HGG Advances

ARTICLE

Neural network classifiers for images of genetic conditions with cutaneous manifestations

Dat Duong,[1] Rebekah L. Waikel,[1] Ping Hu,[1] Cedrik Tekendo-Ngongang,[1] and Benjamin D. Solomon[1,*]

Summary

Neural networks have shown strong potential in research and in healthcare. Mainly due to the need for large datasets, these applications have focused on common medical conditions, where more data are typically available. Leveraging publicly available data, we trained a neural network classifier on images of rare genetic conditions with skin findings. We used approximately 100 images per condition to classify 6 different genetic conditions. We analyzed both preprocessed images that were cropped to show only the skin lesions as well as more complex images showing features such as the entire body segment, the person, and/or the background. The classifier construction process included attribution methods to visualize which pixels were most important for computer-based classification. Our classifier was significantly more accurate than pediatricians or medical geneticists for both types of images and suggests steps for further research involving clinical scenarios and other applications.

Introduction

Neural network models have demonstrated strong potential to improve the practice of healthcare. For example, "artificial intelligence" may help detect breast cancer via mammography or COVID-19 based on computed tomography (CT) scans.[1,2] In this type of computer vision approach, because medical datasets are typically small compared to other types of publicly available datasets, the neural network is first pretrained on a large, general image dataset to help identify major features, such as edges or basic shapes. Next, the neural network is fine-tuned to address a more specific question, such as the recognition of certain diseases. For this approach to work well, the pretrained data must either be similar in type to the medical data or the size of the medical dataset must be relatively large.[3]

Due to these limitations, neural network applications in healthcare have focused on relatively common conditions, where sufficiently large datasets are more readily collected. Genetic conditions, though common in aggregate, are largely individually rare.[4] A recent meta-analysis identified 82 studies comparing deep learning performance to that of healthcare professionals in disease detection using medical imaging. None of the conditions in this meta-analysis were genetic, though some (e.g., breast cancer) involve clear genetic underpinnings in a subset of individuals.[5] Other relatively recent studies have examined skin lesions (especially skin cancer), though they did not focus on genetic conditions.[6–8] A recent scoping review identified a total of 211 papers about specific rare conditions that were analyzed via machine learning; by our count, 59 of these papers focused on genetic conditions (versus other rare conditions).[9]

Despite this lack of representation, neural network approaches have been used in some genetic areas.[10] With efforts to collect adequate training data, these methods could be especially useful in clinical genetics, where there is a lack of trained individuals to help determine whether a person may be affected by a genetic condition, what that condition may be, and what testing strategy and management steps are indicated.[11,12] With the expansion of genomics into diverse fields of medicine,[13] an alternative strategy of training non-geneticist clinicians has not kept pace.[14] Developing computational methods could help geneticists and other clinicians manage the large numbers of affected individuals.

To explore the use of these techniques in proof-of-principle exercises using small datasets, we collected images of a selected group of rare genetic conditions that manifest with characteristic skin findings. We chose clinically impactful conditions that can be nontrivial to diagnose.[15] We built neural network classifiers both for images that were cropped to focus on the lesions of interest, similar to previous studies,[16,17] as well as uncropped images. These uncropped images can be more difficult for a neural network model to analyze but may more closely mimic real-life, unprocessed images, such as those a clinician might encounter or might share with a colleague to request advice. During the classifier construction process, we used an attribution method for image recognition to help visualize how the classifier weighted image pixels. Last, we compared the classifier performance to clinicians.

To summarize, our contributions include: (1) evaluation of a neural network classifier's performance on a small dataset of esoteric genetic conditions, and (2) comparison of how focused and panoramic images affect human and the neural network model's accuracy.

[1]Medical Genomics Unit, Medical Genetics Branch, National Human Genome Research Institute, Bethesda, MD 20892, USA
*Correspondence: solomonb@mail.nih.gov
https://doi.org/10.1016/j.xhgg.2021.100053.

Material and methods

Ethics review

The study was reviewed by National Human Genome Research Institute (NHGRI) bioethicists and the National Institutes of Health (NIH) Institutional Review Board (IRB). The main analyses were considered not human subjects research; a waiver of consent was granted by the NIH IRB (NIH protocol: 000285) for the work involving the surveys of medical professionals, as described below.

Data collection

Using condition and gene names, we searched Google and PubMed to identify publicly available images showing the following six conditions: hypomelanosis of Ito (HMI [MIM: 300337]), incontinentia pigmenti (IP [MIM: 308300]), McCune-Albright syndrome (MA [MIM: 174800]), neurofibromatosis type 1 (NF1 [MIM: 162200]), Noonan syndrome with multiple lentigines (ML; formally known as LEOPARD syndrome [MIM: 151100]), and tuberous sclerosis complex (TSC [MIM: 191100, 613254]) (see Table S1 for more details about these conditions). Our clinician investigators also selected and reviewed images of other skin conditions (e.g., basal cell carcinoma, blue nevus, other congenital nevi, erythema migrans, halo nevus, hemangioma, ichthyosis, melanoma, molluscum contagiosum, piebaldism, port wine stain, psoriasis, tinea versicolor, vitiligo, etc.) that are unrelated to the six aforementioned genetic conditions (the sources for all images are available in Table S2). Some of these other conditions can be clinically relevant and may also be evaluated by clinicians seeing individuals with the genetic conditions we analyzed. For example, a pediatrician might be expected to recognize the presence of a variety of congenital skin findings, including those related to genetic conditions. Though difficult to quantify due to lack of available comprehensive information for many images, we endeavored to collect images from individuals of diverse ancestral backgrounds. This was done by manual review, using images from sources that focus on ancestrally diverse individuals (such as journals devoted to the presentation of medical conditions in diverse geographic locations) and performing searches in multiple languages.

