based on the 7-point checklist (Table III) was lower than that obtained with classifier that pursuit only the final diagnosis results (Table IV).

TABLE IV: Diagnosis performance on direct discrimination of melanomas

	Diagnosis performance		
	Sensitivity [%]	Specificity [%]	AUC
the indicator-labeled dataset A †	77.9	85.6	0.814
the diagnosis-labeled dataset B ‡	84.6	92.6	0.886

† comparable with 1) or 2) in Table III
‡ comparable with 3) or dermatologists in Table III

TABLE V: Example images, and predicted scores and associated gold standard


(a) MAE = 0.117 (Melanoma)
(b) MAE = 0.073 (Nevus)
(c) MAE = 0.479 (Melanoma)

	†	S1	S2	S3	S4	S5	S6	S7	M/N‡
(a)	True	0.00	1.00	0.60	0.00	0.80	0.70	1.00	1.00
	Predicted	0.30	0.94	0.58	0.10	0.91	1.00	0.97	0.99
(b)	True	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00
	Predicted	0.05	0.03	0.03	0.00	0.02	0.17	0.08	0.05
(c)	True	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00
	Predicted	0.73	0.42	0.30	0.08	0.52	0.58	0.30	0.85

† Indicators S1–S7 are from 7-point checklist
‡ M/N : classifier of Melanoma or Nevus

Table V shows three example images, with their associated gold standard and predicted scores. The first two, (a) and (b), give good estimation results, while (c) has a larger error.

V. Discussion

Based on the fact that the average prediction error of our model (0.222 or 0.236) is close to the standard deviation of four dermatologists (0.197), used here as a gold standard, and since the diagnosis performance calculated from those estimated indicators is equal to or better than that of the dermatologists, it can be seen that our proposed method gives good results in terms of estimation of the indicators defined in the 7-point checklist. Some individual cases such as (c) in Table V sometimes show under-estimation for some structures. This is not always the case, but can be seen in the prediction of structures that occur less often in the training dataset. Although these issues remain, we were able to achieve good prediction performance, as a result of an effective combination of semi-supervised VAT training and multi-task learning in a situation where only a limited number of training labels is available.

A comparison between Tables III and IV shows that the model that predicted only the final diagnosis (melanoma or nevus) achieved a much higher diagnosis performance than the models predicting it via human-oriented diagnosis indicators. This demonstrates the limitations of the diagnostic indicators designed for human readability. The explainability of these medical systems is very important, and there are no quantitative alternative measures. Based on these results, we believe that it is desirable to use two models simultaneously, where the first concentrates on diagnosis performance, and the second is specialized for readability based on these clinically proven criteria. We intend to continue our investigation to create better and more explainable models.

VI. Conclusion

In this paper, we developed a prediction model for the dermoscopic structures defined by a 7-point checklist. We demonstrated that our method, based on an effective combination of VAT, semi-supervised learning and the multi-task learning, shows promising performance. We believe these quantified dermoscopic structures can form the grounds for automated diagnosis.

References

[1] G. Argenziano, H. P. Soyer, S. Chimenti, R. Talamini, R. Corona, F. Sara et al., "Dermoscopy of pigmented skin lesions: Results of a consensus meeting via the Internet," Journal of American Academy of Dermatology, vol. 48, no. 5, pp. 679–693, 2003.

[2] K. Hoffmann, T. Gambichler, A. Rick, M. Kreutz, M. Anschuetz, T. Grunendick et al., "Diagnostic and neural analysis of skin cancer (DANAOS). A multicentre study for collection and computer-aided analysis of data from pigmented skin lesions using digital dermoscopy," British Journal of Dermatology, Vol. 149, pp. 801–809, Oct. 2003.

[3] H. Iyatomi, H. Oka, M. E. Celebi, M. Hashimoto, M. Hagiwara, M.Tanaka et al., "An improved Internet-based melanoma screening system with dermatologist-like tumor area extraction algorithm," Computerized Medical Imaging and Graphics, vol. 32, no. 7, pp. 566–579, 2008.

[4] K. Shimizu, H. Iyatomi, M. E. Celebi, K. Norton and M. Tanaka, "Four-class classification of skin lesions with task decomposition strategy," IEEE Trans. on Biomedical Engineering, vol. 62, no. 1, pp. 274–283, 2015.

[5] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, pp. 115–118, 2017.

[6] L. Yu, H. Chen, Q. Dou, J. Qin, and P. Heng, "Automated Melanoma Recognition in Dermoscopy Images via Very Deep Residual Networks," IEEE Trans. on Medical Imaging, vol. 36, no. 4, 2017.

