The role of public challenges and data sets towards algorithm development, trust, and use in clinical practice

Veronica Rotemberg, MD, PhD[1]; Allan Halpern, MD[1]; Steven Dusza, DrPH[1]; and Noel C F Codella, PhD[2]

Abstract

In the past decade, machine learning and artificial intelligence have made significant advancements in pattern analysis, including speech and natural language processing, image recognition, object detection, facial recognition, and action categorization. Indeed, in many of these applications, accuracy has reached or exceeded human levels of performance. Subsequently, a multitude of studies have begun to examine the application of these technologies to health care, and in particular, medical image analysis. Perhaps the most difficult subdomain involves skin imaging because of the lack of standards around imaging hardware, technique, color, and lighting conditions. In addition, unlike radiological images, skin image appearance can be significantly affected by skin tone as well as the broad range of diseases. Furthermore, automated algorithm development relies on large high-quality annotated image data sets that incorporate the breadth of this circumstantial and diagnostic variety. These issues, in combination with unique complexities regarding integrating artificial intelligence systems into a clinical workflow, have led to difficulty in using these systems to improve sensitivity and specificity of skin diagnostics in health care networks around the world. In this article, we summarize recent advancements in machine learning, with a focused perspective on the role of public challenges and data sets on the progression of these technologies in skin imaging. In addition, we highlight the remaining hurdles toward effective implementation of technologies to the clinical workflow and discuss how public challenges and data sets can catalyze the development of solutions.

Semin Cutan Med Surg 38:E38-E42 © 2019 Frontline Medical Communications

Machine learning and artificial intelligence (AI) are quickly permeating and transforming industries and life. Some leaders in the field have drawn the analogy between machine learning and the invention of electric power.[1] Similar to electric power, machine learning has enormous potential not only to improve lives but also to cause harm if it is not properly and safely used. Therefore, a solid understanding of how machine learning works; how it is constructed, trained, and implemented; as well as where and how it can "break" is critical to maximize its beneficial impact on society.

Within dermatology, machine learning provides a unique opportunity to improve diagnostic accuracy, reduce unnecessary biopsies, and provide access to improved care in low-resource areas. However, there remain some significant issues regarding design, training, and integration that still limit its deployment to clinical workflows.

In this article, we summarize some of the most important recent technological advancements in machine learning and how they may be relevant to dermatology. We discuss at a high level how machine learning functions, as well as issues that remain to be solved toward implementation for the clinical dermatological workflow. Public challenges and data sets are an important tool that can help address those issues and provide recommendations for the focus of future research.

In the second section, we will review background of machine learning, and in particular deep learning, lay out the typical design process for applied machine learning systems, and then summarize existing benchmarks, particularly in the field of skin image analysis. In the third section, we will discuss the existing clinical implementation issues and how public benchmarks and data sets are uniquely well suited to help the larger technical and clinical communities address them.

Background

Recent advances in machine learning

Machine learning and AI systems have undergone rapid advancements over the last decade, particularly in applications centered around pattern analysis and recognition, such as image recognition, speech recognition, natural language processing, and others. The 3 primary drivers for the advancement have been (1) the development of large, well-curated, and annotated data sets for training these systems (such as ImageNet[2]); (2) improvements in a subfield of AI and machine learning referred to as "deep learning"; and (3) progression of computer hardware, specifically Graphics Processing Units, that can accelerate deep learning methods as well as open frameworks and application program interfaces to implement deep learning systems on this hardware.

Deep learning is a class of machine learning methods that uses mathematical constructs to approximate a subset of biological neuron functions to create networks of artificial neurons, referred to as "neural networks." At the core, a single artificial neuron consists of a weighted sum of inputs, followed by a non-linear function, such as a sigmoid, hyperbolic tangent, or other piece-wise functions. Networks are then composed of layers of artificial neurons interconnected in a multitude of ways. During training, the weights over the inputs of each of the artificial neurons are adapted to minimize the error of an "objective function," which is meant to represent the system performance for some output task (ie, classification error, or segmentation error). The exact topology, or structural layout, of the neural network can be changed and is commonly defined by a

[1]Dermatology Service, Memorial Sloan-Kettering Cancer Center, New York, NY.
[2]IBM Research AI, T.J. Watson Research Center, Yorktown Heights, NY.
Disclosures: The authors have nothing to disclose.
Correspondence: Noel C F Codella, PhD, nccodell@us.ibm.com

1085-5629/13$-see front matter © 2019 Frontline Medical Communications
DOI: 10.12788/j.sder.2019.013

human engineer or machine learning specialist. However, in more recent years, another class of machine learning methods to automatically learn optimal network topologies has begun to take root, referred to as "meta-learning" or "learning-to-learn."[3] These have become implicitly necessary components of all top-performing systems in modern competitions.

