Original Paper

Issues in Melanoma Detection: Semisupervised Deep Learning Algorithm Development via a Combination of Human and Artificial Intelligence

Xinyuan Zhang[1], PhD; Ziqian Xie[1], PhD; Yang Xiang[1], PhD; Imran Baig[2], BSc; Mena Kozman[2], BSc; Carly Stender[2], BSc; Luca Giancardo[1], PhD; Cui Tao[1], PhD

[1]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, United States

[2]McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, TX, United States

Corresponding Author:
Cui Tao, PhD
School of Biomedical Informatics
The University of Texas Health Science Center at Houston
7000 Fannin St
Suite 600
Houston, TX, 77030
United States
Phone: 1 7135003981
Email: cui.tao@uth.tmc.edu

Abstract

Background: Automatic skin lesion recognition has shown to be effective in increasing access to reliable dermatology evaluation; however, most existing algorithms rely solely on images. Many diagnostic rules, including the 3-point checklist, are not considered by artificial intelligence algorithms, which comprise human knowledge and reflect the diagnosis process of human experts.

Objective: In this paper, we aimed to develop a semisupervised model that can not only integrate the dermoscopic features and scoring rule from the 3-point checklist but also automate the feature-annotation process.

Methods: We first trained the semisupervised model on a small, annotated data set with disease and dermoscopic feature labels and tried to improve the classification accuracy by integrating the 3-point checklist using ranking loss function. We then used a large, unlabeled data set with only disease label to learn from the trained algorithm to automatically classify skin lesions and features.

Results: After adding the 3-point checklist to our model, its performance for melanoma classification improved from a mean of 0.8867 (SD 0.0191) to 0.8943 (SD 0.0115) under 5-fold cross-validation. The trained semisupervised model can automatically detect 3 dermoscopic features from the 3-point checklist, with best performances of 0.80 (area under the curve [AUC] 0.8380), 0.89 (AUC 0.9036), and 0.76 (AUC 0.8444), in some cases outperforming human annotators.

Conclusions: Our proposed semisupervised learning framework can help with the automatic diagnosis of skin disease based on its ability to detect dermoscopic features and automate the label-annotation process. The framework can also help combine semantic knowledge with a computer algorithm to arrive at a more accurate and more interpretable diagnostic result, which can be applied to broader use cases.

(JMIR Dermatol 2022;5(4):e39113) doi: 10.2196/39113

KEYWORDS

deep learning; dermoscopic images; semisupervised learning; 3-point checklist; skin lesion; dermatology; algorithm; melanoma classification; melanoma; automatic diagnosis; skin disease

Introduction

Skin cancer is one of the most common cancers worldwide, with steadily increasing incidence rates of melanoma and nonmelanoma cancers [1]. Early detection of skin cancer is an important prognostic factor that can improve patient survival and overall outcomes [2]. Reliable skin cancer screening, however, may not be readily available to all patients. For

example, individuals who live in rural areas without local dermatology clinics or who face barriers to attending an in-office evaluation may not have an opportunity to have skin cancer detected at an early stage. To address this concern, the use of teledermatology has become increasingly popular, particularly during the COVID-19 pandemic, which has significantly decreased in-person dermatological evaluation [3,4]. Recently, teledermatology has been shown to increase access to reliable dermatology evaluation and to minimize delays in skin cancer management [3,5]. A useful subset of teledermatology is teledermoscopy, whereby digital images of skin lesions are taken using a dermatoscopy or a smartphone with a dermatoscopy attachment [6]. Studies find that the use of dermoscopic images in teledermatology consultations improves the sensitivity and specificity of the diagnosis [3,7]. In this way, teledermoscopy offers itself as a promising tool to increase patient access to reliable skin cancer screening and, thus, the early detection of skin cancer.

The automated classification of dermoscopic images through convolutional neural networks (CNNs) has emerged as a reliable supplement to visual skin examination by on-site specialists in the detection of skin cancer [8-11]. CNNs have the potential to extend reliable skin cancer recognition to clinicians who lack special dermatology training, including nurse practitioners, physician assistants, and primary care physicians. In addition, the use of CNNs enables the evaluation of skin lesions via telemedicine. Images captured on smartphone cameras and analyzed by similar algorithms have been shown to achieve accuracy in identifying melanomas similar to that of board-certified specialists [12]. Some CNN models even exhibit greater sensitivity and specificity in diagnosing early melanoma compared with those of inexperienced clinicians [13,14].

Artificial intelligence (AI) algorithms, however, have some weaknesses. One weakness is interpretability and transparency regarding how the computer arrived at its output, making it difficult for dermatologists to trust the diagnostic results [15-17]. Another is that the current algorithms, such as the deep CNNs used in triaging and classifying suspicious skin lesions, do not provide the reasoning used to arrive at their given result [18]. This is often due to the complexity of the algorithm and hinders their utility due to a lack of the trust in the diagnosis by the patient and the physician [19].

Another limitation of AI algorithms is that a majority rely solely on images as inputs, whereas in a clinical setting, more information can be obtained through, for instance, palpation of the lesion and clinical data on age and family history [20]. The dermatologist also relies on diagnostic rules to make decisions, such as the ABCD rule, pattern analysis, 7-point checklist, and 3-point checklist, which have been developed to standardize the dermoscopic evaluation of melanoma and play a critical role in skin lesion diagnosis [8,9,21-23].

