AI-based detection of erythema migrans and disambiguation against other skin lesions

Philippe M. Burlina, , Neil J. Joshi, Phil A. Mathew, William Paul, Alison W. Rebman, John N. Aucott

PII: S0010-4825(20)30308-5

DOI: https://doi.org/10.1016/j.compbiomed.2020.103977

Reference: CBM 103977

To appear in: *Computers in Biology and Medicine*

Received Date: 24 March 2020

Revised Date: 14 August 2020

Accepted Date: 15 August 2020

Please cite this article as: P.M. Burlina, N.J. Joshi, P.A. Mathew, W. Paul, A.W. Rebman, J.N. Aucott, AI-based detection of erythema migrans and disambiguation against other skin lesions, *Computers in Biology and Medicine* (2020), doi: https://doi.org/10.1016/j.compbiomed.2020.103977.

AI-BASED DETECTION OF ERYTHEMA MIGRANS AND DISAMBIGUATION AGAINST OTHER SKIN LESIONS

Philippe M. Burlina, PHD[1,2], Neil J. Joshi, BS[1], Phil A. Mathew[1],
William Paul[1], Alison W. Rebman, MPH[3], John N. Aucott, MD[3]

[1]Applied Physics Laboratory, Johns Hopkins University

[2]Malone Center for Engineering in Healthcare, Johns Hopkins University

[3]Johns Hopkins Lyme Disease Research Center, Division of Rheumatology, Department of Medicine, Johns Hopkins University School of Medicine

ABSTRACT

This study examines the use of AI methods and deep learning (DL) for prescreening skin lesions and detecting the characteristic erythema migrans rash of acute Lyme disease. Accurate identification of erythema migrans allows for early diagnosis and treatment, which avoids the potential for later neurologic, rheumatologic, and cardiac complications of Lyme disease. We develop and test several deep learning models for detecting erythema migrans versus several other clinically relevant skin conditions, including cellulitis, tinea corporis, herpes zoster, erythema multiforme, lesions due to tick bites and insect bites, as well as non-pathogenic normal skin. We consider a set of clinically-relevant binary and multiclass classification problems of increasing complexity. We train the DL models on a combination of publicly available images and test on public as well as images obtained in the clinical setting. We report performance metrics that measure agreement with a gold standard, as well as a receiver operating characteristic curve and associated area under the curve. On public images, we find that the DL system has an accuracy ranging from 71.58% (and 95% error margin equal to 3.77%) for an 8-class problem of EM versus 7 other classes including other skin pathologies, insect bites

and normal skin, to 94.23% (3.66%) for a binary problem of EM vs. non-pathological skin. On clinical images of affected individuals, the DL system has a sensitivity of 88.55% (2.39%). These results suggest that a DL system can help in prescreening and referring individuals to physicians for earlier diagnosis and treatment, in the presence of clinically relevant confusers, thereby reducing further complications and morbidity.

Introduction

Lyme disease is the most common tick-borne disease in the northern hemisphere, with an estimated 300,000 new cases per year in the United States alone.[1–3] *Borrelia burgdorferi,* the bacterial agent of Lyme disease in North America, is inoculated into the skin through the bite of an infected tick. Between 3 and 30 days later, a round or oval, red, centrifugally expanding skin lesion called erythema migrans (EM) appears in approximately 70-80% of cases.[4,5] EM can present in acute Lyme disease with or without the presence of concomitant flu-like symptoms such as fever, fatigue, myalgia, and arthralgia. Without appropriate antibiotic treatment, EM can persist for several weeks before resolving spontaneously as the host immune response is elicited leaving no cutaneous evidence of the persistent infection.[4]

In early, uncomplicated Lyme disease, treatment with the appropriate oral antibiotics is highly effective at both rapidly resolving the EM lesion and preventing potentially devastating long-term complications.[6,7] If not diagnosed and treated, *Borrelia burgdorferi* infection can persist, advancing from a skin-limited disease to dissemination of the bacteria into the nervous, cardiac, and rheumatologic systems.[7] Consequently, accurate recognition of EM by both patients and clinicians is crucial to early diagnosis and prompt initiation of appropriate treatment. However, timely recognition of EM is often hampered by several factors.

First, although visual identification of EM, the presence of associated symptoms, a history of potential exposure to ticks, and epidemiologic risk remain the primary criteria for diagnosis of early Lyme disease, physicians often continue to rely on serologic test results.[8,9] Antibody based blood tests which are currently available to clinicians are not recommended for diagnosis during the early phase of infection when EM is most likely to be present, due to their low sensitivity in this acute phase of the illness (less than or equal to 40%). Direct detection of *Borrelia burgdorferi* by culture or PCR of blood or skin biopsy samples can be performed, but such tests are generally available only in research settings. In addition, they are not always practical for use by diagnosing clinicians given the extended processing time for results.[10]

Secondly, EM identification remains a challenge because it often takes on a variety of appearances.[11] Notably, only 20% of patients with EM in the United States present with lesions that have the central clearing of a classic target lesion ("ring-within-a-ring" or "bull's eye").[12] The lack of a stereotypical appearance may lead clinicians to make diagnostic errors due to over-reliance on pattern recognition and the assumption that all EM look like the classic target lesion. While the efficiencies of medical practice require the use of heuristics to make efficient clinical diagnosis, such thinking is prone to cognitive biases and error[13]. In contrast to a bull's eye rash, the majority of EM lesions appear uniformly red or bluish-red in color and lack central clearing.[6,12] This may lead to misdiagnosis of cellulitis, another bacterial infection of the skin, that is treated with different types of antibiotics.[14] Antibiotics typically used for cellulitis do not have optimal activity in Lyme disease.[15] Additionally, approximately 4-8% of EM have central blistering, which may lead to a misdiagnosis of conditions such as herpes zoster (HZ).[16]

Finally, approximately 20% of patients have multiple EM at the time of diagnosis, arising from hematogenous dissemination of the bacteria to other areas of the skin. This can be confused with erythema multiforme, urticaria, erythema annulare centrifugum, or other annular skin disorders, resulting in the potential for both over and under-diagnosis of early Lyme disease.[17]

The complexity of the presentations of single and multiple EMs is underappreciated by both patients and general practice clinicians. Among the general public, one internet-based survey found that respondents correctly identified a classic EM with central clearing 73% of the time. In contrast, non-classic skin lesions with vesicular, uniformly red, bluish-purple, or disseminated manifestations were identified less than 30% of the time.[18] Even general practitioners have difficulty recognizing EM. In one study, general practitioners correctly identified non-target EM lesions 64% of the time and classic target lesions 80% of the time.[19]

Across many disciplines within medicine, including dermatology, there has been increased interest in harnessing artificial intelligence and deep learning (DL) to assist doctors with individuals who may have to be routinely monitored (e.g., for skin cancer) and to possibly reduce errors in classification and diagnosis.[20] In the 1990s and early 2000's, image classification in medical image analysis had been largely based on the use of conventional classifiers combined with human-engineered image features.[21,22] In the past years, progress has been made using DL techniques[23–28], and deep convolutional neural networks (DCNNs, for example AlexNet[25] or ResNet[27]). These studies have demonstrated significant increase in image classification performance for computer vision tasks such as classification. Unlike the classical approaches, image features computed via DL techniques are learned from data via an optimization process. This process requires a dataset of images labeled with ground-truth gold standard information.

Applying DL has been successful in part because image features directly learned by neural networks provide a better representation than sub-optimal image features that are hand-designed. Recently, DL approaches have been successfully used for many medical image diagnostic tasks, such as for example skin cancer identification,[29] or diagnosing and estimating prognosis in ophthalmic diseases,[30–33] and have largely replaced classical ML approaches.[34,35]

Application of AI techniques to skin lesions and dermatology have now become an active area of research. Recent examples include work on using semantic segmentation of skin lesions[36]

where a fully convolutional network was coupled with the use of hand-crafted domain specific features. In another study[37] a method for the challenging task of determining lesion borders was developed using a two-stage approach consisting of U-net and a fuzzy edge detection that uses fuzzy intensity features (bright, dark and medium). Much of the work in AI applied to dermatology involves analysis of dermoscopy images and the diagnosis of cancerous skin lesions and melanomas. As one example of this work, a multiclass classification of skin lesions was performed,[38] including lesions such as Actinic Keratosis, Basal Cell Carcinoma, Melanocytic Nevus/Mole, Squamous Cell Carcinoma, Seborrheic Keratosis, Intraepithelial Carcinoma, Pyogenic Granuloma, Hemangioma, Dermatofibroma, and Malignant Melanoma, using pre-trained universal deep features. In another related study,[39] a two-stage method using a very deep residual net used as a fully convolutional network for performing segmentation, combined with other networks for performing classification and was applied to analyze melanomas. Another notable study analyzed melanoma in dermoscopic images using pre-trained networks as well as fine-tuning techniques.[39]

The use of AI for the identification of early Lyme disease with EM – compared to the above cited studies that were mostly focused on skin melanoma and dermoscopic images using AI techniques -- has been relatively less researched. One such study was done using classical machine learning (ML) approaches.[40] By contrast, in a prior study by our group, initial results of AI-based classification were reported, but for simple cases of disease vs confuser detection[41]. The current study's goal is to expand on our prior in work computer-aided EM classification by leveraging DCNNs for a more complex problem, entailing a wider array of confusers and subsequent number of images, and a larger number of clinically collected "gold standard" EM images collected over a wider geographic range which have been used in the analysis[42]. These additions should create a more robust and more generalizable AI-based EM detection algorithm. Machine-based screening of skin lesions for Lyme disease also has the potential to identify a high percentage of both typical and atypical EM. Our ultimate aims are the potential

application of these methods to the prescreening of skin lesions for more informed physician referral, clinical evaluation and early diagnosis of Lyme disease.

