$\mathbf{1}$

\overline{a} Human and computer attention in assessing genetic conditions

Dat Duong**, Anna Rose Johny**, Suzanna Ledgister Hanchard**, Chris Fortney^s*, Fabio Hellmann⁻
Hu¹, Behnam Javanmardi², Shahida Moosa^{5,6}, Tanviben Patel¹, Susan Persky³, Ömer Sümer⁴, Cedrik
Tekendo-Ngongang \overline{a} Hu*, Behnam Javanmardi*, Shahida Moosa^{3,6}, Tanviben Patel*, Susan Persky^o, Omer Sümer*
Tekendo-Ngongang¹, Tzung-Chien Hsieh², Rebekah L. Waikel¹, Elisabeth André⁴, Peter Kraw
Benjamin D. Solomon¹ $\frac{1}{\pi}$ itz², Benjamin D. Solomon 1 , Tzung-Chien Hsieh", Rebekah L. Waikel", Elisabeth André⁴, Peter Krawitz["]
n¹ ,
,
,

Benjamin D. Solomon[.]
¹Medical Genetics Brar Medical Genetica; ²Institute for Genomic Statistics and Bioinformatics, University Hospital Bonn,
-
Meinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany; ³Social and Behavioral Research States of America; "Institute for Genomic Statistics and Bioinformatics, University Hospital Bonn,
Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany; ³Social and Behavioral Research
Branch, National Human G Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany; "Social and Behavioral Research
Branch, National Human Genome Research Institute, Bethesda, Maryland, United States of Americ
⁴Chair for Human-Centered Artif ary
Branch, National Artificial Intelligence, University of Augsburg, Germany; ^SDivision of Anational Artica;
Molecular Biology and Human Genetics, Stellenbosch University, Stellenbosch, South Africa; 4 Chair for Human-Centered Artificial Intelligence, University of Augsburg, Germany; ⁹
Molecular Biology and Human Genetics, Stellenbosch University, Stellenbosch, South
Department of Medical Genetics, Tygerberg Hospital, Molecular Biology and Human Genetics, Stellenbosch University, Stellenbosch, South Africa; Morecular Biology and Human Genetics, Tygerberg Hospital, Tygerberg, South Africa;
⁶Department of Medical Genetics, Tygerberg Hospital, Tygerberg, South Africa;
*Equal contributions Ĭ $\begin{array}{c}\n\text{1.5}\n\text{1.6}\n\end{array}$ $\begin{array}{ccc}\n\text{2.7}\n\text{2.7}\n\end{array}$

*Equal contributions

Benjamin D. Solomon

National Human Genome Research Institute

Building 10 - CRC, Suite 3-2551

galicies
10 Center Drive
Bethesda, MD 20892 Bethesda, MD 2
Phone: 301-402

Bethesda, MD 20892 Email: <u>solomonb@ma</u>
. Email: solomonb@mail.nih.gov

$\overline{2}$

\overline{a}

Abstract
Deep learning (DL) and other types of artificial intelligence (AI) are increasingly used in many biomedical areas, including genetics. One frequent use in medical genetics involves evaluating images of people
with potential genetic conditions to help with diagnosis. A central question involves better areas areas, including the frequent with potential genetic conditions to help with diagnosis. A central question involves better
understanding how AI classifiers assess images compared to humans. To explore this, we perfor with potential generic conditions to help with diagnosity continuity question in our conditions.
understanding how AI classifiers assess images compared to humans. To explore this, we po
eye-tracking analyses of geneticist eye-tracking analyses of geneticist clinicians and non-clinicians. We compared results to DL-based saliency maps. We found that human visual attention when assessing images differs greatly from the parts of images weighted by the DL model. Further, individuals tend to have a specific pattern of image inspection, and clinicians demonstrate different visual attention patterns than non-clinicians. inspection, and clinicians demonstrate different visual attention patterns than non-clinicians.

3

 $\begin{array}{c} \n\blacksquare \quad \blacksquare \quad \$ Introduction
Deep learning (DL) and other types of artificial intelligence (AI) are extensively and increasingly
employed in many biomedical areas, including genetics. In the field of genetics, AI is used for base employed in many biomedical areas, including genetics. In the field of genetics, AI is used for band
calling from genomic sequencing, genetic variant classification, clinical note processing, and to l employed in many biomedical areas, including generation in the field of general, the field of the field
calling from genomic sequencing, genetic variant classification, clinical note processing, and to help
assess radiolog assess radiologic data.¹ Among these applications, Al-based tools that can generate differential widespread in both research and clinical contexts. $2,3$ diagnoses based on facial images of individuals with possible genetic conditions have become

Despite controversies and unsettled questions, healthcare is poised for major near-term changes as AI is widespread in both research and clinical contexts.^{4,3}
Despite controversies and unsettled questions, healt
implemented into workflows.⁴ One crucial issue invo Implemented into workflows.⁴ One crucial issue involves understanding how AI tools perform and how
"choices" made by AI compare with those made by humans, including people with different levels of implemented into workflows." One crucial issue involves understanding how AI tools perform and how
"choices" made by AI compare with those made by humans, including people with different levels of
expertise.⁵⁻⁷ This can expertise.⁵⁻⁷ This can be helpful to test the accuracy of models and to examine whether confounders
present. For example, an analysis of DL to assess chest X-rays for signs of COVID-19 showed that present. For example, an analysis of DL to assess chest X-rays for signs of COVID-19 showed that
classifiers can be highly accurate compared to humans but may rely heavily on confounding information, present. For example, an analysis of DL to assess chest in layer that $\log n$ signs of DL to assess classifiers can be highly accurate compared to humans but may rely heavily on confounding infor
such as radiographic clues o component of the lung fields.⁸ As another example, large-language models have
shown extraordinary recent progress; while these models perform well compared to humans, they may shown extraordinary recent progress; while these models perform well compared to humans, they may
present coherent but incorrect or inconsistent information.⁹ shown extraordinary recent progress; while these models perform well competent compared to humans, they may
present coherent but incorrect or inconsistent information.⁹
To explore these issues further in the context of g

present coherent but incorrect or inconsistent information."
To explore these issues further in the context of genetics, we
attention on images of individuals with or without genetic co ່
ວ To express interact cannot in the context of generally in context of an experiment of visual
attention on images of individuals with or without genetic conditions. We approximate how well
visual attention maps of medical e attention maps of medical experts align with the saliency maps of a neural network model. We
also examine and compare the visual activity of geneticist clinicians and non-clinicians. We find that also examine and compare the visual activity of geneticist clinicians and non-clinicians. We find that
saliency maps differ substantially from regions of human attention and that clinicians and non-clinicia also enamine and compare the visual activity or generator chinemic and that clinicians and non-clinicians.
saliency maps differ substantially from regions of human attention and that clinicians and non-clinicians.
assess i saliency maps differently from regions of the regions of the same attention and that cliential control of the maps of the clinician substantial control of the clinician substantial control of the clinicians and non-clinici

assess images and complete
Methods
Data collection Methods
Data collection Data collection