An important characteristic of deep learning has been its robust ability to generalize: the same network topology that can be used to recognize categories of objects, people, animals, and scenes can be applied to recognize defects in manufacturing,[4] patterns from microscopy, or diseases in other medical images.[5] In addition, neural networks that have first been trained on a source task, such as to recognize entities in natural photographs, and then refined to a new target task, such as recognition of melanoma, tend to perform better than neural networks that have been trained on a single target task alone.[6]

Machine learning applications design process

For any engineering task calling for development of an applied deep learning system, 3 primary groups of complimentary design questions come to the forefront of the process:

1. **What are the input data?** How much of these data are available for training, and how much for evaluation (testing the resulting trained neural network)? What is necessary to ensure the most comprehensive representation of the spectrum of possible inputs the system may encounter in practice?
2. **What are the output data?** What is the network being asked to predict? Have the input data been thoroughly annotated with this output? Are these data deterministic, or is there subjectivity inherent that may be important to quantify?
3. **What network architecture and objective functions will best solve the task?** Will the network architecture learn the task well? Has a proper objective function been selected that will appropriately optimize the best measurement of success? Should the network weights be first pre-trained on another data set?

The cross product of the state space of answers to each of these groups of design questions is effectively infinite; for example, there is no limit to the variety of data sets that can be used to train and test a system and the approaches used to model the task. This makes the design and implementation of deep learning systems incredibly complex; care must be taken to ensure proper development and testing before use in clinical practice.

Importance of testing and evaluation

With the advantages of an unrestricted number of ways to create an AI system to solve a particular task come difficulties in selecting the optimal system. Similar difficulties are observed in other production and educational domains. In the case of health care workers, there are theoretically many ways to educate a doctor, and infinitely many doctors that can be produced. Educational and professional standards, including specialized schools and exams, have been set in place to help curb the number of variations and help enforce a baseline level of quality and safety that can be characterized and understood. In the domain of machine learning development, thoroughly designed standards and benchmarks are also crucial to building and testing AI systems. A comprehensive understanding of the myriad of performance metrics, including various measurements of error, efficiency, integration, dependencies on subpopulations, and more, is critical to controlling the quality of productized technologies and maximizing societal benefit while minimizing potential for harm.

In the third section, we will review the current challenges toward clinical implementation and how public data sets and challenges address these concerns. In particular, public resources for machine learning allow transparency for analysis of training and development biases, an ability to increase diversity of data sets, as well as an increased number of development participants.

Existing medical imaging benchmarks

There has been an increasing opportunity for AI applications in medical imaging, as the advent of electronic medical records and image storage generates ample data that can be used for algorithm development, training, and testing. Radiology has often been the first application area for many types of image analysis approaches, including feature extraction,[7,8] lesion segmentation,[9] and diagnostic classification.[10] Indeed, grand-challenge.org, a website index dedicated to listing public benchmarks in medical imaging, now has over 177 entries on variations of these tasks.

Dermatology provides a natural opportunity for machine learning and computer vision application as well, since clinical images form the basis of dermatology training and practice and are likewise based on visible light photographs. Recent papers have achieved dermatologist-level accuracy for diagnosis of melanoma, highlighting the potential for cutaneous application of AI.[11,12] However, the results of these works do not imply that the developed systems are ready for clinical integration. Most importantly, the data sets used for study do not comprehensively reflect the entire population in which the system would be applied. In addition, many algorithms have been developed and tested with proprietary data sets, which reduces the ability of potential users and consumers to compare automated approaches with one another.