Methods

Data

The sources of images available for use in this study consisted of images available online (in the public domain) as well as clinical images of well-documented patients with EM.

Online images were obtained from scripted searches including Google as well as Bing. These programmatic searches were performed using pre-determined search terms that acted as synonyms for a given condition or lesion. For example, “erythema migrans”, “Lyme”, and “bullseye rash” are synonyms for “EM”. These primary search terms were also combined with secondary search terms that were appended to indicate anatomical locations (e.g. “leg”, “face”), or promote diversity for protected factors, e.g. gender and race/ethnicity (e.g. using terms such as “African American” or “female”).

Skin imaging conditions play an important role in classification success. Skin images in this study that were procured from online public sources were, by virtue of this collection, taken under many different acquisition conditions, including changing viewpoints and illumination. Such conditions are commonly referred to as “in the wild” situations, as opposed to more controlled acquisition situations.

We used AI techniques we developed for machine-based removal of full or near duplicates of online images. This was done by encoding images using a deep neural net and computing proximity in the embedded coded domain representation of these images to find close

duplicates. We also did machine-based removal of unwanted images, including inappropriate, irrelevant, or facial images using deep neural net-based classification.

This machine curation of online images was followed by clinician curation. Clinician-based curation included further excluding images where the classification was uncertain based on visual inspection and further vetting by clinicians to remove duplicates. It also entailed carefully annotating the remaining images with moderate to high probability of accurate group classification based on visual appearance using the estimated size of the skin lesions. This annotation for the main pathologies also entailed a more granular annotation for single vs. multiple EM forms of the lesions. There was no use of additional clinical information attached to these images.

Clinical images used in this study consisted of images acquired at Johns Hopkins University and the Lyme Disease Biobank under conditions where the diagnosis of EM and Lyme disease was confirmed using clinical information available at the time of diagnosis. All images were collected under protocols approved by the Johns Hopkins University or the Advarra Institutional Review Boards, and informed consent was obtained from all participants. Clinical images were cropped to exclude clinician markers on the skin and to exclude rulers that may have been placed in the photographs. Clinical images were then annotated by the study clinician to verify the diagnosis of EM.

Types of classes and lesions considered

The classes of lesions that we studied included: 1) e**rythema migrans** (EM) as well as non-EM conditions: 2) n**ormal skin (NO),** and a host of other confuser skin pathologies or lesions induced by insect bites. These confusers included: 3) **herpes zoster** (or HZ), commonly known as shingles; 4) **tinea corporis** (TC), commonly known as ringworm; 5) i**nsect bites** in general (IB);

6) tick bites (including engorged or non-engorged) (IB-T); 7) **cellulitis** (CELL), a form of skin infection that is often due to a breakage in the skin barrier and may entail bacteria including staphylococcus or streptococcus, and 8) **erythema multiforme** (EMU).

We chose these specific skin conditions and pathologies as confusers for three reasons: a) they represent skin lesions that could be confused with EM because of their circular pattern. Some specific lesions were selected because they required different treatment modalities and could lead to erroneous and sometimes deleterious self-diagnosis and self-treatment. This may occur, for example, in the case of tinea corporis which would require an anti-fungal treatment vs. EM -- that requires antibiotic treatment -- vs. cellulitis, that may necessitate types of antibiotic treatments different from those of Lyme disease. b) some skin lesions could have been contracted in situations similar to those regarding Lyme, i.e. other insect bites, arguably resulting from outdoor activities, and tick bites, which may leave a small skin reaction that is often confused with EM, and c) some of the confounding lesions can manifest in a very short time period, as is the case for EM, such as shingles, and may require quick attention. Indeed, antivirals need to be used in the case of shingles, and these are believed to be effective only if initiated within a 48-hour window. This selection of alternate confuser classes opens the door to other uses of our pre-screener to give recommendations to users to have themselves checked by appropriate clinicians in the case of these alternate conditions.

Classification problems considered

In this study we considered several types of classification problems, including 2-class classification (i.e., binary) as well as M-class (i.e., multiclass) classifications.

The binary classification involved NO vs EM, EM vs CELL, NO vs CELL, and EM vs ALL (all other classes). In addition to EM vs ALL, we ran a ternary (3-class) experiment of EM vs NO vs ALL (all

confusers). We also ran a 4-class experiment of EM vs NO vs HZ vs TC. To further investigate performance among confusers, we ran a large 8-class experiment with all classes versus one another.

While our end goal was the binary classification problem for the pre-screening application for EM, the multiclass experiments open the door to other future uses of our system. Also, solving a more granular multiclass problem allows the machine to be "stressed tested" beyond the binary problem since multiclass classification is fundamentally harder. Finally, it has been found that solving a multiclass problem then fusing the results into a binary problem often confers advantages regarding performance, as opposed to directly solving a binary problem.[29]

Algorithmic methods

The principal method of operation for AI-based approaches currently rests on the utilization of deep learning (DL) via DCNNs. DCNNs serve several purposes; to represent the image via features, and to implement the core of the classification logic. DCNNs generate image feature representations at various levels of abstraction, going from coarse to fine scale representational features. Features are computed from convolutions whose filter parameters are learned directly from data, in contrast to conventional medical imaging methods that were used in the 1990s and early 2000s that relied on human-designed features.[21,22] Features computed via the convolutional steps of DCNNs are then processed via fully connected layers that make up the classification logic. A value in a [0,1] interval often interpreted as a probability (but strictly speaking is not a probability) is computed via SoftMax computations for each class label, in the final layer of the network. This value is taken to be the likelihood for the given lesion class. All network weights in the DCNN network are optimized (learned) from the training data. Training data includes images equipped with gold standard clinician curation. Training is done via a backpropagation process making this approach fully data driven.

In this study we used and compared several state-of-the-art (SOTA) DCNNs, including: variants of ResNet including ResNet50 (Figure 10), and ResNet152[27], InceptionV3 and InceptionResNetV2[43,] and finally, DenseNet121[44]. ResNet50 was used as the reference SOTA algorithm and was employed across all problems, and comparisons with the other networks was done on a subset of problems exploring the two following extreme cases: 1) for the two class EM vs ALL and 2) for the 8-class classification problem. ResNet is a state-of-the-art deep network that has so called "bottleneck" and "skip" connections that make the upstream layer activations available to a down-stream layers, conferring important benefits for good representation as well as backpropagation. We used our own custom framework to train and test our models; this framework itself relies on a full software stack that includes Keras as a front end with TensorFlow as a back end. This framework has various functionalities for efficiently partitioning the data and evaluating results at test time. Transfer learning and fine-tuning is a standard practice in DL and was also used in our study. We used weights initially trained on the ImageNet dataset to classify 1,000 different general object classes, and then modified the network to assume the desired number of class outputs for our skin classification problems. SGD (stochastic gradient descent) with Nesterov momentum=0.9 was used, and the initial learning rate was set to 1E-3.The number of epochs was dictated by patience-based early stopping, in concert with early stopping, to stop training by which training was stopped after 5 epochs of no improvement evaluated on the validation set performance. For the two class EM normal versus EM classification this resulted for example in stopping after 8 epochs and for the 8-class classification training stopped after 9 epochs. After training completion, the model weights yielding the highest performance on the validation set were then used for further testing.

The loss function was a cross entropy loss. Dynamic learning rate scheduling was used, i.e. we multiplied the learning rate by 0.5 when the training loss did not improve for 5 epochs. A batch size of 32 was used. Image preprocessing included rescaling and mean (ImageNet image)

subtraction. Images were resized to conform to ResNet50 input size I.e. 224x224. Data augmentation for the images was also applied and consisted of horizontal flipping, changes to saturation, brightness, contrast, and color balance, as well as random noise and image compression to further expand and generalize the dataset.

Experiments and performance evaluation

Machine performance was done by comparing against the gold standard annotations that were obtained from clinicians in this study and were reported in earlier sections. The experiments we report are for the binary and multiclass test cases described earlier.

For each experiment, in this study, we trained, tested, and reported results on the online data described earlier. We also tested with clinical EM-only data. The online extraction of public domain images generated an initial set of images which were then curated by machine and clinician and split into train/validation/testing subsets (70%, 10%, and 20%, respectively, of the full data). The training/validation/testing datasets were identical for all like-kind classification problems. We also reported results on the clinical set of EM images.