Recent studies have focused on attempts to combine semantic knowledge with the algorithm to arrive at a more accurate diagnosis [20,24-26]. Several studies have suggested that diagnoses derived using more than one source of input are more accurate than are those conceived by one method alone [27-29]. One study showed that nondermatologist physicians were able to improve their accuracy in classifying pigmented lesions when combining their knowledge of age, sex, and localization of the lesion with deep-learning frameworks [24]. Earlier research added factors such as age, body site, proportion of dysplastic nevi, naevus count, and family history of melanoma to a computer image–analysis program and found that the addition of clinical data significantly improved the ability to distinguish between benign and malignant skin lesions [30]. Another study found an improvement in the detection of basal cell carcinoma after adding factors such as lesion size and elevation, age, gender, and location [31]. Kawahara et al [32] conducted a similar work when proposing a multitask deep CNN trained on multimodal data to classify the 7-point melanoma checklist criteria and perform a skin lesion diagnosis. Even though they intergraded each feature from a 7-point checklist using loss blocks, their studies did not integrate the knowledge with the CNN architecture. One major constraint of these studies is the lack of high-quality data related to diagnosis, for example, the dermoscopic features that dermatologists use to diagnose skin lesions. In this study, we address these limitations by developing a semisupervised deep-learning framework that applies the results learned from a small, annotated data set to a larger unlabeled data set as well as by imitating the human diagnosis process in our CNN structure.

In this experiment, we chose the 3-point checklist for melanoma and melanocytic nevus as an illustration of diagnostic rules and disease class. The 3-point checklist is easy to interpret and is highly sensitive for the diagnosis of melanoma by nonexpert clinicians [33]. Melanoma is well known as the most aggressive cutaneous malignancy, accounting for approximately 75% of all skin cancer deaths [24]. It often shares morphology with melanocytic nevi on naked-eye examination, a technique that yields only 60% accuracy in a melanoma diagnosis by expert dermatologists [34]. In this regard, the International Skin Imaging Collaboration (ISIC) organizes data challenges every year, which focus primarily on diagnostic accuracy when distinguishing melanoma from other malignant and benign lesions [35]. Numerous studies that concern the use of the 3-point checklist to help classify melanomas have been conducted [33,36,37]. In these studies, participants with varying experience were able to score proven nonmelanoma and proven melanoma lesions using just the 3-point checklist criteria. A disadvantage of this method, however, is that the checklist tends to miss thinner melanomas [37]. None of the studies related to 3-point checklist has tried to combine visual inspection with CNN-extracted imaging features to arrive at a diagnosis. This is also the major difference in our state-of-the-art methodology as compared to what was seen in previous ISIC data challenges.

Combining diagnostic rules with the 3-point checklist classification algorithm can yield benefits that improve patient access to care and diagnostic accuracy. The proposed algorithms have several potential application scenarios, including the following: (1) they can automatically classify skin disease images and generate feature labels by listing the criteria used to categorize suspicious lesions to improve trust and acceptance of teledermoscopy; (2) they can assist medical students to learn and identify the features in dermoscopic images; given the detailed evaluation of each criterion in the 3-point checklist by

the algorithm, students can use the checklist to learn about the fundamental parameters used to differentiate lesions as a benign nevus or a melanoma; and (3) they can automate the process of feature annotation; thus, fewer human annotators need to be involved, enabling the secondary use of enormous imaging data resources, such as the ISIC archive.

Methods

Data Set

All images from labeled and unlabeled data sets come from the ISIC archive. "Label" here represents the 3-point checklist feature labels, which means both "labeled" and "unlabeled" data sets contain disease type information. For the small, labeled data set, we selected an even distribution of melanoma and melanocytic nevus dermoscopic images from ISIC 2019 to annotate, using the 3-point checklist features. The large unlabeled data set came mainly from ISIC 2020, which contains the 584 melanoma and 5193 melanocytic nevus dermoscopic images. To balance the data set, we added 4062 melanoma images from ISIC 2019, excluding the images in the small, labeled data set. We divided each data set into training and validation sets in an 80/20 ratio and used 5-fold cross-validation, which means the data set was divided equally into 5 subsets and rotating in order to be the training or validation data set. We annotated an additional 400 images as a holdout testing set.

The 3-point checklist is easy to interpret and is highly sensitive for the diagnosis of melanoma versus melanocytic nevus. Our algorithm evaluated dermoscopic images of pigmented lesions based on the 3-point checklist, indicating the presence or absence of (1) asymmetry, (2) atypical pigment network, and (3) blue-white structures. If any one of these features was detected from the skin lesion image, 1 point would be added on top of the scoring for that image. The scoring range per image is 0 to 3. These 3-point automated classification outputs can aid in a provider's decision to biopsy a lesion or to refer to a specialist for a more thorough evaluation. Table 1 presents the number of images for the skin disease categories of melanoma and melanocytic nevus.

Table 1. Number of images for skin disease categories for labeled and unlabeled data sets.

Disease	Unlabeled data set	Labeled data set
Melanoma	4646	450
Melanocytic nevus	5193	450
Total	9839	900

Annotation of the 3-Point Checklist

There are 3 features of the 3-point checklist, which are atypical network, asymmetry, and blue-white structure. For each feature detected, 1 score will be added for that image. The higher the score is (usually higher than 2), the higher the risk of melanoma will be. If the score is lower than 1, according to the 3-point checklist, the lesion is more likely to be benign. Our experiment was developed based on a gold standard whereby each image was rigorously reviewed by at least 2 annotators. If consensus was reached, the resulting diagnosis was annotated. If not, a third annotator would evaluate the image again. We divided the annotation into 2 steps. First, the 3 annotators had training sessions to develop consensus annotation guidelines. We provided the annotators with a small image set annotated by domain experts to annotate and evaluate. During this phase, the annotators are allowed to discuss their different understandings. After interrater agreement reached at least 70%, we moved to the second step, in which they annotated images independently. We divided the whole image data set into 3 subsets, and each annotator was assigned 2 subsets so that every image had at least 2 annotation results. Our final interrater agreement Kappa-Cohen score for the second step was 0.64, which indicated substantial agreement. If any images had different annotation results, we brought in the third annotator, who was not previously assigned to the image, and took a majority vote. Overall, this is a very time-consuming process.