As defined below (see Initial image processing), the images we used included focused images (n = 1,032 [total]: 96 HMI, 116 IP, 122 MA, 105 ML, 230 NF1, 119 TSC, 244 other) and panoramic images (n = 798 [total]: 87 HMI, 85 IP, 100 MA, 83 ML, 120 NF1, 88 TSC, 235 other). All focused images were derived from the panoramic images. The reason that there are more focused images is that some panoramic images (e.g., where there were multiple, discrete lesions) could be split into more than one focused image. Two board-certified clinicians (one medical geneticist and one genetic counselor) reviewed images and data in the source websites to help ensure accuracy of diagnoses based on clinical descriptions and the information described. For example, if a publication showed an image of a person described as having NF1, our team reviewed the image and the description of that person to ensure that there was strong evidence for the diagnosis and that there was not contradictory evidence, such as a statement that the individual, on ultimate genetic testing, had an alternate molecular explanation. All images and URLs used in classification are listed in Table S2.

Initial image processing

First, following the conventions of other large-scale neural network studies on images of skin cancer,[16–18] we cropped images to contain just the lesions of interest. We refer to this dataset as focused images. For conditions with multiple skin or other findings, we focused on the relatively early, main skin manifestations that may first bring the person to clinical attention, before there are other, more obvious signs of the underlying diagnosis. For example, for NF1, we focused on café-au-lait macules (CALMs). This is because, in this condition, CALMs are often the first manifestation, preceding other more obvious signs of the condition, such as cutaneous neurofibromas.[19,20] In HMI and IP, individuals may have other manifestations in addition to skin findings later in life, such as developmental delay, while people with MA may demonstrate precocious puberty later in life. We chose to analyze stage 3 (hyperpigmented stage) of IP, as we hypothesized that this stage may be harder to differentiate from the other conditions. See Table S1 for more details and further references related to these conditions. We also used the photos as they were captured, some of which show an entire person's face or body segment (e.g, an arm or the entire back) with the genetic conditions, and with other features such as clothes or a background, though we cropped out words, such as a heading indicating the image number. We refer to this second dataset as panoramic images.

A single panoramic image can have multiple corresponding focused images. For example (Figure 1), an image of a person with NF1 may include multiple CALMs. We did not want the model to capture anything related to the test images during training. Hence, for the test set, from each of the 7 categories (the 6 genetic conditions and the "other" category), we selected 20 panoramic images and corresponding focused images. Each panoramic test image has exactly one corresponding focused image. In total, the panoramic test set and the corresponding focused test images contained 140 images each. The remaining images were used to train the model.

Classifier

We chose the EfficientNet-B4 classifier, which achieved good performance on the ImageNet data with a relatively low number of parameters.[21] We initialized EfficientNet-B4 with the parameter values pretrained on ImageNet and continued training the entire model, not just the last few fully connected layers.[3] Combining and then jointly training a small dataset of interest with a larger auxiliary dataset often helps the prediction accuracy.[22,23] For the auxiliary dataset, we downloaded the publicly available SIIM-ISIC Melanoma Classification Challenge Dataset from 2018 to 2020.[16,24] This dataset contains 58,459 images of 9 skin cancer diseases: actinic keratosis, basal cell carcinoma, benign keratosis, dermatofibroma, melanoma, melanocytic nevus, squamous cell carcinoma, vascular lesion, and other unknown skin cancer cases. The SIIM-ISIC images are focused images; however, due to our relatively low sample size, we opted to train the models using focused SIIM-ISIC images with both our focused and panoramic images.

We trained our model with the SIIM-ISIC dataset where we classified an image as one of the 16 diseases (7 from our genetic + other disorders dataset and 9 from the SIIM-ISIC dataset). We conducted two experiments, the first with our focused and SIIM-ISIC images and the second with our panoramic and SIIM-ISIC images. Both experiments used 450 by 450 pixel images and the same layers of data augmentation: transposition, vertical and horizontal flip, random brightness and contrast, motion and Gaussian blur, optical and grid distortion, hue saturation, and shift and rescale rotation. Key hyperparameters were learning rates and sample weights for the loss functions. We set learning rates at 0.00003