Performance metrics are reported in Table 1 for ResNet50 and for all problems. The table includes Accuracy, Sensitivity/Recall, Specificity, Positive Predictive Value (PPV/Precision), Negative Predictive Value (NPV), Unweighted Kappa score, F1 Score, Average Precision, and Area under the ROC Curve (AUC). Figure 1 shows example images of the different conditions we considered. Figures 2 through 8 also report confusion matrices for all the different problems.

ROC curves are reported for all the 2-class classification problems for ResNet50 in Figures 9 through 12.

We compared ResNet50 to other DCNNs for two boundary problems including the two class EM vs ALL in Table 2 and for the 8-class problem in Table 3. Table 4 is a characteristic table showing the number of image used for training and testing for each class and shows a comparison of our dataset with the SD-198[45] dataset.

In aggregate, results show promising performance: when testing on public images, we found an accuracy ranging from 71.58% (and 95% error margin equal to 3.77%) for an 8-class problem of EM versus 7 other classes including pathologies, insect bites and normal skin, to 94.23% (3.66%) for a binary problem of EM vs. non-pathological skin. On clinical images of affected individuals, we found a sensitivity of 88.55% (2.39%). Results from Table 2 and Table 3 also show that the results comparing different SOTA networks are – in aggregate -- proving to be within confidence bounds of each other, with the exception of MobileNetV2 which performs much worse for the 8 class problem. More analysis is done in the next section.

Discussion

DL has proven to be very effective for medical imaging and diagnostics and has often achieved performance on par with humans when large datasets are available along with the presence of "gold standard" ground truth annotations. Factors which have made DL effective computationally include the use of graphics processing units during DCNN training time, along with various algorithmic improvements. Inference time computations can be achieved on more modest low power and small form factor devices such as smartphones, which was one of the motivations of our study. Several studies have demonstrated the utility of using DL for medical

diagnostics including for skin. In a prior study we also demonstrated an initial application of DL for Erythema MIgrans analysis, and showed that it could be performed with acceptable sensitivity and specificity[41]. In this study, we significantly expanded the analysis to include more complex use cases entailing EM and 7 other targeted and clinically relevant types of commonly encountered skin conditions that may be confused with EM, resulting in 7 types of classification problems ranging from binary to 8-class classification. In addition, we increased the size of our clinically obtained "gold standard" EM skin lesions of Lyme disease. A very recent paper by Liu et al showed the potential for DL in the general diagnosis of dermatologic conditions[46]. In comparison, our paper remains novel, as it examines primarily acute dermatologic conditions, including manifestations of several acute infectious diseases such as EM, cellulitis, and herpes zoster, which are not the most commonly seen skin lesions in a general dermatology practice setting. Furthermore, our work included the use of images obtained with cell phones or other devices 'in the wild', which are more representative of images that patients would be more likely to generate themselves.

Regarding reliability of the annotations, we did not use the labels obtained online, and all images were re-annotated by a single clinician (JA). We also assessed inter-operator error using another clinician. We had a second clinician perform annotations to compute inter-operator error for EM, tinea corporis, tick bites, cellulitis, and erythema multiforme. We reported these results in Table 3. The error rates ranged from 18.22% (with a 95% confidence interval of 4.61%) for IB-T, to 32.96% (5.61%) for EM. Interpreting those results, the discrepancy between the two clinicians, especially regarding EM, suggests that the high degree of variability in the appearance of skin lesions and additional supporting clinical information may play a significant role in accurate identification of EM by practitioners.

The applications to these problems suggest promising performance. In particular, we observe several interesting findings from these results (looking here at the prototypical performance of the SOTA network ResNet50): Our model performance is very strong in differentiating normal

skin images from either erythema migrans or cellulitis images, and performance remains good when adding in confusers to the mix. Results for the 2-class problem show promising performance. Specifically, we found an accuracy for the problem of normal skin vs. erythema migrans of 94.23 (and 95% confidence interval of 3.66%) for testing on online images, and a sensitivity of 88.55% (2.39%) when testing on erythema migrans-only clinical images. There was a small decrease in performance when training on online and testing on clinical images which is to be expected but it was a graceful degradation considering that this case tested a more difficult situation of domain shift which we also intend to address in a more specific study in the future. Considering this, the performance on clinical images is promising and would be enhanced by training on a mix of online and clinical images when more clinical images become available in the future. For other 2-class problems, we found an accuracy for erythema migrans vs all of 81.51% (6.30%), and for erythema migrans vs cellulitis an accuracy of 79.72% (6.59%). For erythema migrans vs. all we provide in Fig. 9 examples chosen randomly of correct and incorrect prediction, and in that figure, we explain the likely causes of misclassifications. The 2-class normal vs cellulitis gave an accuracy of 95.57% (3.21%). Those results are encouraging considering that erythema migrans and cellulitis could be confused in some circumstances, for instance if image resolution was low, because of the common features of erythema and regional vs. generalized skin involvement . For the 3-class problem of erythema migrans vs normal vs all other classes, we found an accuracy of 83.11% (4.96%).

In addition, when observing the confusion matrix for the 4-class problem, we notice erythema migrans is most likely to be confused with tinea corporis (16% of cases), perhaps because the two skin conditions have a similar round appearance, often with an area of central clearing. Similarly, tinea corporis is most likely to be confused with erythema migrans in 12% of cases. The ‘scaly’ appearance of the advancing border that distinguishes tine corporis may not always be captured in images collected by patients in the wild. In addition, the DL algorithms that we have utilized do not include supporting clinical information in the training or analysis phase,

which would be expected to decrease the performance.[46] Finally, for the 8-class experiment's confusion matrix, it appears insect bites can be confounded with erythema migrans as well (13% of the cases), which should also be expected given the circular nature of the erythema in both conditions. The distinguishing feature of erythema migrans and insect bites in the clinical setting is the ability to observe the expanding size of erythema migrans skin lesion over time which was not possible in the single images that we analyzed. We believe that the loss of specificity in this situation is not clinically harmful however, as the false positive identification of IB as EM would only result in referral for further evaluation and confirmation of the correct diagnosis. This is preferable to the false negative result of calling erythema migrans an insect bite, a scenario which occurred 12% of the time, and could lead to a lost opportunity for further evaluation and treatment. This concern may justify a need to use our detection algorithm at an operating point in the ROC curve with higher sensitivity.

The study also does a sensitivity analysis with regard to network used. For the two classes all performance metrics show results that are generally within confidence bounds suggesting that all SOTA networks perform with similar level of performance on this problem. MobileNetV2 performance is not as good as the other networks for the 8-class problem, likely because it is a shallower network. This is interesting to note since one of the applications of our study is the deployment to a smart phone application and this would guide us to steer away from that network choice for an embedded AI application. In aggregate, our findings echo other studies that found that there is often not a statistically significant difference between these various SOTA networks in terms of performance, when tested on large datasets such as ImageNet[47,48.] Additionally, this difference becomes marginal and not statistically significant most of the time when working with relatively smaller datasets. Also, it is often the case that an ensemble of classifiers can yield a few percentage points improvements, which was not done in our study, but it is notable that ResNet, being a skip connection network, has already some ensembling properties as was shown in this study[49].

One interesting question regards the sensitivity to image resolution and network size. A first order answer is provided by the current experiments as the SOTA networks used offer an insight into the potential sensitivity to input size. Specifically, the input sizes for each network were: For ResNet50:224 X 224 X 3, ResNet152:224 X 224, InceptionV3: 299 X 299, DenseNet121: 224 X 224, InceptionResNetV2: 299 X 299, and MobileNetV2 224 X 224. There does not seem to be a significant benefit in using the larger input size of InceptionResNetV2 (with accuracy of 76.71% for the two class erythema migrans vs. all problem) and InceptionV3 (accuracy of 80.82%) or DenseNet121 (accuracy of 84.25%) when compared to networks with an input size 46% smaller (computed area-wise), e.g. for which ResNet50 had accuracy of 81.51%. Likewise, performance does not seem to depend on size for the 8-class problem comparisons. The likely explanation is given that the lesions – in most cases -- occupy a significant portion of the image as is seen in examples shown in Figure 1 and Figure 9. This should also be our expectation for clinical images as well as images taken from smart phones. Additional work can be directed towards conducting more experiments in the future by altering the architecture of those networks to accept other input sizes.

Our study has the following limitations. The images obtained online for training are highly variable in resolutions and camera viewpoint, which could make the use of semantic segmentation of the location of the lesion or insect bite highly beneficial. This was not done in this study and is left for future work.