 $\overline{4}$

 $\frac{1}{1}$ Similar to our previous methods,^{10,11} we selected publicly available images of individuals between 2 and
18 years of age affected by one of 10 genetic conditions: 22q11.2 deletion syndrome (22q11DS),
Beckwith-Wiedemann 19 years of all the conditions of all the conditions of the Beckwith-Wiedemann syndrome (BWS), Cornelia de Lange syndrome (CdLS), Down Syndrome (D.
10 Kabuki syndrome (KS), Noonan syndrome (NS), Prader-Willi syndrome (PWS) Beckwith-Wiedemann syndrome (EUP), Echtena de Lange syndrome (Cale), EUR) yndrome (C-L),
Kabuki syndrome (KS), Noonan syndrome (NS), Prader-Willi syndrome (PWS), Rubinstein-Taybi
syndrome (RSTS1), Wolf-Hirschhorn syndrome syndrome (RSTS1), Wolf-Hirschhorn syndrome (WHS), and Williams syndrome (WS), as well as images of unaffected individuals similar in age and ancestral diversity to the affected individuals. The genetic
conditions were chosen as they represent relatively common genetic conditions that involve recognizable craniofacial features, and with which geneticists would be expected to be at least conditions were cognizable craniofacial features, and with which geneticists would be expected to be at leating
Somewhat familiar.¹² Two geneticists (BDS, CTN) and a genetic counselor (RLW) on the study recognizable cranic features, and with which generated to be expected to be at least
somewhat familiar.¹² Two geneticists (BDS, CTN) and a genetic counselor (RLW) on the study te
selected the images to represent typical selected the images to represent typical images representing the chosen conditions; images of some
conditions were felt to be more difficult to recognize than others. Sixteen images (one image for each of selected to the interest of the interest of the interest conditions were felt to be more difficult to recognize than others. Sixteen images (one image for each
the above conditions, and six images of unaffected individuals conditions, and six images of unaffected individuals) were selected for the eye-tracking
experiments. The selected images were at least 720-pixel resolution; they were each standardly the above continuous, and six images of unaffected individuals) were concerned to the eye-tracking
experiments. The selected images were at least 720-pixel resolution; they were each standardly
cropped, centered, and align experienced in the selection of the selection of the stopped, centered, and aligned (e.g., the eyes, nose, and mouth were roughly repositioned at the standard resolution; the standard resolution; the standard standard reso coordinates for all images). In our paper to, we include mentions of the figures we prepared; as preprint coordinates for all images). In our paper to, we include mentions of the figures we prepared, as preprint
servers may understandably not allow such images to be shown, even when previously published, we
include an asterisk servers may understandably not allow such images to be shown, even when previously published, we include an asterisk (*) by each figure number to explain why it is not available in this version, or why a portion of the figure is not shown. See **Supplementary Figure 1*** for image sources; images are shown as
the standardized, processed versions used in the eye-tracking experiments. Note that all images shown are versions of previously published and/or publicly available images.
Eye tracking experiments

are versions of previously publications and/or publicly available images.
Eye tracking experiments
The formatted images were embedded in a screen-based eye-tracking Eye tracking experiments
The formatted images we
Lab Pro version 1.194.412 Lab Pro version 1.194.41215; https://www.tobii.com/, Stockholm Sweden). Eye-tracking experiments Lab Pro version 1.194.41215; https://www.tobii.com/, Stockholm Sweden). Eye-tracking experiments

5

t
c took place in the restauration increased in the National (NIH) (Bethesda, Maryland, Partyl States),
and the University of Bonn (Bonn, Germany). The study was considered exempt from formal IRB review
After calibration, each after calibration, each participant viewed the 16 images for seven seconds per image, and answered
questions about each image, including whether the image showed a person affected by a genetic After calibration, each participant viewed the 16 images for seven celebratic per image, and anticipated
questions about each image, including whether the image showed a person affected by a genetic
condition, and, if so, questions about each image, including unsure the image showed a person antested r a generical condition, and, if so, what condition the person had. After extensive initial testing, we chose sever
seconds for the viewing seconds for the viewing time, as subjective feedback and preliminary assessments showed that this amount of time was sufficient to assess an image but minimized visually revisiting areas of the image in
a way that might not further inform the assessment. To minimize head movement or distractions, a way that might not further inform the assessment. To minimize head movement or distractions,
questions were asked verbally during the eye-tracking portions of the experiment. Responses were a way that might not further information as securities in information to minimize the assessment.
questions were asked verbally during the eye-tracking portions of the experiment. Responses were
documented manually by a st documented manually by a study team member.
The NIH cohort included 17 individuals, including physician geneticists (14) and physicians in genetics

subspecialty training (3). The Bonn cohort included 29 total individuals, including physician geneticists Subspecialty training (3). The Bonn cohort included 29 total individuals, including physician geneticists
(2) and physicians in genetics subspecialty training (5), as well as 22 non-clinicians. Some of the nonsubspecialty training (3). The Bonn content included 29 total individuals, including physicial generations.
(2) and physicians in genetics subspecialty training (5), as well as 22 non-clinicians. Some of the non-
clinician example, some are graduate students who study applications of AI in genetics, and thus are familiar example, some are graduate students who study applications of AI in genetics, and thus are familiar wit
genetic conditions, but are not trained clinicians. example, some are graduate students who study applications who genetic, and thus are familiar with
genetic conditions, but are not trained clinicians.
Data extraction and analysis