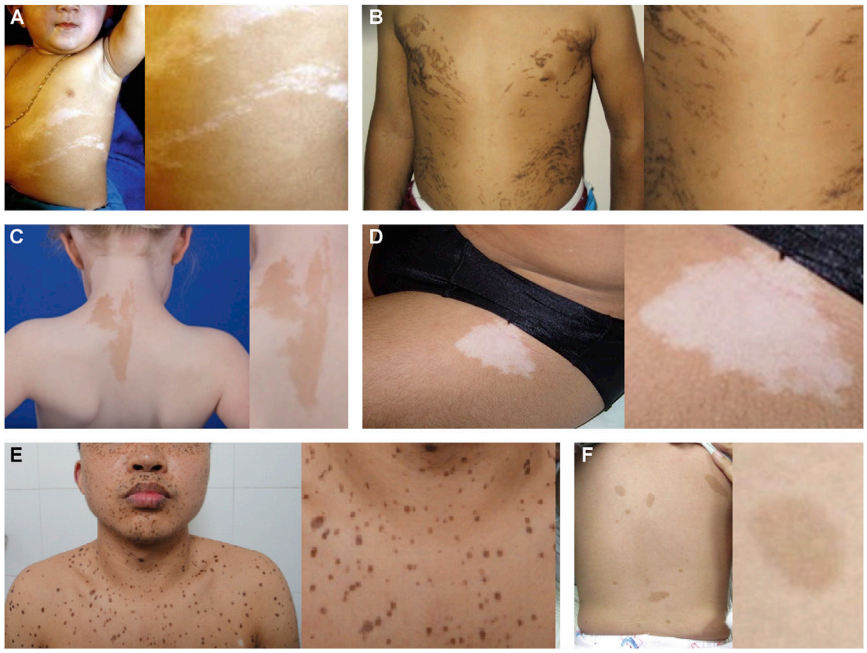

Figure 1. Example panoramic and focused images for each pair of conditions

Hypomelanosis of Ito (HMI) (A), incontinentia pigmenti (IP) (B), McCune-Albright syndrome (MA) (C), tuberous sclerosis complex (TSC) (D), Noonan syndrome with multiple lentigines (ML; formally known as LEOPARD syndrome) (E), and neurofibromatosis type 1 (NF1) (F). See Table S1 for more details on these conditions. Images sources (all are used with appropriate permission) are listed in the Web resources.

and 0.00001 and weighed our datasets 5 and 10 times more than the SIIM-ISIC dataset for our focused and panoramic models, respectively. We used batch size 64 and Adam optimizer with 30 epochs for both focused and panoramic images. We trained all the models on the NIH High Performance Computing Cluster using P100 16 GB Nvidia graphic card. Our code is available at GitHub.

For our focused images, a 5-fold cross-validation was used to build 5 different classifiers (one for each fold). To create an ensemble predictor, we used each classifier to estimate the predicted probabilities for the labels of a test image. The average of these probabilities was calculated for the 5 classifiers. When averaging, we considered only the classifiers that produced a maximum predicted probability (over all the labels) of at least 0.5. The same procedure was used for training the model on our panoramic images. To visualize which parts of an image the classifier considered to be important, we applied Integrated Gradient to identify pixels of an image that most affect the classifier's outcome.[25]

Comparison to clinicians

We compared the classifier to board-certified or board-eligible medical geneticist physicians and pediatricians. We chose these specialties because, in our experience, these types of clinicians more frequently encounter these individuals (versus, for example, dermatologists, who may more often assess other skin conditions).[26] That is, a typical path involves an initial encounter by a pediatrician, followed by referral to a medical geneticist.

We generated surveys using Qualtrics (Provo, UT, USA). Each survey has 4 panoramic images and their corresponding focused versions for each of the 7 conditions (6 genetic conditions + other conditions). As there are 140 panoramic test images, 5 surveys can cover all the test images. We created 6 sets of these 5 surveys, such that each test image would be seen 6 times. In each survey, we showed the focused images first and then the panoramic images. In total, there were 30 unique surveys. For fair comparison, a medical geneticist and a pediatrician completed the same survey, and paired t test was be used to compare their outcomes.

When responding to the survey, participants were directed not to use any external resources for help and to select the genetic condition best represented by the image presented. Although this does not mirror standard medical practice (in which clinicians might use textbooks or web sources when assessing an individual), we hypothesized that these procedures would help with standardization. The surveys also included 3 demographic questions (medical specialty, number of years in practice, and location of current practice), which were only used for verification purposes, rather than for analyses.

Following previous methods,[16,17] we estimated that 30 participants for each clinician type would provide a statistical power of 95% to detect a 10% difference. For each of the 30 surveys, one board-certified or board-eligible medical geneticist physician and one board-certified or board-eligible pediatrician was recruited via e-mail. To identify survey respondents, we obtained e-mail addresses through professional networks, departmental websites, journal publications, and other web-available lists. A total of 105 medical geneticists were contacted; 37 agreed to participate, and 32 completed the survey. A total of 379 pediatricians were contacted; 37 agreed to participate, and 32 completed the survey. Surveys were considered complete if >95% of the multiple-choice questions were answered. If multiple medical geneticists or pediatricians completed the same survey, only the first survey was used for analysis.

Results

Clinician demographics

A range of experience levels was reported for both the medical geneticist and pediatrician participants, with the median participant in both groups reporting greater than 10 years of experience. Of the 30 medical geneticist respondents, 6.7% had less than 1 year of experience, 13.3% had 1 to 5 years of experience, 20% had 5 to 10 years of experience, and 60% had more than 10 years of experience. Of the 30 pediatrician respondents, 10% had less than 1 year of experience, 13.3% had 1 to 5 years of experience, and 76.7% had greater than 10 years of experience. All clinicians in the pediatrician group reported currently practicing in North America. One of the clinicians in the medical geneticist group reported practicing in Asia (though may have trained or practiced elsewhere), whereas

the other 29 reported practicing in North America. See Data S1 for a copy of the survey and table displaying clinician demographic data (Table S4).

Classifier

We first assessed the classifiers' performances when jointly trained on our dataset and the SIIM-ISIC dataset. None of our focused and panoramic test images were classified as one of the cancer conditions in the SIIM-ISIC dataset. This was expected, because SIIM-ISIC diseases are dissimilar to our genetic conditions, and the pose-style and ancestry (SIIM-ISIC largely represents individuals of European descent) in the SIIM-ISIC images are different from those in our images. When evaluated on the same 30 surveys described in Material and methods, under Comparison to clinicians, the classifier trained on focused images and the classifier trained on panoramic images obtained the same average accuracy on all 30 surveys: 0.814 (SD, 0.083) and 0.814 (SD, 0.72), respectively (2-sided paired t test, $p = 1$).