The contribution of public challenges and data sets to inter-algorithm comparison and benchmarking cannot be understated. Benchmarking against validated data sets and metrics is critically important to determine model selection and predict appropriate arenas for clinical implementation.[13] In other fields, large image repositories, such as those available on Kaggle,[14] provide data sets against which to score algorithms and compare them to each other. The public challenges and data sets available for medical image analysis have advanced the field toward more complex problems. Over time, increased number of categories for diagnosis problems as well as multidimensional feature extraction, including the addition of metadata, have become fundamental to these medical imaging problems.

However, the dermatology case in particular is complicated by a lack of coherent guidelines regarding metadata, photographic acquisition technique, image annotation, and what data constitute protected health information. To address the lack of standards in dermatology, the International Skin Imaging Collaboration (ISIC) was formed in 2013. In addition to applying consensus to help establish standards in the field, ISIC hosts the ISIC Archive and annual ISIC Challenges to support the application of computer science to automated melanoma diagnosis. The ISIC Archive[15] is an open-source platform that contains over 23,000 creative-commons licensed images, including over 3,000 melanomas. While there are other public dermatology atlases, the ISIC Archive is, by far, the largest of its type.[16] Because of its public nature, it provides an optimal resource for developers to test automated algorithms and for comparison between AI systems. Unlike other online resources, the ISIC Archive is currently focused on dermoscopy, a dermatologic magnification tool that has been shown to improve diagnostic accuracy for melanoma.[17-19] It is also primarily devoted to lesion images for the purpose of melanoma diagnosis, with currently few inflammatory disease or clinical image examples.

ISIC has sponsored 3 annual grand challenges[20-22] since 2016 in order to provide standardized images and a public online system for evaluation of the accuracy of submitted algorithms. The challenges have included image analysis tasks of segmentation, feature detection, and diagnosis components. In the most recent year, teams were asked to identify 8 different diagnostic classes, including benign lesions, as well as multiple types of malignancy, including basal cell carcinoma and melanoma. This increased granularity is expected to be necessary for clinical application, providing the basis for accuracy evaluation in multiclass problems for dermatology and elsewhere. Over the past 3 years, the challenges have grown in size, complexity, participation, and algorithm performance.

Clinical implementation issues

Clinical implementation problem definition

In addition to quantitative comparative metrics, more qualitative issues arise, specifically in medical applications, regarding integration into the clinical workflow. In particular, something as simple as the user being aware of malfunction is lost due to the complexity of machine learning and AI, an aspect that is not necessarily observed in other domains. For example, if a tire loses its grip in harsh conditions, the driver is immediately aware of the malfunction. Or if a camera malfunctions, the images would become obviously compromised. In both circumstances, the integrity of function can be easily validated. However, AI systems, as they are commonly designed today, have no mechanisms to alert the user of a malfunction.

This "black box" nature of machine learning approaches in dermatology causes the largest barrier to widespread implementation. Without a general understanding of where systems might fail and without communication from the system itself, clinicians might be hesitant to include these tools in their clinical decision-making, and with good reason.

The full set of conditions that lead to failures of AIs system is complex. However, they are commonly tied to artifacts of the training data, usually centered around underrepresentation, as well as algorithmic inability to compensate for deficiencies of training data, or to recognize new data that falls outside of the training distribution.

For example, this type of failure occurs when an AI system is shown a data sample for a disease in which the system has not been trained. In this circumstance, the system may output a positive confidence of a disease class it has been trained on, even though that answer would be incorrect for the as-yet-unseen diagnosis.

In order to safeguard against failures, empower the user to recognize failure, and facilitate trust, there must be improvements to 4 key areas of system design described in the following sections and in Figure 1.

Issues and recommendations for data sets

Demographics

The ideal training scenario for machine learning approach to dermatology includes comprehensive representation of data from an international population. Human demographics are extraordinarily broad and impacted by a variety of factors, some that are readily quantified (such as age, sex, nationality), and some that are not as easily quantified (such as degree of sunlight exposure, diet, exercise, labor environment, etc). Without broad and sufficient representation, algorithms developed within a particular data set may be prone to to inconsistent diagnostic quality,[23] potentially leading to significant harm, especially in population subgroups with low disease prevalence.[24] In addition to concerns about bias, the cross product of all possible demographic scenarios is not feasible to fully represent; therefore, a public repository is necessary to allow quantification of the represented training space and enable measurement of deviations from that space in practice. The availability of an open, public, centralized repository, both for the purposes of training as well as transparent and thorough testing and validation, is critical.