Image Preprocessing

Crop and Resize

Because the training data set came from 3 data sources, each had a different resolution of the images. There could be 1 lesion that took up the entire image or just 1 corner of the graph. Hence, we developed a rule to crop and resize all the training images, which improved the performance of our model.

Color Constancy

Due to the different imaging sources and illuminations, the color of dermoscopic images varied considerably. Therefore, it was important to calibrate the color of the images in the preprocessing stage to reduce possible bias for the deep neural network. Catarina et al [38] compared 4 color-constancy algorithms (Gray World, max-RGB, Shades of Gray, and General Gray World) to calibrate the color of dermoscopic images for the melanoma classification system. These algorithms improved the system performance by increasing sensitivity and specificity, and Shades of Gray achieved better results than did the other color-constancy algorithms. Thus, for the project, we chose Shades of Gray as the color-constancy algorithm to calibrate the color of the dermoscopic images before the training stages. The calibration procedure involves 2 steps. First, the color of the light source in the RGB color space is estimated. Then, the image is transformed, using the estimated illuminant.

Contrast-Limited Adaptive Histogram Equalization

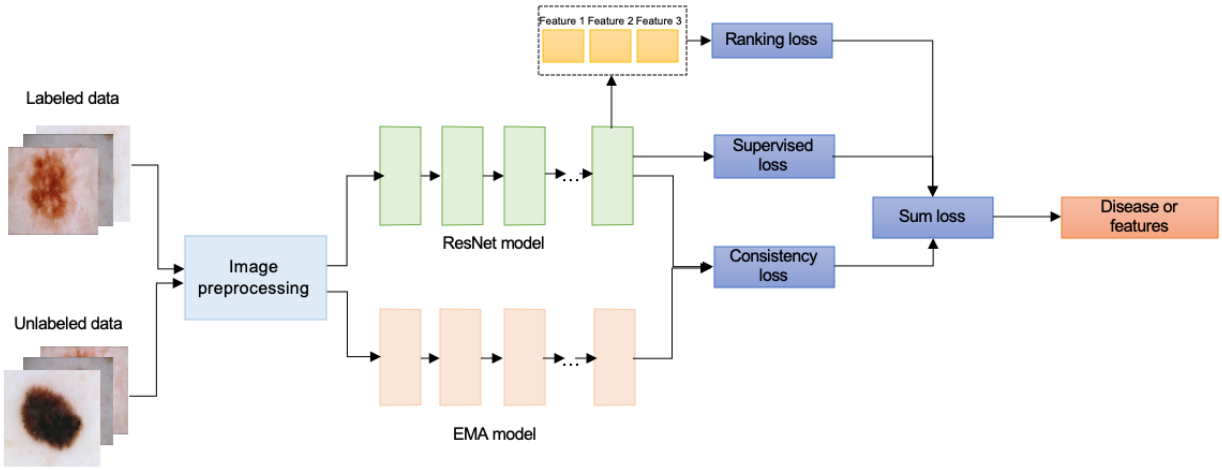
Contrast-Limited Adaptive Histogram Equalization was used to improve the contrast in images. Unlike histogram equalization, it computes several distinct sections of the image

and uses them to redistribute the lightness values of the image. It helps to improve the local contrast and enhance the edges of objects in the image.

Model Architecture

We proposed a semisupervised learning framework for the prediction of skin disease that uses a small set of labeled images and a larger set of unlabeled images. The labeled data set contains 900 images that were labeled with disease tags and the 3-point checklist annotation, while the unlabeled data set contains 9839 images that have only disease tags. The architecture of the proposed classification model is presented in Figure 1 and contains primarily 3 components. The input component involves the preprocessing of both labeled and unlabeled images. The output of the input component is streamed into 2 branches. One branch is the supervised learning component that uses ResNet, inside which the representation of each image is associated with the 3-point labels and the classification tag and with the label-related ranking loss [39] and classification loss, correspondingly. The other is the semisupervised learning component, whereby a consistency loss is optimized using the output from an exponential moving average (EMA) model of the ResNet branch [40]. Finally, the 3 types of losses are combined, and coefficients are used to balance their weights. We provide a detailed description of these 3 components in this section.

Figure 1. Architecture of the proposed semisupervised learning framework. EMA: exponential moving average; ResNet: residual neural network.



Supervised Learning + Ranking Loss

The supervised learning consists of 2 tasks, which are jointly learned during training. One task is the classification of the skin disease, and the other is the classification of each feature in the 3-point checklist. Using the 3-point checklist, each feature is given a binary score of 0 or 1 in the training phase, indicating whether it exists in the image. A total score higher than 2 suggests that the lesion is more likely to be malignant. We incorporated the traditional cross-entropy loss to optimize the skin disease classification part and used ranking loss to represent the 3-point checklist knowledge. The hyperparameters for our training models are as follows: a batch size of 128, stochastic gradient descent optimizer, and ReduceLROnPlateau learning rate decay (mode="min," factor=0.5, threshold=0.01, patience=7, verbose=True).