We evaluated the reliability of our surveys using the intraclass (within-class) correlation coefficient (ICC) in two ways. First, we computed the agreement ICC score for the three groups: classifier, medical geneticists, pediatricians. Because we have different participants for each survey, we used a one-way random effects model to compute the agreement ICC score. For surveys with focused images, the ICC score for these three groups is 0.315 (95% confidence interval [CI]: −0.393, −0.176). The ICC score for surveys with panoramic images is −0.090 (95% CI: −0.244, 0.137), indicating that the three groups do not agree well.[27] Second, we computed the clinicians' agreement and consistency ICC scores across all surveys for focused and panoramic images. We computed the 2-way random effects ICC scores on agreement and consistency on the accuracy of focused and panoramic images for the medical geneticist and pediatrician group separately. For the geneticists, the agreement ICC is 0.543 (95% CI: 0.034, 0.794), indicating that the medical geneticists do not often give the same answers for focused and panoramic images.[27] However, the consistency ICC is 0.682 (95% CI: 0.431, 0.835), which indicates that, over all the surveys, the geneticists are often better at classifying panoramic versus focused images. For the pediatricians, the agreement and consistency ICC scores are 0.693 (95% CI: 0.194, 0.874) and 0.792 (95% CI: 0.608, 0.895), respectively, indicating that pediatricians are more likely to obtain similar accuracy for both focused images and their corresponding panoramic versions.[27]

ICC scores do not compare whether one group (e.g., the classifier) is better than another (e.g., the medical geneticists or pediatricians) at classifying the diseases. Following previous studies,[17] we used 2-sided paired t test to compare our classifier's accuracy to that of medical geneticists and pediatricians. Group results for accuracy for classification are shown in Figure 2. Overall, the computer classifier performed 37.3% ($p = 3.87 \times 10^{-13}$) and 28.7% ($p = 2.81 \times 10^{-11}$) better than pediatricians for focused and panoramic images, respectively. Overall, the computer classifier performed 21.8% ($p = 1.52 \times 10^{-10}$) and 12.7% ($p = 4.57 \times 10^{-7}$) better than medical geneticists for focused and panoramic images, respectively.

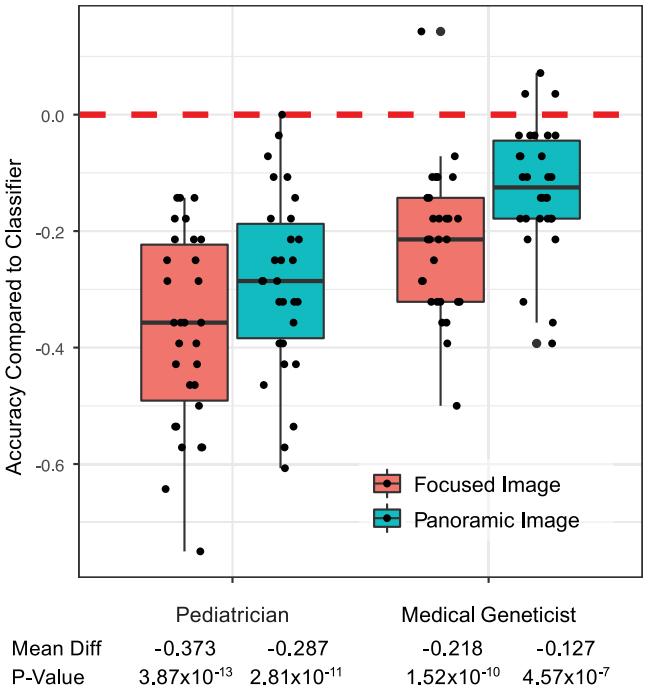

Figure 2. Performance of physicians compared to deep learning classifier

We trained two classifiers, one on focused images and the other on panoramic images. We compared the performance of the classifiers to that of pediatricians and medical geneticists. In the boxplots, each point represents the accuracy difference between the classifier and the human performance for a single survey, with the ranges for each group of respondents shown by the lines extending from each boxplot. The red line indicates the baseline accuracy for the classifier.

Medical geneticists performed better than pediatricians on focused and panoramic images by 15.6% ($p = 3.63 \times 10^{-4}$) and 16.0% ($p = 1.44 \times 10^{-4}$), respectively. On average, humans performed better with panoramic than focused images. For both the medical geneticist and pediatrician groups, the accuracy for panoramic images was higher than for focused images by 9.05% ($p = 2.55 \times 10^{-5}$) and 8.57% ($p = 6.60 \times 10^{-5}$), respectively. The results for each of the individual genetic conditions are shown in Figure 3. For the panoramic images, the condition with the lowest accuracy (80%) for the classifier was NF1, still higher than the average for the physicians (77.5% and 61.7% for medical geneticists and pediatricians, respectively). The most difficult condition for both pediatricians and medical geneticists to classify based on panoramic images was MA, with 28.3% and 49.2% accuracy, respectively. The classifier identified 75% of MA panoramic images accurately. As shown in Figure 3, the clinicians had more difficulty differentiating conditions that have similar skin manifestations than the classifier. For