Additional datasets of skin lesions exist online and could be considered for mixing with our dataset. One possibility would be to use the ISIC2018[45], however those images only contain dermoscopic imagery of skin cancer lesions. As we explicitly focused on the use case of EM against confuser lesions, this dataset is therefore of no use for the current work. Another possible dataset is SD-198[45], but unfortunately that dataset is also not adequate for our study, as it contains images of various diseases and various erythema but not erythema migrans. That

dataset does contain some classes that may be of interest as direct confusers of erythema migrans, but those classes are minimally represented in numbers which would be of little additional benefit for training as well as for evaluation. For perspective characteristic table 3 showed the number of examples of images for all conditions considered in our study and as can be seen our dataset exceeds by one or two factors of magnitude the number of exemplars found in SD-198 dataset. The characteristic Table also shows that for SD-198 dataset there is a strong paucity of examples for all classes that this study is focused on (i.e. EM and strong confusers), it contains no examples of erythema migrans or insect or tick bites which are an important confuser use case for our study. Also, a close inspection of SD-198 images shows that some images are near replica or cropped versions of each other (by contrast in our curate images replica were removed via manual and machine learning methods) which would put the estimate of hthe number of images available even lower. Another larger variation of SD-260 is available but at the time this study was conducted this dataset was no longer available for public download and we were not able to evaluate it.

Lastly, one important limitation of our study re. bias in AI: our dataset did not include a significant number of individuals with dark skin, and we could not ascertain that our dataset was diverse with regard to other protected factors such as gender and age, despite the fact that we attempted to target racial, ethnic, and gender diversity in our search terms. Recently the important issue of bias in AI and fairness regarding protected factors has gained much scrutiny[45] and this is an issue we intend to address in future studies with regard to skin lesions. We also will address issues related to having other types of imbalance in the data, such as paucity of data (so-called low shot learning) and domain shift and domain generalization.

Normal Skin (NO) **Erythema Migrans (EM)**

Herpes Zoster (HZ)

Tinea Corporis (TC)

Insect Bite (IB)

Tick Bite (IB-T)

Cellulitis (CELL)

Erythema Multiforme (EMU)

Figure 1 examples of the type of skin conditions investigated in this study in addition to EM

	Accuracy	Sensitivity/Recall	Specificity	PPV/Precision	NPV	Unweighted Kappa	F1 Score	Average Precision	AUC	Epochs Ran
NO vs EM	94.23 (3.66)	98.72 (2.50)	89.74 (6.73)	90.59 (6.21)	98.59 (2.74)	0.8846	0.9448	0.9850 (0.0191)	0.9854 (0.0188)	8
EM vs All	81.51 (6.30)	84.93 (8.21)	78.08 (9.49)	79.49 (8.96)	83.82 (8.75)	0.6301	0.8212	0.9022 (0.0482)	0.8887 (0.051)	14
EM vs NO vs All	83.11 (4.96)	83.11 (4.96)	91.55 (3.68)	83.08 (4.97)	91.57 (3.68)	0.7466	0.8311	0.8955 (0.0405)	0.9383 (0.0319)	13
EM vs CELL	79.72 (6.59)	81.82 (9.31)	77.92 (9.26)	76.06 (9.93)	83.33 (8.61)	0.5942	0.7883	0.8744 (0.0543)	0.8941 (0.0504)	6
NO vs CELL	95.57 (3.21)	94.59 (5.15)	96.43 (3.97)	95.89 (4.55)	95.29 (4.50)	0.9110	0.9524	0.9951 (0.0109)	0.9955 (0.0104)	9
NO vs EM vs HZ vs TC	81.58 (4.36)	80.86 (4.42)	93.82 (2.71)	81.3 (4.38)	94.02 (2.66)	0.7522	0.8158	0.8990 (0.0339)	0.9531 (0.0238)	10
NO vs	71.58 (3.77)	70.18 (3.83)	95.88 (1.66)	71.48 (3.78)	95.91	0.6709	0.7158	0.7973 (0.033	0.9385	9

EM vs HZ vs TC vs IB vs IB-T vs CELL vs EMU					(1.66)			6)	(0.0201)	
Clinical (EM Only)	N/A	88.55 (2.39)	N/A	N/A	N/A	N/A	0.9393	N/A	N/A	N/A

Table 1: performance results for various test cases of binary and multiclass classification. Reported values include the metric and the 95% error margin in parenthesis. See table 2 for acronym translation. Termination epochs are reported in the last column.

models	**Accuracy**	**Sensitivity/ Recall**	**Specificity**	**PPV/ Precision**	**NPV**	**Unweighted Kappa**	**F1 Score**	**Average Precision**	**AUC**	**Epochs Ran**
ResNet50	81.51 (6.30)	84.93 (8.21)	78.08 (9.49)	79.49 (8.96)	83.82 (8.75)	0.6301	0.8212	0.9022 (0.0482)	0.8887 (0.051)	14
InceptionV3	80.82 (6.39)	89.04 (7.17)	72.6 (10.23)	76.47 (9.02)	86.89 (8.47)	0.6164	0.8228	0.8529 (0.05	0.8707 (0.0	9

								74)	544)	
MobileNetV2	80.82 (6.39)	86.3 (7.89)	75.34 (9.89)	77.78 (9.05)	84.62 (8.77)	0.6164	0.8182	0.8853 (0.0517)	0.8898 (0.0508)	10
DenseNet121	84.25 (5.91)	84.93 (8.21)	83.56 (8.5)	83.78 (8.4)	84.72 (8.31)	0.6849	0.8435	0.8895 (0.0509)	0.8994 (0.0488)	9
InceptionResNetV2	76.71 (6.86)	78.08 (9.49)	75.34 (9.89)	76.0 (9.67)	77.46 (9.72)	0.5342	0.7703	0.8921 (0.0503)	0.8913 (0.0505)	12
ResNet152	82.88 (6.11)	86.3 (7.89)	79.45 (9.27)	80.77 (8.75)	85.29 (8.42)	0.6575	0.8344	0.8809 (0.0525)	0.884 (0.0519)	15

Table 2 EM vs ALL performance comparison for all other DCNNs

models	**Accuracy**	**Sensitivity/ Recall**	**Specificity**	**PPV/Precision**	**NPV**	**Unweighted Kappa**	**F1 Score**	**Average Precision**	**AUC**	**Epochs Ran**
ResNet50	71.58 (3.77)	70.18 (3.83)	95.88 (1.66)	71.48 (3.78)	95.91 (1.66)	0.6709	0.7158	0.7973 (0.0336)	0.9385 (0.0201)	9
MobileNetV2	59.38 (4.11)	58.07 (4.13)	94.07 (1.98)	64.34 (4.01)	94.15 (1.96)	0.5280	0.5938	0.6637 (0.0395)	0.8952 (0.0256)	7

DenseNet121	71.58 (3.77)	70.2 (3.83)	95.86 (1.67)	73.35 (3.7)	95.91 (1.66)	0.6704	0.7158	0.7765 (0.0348)	0.936 (0.0205)	9
InceptionResNetV2	67.94 (3.9)	67.56 (3.92)	95.38 (1.76)	67.12 (3.93)	95.39 (1.75)	0.6297	0.6794	0.745 (0.0365)	0.9252 (0.022)	10
InceptionV3	67.94 (3.9)	67.86 (3.91)	95.37 (1.76)	66.88 (3.94)	95.39 (1.75)	0.6296	0.6794	0.7529 (0.0361)	0.9227 (0.0223)	10
ResNet152	70.86 (3.8)	71.04 (3.79)	95.8 (1.68)	70.08 (3.83)	95.81 (1.68)	0.6636	0.7086	0.7809 (0.0346)	0.9396 (0.0199)	8

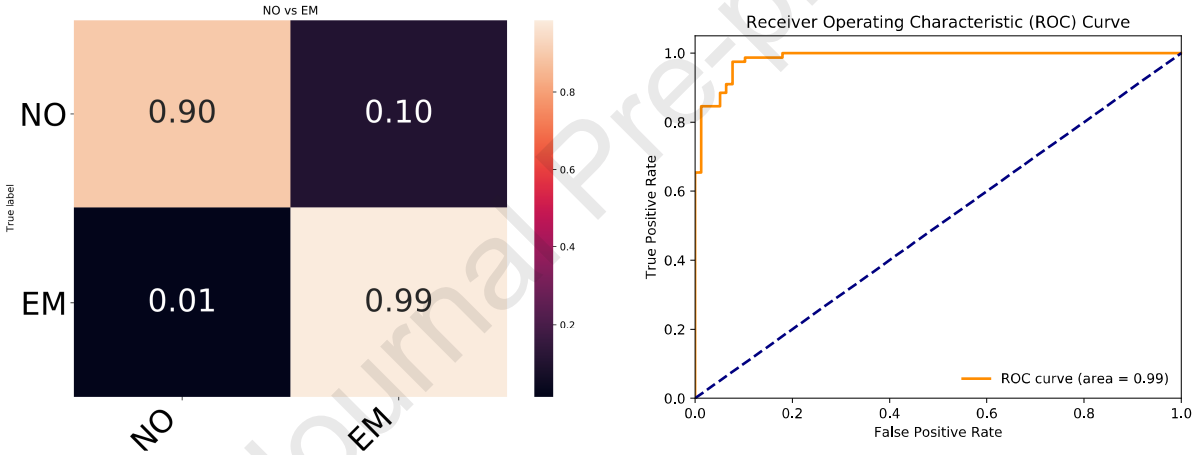
Table 3 For the problem NO vs EM vs HZ vs TC vs IB vs IB-T vs CELL vs EMU performance comparison is reported.