Semisupervised Learning

Image annotation requires not only extensive time investment but also domain expertise of human annotators. Inspired by the research of Tarvainen and Valpola [40], we developed a semisupervised scheme based on their "mean teacher" framework to automate the feature annotation process of skin lesion images. This model can use the information from small-scaled labeled images and make skin feature and disease predictions on larger unlabeled image data sets. On top of that, we developed and integrated disease- or feature-specific loss functions to combine knowledge from human expertise into the model. The predicted features can be used simultaneously in the training phase to improve the disease classification accuracy. The supervised loss is associated with the disease label of each image and denoted by the cross-entropy function. In the semisupervised learning component, the mean-teacher strategy was adopted to minimize the consistency loss between labeled and unlabeled data sets and to average the model weights from supervised and unsupervised learning.

Theory and Calculation

Supervised Learning + Ranking Loss

Using the ranking loss, we enforce the model to learn a predefined diagnostic rule—the samples with higher scores are more likely to have melanoma. The ranking loss is computed from each pair of samples in a batch. We denote $o_{ij} \equiv f(x_i) - f(x_j)$, where f is the logit corresponding to the disease class, the posterior P_{ij}, and the desired target values $\bar{P}_{ij}$:

$$\bar{P}_{ij} = \begin{cases} 0 & score(i) < score(j) \\ 1/2 & score(i) = score(j) \\ 1 & score(i) > score(j) \end{cases} \quad (1)$$

Then, the cross-entropy loss function can be represented as

$$C_{ij} \equiv C(o_{ij}) = -\bar{P}_{ij} log P_{ij} - (1 - \bar{P}_{ij}) \log (1 - P_{ij}) \quad (2)$$

We compute P_{ij} from o_{ij} using the sigmoid function as follows; the loss function can be further rewritten as:

$$P_{ij} \equiv \frac{e^{o_{ij}}}{1 + e^{o_{ij}}} \quad (3)$$

$$C_{ij} = -\bar{P}_{ij} o_{ij} + \log (1 + e^{o_{ij}}) \quad (4)$$

Semisupervised Learning

The EMA model behaves as the teacher model on the unlabeled. This method constrains the model to behave similarly to the past models during the update so it can potentially find flatter local minima and avoid singularity points where a small update would result in large behavior change in the model. The mean-teacher strategy proved useful in previous works, and the consistency cost is defined as follows, where is updated based on EMA parameters:

$$J(\theta) = E_{x,\eta',\eta}[\| f(x,\theta',\eta') - f(x,\theta,\eta) \|]^2 \quad \text{where } \theta'_t = \alpha\,\theta'_{t-1} + (1-\alpha)\,\theta_t. \quad (5)$$

Finally, the ranking loss, disease supervised loss, feature supervised loss (FSL), and consistency loss were added together to train the model.

$$L_{Sum} = \alpha_1 \cdot L_R + \alpha_2 \cdot L_{DS} + \alpha_3 \cdot L_{FS} + \alpha_4 \cdot L_C \quad (6)$$

Results

Our models were built based on the state-of-the-art ResNet model. We tried ResNet-18, ResNet-50, ResNet-152, and Resnext50_32x4d, and there was no significant difference in classification accuracies. To facilitate the training process, we used a relatively light architecture, ResNet-18, as our baseline.

The first task is to test whether the model will increase the classification accuracy after adding human knowledge, which is transformed and represented in the Ranking Loss format. Many state-of-the-art CNN model architectures have been developed for image recognition task, some of which achieved great performance on the skin lesion recognition task on ISIC data sets. In a 2021 paper published by Yiming Zhang et al [41], they reported that DenseNet [42] achieved superior performance over other deep learning approaches on the melanoma classification task using ISIC 2020 data set. MobileNet [43] is another CNN model developed in the recent years, and it has been adapted to ISIC image classification tasks in many cases [44,45]. To choose a CNN architecture as our baseline model and show the improvement of accuracy after combining the human knowledge in the ranking loss format, we compared accuracy results of the state-of-the-art CNN models mentioned above. The comparison outcomes are shown in Table 2. We chose ResNet as our baseline model for its better performance. All the models were trained using a 900 labeled data set (from Table 1). We tested the performance of pretrained baseline model on our larger 9000-image data set using 80/20 data split. The results are shown in Table 2. We used 5-fold cross-validation to calculate the mean and standard deviation of the validation accuracy.

As can be seen from the table, the pretrained baseline model reached the same level of accuracy on the large 9000-image data set. After adding the human knowledge of the 3-point checklist rule, the average accuracy even improved on this basis.

The previous experiment was based on human-annotated, 3-point feature labels. The entire process, from recruiting annotators to finally reaching agreement, took more than 2 months. Hence, we developed the semisupervised model to automate the feature-annotation process. We combined the generated features as human knowledge to test whether such knowledge can help to improve the disease classification accuracy.

To evaluate the performance of the 3-point feature classification for our semisupervised model, we calculated the testing accuracy and area under the receiving operating characteristic curve (AUC) on a separate holdout testing data set that contains 100 images with annotated 3-point features and disease type. We tested the performance for feature and disease classification on the models shown in Table 3, for which “baseline” is the labeled 900-image data set for supervised training, followed by different combinations of loss functions.

As seen in Table 3, the semisupervised model that combined all 3 loss functions achieved the best accuracy for disease classification. Adding FSL improved the performance of disease classification by 2%. The result shows that emphasizing the weight of “Asymmetry” feature improved the testing accuracy of “asymmetry” by 2% and improved the classification of the “atypical network” by 3%. Nevertheless, the accuracy of “Blue-white structure” and disease classification has a significant decrease.

Table 2. Five-fold cross validation results for the disease classification task.