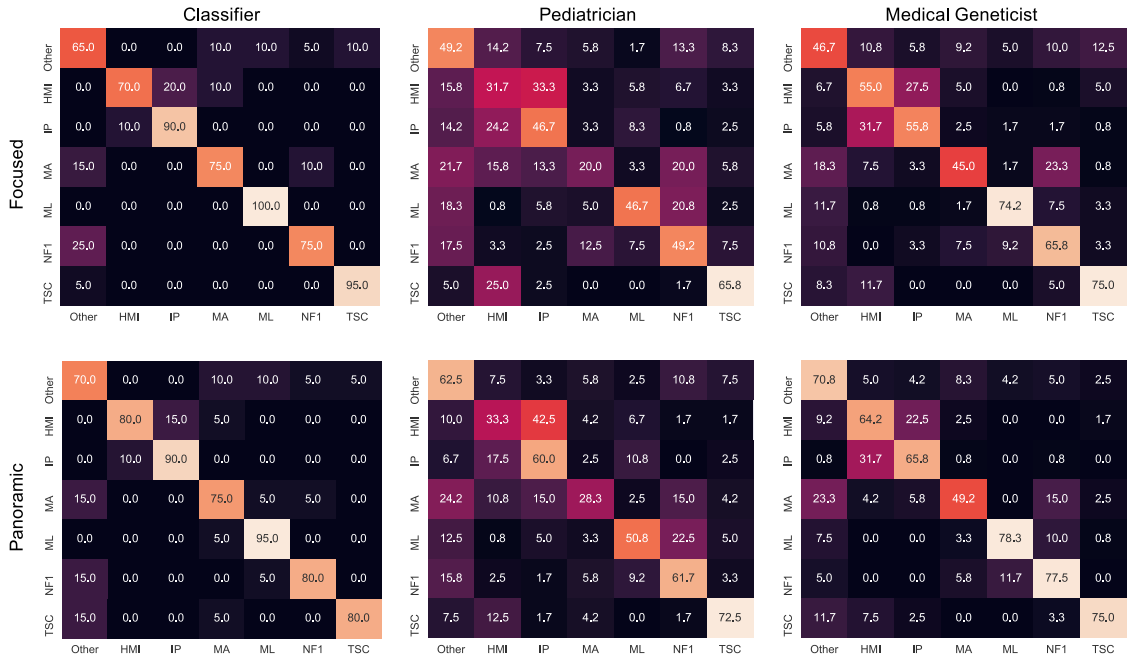

Figure 3. Confusion matrices

Confusion matrix comparing our classifier (left) versus the two different clinician types (middle and right) for classification of focused and panoramic images. Rows represent the correct label, while columns represent the label chosen by the classifier or the clinicians. The diagonal numbers represent the percent accuracy for each category (the percentage of time the correct label was identified), while the off-diagonal numbers represent misclassifications, with the number corresponding to the percentage of time the label for a given image type was ascribed to another, incorrect category.

example, the clinicians had more difficulty differentiating HMI from IP and MA from ML or NF1.

We validated that our classifier obtains adequate accuracy because it weights important parts of an image in the classification process. That is, the attribution method (Figure 4) helped us determine which pixels most affect the classifier's decision making. This helped ensure that intuitively important pixels were being weighted more frequently than potential common artifacts, such as recurrent background types or articles of clothing in the panoramic images.[29] This approach was used in our classifier development process to subjectively examine which combinations of training datasets provided the best output.

Discussion

Our overarching goal was to use neural networks to demonstrate how these and related methods can be leveraged in potentially useful and interesting ways with datasets involving genetic conditions. Our aim was not to build the most accurate classifier possible. We could achieve better accuracy with additional (computationally expensive) modifications, such as further modifying the model's hyperparameters or by incorporating larger or different datasets for training. We note that the availability of larger, centralized, and freely available datasets relevant to genetic diseases such as those we analyzed is currently lacking compared to more common conditions.

Our classifier outperformed both pediatricians and geneticists for both focused and panoramic images. This does not imply that the classifier can or should replace human experts or that our methods represent clinical practice. To help reduce potential bias in our surveys, and to allow us to efficiently gather more information from respondents, we did not allow clinicians to access materials (such as textbooks or the internet) that they might employ in real-life scenarios. We also asked clinicians to classify images without incorporating other information that is often important in clinical practice, such as family history and other clinical manifestations (e.g., the presence or absence of developmental delay or certain types of cancer). These data could be incorporated into a computer-based classifier, though such a model was not part of our objectives, in part because we did not have uniform access to these data. Despite this, our experimental set-up did allow us to estimate how well the physicians perform when provided with the more holistic panoramic images versus the focused images. This estimation was not done in some other studies involving skin images.[6,24]

The fact that our classifier worked relatively well may demonstrate possible use cases. For example, this type of approach could help primary care doctors determine which individuals should be prioritized for evaluation by subspecialists like geneticists. In settings—both in the United States and in other countries—with less access to specialists, these tools could help identify the most efficient genetic testing strategies. Even for experienced geneticists, these classifiers could be used as a back-up or