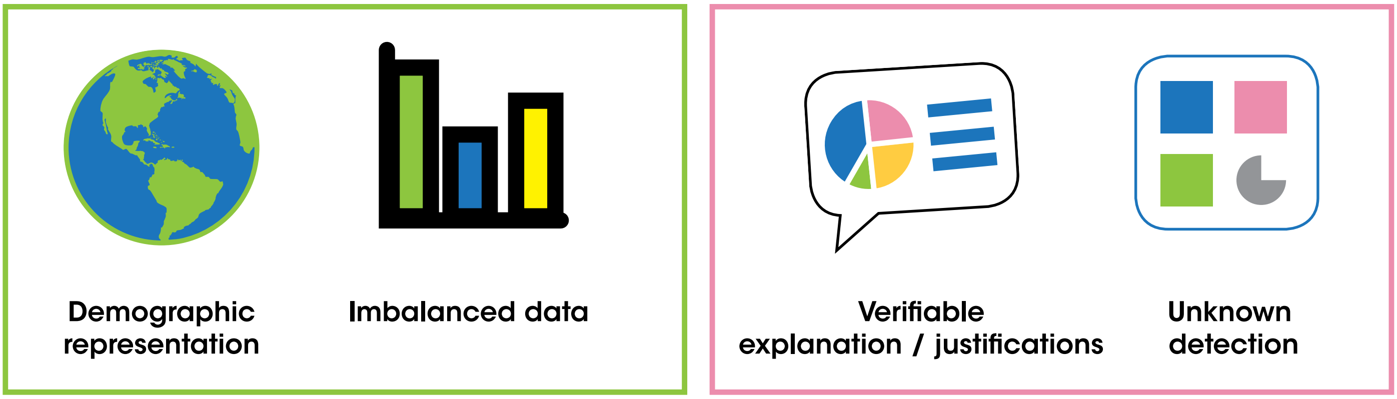
Demographic representation
Imbalanced data
Verifiable explanation / justifications
Unknown detection

FIGURE 1. Four categories of improvement in automated clinical decision-support systems that can be addressed by public challenges and data sets. These are subdivided into challenges with input data (left, green) and with algorithm design (right, red).

In addition to the enhanced assessments of algorithms that can be performed, a public archive such as ISIC also provides necessary tools for increasing demographic representation.[15] While difficulties in organizing public data sets exist for dermatology imaging, such as the ethical and legal requirements to protect patients' privacy, these are worthwhile to overcome for the clear advantages that publicly available images provide. An easily accessible public interface allows clinicians from all over the world to contribute their own images and subsequently grow and improve demographic representation for machine learning development in an organic and exponential fashion.

Imbalanced data

In addition to challenges with respect to demographic representation, breadth of an image archive used for training and testing is challenging to achieve in a balanced manner. It is inevitable that uncommon diagnoses, for example melanoma, will be overrepresented in a cutaneous image archive, as they will be more commonly photographed in clinical practice. Imbalances may come from all types of variances in distribution of human stratified metadata, such as disease state, clinical attributes, age, gender, etc. Some may be unquantified, ie, there may be more "rulers" present in nevi versus melanomas. The proper correction of artefactual prior distributions in training data sets is still being researched but, as with demographic representation, relies on the public nature of the images used for training and testing to continue to develop appropriate techniques. Public challenges can also contribute to corrections for this imbalance by providing state-of-the-art metrics for scoring and measuring diagnostic performance, such balanced multiclass accuracy, which was used in the 2018 ISIC Challenge for melanoma diagnosis.[13,25]

Issues and recommendations for algorithms

There are various implementation issues that relate to transparency and interpretability of AI algorithms for dermatology application. Public challenges are especially valuable for these types of hurdles because they can target data sets and scoring toward particular implementation approaches and define problems very specifically toward advancing the field forward.[26]

Explanations/justification

Clinicians will require an understanding about the factors that influence AI output, especially for lesion classification tasks in dermatology, in the form of an interpretable explanation or justification of a system decision. Such insight will help users catch aspects of images such as lighting, a ruler, or other features that could contribute towards erroneous classification. Explanations may also provide insight into signals that an AI system relies on that may relate to imperfections in the data set and have no bearing in logic or reason (lack of reason regardless of correctness of decision).