Model	Five-fold accuracy, mean (SD)
MobileNetV3 (Pretrain=True)	0.8733 (0.0113)
DenseNet (Pretrain=True)	0.8856 (0.0114)
Baseline (ResNet-18, Pretrain=True)	0.8867 (0.0191)
Baseline + Human Knowledge (RL[a])	0.8943 (0.0115)

[a]RL: ranking loss.

Table 3. Results for semisupervised model for disease or feature classification tasks with different loss functions—disease supervised loss (DSL), feature supervised loss (FSL), and consistency loss (CL).

Model	Asymmetry, accuracy (AUC[a])	Atypical network, accuracy (AUC)	Blue-white structure, accuracy (AUC)	Disease, accuracy (AUC)
CL	0.51 (0.5760)	0.53 (0.5021)	0.54 (0.5620)	0.54 (0.5648)
DSL	0.51 (0.5480)	0.76 (0.6480)	0.58 (0.5285)	0.76 (0.8690)
FSL	0.80 (0.8380)	0.89 (0.9036)	0.74 (0.8036)	0.51 (0.5339)
FSL+CL	0.68 (0.7816)	0.87 (0.8752)	0.75 (0.8137)	0.53 (0.5402)
DSL+FSL	0.76 (0.7892)	0.86 (0.8602)	0.76 (0.8133)	0.74 (0.8418)
DSL+CL	0.53 (0.5448)	0.79 (0.4340)	0.47 (0.5943)	0.77 (0.8389)
DSL+FSL+CL	0.73 (0.8036)	0.85 (0.8474)	0.76 (0.8444)	0.79 (0.8402)
DSL+FSL[b]+CL	0.75 (0.7932)	0.88 (0.8752)	0.71 (0.7951)	0.69 (0.7971)

[a]AUC: area under the receiving operating characteristic curve.

[b]We emphasized the weight of the "Asymmetry" feature in the loss function.

Discussion

Annotation Process

Annotators in this study were medical students with no expert training in dermatology. They evaluated images based solely on tutorials from web-based resources and textbooks. Without any designated training, using example images, each of the annotators initially had a different idea of what each feature looked like. Preliminary agreement scores may have been improved if annotators had been given reference images from which to learn the dermoscopic features. This finding highlights the potential value of our algorithm as an educational tool. If medical students can evaluate a dermoscopic image and check their 3-point annotation against the algorithm's validated output, it will help them develop their ability to visually identify each dermoscopic feature.

During the image-annotation process, there were some uncertainties for annotators. First, the vague definition of dermoscopic features, especially "atypical network" posed an issue, as each annotator had a different idea of what that looks like. This resulted in initial low agreement scores. We address this concern by proposing an ontology that can integrate the domain knowledge on dermoscopic features and represent the features in a more standardized, computer-readable format.

Another uncertainty in analyzing the images was the use of different screens with various color-display settings. One common error that was encountered was the inability to properly characterize blue structures when night light or blue light filters were activated. As such options can be automatically engaged on a schedule, however, this could lead to annotation errors. The use of different screens led to initial disagreement among the annotators but can be corrected by proper calibration and ensuring that no color filter is on.

One limitation of this study was that most of the images are taken from White skin. This has implications for whether the algorithm can be effective in detecting melanoma in colored skin. Training the algorithm to identify lesions in more than just one group of skin colors would be valuable in helping to screen a larger population of patients at risk of melanoma. Another limitation was that the image quality could have been decreased due to shadows, hairs, reflections, and noise, leading to an inadequate lesion analysis, as discussed in an earlier study [46].

Classification Models

For the first task, after combining the 3-point checklist human knowledge, the loaded model weights from the large data set improved the classification accuracy from an average of 0.8867 to 0.8943. This shows that the ranking loss has a positive impact on classification accuracy. We plan to continue to work on expanding human knowledge to develop more complicated diagnostic rules to test their impacts on computer algorithms.

For the feature- and disease-classification task that used semisupervised architecture, interesting findings were discovered in Table 3. The improvement of the classification accuracy for certain feature labels can be accomplished by assigning a heavier weight on the corresponding feature's loss function, however, at the cost of scarifying the accuracy for disease classification. Among the 3 features, blue-white structure has a relatively low accuracy when classified without feature-supervised loss function, the potential reason being the unbalance of blue-white structure data set where most of them are negative. While adding FSL is helpful for the feature classification task, adding disease-supervised loss function could bring down the performance of feature classification. For the disease classification, adding FSL alone did not improve the accuracy; however, combining consistency loss with FSL is showing a positive effect on disease classification.

We also noticed that, during the human annotation process for the 3-point checklist, the atypical network had the lowest interagreement rate among the 3 annotators. For the computer feature-classification task, however, the atypical network had the highest classification accuracy. This suggests that the algorithm has the advantage of learning certain image features that might be a challenge for human experts. This shows that human intelligence and AI can complement each other.

Because our image data set is from the ISIC archives, we also compared the performance of our algorithm with the winner of the ISIC 2020 leaderboard [47]. The current best performance

has an AUC of 0.949. The AUC of the proposed model on the 400 unlabeled-image testing set (from ISIC 2020) is 0.9848 with different settings of disease category. Our 0.9848 AUC, however, cannot be directly compared with the results from the ISIC leaderboard, as our classification task includes only melanoma and melanocytic nevus, whereas the ISIC challenge has some "unknown" images. The remainder of the results in this regard are calculated on the small 100 labeled-image testing set, which has significant improvement over the application of the student-teacher framework, indicating the power of semisupervised learning.