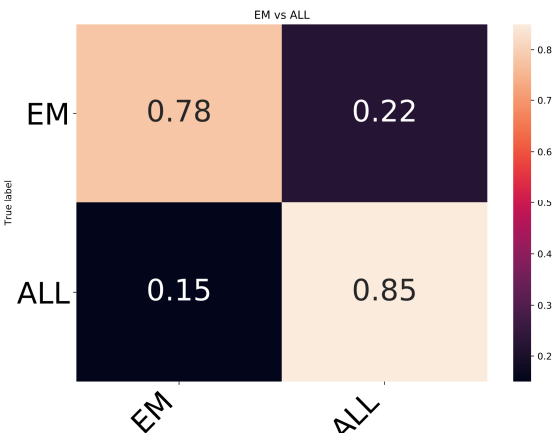
class	**Normal Skin (NO)**	**Erythema Migrans (EM)**	**Herpes Zoster (HZ)**	**Tinea Corporis (TC)**	**Insect Bite (IB)**	**Tick Bite (IB-T)**	**Cellulitis (CELL)**	**Erythema Multiforme (EMU)**
Train	777	723	594	612	708	261	599	400
Validation	42	40	36	35	41	16	30	34
Test (online)	72	91	83	73	87	31	67	45
Test (clinical)	N/A	681	N/A	N/A	N/A	N/A	N/A	N/A
Total for this study dataset	891	1535	713	720	836	308	696	479
Comparison with SD-198	0	0	25	61	0	0	36	25

dataset								

Table 4: characteristic table showing the number of images per class among training, validation, and test sets (for the 8-class experiment)

Figure 2: confusion matrix and ROC curve for the 2-class problem normal skin (NO) vs. EM





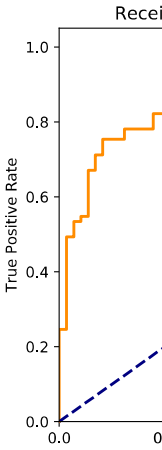


Figure 3: confusion matrix and ROC curve for the 2-class EM vs. all other (ALL) problem

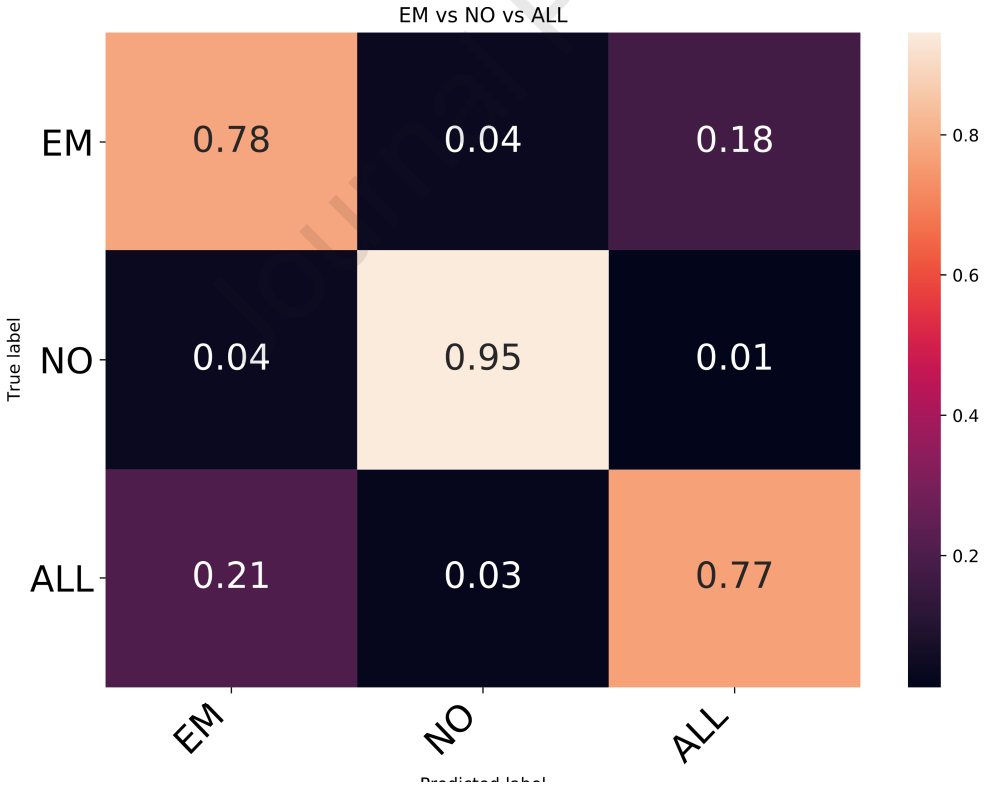


Figure 4: *confusion matrix for the 3-class EM vs. normal skin (NO) vs. all (ALL) other classes*

problem

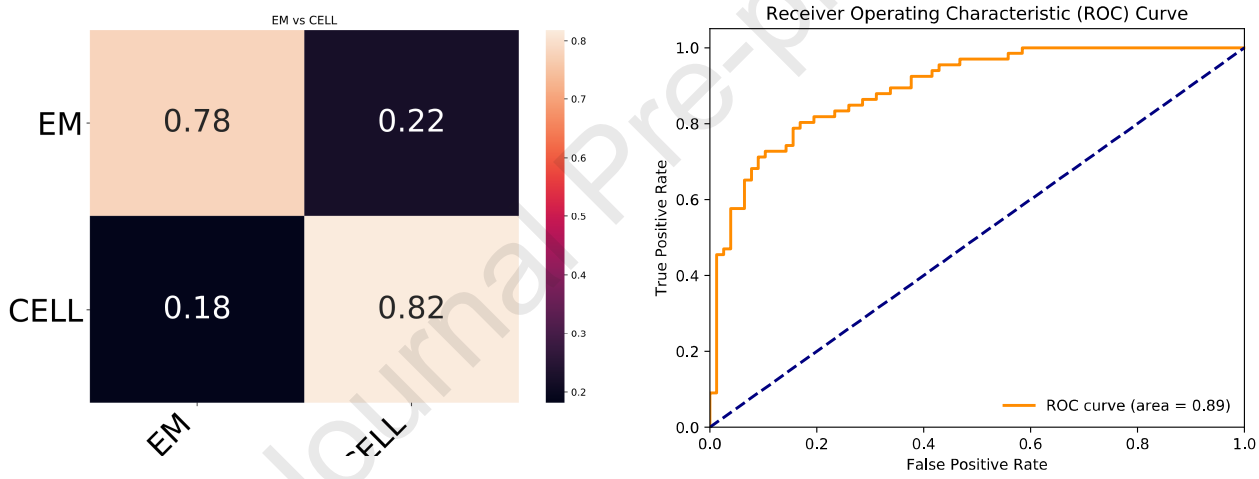


Figure 5: confusion matrix and ROC curve for the *2-class EM vs. Cellulitis (CELL) problem*

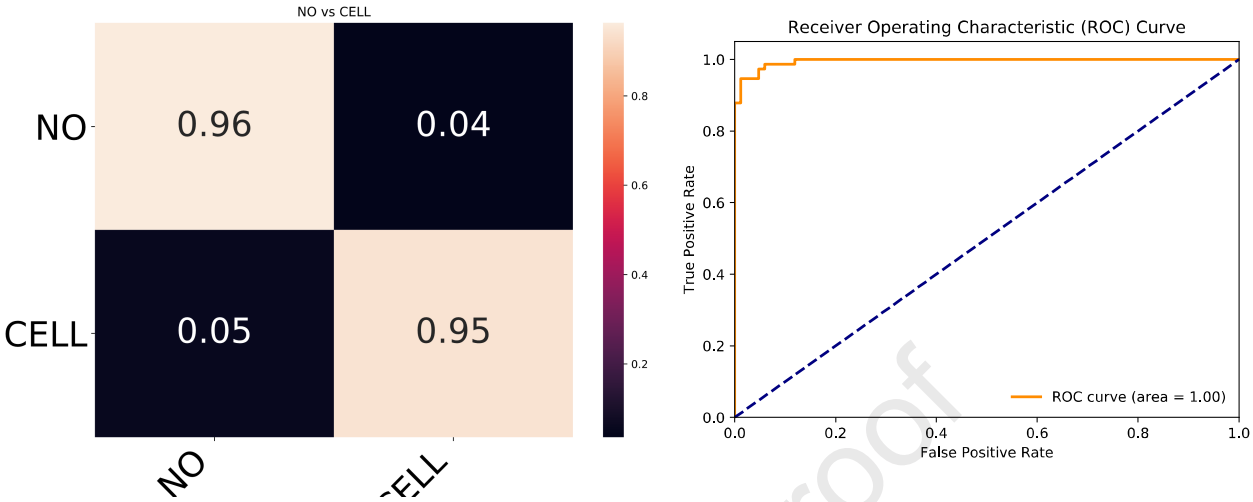


Figure 6: confusion matrix and ROC curve for the 2-class normal vs. Cellulitis problem