This requirement contributes to the development of interpretable or explainable AI systems that provide evidence or justification for decisions and recommendations in order to enable clinical staff of appropriate skill to verify the validity of decisions. The goal is to encode features that enable human operators to have some better-than-random performance at catching instances where the system prediction might be incorrect (for any reason). Some examples of features that are helpful toward this process include saliency maps and other methods that can lead to improved transparency overall of automated clinical decision-support systems.[27-30]

Public challenges and data sets can contribute to the development of explanations as the data sets can be annotated to include salient features used for diagnosis and understandable by clinicians.

Unknown detection

The diagnostic breadth in dermatology is unparalleled in other medical disciplines.[31] In order to ensure clinical utility, the development of AI systems that are able to recognize lesions or disease states for which either sufficient training data were not available to render a reliable decision or for which no training data exist at all will be necessary, especially for applications to low-resource areas where as-yet-unrecorded diagnoses may be more common. In the machine learning literature, this is referred to as "Out-of-Distribution" detection.[32] Because current machine learning systems by nature tend to misclassify out-of-distribution images, alternative approaches for testing and design are needed. This includes holding out some subset of diagnostic conditions from training data sets but including those same subsets in test data sets and measuring how well algorithms can detect those disease conditions for which they have not been trained.

Conclusion

Machine learning has recently undergone rapid advancements, leading to broad study of the application of these technologies to medical imaging. Public benchmarks provided by challenges and data sets have played a key role in improving comparisons between systems and setting clinically relevant standards. However, several issues remain that impede its deployment to clinical practice. These include a lack of comprehensive demographic representation and imbalance in existing data sets, as well as a lack of prediction explanation or justification, and detection of unknown disease states in automated systems. We have outlined how public resources are uniquely suited to address these remaining challenges as well as discussed several strategies to maximize their impact. The ISIC Archive has implemented many of these strategies for the use cases in dermoscopic imaging, but the concepts generalize across all medical imaging domains.

References

1. Li O. Artificial Intelligence is the New Electricity—Andrew Ng. https://medium.com/syncedreview/artificial-intelligence-is-the-new-electricity-andrew-ng-cc132ea6264. Published 2017. Accessed December 10, 2018.
2. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. Paper presented at: 2009 IEEE Conference on Computer Vision and Pattern Recognition; June 20–25, 2009; Miami, FL, pp. 248–255.
3. Finn C, Abbeel P, Levine S. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. https://arxiv.org/abs/1703.03400. 2017.
4. Wang J, Ma Y, Zhang L, Gao RX, Wu D. Deep learning for smart manufacturing: Methods and applications. *J Manuf Syst*. 2018;48:144–156. doi:10.1016/j.jmsy.2018.01.003
5. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42:60–88. doi:10.1016/j.media.2017.07.005.
6. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? https://arxiv.org/abs/1411.1792. 2014.