genetic conditions
Data extraction and analysis
We extracted eye-tracking data for analysis in two *Data extraction and analysis*
We extracted eye-tracking da
for each participant and imag for each participant and image. We used the default Tobii software settings, but with two changes to
enable our analyses: a) we changed the output eye-tracking radius to 25 pixels, as our preliminary analyses showed that the default 50 pixel radius was insufficiently precise for comparisons; b) we used a enable our analyses: a) we changed the output eye-tracking radius to 25 pixels, as our preliminary
analyses showed that the default 50 pixel radius was insufficiently precise for comparisons; b) we i
more homogeneous color analysis of the default of the default of the default field in the default of the default of the heat maps (with the following hexadecimal color codes: high:
#943126, medium: #B03A2E, low: #CB4335), which enabled our quant #943126, medium: #B03A2E, low: #CB4335), which enabled our quantitative analyses to take into $\frac{m}{2}$

6

E
T account all captured heat map data, including to compare eye-tracking data with classifier saliency
maps.
Second, to enable additional analyses that took into account gaze trajectory behavior to analyze th

r
Secono
timin<mark>g</mark> timing of participant gaze,¹³ a dysmorphologist (BDS) manually drew areas-of-interest (AOIs) for each image (see **Supplementary Figures 2a-j***). The AOI set included only those features that were listed as
having dysmorphic manifestations based on the clinical synopses in OMIM (https://www.omim.org/) image (see Supplementary Figures 2a-j*). The AOI set included only those reatures that were listed as
having dysmorphic manifestations based on the clinical synopses in OMIM (https://www.omim.org/)
and which were also pres having dysinorphic manifestations above the clinical synopter in China (<u>https://www.manifestation</u>)
and which were also present in the images. For example, if a condition were listed in the clinical synop
section of OMIM section of OMIM as having a dysmorphic manifestation affecting the eyebrows and that manifestation
was present in the image of a person with that condition used in the survey, an AOI was drawn around section of OMIM as having a dysinorphic manifestation affecting the eyed and that manifestation
was present in the image of a person with that condition used in the survey, an AOI was drawn around
the eyebrows. Using the T was present in the image or a person with that condition used in the survey, and that such that the international
the eyebrows. Using the Tobii software, we extracted tabular data for analysis based on these defined
AOIs.

AOIs.
Prior to analysis, we manually reviewed results and excluded heat maps where eye-tracking data was not recorded (this occurred for all results for one NIH participant, which may have been due to Prior to analysis, it entimality results for and exclude the changed maps where eye-tracking data map
not recorded (this occurred for all results for one NIH participant, which may have been due to
ophthalmologic issues li ophthalmologic issues like severe myopia, and seven total other isolated eye-tracking results fo
reasons). ophthalmologic issues is that standard heat maps primarily showed that the areas around the eyes, nose, and we observed that standard heat maps primarily showed that the areas around the eyes, nose, and

reasons).
We obser
mouth are mouth are the regions with the highest attention. These common visual attentions (Supplementary Figure 3), which likely reflect standard human behavior, make it difficult to quantitatively compare the
visual activity differences between groups unless accounted for. To mitigate this issue, we computed the average heat map for all the clinicians and non-clinicians (separately) over all of the images. We then average heat map for all the clinicians and non-clinicians (separately) over all of the images. We then
subtracted this common average gaze pattern (separately for clinicians and non-clinicians) from each average heat map for all the clinicians and non-clinicians and non-clinicians) from each
subtracted this common average gaze pattern (separately for clinicians and non-clinicians) from each
heat map. This does not cause us subtracted this common facial attention in our analyses but
helps account for typical human behavior when viewing faces. **Figure 1*** explains our data prehelps account for typical human behavior when viewing faces. **Figure 1*** explains our data pre-
processing approach. helps account for typical human behavior when viewing faces. Figure 1 explains our data pre-
processing approach. processing approach.

 $\overline{7}$

 $\frac{1}{1}$ In a later, and your analyses, we focused on four subgroups: clinicians and non-clinician (e.g., see the
Categories in Table 1). One interest is whether the untrained intuition of a non-clinician aligns with or incorrectly ideas.
Categories in Table 1). One interest is whether the untrained intuition of a non-clinician aligns with a clinician behavior. In the later sections of our analyses, we use the terms successful and categories in Table 1). One interest is whether the untrained intuition or a non-clinician aligns with
clinician behavior. In the later sections of our analyses, we use the terms successful and
underperforming clinicians a clinician behavior. In the later sections of our analyses, we use the terms successful and
underperforming clinicians and non-clinicians to refer to the clinicians and non-clinicians
and incorrectly recognize the presence and incorrectly recognize the presence of a genetic disease for a given test image, respectively. For two different test images, the same participant may fail to recognize the disorders in both images. Hence, for two different images, the groups of successful clinicians (and likewise non-clinicians) may not have the
same participants. the annormanges, the groups of succession annormality (and likewise non-clinicians) may no may no
same participants.
To understand the general behavior of a specific participant group (e.g., successful clinician group), af

same participants.
To understand the
accounting for the accounting for the areas of common visual attention, we computed the average heat map for each test
image over all the participants. Next, we applied two different thresholds to binarize an average heat accomming for the areas of common visual attention, we compute the average measuring for each test
image over all the participants. Next, we applied two different thresholds to binarize an average heat
map so that the valu image over all the participants. Next, we applied the all thresholds to binarize an average heat
map so that the values in this heat map are either 0 or 1.¹⁴ The first, lower threshold is meant to remo
spurious noise. Th spurious noise. The second, higher threshold removes a large proportion of the signal; here, we would
analyze only facial regions with the highest visual attention. At each of these thresholds, we use the spurses in the second, might intermediations and gap proportion of these thresholds, we use the
analyze only facial regions with the highest visual attention. At each of these thresholds, we use the
intersection-over-union analyze only intersection-over-union metric (IoU) to compare average heat maps.¹⁴ Standard deviation for the
highest visual attention. At each other the observed IoU was computed via bootstrapping. Finally, we applied ra observed IoU was computed via bootstrapping. Finally, we applied random effects meta-analysis (we refer to this as "RE Model" in the Figures) to summarize the behavioral differences over many images between the two groups of participants (e.g., successful clinicians versus successful non-clinicians).
Classifier and saliency maps

between the two groups of participants (e.g., successful clinicians versus successive classifier and saliency maps

Our test set is the 16 images selected in the eye-tracking experiment. With the remaining images of the
images collected (but which were not used for eye-tracking experiments), we trained EfficientNet-B4 our test images collected (but which were not used for eye-tracking experiments), we trained EfficientNet-B4
using 5-fold cross- validation to build an ensemble model.¹⁵ Each image was resized to 448 x 448 pixels, images concreat (but which were not used for eye-tracking experiments), we change an entired for every collect
using 5-fold cross-validation to build an ensemble model.¹⁵ Each image was resized to 448 x 448 pixel using 5-fold cross- validation to build an ensemble model.¹⁵ Each image was resized to 448 x 448 pixels,
,
,

8

E
C and cross-entropy loss with equal weights for the diseases was used. Training process follows standard
convention; for example, we tried different values for batch size and dropout rate and found that batch
size 64 and dro convertion; for example is the size 64 and dropout rate 0.2 worked best. We also found that a learning rate 0.00003 and the Adam
optimizer were most effective for our analyses.¹⁶ For a test image, to combine the predicti size 64 and dropout also found a learning term of the and dropout rate of the continue that and the that and o
poptimizer were most effective for our analyses.¹⁶ For a test image, to combine the prediction of the
models models from 5-fold cross-validation, we average the classifiers' outputs for this image.

The occlusion method with box size 20 x 20 pixels and stride 10 x 10 was used to produce the saliency maps.¹⁷ The intuition is to remove a 20 x 20 pixel box from the image and then measure how much the prediction accuracy drops. This idea allows us to identify key regions affecting the classifier's output. We average the saliency map produced by the models from 5-fold cross-validation for each test image. We then apply a low and a high threshold to respectively remove just spurious signals and retain only the
most important regions. Finally, we compute the loU score for each image comparing the participant then are somewhat the soles are threshold to remove then a high threshold to remove important regions. Finally, we compute the loU score for each image comparing the participant heat maps and the saliency map at the same f most maps and the saliency map at the same filtering threshold (e.g., low or high for both human and
model). We acknowledge that there are other models besides EfficientNet-B4 that have been studied head maps and the salidity map at the same mising incentive (e.g., low or high for both hamman and
model). We acknowledge that there are other models besides EfficientNet-B4 that have been studied
the context of genetic an the context of genetic and other congenital diseases.^{2,3} Moreover, there are other saliency approaches
besides the occlusion method. This paper's combination of EfficientNet-B4 and the occlusion besides the occlusion method. This paper's combination of EfficientNet-B4 and the occlusion besides the occlusion method. This paper's combination of EfficientNet-B4 and the occlusion
method seems to return reasonable results. In future studies, we plan to evaluate other ty
of image classifiers and saliency appro method seems to return reasonable results. In future studies, we plan to evaluate other types of image classifiers and saliency approaches.

Code is available at: <u>https://github.com/datduong/tobii-eye-track-syndromic-faces</u> and
https://github.com/datduong/classify-syndromic-faces
Results

https://github.com/data.com/data-faces.com/data-faces.com/data-faces.com/data-faces.com/data-faces.com/data-faces.com/data-faces.com/data-faces.com/data-faces.com/data-faces.com/data-faces.com/data-faces.com/data-faces.com

nesuns
General
We divid General accuracy and results
We divided the human partic
Bonn) and 22 non-clinicians. Bonn) and 22 non-clinicians. Table 1 compares the accuracy of the two groups at identifying whether an Bonn) and 22 non-clinicians. Table 1 compares the accuracy of the two groups at identifying whether an

9

 $\frac{1}{2}$ i
I image shows a person an extensive mannerical maning group is condition. The climician group is a
more accurate than the non-clinician group at recognizing whether an image showed an affected perso
(correct identification w %
(correct identification was 85.6% for clinicians versus 76.9% for non-clinicians, p = 0.0032 by Chi-square).
In further responding to a question about what specific genetic condition the image might depict, non-In further responding to a question about what specific genetic condition the image might depict, non-
clinicians were, unsurprisingly, usually unable to identify the disease names. However, some nonclinicians are familiar with specific conditions based on their work, such as involvement in genetic clinicians were, usually usually usually usually understanding the disease names. However, some non-transformation of the disease names. However, some names is not the disease names. However, some names is not the disease clinicians are familiar with specific conditions based on their with specific conditions generic
research.
Although our neural net (NN) classifier model was trained to predict 11 different labels (10 disease

Although
unaffecte unaffected) for the conditions included in this study, when looking at unaffected and affected as the only two label choices, our classifier achieves 100% accuracy (see more details below).

Evaluating human visual attention versus saliency map

NN models perform at least as well as humans at many image classification tasks, including the identification of genetic and other congenital diseases from facial images.^{2,9,11} For a specific input image,
one can extract the corresponding saliency map from a NN image classifier. This saliency map indicates identification of genetic and other congenital diseases from facial images.^{2,9,11} For a specific input image,
one can extract the corresponding saliency map from a NN image classifier. This saliency map indicates
the reg one can extract the cagions of the image that the NN model deems to be relevant for the prediction of the ground truth
label. Our NN classifier correctly labeled 13 out of the 16 test images with the specific genetic condi label. Our NN classifier correctly labeled 13 out of the 16 test images with the specific genetic condition
that person has; averaging over these 13 images, the prediction probability of the ground truth label is label. Our Nat person has; averaging over these 13 images, the prediction probability of the ground truth label is
1.7853. The three misclassified affected images are CdLS, PWS, and RSTS1. For these three images, the that person has; as enging over these 13 images, the prediction probability of the ground truth label
0.7853. The three misclassified affected images are CdLS, PWS, and RSTS1. For these three images, the
predicted probabil predicted probabilities of the ground truth label and incorrect label were: CdLS (0.3196 vs 0.3285
misclassified as NS), PWS (0.3383 vs 0.4690 misclassified as WS), and RSTS1 (0.0005 vs 0.7976 misclassified as NS). Because the classification accuracy of RSTS1 is low, we excluded this image in the misclassified as NS). Because the classification accuracy of RSTS1 is low, we excluded this imag
analysis of this section (Supplementary Figure 4 contains the analysis for all the affected image analysis of this section (**Supplementary Figure 4** contains the analysis for all the affected images). We
extracted the saliency maps for our NN classifiers to observe regions of the images that affect the extracted the saliency maps for our NN classifiers to observe regions of the images that affect the extracted the saliency maps for our NN classifiers to observe regions of the images that affect the

 $\overline{1}$ $\frac{1}{2}$ model spredicted probability for their ground truth labels. In other words, this approach allowed us to
create saliency maps that are specific to each condition.
In this section, we evaluate whether the clinicians' visual

create to the clinicians' visual
In this section, we evaluate whether the clinicians' visual
Classifier's saliency maps. We do not imply that humans classifier's saliency maps. We do not imply that humans and computers are doing the same thing when classifying an image, but this method allows a means of comparison. For example, a heat map shows
how a single human participant assesses a single image. Although we can combine heat maps from classify the case of controls in this method as single image. Although we can combine heat maps from
different participants or of different images, generalizations and comparisons to the saliency map bas how a single human participant assesses a single image. Although we can communicate inacting them.
different participants or of different images, generalizations and comparisons to the saliency map b
on a NN may be difficu different participants or of difficult.
For the analyses, we treat the saliency maps as if they were heat maps, and then find the most

on a normal parameter.
For the analyses, we treat
reasonable threshold to b reasonable threshold to binarize the saliency maps. Next, for a specific test image of an affected individual, we compute the IoU between the saliency map and the average heat maps of successful
clinicians; examples are shown in Figure 2* (see also Supplementary Figures 5a-j*).

Figure 2^* also shows the IoU metrics averaged over the 9 affected images (excluding RSTS1 for reason clinicians, examples are shown in Figure 2 and Supplementary Figures Sa-j*).
Figure 2* also shows the loU metrics averaged over the 9 affected images (excluding
mentioned above). We observed significant differences between Figure 2* also shows the IoO metrics averaged over the 9 affected images (excluding RSTS1 for reason
mentioned above). We observed significant differences between human visual attention and what a NI
model considers import model considers important. For example, the average loU metric for successful clinicians versus the
saliency map we generated was 0.11 for both low and high threshold heat maps. This loU metric is much model considers important. For example, the average Islam is considered as antisomer consider the succession
saliency map we generated was 0.11 for both low and high threshold heat maps. This loU metric is r
less than the saliency map we generate ware in the metric comparing our clinicians and non-clinician groups (see below). In other words,
human heat maps show that our participants tended to look at very different regions than the classi less than the metric comparing our clinicians and non-clinician groups (see below). In other words,
human heat maps show that our participants tended to look at very different regions than the class
per the models we exami heat the models we examined.
Analyses comparing groups of human participants

r
Analyses comparing groups of
Using the methods described a Analyses comparing groups of human participants

Using the methods described above, we compared the successful clinicians and non-clinicians for each
of the 16 test images, and then summarized the visual attention differences over all these images. In
Figure 3*, qualitat of the 16 test images, and then summarized the visual attention anti-tention of the visition of **Figure 3***, qualitatively speaking, the visual attentions of these clinicians and non-clinicians are more similar at a lower Figure 3*, qualitatively speaking, the visual attentions of these clinicians and non-clinicians are more
similar at a lower threshold, where only spurious visual signals were removed, and later become mor similar at a lower threshold, where only spurious visual signals were removed, and later become more

 $\overline{1}$ different at a higher threshold, where only regions with high visual interests were retained (Figure 3*).
Thus, as shown in Figure 3*, although these clinicians and non-clinicians correctly recognize affected
versus unaffe Thus, as shown in Figure 3*, although these clinicians and non-clinicians correctly recognize affected
versus unaffected individuals, on average these two participant groups do not show similar visual
attention to the same versus unancesses unaffectives in average these the participant groups do not shown similar visual
attention to the same facial regions; the IoU is 0.47 and 0.34 for low and high thresholds, respectiv
Figure 4 quantitively Figure 4 quantitively compares the visual interests of successful clinicians versus successful nonclinicians. Here, the result aligns with our qualitative inspection in Figure 3. When excluding low spurious signals, these two subgroups show similar eye-gaze interests (IoU 0.47). However, these two
subgroups become more different when considering only the greatest visual attention (the IoU dropped subgroups become more different when considering only the greatest visual attention (the IoU droppe
to 0.34). This same trend is observed when we analyzed on just the set of affected and unaffected subsetting only to 0.34). This same trend is observed when we analyzed on just the set of affected and unaffected
Images (Supplementary Figure 6). to be any this same trend is observed when we analyzed on just the set of affected and unaffected
images (Supplementary Figure 6).
We affect the set of a few of a few of a few orders on the set of a few orders of a few ord images (Supplementary Figure 6).
We further compare successful clinicians and underperforming clinicians. Figure 5a shows that, on

average, these two subgroups of clinicians do not have similar visual interests. The same trend is seen
when stratifying the non-clinicians according to their accuracy (Figure 5b). Conditioned only on average, these the subgroups of clinicians are the that community clinical interestion in statute as the second
when stratifying the non-clinicians according to their accuracy (Figure 5b). Conditioned only on
participants when stratifying the non-clinicians according to their accuracy (Figure 5b). Conditioned only only
participants who misclassified the images, **Figure 5c** also shows visual attention differences bety
clinicians and non-clin participants who misclassified the images, Figure 5c also shows visual attention differences between
clinicians and non-clinicians. Hence, when participants misclassify images, they have distinct ways of
inspecting the ima constants and non-clinicians critici_s, when participants misclassify images, they have distinct may be distin
inspecting the images that do not appear to be similarly influenced by some common confounders in
the images (the images (e.g., hairstyles, facial expressions, or earrings that could affect a participant's decision).
Overall, there seems to be more similar visual attention when a participant (whether a clinician or not) the images (e.g., hairstyles, facial expression) of earrings that could affect a paracipant could be
Overall, there seems to be more similar visual attention when a participant (whether a clinician or n
can correctly ident Can correctly identify that a person is affected by a genetic condition; for example, the IoU values in

Figure 4 are higher than those in Figure 5. correctly identify identify the person correction is a person in the India
The person is a person is a person is a generic condition; for example, the India
We also examined the heat maps from the clinicians and non-clinic

Figure 4 are higher than those in Figure 5.
We also examined the heat maps from the
whether there were different visual patter whether there were different visual patterns when clinicians examine affected or unaffected images. As shown in Supplementary Figures 7 and 8, participants had different individual patterns of assessing shown in Supplementary Figures 7 and 8, participants had different individual patterns of assessing

12
images. However, this individual pattern appeared consistent whether the participant observed affected $\frac{1}{\sqrt{2}}$ versus unaffected images (see also **Supplementary Figure 9***).
Finally, we examined the AOIs that corresponded to specific dysmorphic features in the images of

versus unaffected images (see also Supplementary Figure 9*).
Finally, we examined the AOIs that corresponded to specific dy
affected individuals (Supplementary Figure 2a-j*). As each test Finally, we examined the AOIS that corresponded to specific ay interpresentation in the images of
affected individuals (Supplementary Figure 2a-j*). As each test image has more than one AOI that
be important to the underly be important to the underlying conditions, metrics like duration-of-fixation and time-to-first-whole-
fixation for any single AOI can have high standard deviations, making it difficult to observe statistical ${\rm sgnr}$ icance. Overall, we did not detect a specific pattern for these AOIs that differentiated the $$ fixation for any single AOI can have inglued and the deviations, making it anticial to observe statistical
significance. Overall, we did not detect a specific pattern for these AOIs that differentiated the
categories of pa significance. Overall, we did not detect a specific pattern for the specific pattern for the AOIs that differentiated the AOIs that differentiated the AOIs that differentiate the AOIs that differentiate the AOIs that diffe

categories of participants.
Discussion
This study has two key resi **Discussion**
This study ł
evaluating i evaluating images of individuals with potential genetic conditions. This does not mean that one is superior to the other, but these types of analyses and metrics may be helpful in future studies. For
example, methods to compare human and computer attention can be used to explore potential superior to the other, but these types of analyses and metrics may be helpful in future studies.
Example, methods to compare human and computer attention can be used to explore potential
confounders in Al-based analyses, o example, methods that improve the accuracy or applicabi
tools. As AI is increasingly adopted in clinical scenarios, such studies will be critical to assess mod cools. As AI is increasingly adopted in clinical scenarios, such studies will be critical to assess model
performance. For generalizability, such future studies would need to be much larger, both in terms of to the the method increasingly adopted in clinical scenarios, such statute in a scenario ansetted in and
performance. For generalizability, such future studies would need to be much larger, both in terms
the number of part performance. For generalizability, such the number of images and genetic conditions included.
Second, clinicians and non-clinicians exhibit different gaze behaviors when assessing images. This is no the number of participants and the number of images and generic conditions when assessing im
Second, clinicians and non-clinicians exhibit different gaze behaviors when assessing im
surprising, but quantifying these behavi surprising, but quantifying these behaviors using methods like these may be helpful for activities such as
ascertaining which phenotypic characteristics may be diagnostically important but which are frequently surprising, but any surprision of these behaviors of these behaviors ascertaining which phenotypic characteristics may be diagnostically important but which are frequently
overlooked. Again, as AI support enters more and m as secure, then there system and actually in the single characteristics may be discussed to specific again, as
Noverlooked. Again, as AI support enters more and more clinical areas, these types of analyses may point
to spe to specific ways to augment the relationship between clinicians and AI tools. For example, data from extremely high-performing clinicians in human/classifier comparison experiments may be useful to extremely high-performal of the sextremely herein in the education of less experienced clinicians or trainees, as well as AI tools.⁹ design the education of less experienced clinicians or trainees, as well as AI tools.⁹

--
|
V $\frac{1}{1}$ This study has several limitations. These include the number of participants and images viewed. The
images also represent heterogeneous genetic conditions, and eye-tracking behavior may be affected by
certain aspects of a images are represent of a particular image. While we grouped clinicians and non-clinicians into separate
categories, there is varying experience and expertise within these groups, and differences in behavior categories, there is varying experience and expertise within these groups, and differences in behave
between individuals. Additionally, our analyses focused on metrics like human visual attention, wh between individuals. Additionally, our analyses focused on metrics like human visual attention, which may not reflect what is most important to a person evaluating an image. For example, an expert clinician
may immediately perceive a key visual clue to a diagnosis and may then move on to spend more time may immediately perceive a key visual clue to a diagnosis and may then move on to spend more time
observing less obvious features or searching the image for subtle clues. Overall, while our results show may interpret and the set of the solvent observing less obvious features or searching the image for subtle clues. Overall, while our results show
interesting trends, caution should be taken that they will generalize to oth interesting trends, caution should be taken that they will generalize to other or all genetic conditions, to
larger groups of individuals, or to other Al-based approaches.

As mentioned above, the results here point to multiple future possibilities, including involving larger numbers of participants and datasets to analyze further how well the results can be extrapolated. numbers of participants and datasets to analyze further how well the results can be extrapolated.
Related eye-tracking experiments could be used to explore multiple questions germane to genetics, s numbers of participants and datasets to analyze further how well the results can be entanglemental.
Related eye-tracking experiments could be used to explore multiple questions germane to genetic
as with different data typ relations of the setted examination features as with different data types (e.g., radiological studies or other physical examination features)
encountered in clinical practice. Additional work could be done specifically qua an as with different data types (e.g., radiology)
and the done specifically quantifying differe
classifiers and saliency methods. encountered in clinical practice. Additional work could be done specifically quantifying different
classifiers and saliency methods.
Acknowledgements

Acknowledgements

classifiers and saliency methods.
 Acknowledgements

This research was supported in part 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.
References

- of the NIH HPC Biowulf cluster.
 References

1 Ledgister Hanchard, S. E., Dwyer, M.C., Liu, S., Hu, P., Tekendo-Ngongang, C., Waikel, R.L., Duong References
1 Led_i
D. S
Mec 1. Solomon, B.D. Scoping review and classification of deep learning in medical genetics. Genet
12 Med (2022).
12 Hsieh, T. C. et al. GestaltMatcher facilitates rare disease matching using facial phenotype
- D. Solomon, B.D. Scoping review and classification of deep learning in medical genetics. Oenet
Med (2022).
Hsieh, T. C. *et al.* GestaltMatcher facilitates rare disease matching using facial phenotype
descriptors. Nat Gene Med (2022).
Hsieh, T. C. *e*
descriptors. *I* descriptors. Nat Genet 54, 349-357, doi:10.1038/s41588-021-01010-x (2022). descriptors. Nat Genet 54, 349-357, doi:10.1038/s41588-021-01010-x (2022).

- \ddot{a} 3 Portal and The Particulary 1999 and 1

3 Children: a multinational retrospective study. *Lancet Digit Health*, doi:10.1016/S2589-

3 Porras, D. Deve evaluation of a multinational retrospective study. Lancet Digit Health, doi:10.1016/S2589-
1980 (21)00137-0 (2021).
Lee, P., Bubeck, S. & Petro, J. Benefits, Limits, and Risks of GPT-4 as an Al Chatbot for Medicine
N Engl
- children: a mahmational retrospective study. *Lancet Digit Health, doi:10:1010/32589*
7500(21)00137-0 (2021).
Lee, P., Bubeck, S. & Petro, J. Benefits, Limits, and Risks of GPT-4 as an Al Chatbot for I
N Engl J Med **388** Lee, P., Bubeck, S. & Petro
N Engl J Med **388**, 1233-1
Krupinski, E. A. Visual scar
137-144, doi:10.1016/s10
- 4 Lengl J Med 388, 1233-1239, doi:10.1056/NEJMsr2214184 (2023).

4 Lengl J Med 388, 1233-1239, doi:10.1056/NEJMsr2214184 (2023).

4 Engl J Med 388, 1233-1239, doi:10.1056/NEJMsr2214184 (2023).

4 Engl J Med 388, 1233-1239,
- N Engra Med 388, 1233-1235, doi:10:1056/NEJWSI2214184 (2023).
Krupinski, E. A. Visual scanning patterns of radiologists searching ma
137-144, doi:10.1016/s1076-6332(05)80381-2 (1996).
Bond, R. R. *et al.* Assessing compute 137-144, doi:10.1016/s1076-6332(05)80381-2 (1996).

Bond, R. R. *et al.* Assessing computerized eye tracking technology for gaining insight into expert

interpretation of the 12-lead electrocardiogram: an objective quantit Bond, R. R. *et al.* Assessing computerized eye tracking interpretation of the 12-lead electrocardiogram: an ob
Electrocardiol **47**, 895-906, doi:10.1016/j.jelectrocard..
Sudin, E. *et al.* Digital pathology: the effect of 6 Bond, R. R. et al. Assessing computerized eye tracking technology for gaining insight into expert
interpretation of the 12-lead electrocardiogram: an objective quantitative approach. J
Electrocardiol 47, 895-906, doi:10.
- Flectrocardiol 47, 895-906, doi:10.1016/j.jelectrocard.2014.07.011 (2014).
Sudin, E. *et al.* Digital pathology: the effect of experience on visual search behavior. *J N*
Imaging (Bellingham) 9, 035501, doi:10.1117/1.JMI.9 Electrocardiol 47, 893-900, doi:10.10107j.jelectrocard.2014.07.011 (2014).
Sudin, E. *et al.* Digital pathology: the effect of experience on visual search k
Imaging (Bellingham) 9, 035501, doi:10.1117/1.JMI.9.3.035501 (202
- Franchine Internation of the effect of experience on visual search behavior. J Med
Imaging (Bellingham) 9, 035501, doi:10.1117/1.JMl.9.3.035501 (2022).
DeGrave, A. J., Janizek, J. D. & Lee, S.-I. Al for radiographic COVID-
- Imaging (Bellingham) 9, 033501, doi:10.1117/1.JMI.5.3.033501 (2022).
DeGrave, A. J., Janizek, J. D. & Lee, S.-l. Al for radiographic COVID-19 det
over signal. Nature Machine Intelligence 3, 610-619, doi:10.1038/s4225
Duong over signal. *Nature Machine Intelligence* 3, 610-619, doi:10.1038/s42256-021-00338-7 (2021).
Duong, D., Waikel, R. L., Hu, P., Tekendo-Ngongang, C. & Solomon, B. D. Neural network
classifiers for images of genetic conditi over signal. Nature Machine Intelligence 3, 610-613, doi:10:1038/342230-021-00338-7 (2021).
Duong, D., Waikel, R. L., Hu, P., Tekendo-Ngongang, C. & Solomon, B. D. Neural network
dassifiers for images of genetic conditions
- examples of genetic conditions with cutaneous manifestations. HGG Adv 3, 11
doi:10.1016/j.xhgg.2021.100053 (2022).
Duong, D., Waikel, R. L., Hu, P., Tekendo-Ngongang, C. & Solomon, B. D. Neural network
classifiers for imag classifiers for images of genetic conditions with cutaneous manifestations. Hoo Adv 3, 100053,
doi:10.1016/j.xhgg.2021.100053 (2022).
Duong, D., Waikel, R. L., Hu, P., Tekendo-Ngongang, C. & Solomon, B. D. Neural network
c Duong, D., Waikel, R. L., Hu, P., Tekendo-
classifiers for images of genetic conditior
Genomics Advances, 100053 (2021).
Duong, D. *et al.* Neural Networks for Clas 10 classifiers for images of genetic conditions with cutaneous manifestations. Human Genet
11 Genomics Advances, 100053 (2021).
11 Duong, D. et al. Neural Networks for Classification and Image Generation of Aging in Gen
13
- Genomics Advances, 100053 (2021).
Duong, D. *et al.* Neural Networks for Classification and Image Generation of Aging in Genetic
Syndromes. *Front Genet* **13**, 864092, doi:10.3389/fgene.2022.864092 (2022).
Nguengang Wakap,
- Syndromes. Front Genet 13, 864092, doi:10.3389/fgene.2022.864092 (2022).
Nguengang Wakap, S. et al. Estimating cumulative point prevalence of rare diseases: an
the Orphanet database. *Eur J Hum Genet* 28, 165-173, doi:10.1 Buong, D. et al. Neural Networks for Classification and Image Generation of Aging in Genetic
Syndromes. Front Genet 13, 864092, doi:10.3389/fgene.2022.864092 (2022).
12 Nguengang Wakap, S. *et al.* Estimating cumulative po Syndromes. From Genet 13, 884092, doi:10.33897igene.2022.884092 (2022).
Nguengang Wakap, S. *et al.* Estimating cumulative point prevalence of rare dist
the Orphanet database. *Eur J Hum Genet* 28, 165-173, doi:10.1038/s41
- 12 Nguengang Wakap, S. et al. Estimating cumulative point prevalence of rare diseases. analysis of
the Orphanet database. *Eur J Hum Genet* 28, 165-173, doi:10.1038/s41431-019-0508-0 (2020).
Brams, S. *et al.* The relation
- systematic review. Psychol Bull 145, 980-1027, doi:10.1037/bul0000207 (2019).
Saporta, A. et al. Benchmarking saliency methods for chest X-ray interpretation. Nature Machine
Intelligence 4, 867-878 (2022). 13 Brams, S. et al. The relationship between gaze behavior, expertise, and performance: A
systematic review. Psychol Bull 145, 980-1027, doi:10.1037/bul0000207 (2019).
Saporta, A. *et al.* Benchmarking saliency methods for systematic review. Psychol Bull 145, 380-1027, doi:10.1037/bul0000207 (2013).
Saporta, A. *et al.* Benchmarking saliency methods for chest X-ray interpretation.
Intelligence 4, 867-878 (2022).
Tan, M. & Le, Q. in *Internat*
-
- 14 Saporta, A. et al. Benchmarking saliency methods for chest X-ray interpretation. Nature Machine
Intelligence 4, 867-878 (2022).
16 Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint*
arX Tan, M. & Le, Q. in *Internatione*
Tan, M. & Le, Q. in *Internatione*
Kingma, D. P. & Ba, J. Adam: A
arXiv:1412.6980 (2014).
Zeiler, M. D. & Fergus, R. in *Cor*
- 15 Tan, M. & Le, Q. in *international conference on Machine Learning*. 6105-0114 (PMLR, 2015).
16 Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint*
arXiv:1412.6980 (2014).
2014, *Proceed* 16 Kingma, D. P. & Ba, J. Adam. A method for stochastic optimization. arxiv preprint
arXiv:1412.6980 (2014).
The Zeiler, M. D. & Fergus, R. in Computer Vision–ECCV 2014: 13th European Conferent
Switzerland, September 6-12, arxiv:1412.6980 (2014).
Zeiler, M. D. & Fergus, R.
Switzerland, September (17 Zeiler, M. D. & Fergus, R. in Computer Vision–ECCV 2014: 13th European Confer
Switzerland, September 6-12, 2014, Proceedings, Part I 13. 818-833 (Springer). Switzerland, September 6-12, 2014, Proceedings, Part I 13. 818-833 (Springer).

Table 1: For a specific image, we count the number of clinicians and non-clinicians who correctly and incorrectly identified that this image is of an affected or unaffected individual. Column Disease name indicates the number of clinicians and non-clinicians who correctly identified the disease name. As seven individual results had to be discarded for technical reasons, not every image has a complete data from 22 clinicians and 22 non-clinicians. The images are listed in the order shown to the participants.

Figure 1: An example illustrating how we pre-process Tobii visual heat maps for subsequent analyses. (a) Original image of an individual with Kabuki syndrome (KS) (not shown due to image restrictions for this preprint server). (b) For this KS image, conditioned on the group of successful clinicians, we average the default attention heat maps. (c) We removed the common visual signals (See Supplementary Figure 3). (d) Finally, we smooth the image with cv2.boxFilter and increase the color intensity. Here, comparing to (b), in (d) we can better observe the high visual interest at the lower part of the left eye in KS. ALL IMAGES OF INDIVIDUALS ARE PREVIOUSLY PUBLISHED AND/OR ARE PUBLICLY AVAILABLE - SEE SUPPLEMENTARY FIGURE 1 FOR REFERENCES FOR EACH IMAGE USED.

 (a)

Successful clinicians vs model loU threshold low

Successful clinicians vs model loU threshold high

Figure 2: (a) Saliency maps of 22q11DS (top) and KS (bottom) test image highlight key regions that affect the classifier accuracy. Like the heat maps generated through eye-tracking experiments, we also apply a low (left) and high threshold (right) to the saliency maps. Visually, there are differences compared to the visual attention of human clinicians (not shown due to image restrictions for this preprint server). (b-c) loU metric compares the visual attention of successful clinicians and model saliency maps over the 9 test diseases (removing RSTS1) at the same filtering threshold (e.g., low and then high thresholds for both saliency maps and Tobii heat maps). ALL IMAGES OF INDIVIDUALS ARE PREVIOUSLY PUBLISHED AND/OR ARE PUBLICLY AVAILABLE - SEE SUPPLEMENTARY FIGURE 1 FOR REFERENCES FOR EACH IMAGE USED.

 (a)

Successful clinicians vs nonclinicians loU threshold low

Successful clinicians vs nonclinicians loU threshold high

Figure 4: IoU metric compares the heat maps of successful clinicians versus non-clinicians. When removing only low spurious signal (left), the two groups display more similar visual interests. However, when looking at only regions with high visual interests (right), the two groups differ more drastically. The images are listed in the order shown to the participants.

Successful vs underperforming **Clinicians IoU threshold low**

Successful vs underperforming Nonclinicians IoU threshold low

Figure 5: Images are sorted based on the IoU metric, which compares the heat maps between different groups of participants. Here, we applied low filtering threshold to remove only low spurious signal from the average heat map of each group. These three comparisons show lower similarity scores than that of successful clinicians versus non-clinicians. Left: successful versus underperforming clinicians. Middle: successful versus underperforming non-clinicians. Clinicians tend to show more similar visual interests for PWS, BWS, and unaffected images. This trend is not seen for non-clinicians. Right: underperforming clinicians versus non-clinicians.