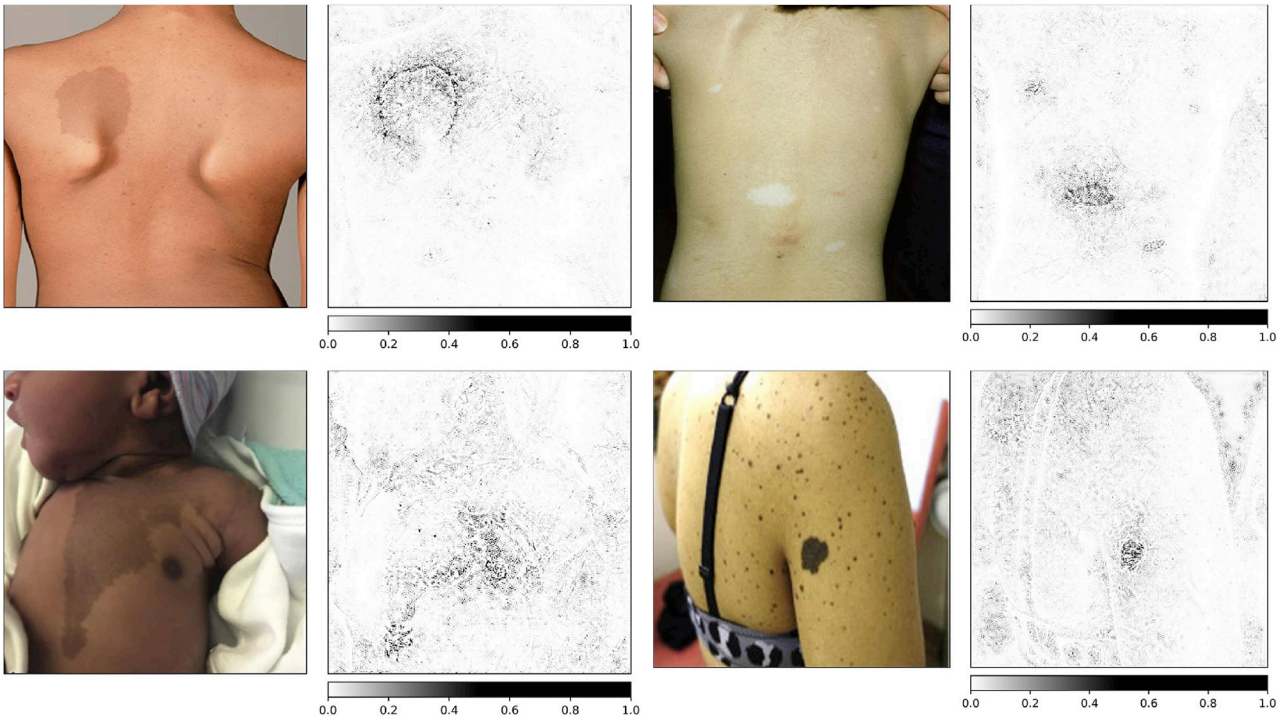

Figure 4. Attribution images

The attribution images show which pixels the classifier weights when "deciding" how to categorize. As shown, the classifier uses pixels involved in the skin finding but may also use other pixels as well, some of which may represent confounders.[28] Our research team examined these attribution methods during stages of classifier training and testing to determine how to improve performance, such as by incorporating other datasets for training, or when adjusting the neural network hyperparameters. Clockwise, from top left: NF1, TSC, ML, and MA. Images sources (all are used with appropriate permission) are listed in the Web resources.

additional tool to augment their assessment of certain characteristics of an individual they see.

We observed a range of accuracy for the clinicians. This is logical: some clinicians may be more experienced with these specific conditions or may simply be more gifted at this type of task. The fact that the classifier does relatively well across different conditions, in addition to the overall accuracy, is notable. As an example, the confusion matrices (Figures 3 and S1) show that the clinicians classify certain conditions better than others. This may relate to the rarity of certain conditions that may be clinically important but are rarely encountered in training or clinical practice. This is one advantage of computerized methods, where a very rare condition can be included if adequate overall data can be gathered. Interestingly, the clinicians had more difficulty discerning conditions that can appear similar to each other, such as HMI and IP, than the classifier.

We built a potentially useful method using small datasets; we were able to build the classifying algorithm using a minimum of about 100 images per condition. This is important when considering methods for conditions that are relatively rare. We also endeavored to include individuals of diverse ancestral backgrounds when collecting our training datasets. As the ancestral background of most individuals whose images we used was not described in the primary literature, this was difficult to quantify and requires further testing and attention in this type of work.[30,31] However, we do not want to overstate our ability to quantify the diversity of the images in this dataset, as these data were often not available. We also did not want to assume the ancestry of an individual where that was not specifically mentioned. We plan to pursue this important question in prospective studies.

The methods we built can also be readily modified. For example, other conditions could be incorporated by collecting additional images and retraining our classifier, which can be done quickly using the code we provide. We also applied related techniques on our current dataset to generate new images via generative adversarial networks (see Supplemental methods), which may be useful for improving classifier performance.

One concern about neural network and related methods is that they are a "black box" that is opaque to human intuition or explanation. Our attribution methods show that one can correlate which features are important for the computer classifier to make decisions. This has recently been used to explore confounders in analysis of X-rays from individuals with COVID-19.[28] This was useful during the classifier building process to ensure that pixels were weighted in what would be considered a logical fashion. This is not dissimilar to how a human might identify which condition a person has. That is, the human may pay more attention to certain informative features, such as the shape of a skin lesion or the angle of a bone on an

X-ray. A key clinical skill learn is learning which features deserve attention and which are less important. Using computer-based attribution methods can similarly help understand which features a model uses. We plan to explore objective quantification of these methods in future studies.

