nature research

Corresponding author(s): Yaron Gurovich

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.
For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

Experimental design

1. Sample size

Describe how sample size was determined.

Multiple experiments are described in this paper:

Binary Gestalt Model, CDLS -
the test set size of N=32 frontal facial images is based on the publication (reference 4) in order to compare to the same benchmark, as published in previous work;

Binary Gestalt Model, Angelman -
test set size of N=25 is based on publication (reference 56) in order to compare to the same benchmark, as published in previous work;

Specialized Gestalt Model, Noonan -
test set size N=25 sampled from references (57, 58, 59, 60 , 61, 62) making sure samples were not used in the train sets; considering the limited available data in previous publications (57, 58, 59, 60 , 61, 62) we allocated 5 representative images per class.

Multi-class Gestalt model -
test set size of N=502, was sampled from real clinical cases submitted to the Face2Gene application, described in the Methods section. Within a certain period of time, we sampled all real diagnosed clinical cases of any of the syndromes supported at the time by DeepGestalt in Face2Gene. We removed images that were part of our training set and ignored duplicate images. In order to maintain similarity to clinical usage, no exclusions based on age or ethnicity were performed. When building the test set, we made sure that all images of each subject are either in the training set or in the test set. We ended up with 502 images covering 92 different syndromes.
In addition we sampled N=329 images from the London Medical Database, as described in the supplementary materials. In order to create a high quality test set, we applied a set of data pruning rules on the full LMD dataset of thousands of images. We excluded images with no frontal face, images of bad quality, or where the subject is under 1 or over 18 years old, images where the subject is occluded (wearing glasses for example), etc. Additional information can be found in the Methods section.

2. Data exclusions

Describe any data exclusions.

Exclusion criteria 1 - All data that was used to test the system in the different experiments, was excluded from the training sets.
Exclusion criteria 2 - We use only cases that have been either clinically or molecularly diagnosed by relevant healthcare professionals
Exclusion criteria 3 - automatically exclude images of low resolution and images where no frontal face was detected.

3. Replication

Describe the measures taken to verify the reproducibility of the experimental findings.

In order to reproduce all experiments described in this paper we created a snapshot of the data, code and models used, including instructions for the evaluation protocol. More specifically, we use version control tools (Git) and docker images to make sure that our experiments are reproducible. In addition, to allow a reproducible research, we composed a new test set of 329 images covering 93 syndromes, published in the London Medical Database. All attempts at replication were successful.

4. Randomization

Describe how samples/organisms/participants were

Multi-class Gestalt model - During a period of several weeks, we sampled all diagnosed real

Describe how samples/organisms/participants were allocated into experimental groups.

clinical cases of any of the 216 syndromes supported at the time by DeepGestalt in the Face2Gene application. This process included verification that the sampled images were not part of the training images , remove duplicates etc. As described in sub section C (Datasets) within the Online Methods section. The test set is skewed towards ultra-rare syndromes, 65% of the syndromes are present in only 1 to 5 images and 35% in 6 to 42 images. This results in a median value of 4 and average of 5.46 images per syndrome. This distribution of patients and syndromes mirrors the prevalence of rare syndromes and is, therefore, a representative test set for genetic counseling.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

To evaluate our machine learning algorithms in each experiment , we defined a blind test set. Where possible we used external test sets from publications, as described in three experiments (CDLS, AS and Noonan Syndrome). In the Multi-class Gestalt model, we used a blind test set composed of images submitted to the Face2Gene application. The training and optimization processes were blind to the test sets.

Note: all in vivo studies must report how sample size was determined and whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ Test values indicating whether an effect is present *Provide confidence intervals or give results of significance tests (e.g. P values) as exact values whenever appropriate and with effect sizes noted.*
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars in all relevant figure captions (with explicit mention of central tendency and variation)

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

The DeepGestalt model, used in this study, is available through the Face2Gene application, http://face2gene.com. The access to the published dataset is available through the same application, as described in the supplementary materials.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party.

No unique materials were used.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

No antibodies were used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

No eukaryotic cell lines were used.

b. Describe the method of cell line authentication used.

No eukaryotic cell lines were used.

c. Report whether the cell lines were tested for mycoplasma contamination.

No eukaryotic cell lines were used.

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

No commonly misidentified cell lines were used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide all relevant details on animals and/or animal-derived materials used in the study.

No animals were used.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

The covariate-relevant population description for three of the four experiments we used data published by others and thus appears in the relevant references. For the Multi-class Gestalt model experiment, the data was sampled from real clinical cases submitted to the Face2Gene application and used in a blind manner. The covariate-relevant population description for three out of the four experiments were published by others and appears in the relevant references. For the Multi-class Gestalt model experiment, the data was sampled from real clinical cases submitted to the Face2Gene application and used in a blind manner. Covariate information, when available, can be found in the supplemental materials. Following is a brief description of subjects used for training: Age-group: 0-12 (~47%), 12-above (~15%), the remainder, unreported. Sex: males (~50%), women (~40%) the remainder unreported. Diagnosis type: ~42% molecularly diagnosed. Ethnicity: Caucasian (~43%), different ethnicities (~16%), the remainder unreported.