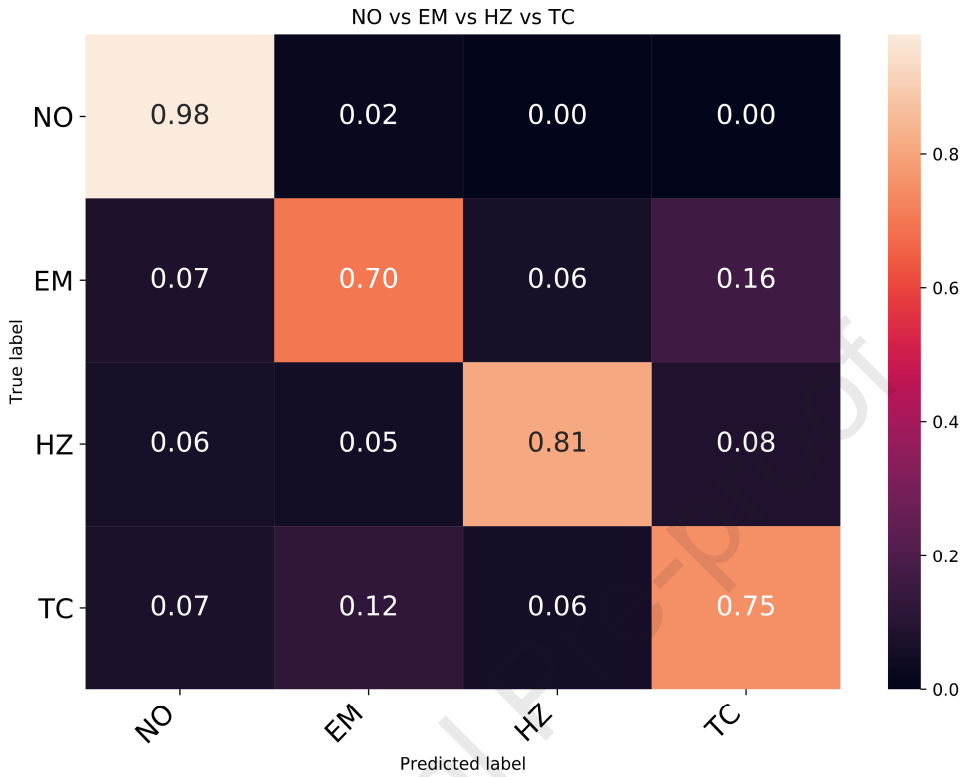


Figure 7: confusion matrix for the 4-class normal vs. EM vs. HZ vs TC problem

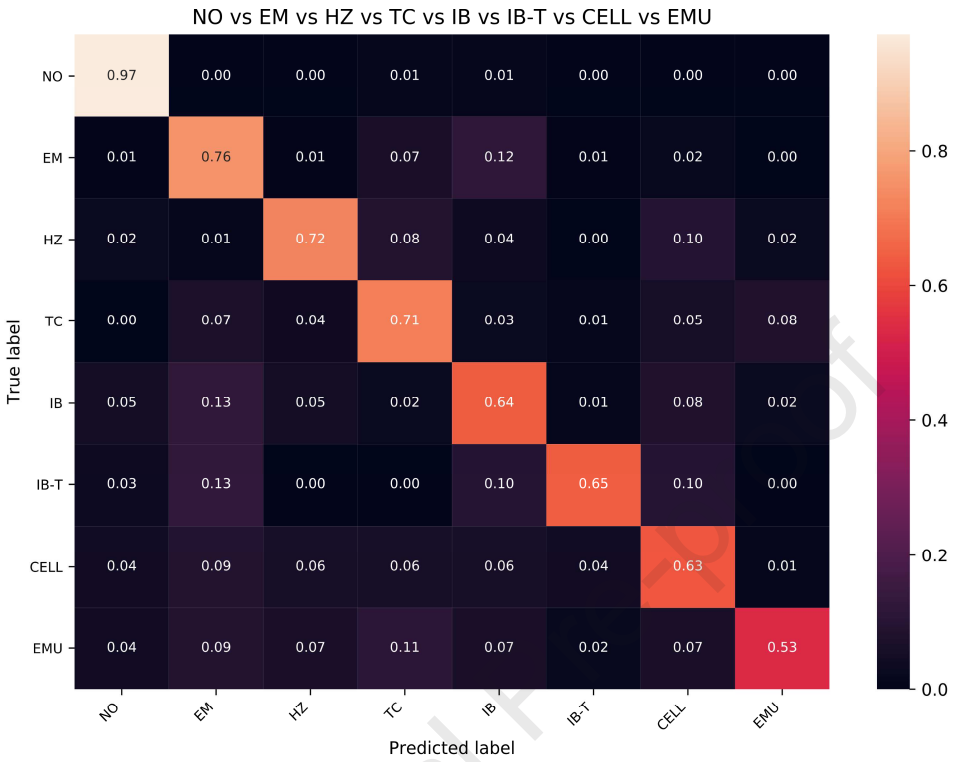


Figure 8: confusion matrix for the full 8 class problem

class	Erythema Migrans (EM)	Tinea Corporis (TC)	Tick Bite (IB-T)	Cellulitis (CELL)	Erythema Multiforme (EMU)
Error % (95% CI)	32.96 (5.61)	30.77 (5.47)	18.71 (4.71)	25.62 (5.10)	18.22 (4.61)
Number of Images	270	273	274	281	269

Table 3: Results of inter-operator error between two graders (JA & CN for several skin conditions

Conclusion

We used AI and DL approaches for erythema migrans classification and Lyme disease diagnostics , looking at several complex training/testing use cases that focused on the problem of EM classification against clinically relevant confusers. Based on the performance results we reported for both public domain and the generalization to clinical image testing, this study appears to show substantial potential for possible future applications of pre-screening for clinician referral. These applications would have the benefit of increasing the likelihood that patients who need further medical assessment see a physician for further examination. This would help address morbidity by avoiding unevaluated and undiagnosed patients which could evolve into more serious long-term complications resulting from late-stage Lyme disease, should these patients turn out to be affected by the disease.

Acknowledgements:

The funding of the Lyme Disease Research Foundation, the Johns Hopkins University Applied Physics Laboratory Research and Development Funds, and the support of the Johns Hopkins Institute for Assured Autonomy are gratefully acknowledged. The views expressed here are those of the authors and not of the funding entities. We thank Dr. Elizabeth Horn from the Lyme Disease Biobank for the procurement of additional EM images, and Cheryl Novak for assistance with annotation of images.