Future Steps

We plan to implement more fine-tuned model architectures trained from scratch so that a more advanced ensemble can be applied by integrating architectures from submodels. Our current experimental setting for the disease classes and rules of the 3-point checklist is only a demonstration of how we can integrate the human thinking process into the structure of CNNs. There are numerous diagnostic rules that are being developed, as dermatology is thriving, and we plan to summarize all the diagnostic rules and dermoscopic features mentioned, as well as their relationship with skin diseases, into ontology and to further accelerate the automation process of clinical decision support by computer algorithms. With our trained algorithm, we can already automate the 3-point checklist annotation process and apply it to a wider range of image databases.

Conclusions

This study is distinctive because it combines the semantic knowledge from the 3-point checklist with a computer algorithm (CNN) to arrive at a more accurate and more interpretable diagnosis. The CNN classification was conducted based on more information than just the imaging pixels. Due to the time and labor consumption of the image-annotation process, there are vast imaging data sets that remain undiscovered. Our proposed semisupervised learning framework can help automate the annotation process, enabling the reuse of many skin-imaging data sets, which is also beneficial to the robustness and domain adaptation of the deep-learning model.

Acknowledgments

XZ conducted the experiments and led the writing of the manuscript. ZX and YX helped with the design of the model and the writing of methodology. IB, MK, and CS conducted the annotation and contributed to the writing of the manuscript from the clinician's perspective. LG and CT supervised the project. All authors participated in the design of this study.

This work was supported by UTHealth Innovation for Cancer Prevention Research Training Program Pre-doctoral Fellowship (Cancer Prevention and Research Institute of Texas Grant No. RP160015 and No. RP210042).

Conflicts of Interest

None declared.

References

1. Guy GP, Thomas C, Thompson T, Watson M, Massetti GM, Richardson LC, Centers for Disease ControlPrevention (CDC). Vital signs: melanoma incidence and mortality trends and projections - United States, 1982-2030. MMWR Morb Mortal Wkly Rep 2015 Jun 05;64(21):591-596 [FREE Full text] [Medline: 26042651]
2. Jerant A, Johnson J, Sheridan C, Caffrey T. Early detection and treatment of skin cancer. Am Fam Physician 2000 Jul 15;62(2):357-68, 375 [FREE Full text] [Medline: 10929700]
3. Giavina-Bianchi M, Santos AP, Cordioli E. Teledermatology reduces dermatology referrals and improves access to specialists. EClinicalMedicine 2020 Dec;29-30:100641 [FREE Full text] [doi: 10.1016/j.eclinm.2020.100641] [Medline: 33437950]
4. Conforti C, Lallas A, Argenziano G, Dianzani C, Di Meo N, Giuffrida R, et al. Impact of the COVID-19 Pandemic on Dermatology Practice Worldwide: Results of a Survey Promoted by the International Dermoscopy Society (IDS). Dermatol Pract Concept 2021 Jan;11(1):e2021153 [FREE Full text] [doi: 10.5826/dpc.1101a153] [Medline: 33614221]
5. Duong TA, Lamé G, Zehou O, Skayem C, Monnet P, El Khemiri M, et al. A process modelling approach to assess the impact of teledermatology deployment onto the skin tumor care pathway. Int J Med Inform 2021 Feb;146:104361. [doi: 10.1016/j.ijmedinf.2020.104361] [Medline: 33348274]
6. Lee K, Finnane A, Soyer HP. Recent trends in teledermatology and teledermoscopy. Dermatol Pract Concept 2018 Oct 31;8(3):214-223. [doi: 10.5826/dpc.0803a013]
7. Ferrándiz L, Ojeda-Vila T, Corrales A, Martín-Gutiérrez FJ, Ruíz-de-Casas A, Galdeano R, et al. Internet-based skin cancer screening using clinical images alone or in conjunction with dermoscopic images: A randomized teledermoscopy trial. J Am Acad Dermatol 2017 Apr;76(4):676-682. [doi: 10.1016/j.jaad.2016.10.041] [Medline: 28089728]
8. Brinker TJ, Hekler A, Enk AH, Klode J, Hauschild A, Berking C, Collaborators. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. Eur J Cancer 2019 May;113:47-54 [FREE Full text] [doi: 10.1016/j.ejca.2019.04.001] [Medline: 30981091]
9. Haenssle H, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A, Reader study level-Ilevel-II Groups, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol 2018 Aug 01;29(8):1836-1842 [FREE Full text] [doi: 10.1093/annonc/mdy166] [Medline: 29846502]

10. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017 Feb 02;542(7639):115-118 [FREE Full text] [doi: 10.1038/nature21056] [Medline: 28117445]
11. Zhang X, Wang S, Liu J, Tao C. Computer-aided diagnosis of four common cutaneous diseases using deep learning algorithm. 2017 Presented at: IEEE Int Conf Bioinforma Biomed; November 13-16, 2017; Kansas City, MO, USA. [doi: 10.1109/BIBM.2017.8217850]
12. Phillips M, Marsden H, Jaffe W, Matin RN, Wali GN, Greenhalgh J, et al. Assessment of Accuracy of an Artificial Intelligence Algorithm to Detect Melanoma in Images of Skin Lesions. JAMA Netw Open 2019 Oct 02;2(10):e1913436 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.13436] [Medline: 31617929]
13. Tran K, Ayad M, Weinberg J, Cherng A, Chowdhury M, Monir S, et al. Mobile teledermatology in the developing world: implications of a feasibility study on 30 Egyptian patients with common skin diseases. J Am Acad Dermatol 2011 Feb;64(2):302-309. [doi: 10.1016/j.jaad.2010.01.010] [Medline: 21094560]
14. Thomas L, Puig S. Dermoscopy, Digital Dermoscopy and Other Diagnostic Tools in the Early Detection of Melanoma and Follow-up of High-risk Skin Cancer Patients. Acta Derm Venereol 2017 Jul;Suppl 218:14-21 [FREE Full text] [doi: 10.2340/00015555-2719] [Medline: 28676882]
15. Schwartz A, Elstein AS. Clinical problem solving and diagnostic decision making: A selective review of the cognitive research literature. In: The Evidence Base of Clinical Diagnosis: Theory and Methods of Diagnostic Research, Second Edition. Chichester (UK): Blackwell Publishing Ltd; 2009:237-255.
16. Asan O, Bayrak A, Choudhury A. Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. J Med Internet Res 2020 Jun 19;22(6):e15154 [FREE Full text] [doi: 10.2196/15154] [Medline: 32558657]
17. Hauschild A, Chen SC, Weichenthal M, Blum A, King HC, Goldsmith J, et al. To excise or not: impact of MelaFind on German dermatologists' decisions to biopsy atypical lesions. J Dtsch Dermatol Ges 2014 Jul;12(7):606-614. [doi: 10.1111/ddg.12362] [Medline: 24944011]
18. Demyanov S, Chakravorty R, Abedini M, Halpern A, Garnavi R. Classification of dermoscopy patterns using deep convolutional neural networks. 2016 Presented at: IEEE 13th International Symposium on Biomedical Imaging (ISBI); Prague, Czech Republic; April 13-16, 2016 p. 364. [doi: 10.1109/isbi.2016.7493284]
19. Jones OT, Jurascheck LC, van Melle MA, Hickman S, Burrows NP, Hall PN, et al. Dermoscopy for melanoma detection and triage in primary care: a systematic review. BMJ Open 2019 Aug 20;9(8):e027529 [FREE Full text] [doi: 10.1136/bmjopen-2018-027529] [Medline: 31434767]
20. Hekler A, Utikal JS, Enk AH, Hauschild A, Weichenthal M, Maron RC, Collaborators. Superior skin cancer classification by the combination of human and artificial intelligence. Eur J Cancer 2019 Oct;120:114-121 [FREE Full text] [doi: 10.1016/j.ejca.2019.07.019] [Medline: 31518967]
21. Xie F, Yang J, Liu J, Jiang Z, Zheng Y, Wang Y. Skin lesion segmentation using high-resolution convolutional neural network. Comput Methods Programs Biomed 2020 Apr;186:105241. [doi: 10.1016/j.cmpb.2019.105241] [Medline: 31837637]
22. Rasul M, Dey N, Hashem M. A Comparative Study of Neural Network Architectures for Lesion Segmentation and Melanoma Detection. 2020 Presented at: 2020 IEEE Region 10 Symposium (TENSYMP); June 05-07, 2020; Dhaka, Bangladesh p. 1572-1575. [doi: 10.1109/tensymp50017.2020.9230969]
23. Shen J, Zhang C, Jiang B, Chen J, Song J, Liu Z, et al. Artificial Intelligence Versus Clinicians in Disease Diagnosis: Systematic Review. JMIR Med Inform 2019 Aug 16;7(3):e10010 [FREE Full text] [doi: 10.2196/10010] [Medline: 31420959]
24. Lucius M, De All J, De All JA, Belvisi M, Radizza L, Lanfranconi M, et al. Deep Neural Frameworks Improve the Accuracy of General Practitioners in the Classification of Pigmented Skin Lesions. Diagnostics (Basel) 2020 Nov 18;10(11):969 [FREE Full text] [doi: 10.3390/diagnostics10110969] [Medline: 33218060]
25. Felmingham CM, Adler NR, Ge Z, Morton RL, Janda M, Mar VJ. The Importance of Incorporating Human Factors in the Design and Implementation of Artificial Intelligence for Skin Cancer Diagnosis in the Real World. Am J Clin Dermatol 2021 Mar 22;22(2):233-242. [doi: 10.1007/s40257-020-00574-4] [Medline: 33354741]
26. Yap J, Yolland W, Tschandl P. Multimodal skin lesion classification using deep learning. Exp Dermatol 2018 Nov;27(11):1261-1267. [doi: 10.1111/exd.13777] [Medline: 30187575]
27. Barnett ML, Boddupalli D, Nundy S, Bates DW. Comparative Accuracy of Diagnosis by Collective Intelligence of Multiple Physicians vs Individual Physicians. JAMA Netw Open 2019 Mar 01;2(3):e190096 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.0096] [Medline: 30821822]
28. Kurvers RHJM, Krause J, Argenziano G, Zalaudek I, Wolf M. Detection Accuracy of Collective Intelligence Assessments for Skin Cancer Diagnosis. JAMA Dermatol 2015 Dec 01;151(12):1346-1353. [doi: 10.1001/jamadermatol.2015.3149] [Medline: 26501400]
29. Kämmer JE, Hautz WE, Herzog SM, Kunina-Habenicht O, Kurvers RHJM. The Potential of Collective Intelligence in Emergency Medicine: Pooling Medical Students' Independent Decisions Improves Diagnostic Performance. Med Decis Making 2017 Aug;37(6):715-724. [doi: 10.1177/0272989X17696998] [Medline: 28355975]