While our work provides insights into the use of advanced computational approaches in the study of rare diseases, our study has limitations. In collecting a large enough dataset, we relied on publicly available data. Our clinical team vetted each image, but it is possible that some data were inaccurate. For example, some depicted images could derive from people affected by more than one genetic condition,[32] which could complicate the phenotype. The conditions analyzed can also have genetic heterogeneity (can occur due to different genetic causes) or can involve distinct genotype-phenotype correlations.[33,34] As we treated each condition collectively (as a single entity), we were not able to parse out unique attributes to a given genetic variant or subset of a condition. Additionally, our approach did not account for possible overlaps between the conditions, such as might occur in the two RASopathies (NF1 and ML).[35] As our major focus of our approach is to build methods that can be useful with smaller datasets, our accuracy was not as high as it would be with a larger dataset. We anticipate that as publicly available datasets are established for rare diseases, work in this area will approach the accuracy of that described for more common conditions. Our work with morphing and style mixing is highly exploratory. We plan to study how these and other techniques can aid trainees and healthcare practitioners. We compared our classifier to two types of physicians, clinicians who most frequently encounter these types of conditions collectively. However, other clinicians may have different (better or worse) abilities to classify some conditions. For example, dermatologists, family practitioners, neurologists, and other specialists may yield different results. Finally, we emphasize that these results do not predict real-life performance, where clinicians often have access to more information or have access to additional resources. The point of our classification comparison was not to devise a head-to-head competition but rather to use our approaches to explore the utility of how advanced analytics and large publicly available datasets may augment the identification of rare genetic diseases.

Data and code availability

Source data are available in files deposited at Figshare (see web resources) and described in Tables S2 and S3. Code to classify labels and generate images (see also supplemental information for additional explanations and examples) is available on GitHub (see web resources).

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.xhgg.2021.100053.

Acknowledgments

This research was supported by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health. This work utilized the computational resources of the NIH HPC Biowulf cluster. The authors thank Vence L. Bonham, Jr., JD, Senior Advisor to the NHGRI Director on Genomics and Health Disparities and Head of the NHGRI Health Disparities Unit, for his comments regarding considerations related to ancestral diversity and Daniel L. Kastner, MD, PhD, for his mentorship and guidance regarding research plans.

Declaration of interests

B.D.S.: Editor-in-Chief of the American Journal of Medical Genetics; previously (until 2019) employee of GeneDx, Inc., a genetic/genomic testing laboratory that is a subsidiary of Opko Health, and previous holder of Opko Health stock options; previously (until 2019) member of Scientific Advisory Board for FDNA. All other authors declare no competing interests.

Received: May 18, 2021
Accepted: August 6, 2021

Web resources

Figure 1 sources: HMI: https://link.springer.com/referenceworkentry/10.1007%2F978-3-642-02202-9_148 IP: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3103180/ MA: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5292317/ TSC: http://www.atlasdermatologico.com.br/disease.jsf?diseaseId=477 ML: https://www.sciencedirect.com/science/article/pii/S0167527314008419?via%3Dihub NF1: https://link.springer.com/chapter/10.1007/978-3-030-50823-4_11

Figure 4 sources: NF1: https://pubmed.ncbi.nlm.nih.gov/24432075/#&gid=article-figures&pid=fig-2-uid-1TSC: https://pubmed.ncbi.nlm.nih.gov/24143074/#&gid=article-figures&pid=figure-3-uid-2ML: https://ars.els-cdn.com/content/image/1-s2.0-S000293431930347X-gr1_lrg.jpgMA: https://www.sciencedirect.com/science/article/pii/S235251261930044X?via%3Dihub#fig1

GitHub, https://github.com/datduong/ClassifyNF1 and github.com/datduong/stylegan2-ada-MorphNF1

NF1 and related diseases on Figshare, https://figshare.com/articles/figure/NF1-and-related-diseases-images_zip/14721051

NIH HPC Biowulf Cluster, https://hpc.nih.gov/

OMIM, https://www.omim.org

References

1. McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G.S., Darzi, A., et al. (2020). International evaluation of an AI system for breast cancer screening. Nature *577*, 89–94.
2. Mei, X., Lee, H.C., Diao, K.Y., Huang, M., Lin, B., Liu, C., Xie, Z., Ma, Y., Robson, P.M., Chung, M., et al. (2020). Artificial intelligence-enabled rapid diagnosis of patients with COVID-19. Nat. Med *26*, 1224–1228.
3. Morid, M.A., Borjali, A., and Del Fiol, G. (2021). A scoping review of transfer learning research on medical image analysis using ImageNet. Comput. Biol. Med. *128*, 104115.
4. Ferreira, C.R. (2019). The burden of rare diseases. Am. J. Med. Genet. A. *179*, 885–892.

5. Liu, X., Faes, L., Kale, A.U., Wagner, S.K., Fu, D.J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M., Kern, C., et al. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit. Health *1*, e271–e297.
6. Yang, Y., Ge, Y., Guo, L., Wu, Q., Peng, L., Zhang, E., Xie, J., Li, Y., and Lin, T. (2021). Development and validation of two artificial intelligence models for diagnosing benign, pigmented facial skin lesions. Skin Res. Technol. *27*, 74–79.
7. Srinivasu, P.N., SivaSai, J.G., Ijaz, M.F., Bhoi, A.K., Kim, W., and Kang, J.J. (2021). Classification of Skin Disease Using Deep Learning Neural Networks with MobileNet V2 and LSTM. Sensors (Basel) *21*, 2852.
8. Brinker, T.J., Hekler, A., Utikal, J.S., Grabe, N., Schadendorf, D., Klode, J., Berking, C., Steeb, T., Enk, A.H., and von Kalle, C. (2018). Skin Cancer Classification Using Convolutional Neural Networks: Systematic Review. J. Med. Internet Res. *20*, e11936.
9. Schaefer, J., Lehne, M., Schepers, J., Prasser, F., and Thun, S. (2020). The use of machine learning in rare diseases: a scoping review. Orphanet J. Rare Dis. *15*, 145.
10. Dias, R., and Torkamani, A. (2019). Artificial intelligence in clinical and genomic diagnostics. Genome Med. *11*, 70.
11. Maiese, D.R., Keehn, A., Lyon, M., Flannery, D., Watson, M.; and Working Groups of the National Coordinating Center for Seven Regional Genetics Service Collaboratives (2019). Current conditions in medical genetics practice. Genet. Med. *21*, 1874–1877.
12. Abacan, M., Alsubaie, L., Barlow-Stewart, K., Caanen, B., Cordier, C., Courtney, E., Davoine, E., Edwards, J., Elackatt, N.J., Gardiner, K., et al. (2019). The Global State of the Genetic Counseling Profession. Eur. J. Hum. Genet. *27*, 183–197.
13. Green, E.D., Gunter, C., Biesecker, L.G., Di Francesco, V., Easter, C.L., Feingold, E.A., Felsenfeld, A.L., Kaufman, D.J., Ostrander, E.A., Pavan, W.J., et al. (2020). Strategic vision for improving human health at The Forefront of Genomics. Nature *586*, 683–692.
14. Plunkett-Rondeau, J., Hyland, K., and Dasgupta, S. (2015). Training future physicians in the era of genomic medicine: trends in undergraduate medical genetics education. Genet. Med. *17*, 927–934.
15. Korf, B.R., and Bebin, E.M. (2017). Neurocutaneous Disorders in Children. Pediatr. Rev. *38*, 119–128.
16. Tschandl, P., Rosendahl, C., and Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data *5*, 180161.
17. Tschandl, P., Codella, N., Akay, B.N., Argenziano, G., Braun, R.P., Cabo, H., Gutman, D., Halpern, A., Helba, B., Hofmann-Wellenhof, R., et al. (2019). Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. Lancet Oncol. *20*, 938–947.
18. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature *542*, 115–118.
19. Williams, V.C., Lucas, J., Babcock, M.A., Gutmann, D.H., Korf, B., and Maria, B.L. (2009). Neurofibromatosis type 1 revisited. Pediatrics *123*, 124–133.
20. Gutmann, D.H., Ferner, R.E., Listernick, R.H., Korf, B.R., Wolters, P.L., and Johnson, K.J. (2017). Neurofibromatosis type 1. Nat. Rev. Dis. Primers *3*, 17004.
21. Tan, M., and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In International Conference on Machine Learning. (Proceedings of Machine Learning Research *97*, 6105–6114.
22. Meftah, S., Semmar, N., Tahiri, M.-A., Tamaazousti, Y., Essafi, H., and Sadat, F. (2020). Multi-Task Supervised Pretraining for Neural Domain Adaptation. In Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media (Association for Computational Linguistics), pp. 61–71.
23. Ahmad, W.U., Bai, X., Huang, Z., Jiang, C., Peng, N., and Chang, K.-W. (2018). Multi-task Learning for Universal Sentence Embeddings: A Thorough Evaluation using Transfer and Auxiliary Tasks. arXiv.
24. Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., Combalia, M., Dusza, S., Guitera, P., Gutman, D., et al. (2021). A patient-centric dataset of images and metadata for identifying melanomas using clinical context. Sci. Data *8*, 34.
25. Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In International Conference on Machine Learning. Proceedings of Machine Learning Research, pp. 3319–3328.
26. Liu, Y., Jain, A., Eng, C., Way, D.H., Lee, K., Bui, P., Kanada, K., de Oliveira Marinho, G., Gallegos, J., Gabriele, S., et al. (2020). A deep learning system for differential diagnosis of skin diseases. Nat. Med. *26*, 900–908.
27. Koo, T.K., and Li, M.Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J. Chiropr. Med. *15*, 155–163.
28. DeGrave, A.J., Janizek, J.D., and Lee, S.I. (2020). AI for radiographic COVID-19 detection selects shortcuts over signal. medRxiv, 2020.09.13.20193565.
29. Samek, W., Montavon, G., Lapuschkin, S., Anders, C.J., and Müller, K.-R. (2020). Toward interpretable machine learning: Transparent deep neural networks and beyond. arXiv.
30. Kruszka, P., Tekendo-Ngongang, C., and Muenke, M. (2019). Diversity and dysmorphology. Curr. Opin. Pediatr. *31*, 702–707.
31. Muenke, M., Adeyemo, A., and Kruszka, P. (2016). An electronic atlas of human malformation syndromes in diverse populations. Genet. Med. *18*, 1085–1087.
32. Smith, E.D., Blanco, K., Sajan, S.A., Hunter, J.M., Shinde, D.N., Wayburn, B., Rossi, M., Huang, J., Stevens, C.A., Muss, C., et al. (2019). A retrospective review of multiple findings in diagnostic exome sequencing: half are distinct and half are overlapping diagnoses. Genet. Med. *21*, 2199–2207.
33. Koczkowska, M., Callens, T., Gomes, A., Sharp, A., Chen, Y., Hicks, A.D., Aylsworth, A.S., Azizi, A.A., Basel, D.G., Bellus, G., et al. (2019). Expanding the clinical phenotype of individuals with a 3-bp in-frame deletion of the NF1 gene (c.2970_2972del): an update of genotype-phenotype correlation. Genet. Med. *21*, 867–876.
34. Koczkowska, M., Callens, T., Chen, Y., Gomes, A., Hicks, A.D., Sharp, A., Johns, E., Uhas, K.A., Armstrong, L., Bosanko, K.A., et al. (2020). Clinical spectrum of individuals with pathogenic NF1 missense variants affecting p.Met1149, p.Arg1276, and p.Lys1423: genotype-phenotype study in neurofibromatosis type 1. Hum. Mutat. *41*, 299–315.
35. Jafry, M., and Sidbury, R. (2020). RASopathies. Clin. Dermatol. *38*, 455–461.