[7] H. A. Haenssle, C. Fink, R. Schneiderbauer, F. Toberer, T. Buhl, A. Blum et al., "Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists," Annals of Oncology, vol. 29, no. 8, pp. 1836–1842, 2018

[8] H. Iyatomi, H. oka, M. Saito, A. Miyake, M. Kimoto, J. Yamagami et al., "Quantitative assessment of tumor extraction from dermoscopy images and evaluation of computer-based extraction methods for automatic melanoma diagnostic system," Melanoma Research, vol. 16, no. 2, pp. 183–190, 2006.

[9] M. E. Celebi, Q. Wen, H. Iyatomi, K. Shimizu, H. Zhou, and G. Schaefer, "A State-of-the-Art Survey on Lesion Border Detection in Dermoscopy Images," Digital Imaging and Computer Vision, pp. 97–129, 2015. doi:10.1201/b19107-510.1201/b19107-5.

[10] H. Iyatomi, M. E. Celebi, G. Schaefer, and M. Tanaka, "Automated color calibration method for dermoscopy images," Computerized Medical Imaging and Graphics, vol. 35, no. 2, pp. 89–98, 2011.

[11] M. G. Fleming, C. Steger, J. Zhang, J. Gao, A. B. Cognetta, I. Pollak et al, "Techniques for a structural analysis of dermatoscopic imagery," Computerized Medical Imaging and Graphics, vol. 22, no. 5, pp. 375–389, 1998.

[12] M. E. Celebi, H. Iyatomi, W. V. Stoecker, R. H. Moss, H. S. Rabinovitz, G. Argenziano, and H. P. Soyer, "Automatic Detection of Blue-White Veil and Related Structures in Dermoscopy Images," Computerized Medical Imaging and Graphics, vol. 32, no. 8, pp. 670–677, 2008.

[13] B. Shrestha, J. Bishop, K. Kam, R. H. Moss, W. V. Stoecker, S. Umbaugh et al., "Detection of atypical texture features in early malignant melanoma," Skin Research and Technology, vol. 16, pp. 60–65, 2010.

[14] W. Barhoumi, and A. Baazaoui, "Pigment network detection in dermatoscopic images for melanoma diagnosis," IRBM, vol. 35, no. 3, pp. 128–138, 2014.

[15] N. C. F. Codella, C. Lin, A. Halpern, M. Hind, R. Feris, and J. R. Smith, "Collaborative Human-AI (CHAI): Evidence-Based Interpretable Melanoma Classification in Dermoscopic Images," in IMIMIC 2018: Understanding and Interpreting Machine Learning in Medical Image Computing Applications, 2018, pp. 97–105.

[16] P. V. Molle, M. D. Strooper, T. Verbelen, Bert Vankeirsbilck, P. Simoens, and B. Dhoedt, "Visualizing Convolutional Neural Networks to Improve Decision Support for Skin Lesion Classification," in Workshop on Interpretability of Machine Intelligence in Medical Image Computing at MICCAI 2018, pp. 1–8, 2018.

[17] H. Iyatomi, H. Oka, M. E. Celebi, M. Tanaka and K. Ogawa, "Parameterization of Dermoscopic Findings for the Internet-based Melanoma Diagnostic System," in IEEE Proc. CIISP 2007, pp. 183–193, 2007.

[18] G. Argenziano, G. Fabbrocini, P. Carli, V. D. Giorgi, E. Sammarco, M. Delfino, "Epiluminescence microscopy for the diagnosis of ABCD rule of dermatoscopy and a new 7-point checklist based on pattern analysis," Archives of Dermatology, no. 134, pp. 1536–1570, 1998.

[19] J. Bagherzadeh, and H. Asil, "A review of various semi-supervised learning models with a deep learning and memory approach," Iran Journal of Computer Science, vol. 2, no. 2, pp. 65–80, 2019.

[20] E. Hoffer, and N. Ailon, "Deep metric learning using triplet network," In International Workshop on Similarity-Based Pattern Recognition, Springer, 2015, pp. 84–92.

[21] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," In Advances in Neural Information Processing Systems, 2016, pp. 1857–1865.

[22] S. Chen, C. Gong, J. Yang, X. Li, Y. Wei, and J. Li, "Adversarial metric learning," in Proc. IJCAI 2018.

[23] S. Kitada and H. Iyatomi, "Skin lesion classification with ensemble of squeeze-and-excitation networks and semi-supervised learning," arXiv:1809.02568, 2018

[24] T. Miyato, S. Maeda, M. Koyama, and S. Ishii, "Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 41, no. 8, pp. 1979–1993, 2019.

[25] R. Caruana, "Multitask learning," Machine Learning, vol. 28, no. 1, pp. 41–75, 1997.

[26] P. Tschandl, C. Rosendahl, H. Kittler, "The HAM10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions," Scientific Data, doi: 10.1038/sdata.2018.161, 2018.

[27] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," arXiv preprint, arXiv:1412.6572, Dec. 2014.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," IEEE Proc. on Computer Vision and Pattern Recognition, pp. 770–778, 2016.