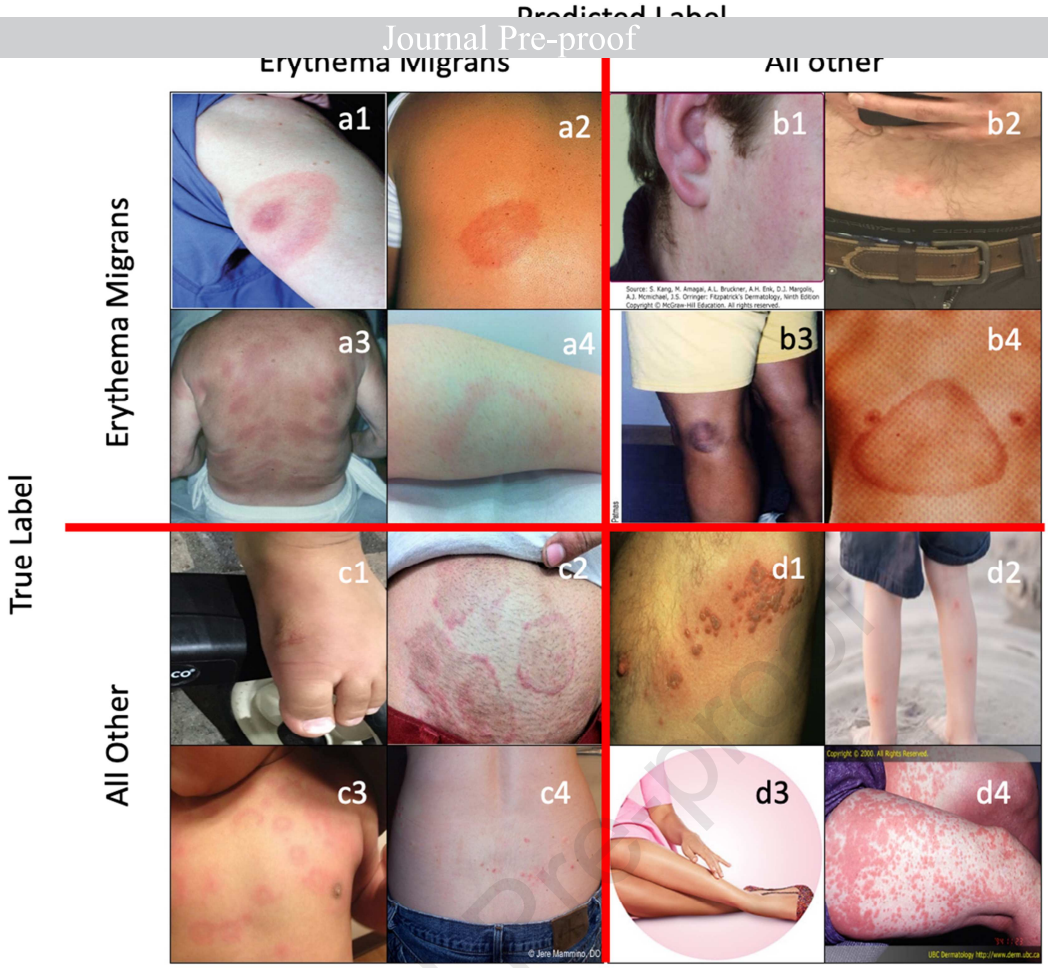


Figure 9: examples selected randomly of correct and incorrect predictions for the Erythema Migrans vs All classification, displayed as a 2X2 confusion matrix. Top left quadrant show examples of true positives and lower right quadrant, of true negatives. Examples in the upper right quadrant are of false negatives: a1 is a typical difficult example of EM on a neck area; a2 has very faint and small erythema that could be confused for an insect bite; a3 shows an example on a darker skin individual for which there are few training examples in the dataset, a challenge which is discussed in the discussion section and should be addressed in future work via AI debiasing methods; a4 has a triangular-shaped erythema which is atypical for EM. Lower left quadrant shows examples of false positives: c1 is ambiguous, c2 and c3 are cases of erythema multiforme which are easily confused with EM because of the circular shape and central clearing; c4 is likely herpes zoster but the patch of erythema areas on the right hip likely confused the prediction.

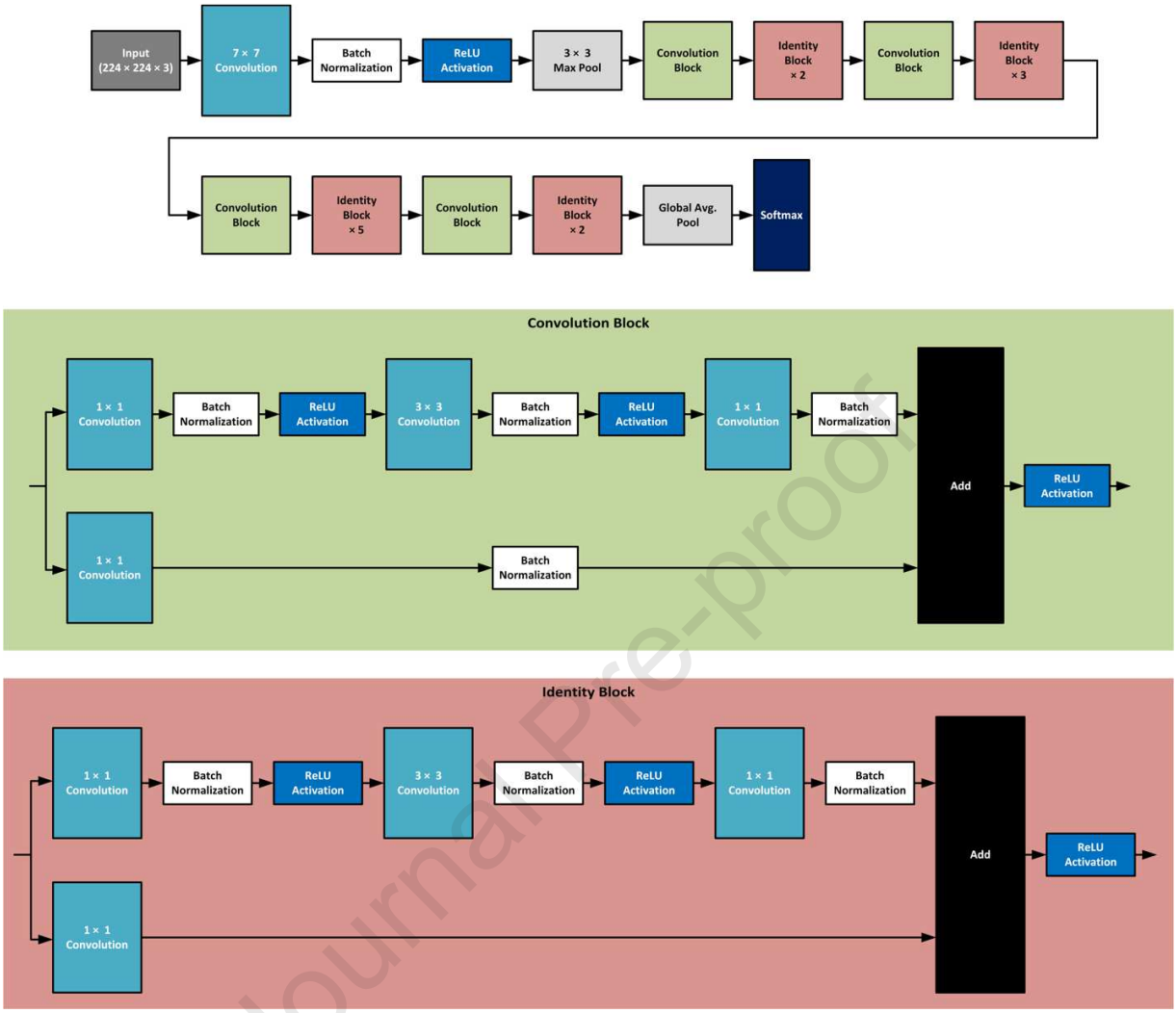


Fig 10: Diagram for ResNet50 deep neural net, one of the several deep convolutional networks used in this study.

References

1. Kuehn BM. CDC estimates 300,000 US cases of Lyme disease annually. *JAMA*. 2013;310(11):1110. doi:10.1001/jama.2013.278331

2. Hinckley AF, Connally NP, Meek JI, et al. Lyme disease testing by large commercial laboratories in the United States. *Clin Infect Dis*. 2014;59(5):676-681. doi:10.1093/cid/ciu397

3. Stanek G, Wormser GP, Gray J, Strle F. Lyme borreliosis. *Lancet*. 2012;379(9814):461-473. doi:10.1016/S0140-6736(11)60103-7

4. Nadelman RB. Erythema migrans. *Infect Dis Clin North Am*. 2015;29(2):211-239. doi:10.1016/j.idc.2015.02.001

5. Steere AC, Sikand VK. The presenting manifestations of Lyme disease and the outcomes of treatment. *N Engl J Med*. 2003;348(24):2472-2474. doi:10.1056/NEJM200306123482423

6. Smith RP, Schoen RT, Rahn DW, et al. Clinical characteristics and treatment outcome of early Lyme disease in patients with microbiologically confirmed erythema migrans. *Ann Intern Med*. 2002;136(6):421-428. http://www.ncbi.nlm.nih.gov/pubmed/11900494

7. Steere AC, Strle F, Wormser GP, et al. Lyme borreliosis. *Nat Rev Dis Prim*. 2016;2:16090. doi:10.1038/nrdp.2016.90

8. Wormser GP, Dattwyler RJ, Shapiro ED, et al. The clinical assessment, treatment, and prevention of Lyme disease, human granulocytic anaplasmosis, and babesiosis: clinical practice guidelines by the Infectious Diseases Society of America. *Clin Infect Dis*. 2006;43(9):1089-1134. doi:10.1086/508667

9. Centers for Disease Control and Prevention. Lyme Disease (Borrelia burgdorferi) 2017 Case Definition. Published 2017. https://wwwn.cdc.gov/nndss/conditions/lyme-disease/case-definition/2017/

10. Shapiro ED. Clinical practice. Lyme disease. *N Engl J Med*. 2014;370(18):1724-1731. doi:10.1056/NEJMcp1314325

11. Bhate C, Schwartz RA. Lyme disease: Part I. Advances and perspectives. *J Am Acad Dermatol*. 2011;64(4):618-619. doi:10.1016/j.jaad.2010.03.046

12. Tibbles CD, Edlow JA. Does this patient have erythema migrans? *JAMA*. 2007;297(23):2617-2627. doi:10.1001/jama.297.23.2617

13. Soloman S, Tanael M. Rash, Radiculopathy, and Cognitive Biases. In: *Aerospace Medicine and Human Performance*. Aerospace Medical Association; 2019:652-654. doi:10.3357/AMHP.5339.2019

14. Li TH, Shih CM, Lin WJ, Lu CW, Chao LL, Wang CC. Erythema migrans mimicking cervical cellulitis with deep neck infection in a child with lyme disease. *J Formos Med Assoc*. 2007;106(7):577-581. doi:10.1016/S0929-6646(07)60009-6

15. Nowakowski J, McKenna D, Nadelman RB, et al. Failure of treatment with cephalexin for Lyme disease. *Arch Fam Med*. 2000;9(6):563-567. doi:doi:10.1001/archfami.9.6.563

16. Mazori DR, Orme CM, Mir A, Meehan SA, Neimann AL. Vesicular erythema migrans: an atypical and easily misdiagnosed form of Lyme disease. *Dermatol Online J*. 2015;21(8). http://dx.doi.org/

17. Mullegger RR, Glatz M. Skin manifestations of lyme borreliosis: diagnosis and management. *Am J Clin Dermatol*. 2008;9(6):355-368. doi:10.2165/0128071-200809060-00002

18. Aucott JN, Crowder LA, Yedlin V, Kortte KB. Bull's-Eye and nontarget skin lesions of Lyme disease: an internet survey of identification of erythema migrans. *Dermatol Res Pr*. 2012;2012:451727. doi:10.1155/2012/451727

19. Lipsker D, Lieber-Mbomeyo A, Hedelin G. How accurate is a clinical diagnosis of erythema

chronicum migrans? Prospective study comparing the diagnostic accuracy of general practitioners and dermatologists in an area where lyme borreliosis is endemic. *Arch Dermatol*. 2004;140(5):620-621. doi:10.1001/archderm.140.5.620

20. Fujisawa Y, Otomo Y, Ogata Y, et al. Deep learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumor diagnosis. *Br J Dermatol*. Published online 2018. doi:10.1111/bjd.16924

21. Burlina P, Billings S, Joshi N, Albayda J. Automated diagnosis of myositis from muscle ultrasound: Exploring the use of machine learning and deep learning methods. *PLoS One*. 2017;12(8). doi:10.1371/journal.pone.0184059

22. Feeny A, Tadarati M, Freund D, Bressler N, Burlina P. Automated segmentation of geographic atrophy of the retinal epithelium via random forests in AREDS color fundus images. *Comput Biol Med*. 2015;65:124-136. doi:10.1016/j.compbiomed.2015.06.018

23. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *Int Conf Learn Represent*. Published online 2015. http://arxiv.org/abs/1409.1556

24. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. *IEEE Conf Comput Vis Pattern Recognit*. Published online 2015:1-9. doi:10.1109/CVPR.2015.7298594

25. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Proc 25th Int Conf Neural Inf Process Syst - Vol 1*. Published online 2012:1097-1105.

26. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444. doi:10.1038/nature14539

27. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. ; 2016:770-778. doi:10.1109/CVPR.2016.90

28. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. Vol 1. MIT Press; 2016.

29. 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

30. Burlina PM, Joshi N, Pekala M, Pacheco KD, Freund DE, Bressler NM. Automated Grading of Age-Related Macular Degeneration From Color Fundus Images Using Deep Convolutional Neural Networks. *JAMA Ophthalmol*. 2017;135(11):1170-1176. doi:10.1001/jamaophthalmol.2017.3782

31. Burlina P, Pacheco KD, Joshi N, Freund DE, Bressler NM. Comparing humans and deep learning performance for grading AMD: A study in using universal deep features and transfer learning for automated AMD analysis. *Comput Biol Med*. 2017;82:80-86. doi:10.1016/j.compbiomed.2017.01.018

32. Burlina P, Joshi N, Pacheco K, Freund D, Kong J, Bressler N. Utility of Deep Learning Methods for Referability Classification of Age-Related Macular Degeneration. *Jama Opthalmology*. 2018;136(11):1305-1307. doi:10.1001/jamaophthalmol.2018.3799

33. Burlina P, Joshi N, Pacheco K, Freund D, Kong J, Berssler N. Use of Deep Learning for Detailed Severity Characterization and Estimation of 5-Year Risk Among Patients With Age-Related Macular Degeneration. *JAMA Opthalmology*. 2018;136(12):1359-1366. doi:10.1001/jamaophthalmol.2018.4118

34. Kankanahalli S, Burlina PM, Wolfson Y, Freund DE, Bressler NM. Automated classification of severity of age-related macular degeneration from fundus photographs. *Investig Ophthalmol Vis Sci*. 2013;54(3):1789-1796. doi:10.1167/iovs.12-10928

35. Burlina P, Freund D, Dupas B, Bressler N. Automatic Screening of Age-Related Macular Degeneration and Retinal Abnormalities. In: *IEEE Engineering in Medicine and Biology Society*. ; 2011:3962-3966. doi:10.1109/IEMBS.2011.6090984

36. Zhang L, Yang G, Ye X. Automatic skin lesion segmentation by coupling deep fully convolutional networks and shallow network with textons. *J Med Imaging*. 2019;6(02):1. doi:10.1117/1.jmi.6.2.024001

37. Ali A-R, Li J, O'shea SJ, Yang G, Trappenberg T, Ye X. *A Deep Learning Based Approach to Skin Lesion Border Extraction With a Novel Edge Detector in Dermoscopy Images*.; 2019. https://github.com/abderhasan/fuzzedge.

38. Kawahara J, Bentaieb AA, Hamarneh G. *Deep Features to Classify Skin Lesions*. https://licensing.eri.ed.ac.uk/i/software/

39. Yu L, Chen H, Dou Q, Qin J, Heng P-A. Automated Melanoma Recognition in Dermoscopy Images via Very Deep Residual Networks. *IEEE Trans Med Imaging*. 2017;36(4):994-1004. doi:10.1109/TMI.2016.2642839

40. Čuk E, Gams M, Možek M, Strle F, Čarman VM, Tasič JT. Supervised visual system for recognition of Erythema Migrans, an early skin manifestation of Lyme Borreliosis. *Strojniški Vestn - J Mech Eng*. 2014;60(2):115-123. doi:10.5545/sv-jme.2013.1046

41. Burlina PM, Joshi NJ, Ng E, Billings SD, Rebman AW, Aucott JN. Automated detection of erythema migrans and other confounding skin lesions via deep learning. *Comput Biol Med*. 2019;105. doi:10.1016/j.compbiomed.2018.12.007

42. Horn EJ, Dempsey G, Schotthoefer AM, et al. The Lyme Disease Biobank: Characterization of 550 Patient and Control Samples from the East Coast and Upper Midwest of the United States. *J Clin Microbiol*. 2020;58(6):1-12. doi:10.1128/JCM.00032-20

43. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-ResNet and the impact of residual connections on learning. *31st AAAI Conf Artif Intell AAAI 2017*. Published online 2017:4278-4284.

44. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*. Vol 2017-Janua. ; 2017:2261-2269. doi:10.1109/CVPR.2017.243

45. Kinyanjui NM, Odonga T, Cintas C, et al. Estimating Skin Tone and Effects on Classification Performance in Dermatology Datasets. Published online October 29, 2019. http://arxiv.org/abs/1910.13268

46. Liu Y, Jain A, Eng C, et al. A deep learning system for differential diagnosis of skin diseases. *Nat Med*. 2020;26:900-908. doi:10.1038/s41591-020-0842-3

47. Khan A, Sohail A, Zahoora U, Qureshi AS. A survey of the recent architectures of deep convolutional neural networks. *Artif Intell Rev*. Published online 2020. doi:10.1007/s10462-020-09825-6

48. Zoph B, Vasudevan V, Shlens J, Le Q V. Learning Transferable Architectures for Scalable Image Recognition. Published online July 21, 2017. http://arxiv.org/abs/1707.07012

49. Veit A, Wilber M, Belongie S. Residual networks behave like ensembles of relatively shallow networks. *Adv Neural Inf Process Syst*. Published online 2016:550-558.

Highlights

- This study examines the use AI methods for detecting erythema migrans (EM) against the most clinically relevant skin conditions that may be "confusers."
- Accurate identification of erythema migrans allows for early diagnosis and treatment of Lyme disease, which avoids the potential for later neurologic, rheumatologic, and cardiac complications.
- We develop the most extensively curated dataset thus far for this challenging problem.
- We develop and test several deep learning models against various problems of growing complexity and test on a combination of public domain and clinical images.
- The DL system has accuracy ranging from 71.58% (and 95% error margin equal to 3.77%) for an 8-class problem of EM versus 7 other confusers, to 94.23% (3.66%) for a binary problem of EM vs. non-pathological skin.
- We test generalization on clinical images of affected individuals and obtain a sensitivity of 88.55% (2.39%).
- These results suggest that AI can help in prescreening and referring individuals to physicians for earlier diagnosis and treatment, in the presence of clinically relevant confusers, thereby reducing further complications and morbidity of Lyme disease.

AI-BASED DETECTION OF ERYTHEMA MIGRANS AND DISAMBIGUATION AGAINST OTHER SKIN LESIONS

Philippe M. Burlina, PHD[1,2], Neil J. Joshi, BS[1], Phil A. Mathew[1],

William Paul[1], Alison W. Rebman, MPH[3], John N. Aucott, MD[3]

[1]Applied Physics Laboratory, Johns Hopkins University

[2]Malone Center for Engineering in Healthcare, Johns Hopkins University

[3]Johns Hopkins Lyme Disease Research Center, Division of Rheumatology, Department of Medicine, Johns Hopkins University School of Medicine

The authors declare no conflict of interest.