7. Banerjee I, Beaulieu CF, Rubin DL. Computerized Prediction of Radiological Observations Based on Quantitative Feature Analysis: Initial Experience in Liver Lesions. *J Digit Imaging*. 2017;30(4):506–518. doi:10.1007/s10278-017-9987-0.
8. Zhou C, Chan HP, Wei J, Hadjiiski LM, Chughtai A, Kazerooni EA. Quantitative analysis of CT attenuation distribution patterns of nodule components for pathologic categorization of lung nodules. *Proc SPIE Med Imaging*. 2017;10134. doi:10.1117/12.2254155.
9. Lee H, Troschel FM, Tajmir S, et al. Pixel-Level Deep Segmentation: Artificial Intelligence Quantifies Muscle on Computed Tomography for Body Morphometric Analysis. *J Digit Imaging.* 2017;30(4):487–498. doi:10.1007/s10278-017-9988-z.
10. Lindsey R, Daluiski A, Chopra S, et al. Deep neural network improves fracture detection by clinicians. *Proc Natl Acad Sci USA*. 2018;115(45):11591–11596. doi:10.1073/pnas.1806905115.
11. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115–118. doi:10.1038/nature21056.
12. Codella N, Nguyen QB, Pankanti S, et al. Deep Learning Ensembles for Melanoma Recognition in Dermoscopy Images. https://arxiv.org/abs/1610.04662. 2016.
13. Olson RS, La Cava W, Orzechowski P, Urbanowicz RJ, Moore JH. PMLB: a large benchmark suite for machine learning evaluation and comparison. *BioData Min*. 2017;10(1):36. doi:10.1186/s13040-017-0154-4.
14. Kaggle: your home for data management. https://www.kaggle.com. Accessed December 1, 2018.
15. The International Skin Imaging Collaboration (ISIC). https://www.isic-archive.com/. Accessed September 20, 2018.
16. DermNet NZ. https://http://www.dermnetnz.org/about-us. Accessed November 26, 2018.
17. Kittler H, Marghoob AA, Argenziano G, et al. Standardization of terminology in dermoscopy/dermatoscopy: Results of the third consensus conference of the International Society of Dermoscopy. *J Am Acad Dermatol*. 2016;74(6):1093–1106. doi:10.1016/j.jaad.2015.12.038.
18. Bafounta ML, Beauchet A, Aegerter P, Saiag P. Is dermoscopy (epiluminescence microscopy) useful for the diagnosis of melanoma? Results of a meta-analysis using techniques adapted to the evaluation of diagnostic tests. *Arch Dermatol*. 2001;137(10):1343–1350.
19. Vestergaard ME, Macaskill P, Holt PE, Menzies SW. 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*. 2008;159(3):669–676. doi: 10.1111/j.1365-2133.2008.08713.x.
20. Marchetti MA, Codella NCF, Dusza SW, 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. *J Am Acad Dermatol.* 2018;78(2):270–277.e1. doi:10.1016/j.jaad.2017.08.016.
21. Codella N, Gutman D, Celebi ME, et al. Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC). https://arxiv.org/abs/1710.05006. 2017.
22. Gutman D, Codella N, Celebi E, et al. Skin Lesion Analysis toward Melanoma Detection: A Challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC). https://arxiv.org/abs/1605.01397. 2016.
23. Adamson AS, Smith A. Machine learning and health care disparities in dermatology. *JAMA Dermatol*. 2018;154(11):1247–1248. doi:10.1001/jamadermatol.2018.2348.
24. Marchetti MA, Chung E, Halpern AC. Screening for Acral Lentiginous Melanoma in Dark-Skinned Individuals. *JAMA Dermatol.* 2015;151(10):1055-1056. doi:10.1001/jamadermatol.2015.1347
25. Buda M, Maki A, Mazurowski MA. A systematic study of the class imbalance problem in convolutional neural networks. https://arxiv.org/abs/1710.05381. 2017.
26. Kohli MD, Summers RM, Geis JR. Medical Image Data and Datasets in the Era of Machine Learning-Whitepaper from the 2016 C-MIMI Meeting Dataset Session. *J Digit Imaging*. 2017;30(4):392–399.
27. Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: Addressing ethical challenges. *PLoS Med*. 2018;15(11):e1002689. doi: 10.1371/journal.pmed.1002689.
28. Codella N, Lin CC, Halpern A, Hind M, Feris R, Smith JR. Collaborative Human-AI (CHAI): Evidence-Based Interpretable Melanoma Classification in Dermoscopic Images. Paper presented at: First International Workshops, MLCN 2018, DLF 2018, and iMIMIC 2018, Held in Conjunction with MICCAI 2018; September 16–20, 2018; Granada, Spain. https://arxiv.org/abs/1805.12234.
29. Sadeghi M, Chilana PK, Atkins MS. *How Users Perceive Content-Based Image Retrieval for Identifying Skin Images*. in *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*. Cham, Switzerland: Springer International Publishing; 2018.
30. Ge Z, Demyanov S, Chakravorty R, Bowling A, Garnavi R. Skin Disease Recognition Using Deep Saliency Features and Multimodal Learning of Dermoscopy and Clinical Images. In: Descoteaux M, Maier-Hein L, Franz A, Jannin P, Collins D, Duchesne S, eds. *Medical Image Computing and Computer-Assisted Intervention − MICCAI 2017.* Cham, Switzerland: Springer International Publishing; 2017; 2017:250–258.
31. Bolognia J, Jorizzo J, Schaffer J. *Dermatology*. Philadelphia, PA: Elsevier Saunders; 2012.
32. Hendrycks D, Gimpel K. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. https://arxiv.org/abs/1610.02136. 2016.