30. Binder M, Kittler H, Dreiseitl S, Ganster H, Wolff K, Pehamberger H. Computer-aided epiluminescence microscopy of pigmented skin lesions: the value of clinical data for the classification process. Melanoma Res 2000 Dec;10(6):556-561. [doi: 10.1097/00008390-200012000-00007] [Medline: 11198477]
31. Kharazmi P, Kalia S, Lui H, Wang ZJ, Lee TK. A feature fusion system for basal cell carcinoma detection through data-driven feature learning and patient profile. Skin Res Technol 2018 May;24(2):256-264. [doi: 10.1111/srt.12422] [Medline: 29057507]
32. Kawahara J, Daneshvar S, Argenziano G, Hamarneh G. Seven-Point Checklist and Skin Lesion Classification Using Multitask Multimodal Neural Nets. IEEE J. Biomed. Health Inform 2019 Mar;23(2):538-546 [FREE Full text] [doi: 10.1109/jbhi.2018.2824327]
33. Soyer HP, Argenziano G, Zalaudek I, Corona R, Sera F, Talamini R, et al. Three-point checklist of dermoscopy. A new screening method for early detection of melanoma. Dermatology 2004 Feb 3;208(1):27-31. [doi: 10.1159/000075042] [Medline: 14730233]
34. Kittler H, Pehamberger H, Wolff K, Binder M. Diagnostic accuracy of dermoscopy. The Lancet Oncology 2002 Mar;3(3):159-165. [doi: 10.1016/s1470-2045(02)00679-4]
35. ISIC 2022: International Semantic Intelligence Conference (ISIC 2022). WikiCFP. URL: http://www.wikicfp.com/cfp/servlet/event.showcfp?eventid=133712&copyownerid=169171 [accessed 2021-10-28]
36. Rogers T, Marino M, Dusza S, Bajaj S, Marchetti M, Marghoob A. Triage amalgamated dermoscopic algorithm (TADA) for skin cancer screening. Dermatol Pract Concept 2017 Apr 30;7(2):39-46 [FREE Full text] [doi: 10.5826/dpc.0702a09] [Medline: 28515993]
37. Gereli MC, Onsun N, Atilganoglu U, Demirkesen C. Comparison of two dermoscopic techniques in the diagnosis of clinically atypical pigmented skin lesions and melanoma: seven-point and three-point checklists. Int J Dermatol 2010 Jan;49(1):33-38. [doi: 10.1111/j.1365-4632.2009.04152.x] [Medline: 20465608]
38. Fidalgo Barata A, Celebi E, Marques J. Improving Dermoscopy Image Classification Using Color Constancy. IEEE J. Biomed. Health Inform 2014:1-1. [doi: 10.1109/jbhi.2014.2336473]
39. Burges C, Shaked T, Renshaw E, Hamilton N, Hullender G. Learning to rank using gradient descent. 2005 Presented at: ICML '05: Proceedings of the 22nd international conference on Machine learning; August 7-11, 2005; Bonn, Germany p. 89-96. [doi: 10.1145/1102351.1102363]
40. Tarvainen A, Valpola H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. NeurIPS Proceedings. URL: https://proceedings.neurips.cc/paper/2017/file/68053af2923e00204c3ca7c6a3150cf7-Paper.pdf [accessed 2022-10-14]
41. Zhang Y, Wang C. SIIM-ISIC melanoma classification with DenseNet. 2021 Presented at: IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE); March 26-28, 2021; Nanchang, China. [doi: 10.1109/icbaie52039.2021.9389983]
42. Huang G, Liu S, van DML, Weinberger K. CondenseNet: An efficient DenseNet using learned group convolutions. 2018 Presented at: IEEE/CVF Conference on Computer Vision and Pattern Recognition; June 18-23, 2018; Salt Lake City, UT, USA. [doi: 10.1109/cvpr.2018.00291]
43. Howard A, Sandler M, Chen B, Wang W, Chen LC, Tan M, et al. Searching for MobileNetV3. 2019 Presented at: IEEE/CVF International Conference on Computer Vision (ICCV); October 27-November 02, 2019; Seoul, Korea (South). [doi: 10.1109/iccv.2019.00140]
44. Mohamed E, El-Behaidy W. Enhanced skin lesions classification using deep convolutional networks. 2019 Presented at: Ninth International Conference on Intelligent Computing and Information Systems (ICICIS); December 08-10, 2019; Cairo, Egypt. [doi: 10.1109/icicis46948.2019.9014823]
45. Widiansyah M, Rasyid S, Wisnu P, Wibowo A. Image segmentation of skin cancer using MobileNet as an encoder and linknet as a decoder. J. Phys.: Conf. Ser 2021 Jul 01;1943(1):012113. [doi: 10.1088/1742-6596/1943/1/012113]
46. Zhang X, Wang S, Liu J, Tao C. Towards improving diagnosis of skin diseases by combining deep neural network and human knowledge. BMC Med Inform Decis Mak 2018 Jul 23;18(Suppl 2):59 [FREE Full text] [doi: 10.1186/s12911-018-0631-9] [Medline: 30066649]
47. SIIM-ISIC melanoma classification: Identify melanoma in lesion images. kaggle. URL: https://www.kaggle.com/c/siim-isic-melanoma-classification/leaderboard [accessed 2022-10-14]

Abbreviations

AI: artificial intelligence
AUC: area under the receiving operating characteristic curve
CNN: convolutional neural network
EMA: exponential moving average
FSL: feature supervised loss
ISIC: International Skin Imaging Collaboration

Edited by R Dellavalle; submitted 28.04.22; peer-reviewed by V Singh; comments to author 24.05.22; revised version received 01.09.22; accepted 12.10.22; published 12.12.22

Please cite as:
Zhang X, Xie Z, Xiang Y, Baig I, Kozman M, Stender C, Giancardo L, Tao C
Issues in Melanoma Detection: Semisupervised Deep Learning Algorithm Development via a Combination of Human and Artificial Intelligence
JMIR Dermatol 2022;5(4):e39113
URL: https://derma.jmir.org/2022/4/e39113
doi: 10.2196/39113
PMID: