https://doi.org/10.1038/s41591-018-0279-0

Identifying facial phenotypes of genetic disorders using deep learning

Yaron Gurovich[1]*, Yair Hanani[1], Omri Bar[1], Guy Nadav[1], Nicole Fleischer[1], Dekel Gelbman[1], Lina Basel-Salmon[2,3], Peter M. Krawitz[4], Susanne B. Kamphausen[5], Martin Zenker[5], Lynne M. Bird[6,7] and Karen W. Gripp[8]

Syndromic genetic conditions, in aggregate, affect 8% of the population[1]. Many syndromes have recognizable facial features[2] that are highly informative to clinical geneticists[3-5]. Recent studies show that facial analysis technologies measured up to the capabilities of expert clinicians in syndrome identification[6-9]. However, these technologies identified only a few disease phenotypes, limiting their role in clinical settings, where hundreds of diagnoses must be considered. Here we present a facial image analysis framework, DeepGestalt, using computer vision and deep-learning algorithms, that quantifies similarities to hundreds of syndromes. DeepGestalt outperformed clinicians in three initial experiments, two with the goal of distinguishing subjects with a target syndrome from other syndromes, and one of separating different genetic subtypes in Noonan syndrome. On the final experiment reflecting a real clinical setting problem, DeepGestalt achieved 91% top-10 accuracy in identifying the correct syndrome on 502 different images. The model was trained on a dataset of over 17,000 images representing more than 200 syndromes, curated through a community-driven phenotyping platform. DeepGestalt potentially adds considerable value to phenotypic evaluations in clinical genetics, genetic testing, research and precision medicine.

Timely diagnosis of genetic syndromes improves outcomes[10]. Due to the large number of possible syndromes and their rarity, achieving the correct diagnosis involves a lengthy and expensive process (the diagnostic odyssey)[11]. Recognition of nonclassical presentations or ultrarare syndromes is constrained by the individual expert's prior experience, making computerized systems as a reference increasingly important.

Computer vision research has long been dealing with facial analysis–related problems. DeepFace[12] showed how deep convolutional neural networks (DCNNs) achieved human-level performance on the task of person verification on the dataset Labeled Faces in the Wild[13]. Current state-of-the-art systems are trained on large-scale datasets, ranging from 0.5 million images[14] to 260 million images[15]. Computer-aided recognition of a genetic syndrome with a facial phenotype is closely related to facial recognition, but with additional challenges, such as the difficulty of data collection and the subtle phenotypic patterns of many syndromes. Earlier computer-aided syndrome recognition technologies showed promise in assisting clinicians through analysis of patients' facial images[4,7,8]. Use in clinical settings, in combination with molecular analysis, suggests that such technologies complement next-generation sequencing (NGS) analysis by inferring causative genetic variants from sequencing data[9]. However, most studies focus on distinguishing unaffected from affected individuals or recognizing a few syndromes[5] using photos captured in a constrained manner, rather than addressing the real-world problem of classifying hundreds of syndromes from unconstrained images. Additionally, previous studies have used small-scale data for training, typically up to 200 images, which are small for deep-learning models. Since no public benchmark for comparison exists, it is impossible to compare the performance or accuracy of various methods. Supplementary Table 1 compares previous studies in terms of number of syndromes and training samples, evaluation methods and accuracy.

Here we report on DeepGestalt, the technology powering Face2Gene (FDNA Inc.), a community-driven phenotyping platform trained on tens of thousands of patient images and used to analyze hundreds of syndromes. It directly uses DCNNs for classification and is based on a knowledge transfer model from an adjacent domain. DeepGestalt was evaluated on test sets collected from clinical cases and publications. Comparison to human experts was done in three different experiments where reference results are available.

Results

Methodological development of DeepGestalt. Given an input image, the first step is face detection using a cascaded DCNN-based method[16]. Facial landmarks (Fig. 1a) are detected[17] and used to geometrically normalize the face (Supplementary Fig. 1a) and to crop it into multiple regions (Fig. 1a). Each region is scaled to a fixed size (100 × 100 pixels) and converted to grayscale. Specialized DCNNs process the facial regions, predict the probability for each syndrome per region and aggregate a Gestalt model for syndrome classification. Gestalt refers to the information contained in the facial morphology. All specialized DCNNs were trained in the same manner, using the same architecture (Fig. 1b) and optimization procedure. The model was initially trained on the Casia-WebFace dataset[14] for face identification and fine-tuned to the syndromes domain using validated patient images (Supplementary Table 2) (Fig. 1a).

DeepGestalt's performance is evaluated by measuring the top-1, top-5 and top-10 accuracy. Top-10 accuracy evaluation emphasizes the clinical use of DeepGestalt as a reference tool, where all top syndromes are considered. Where applicable, we report sensitivity and specificity. Each of the above is reported with its 95% confidence interval (CI) and P value.

Binary classification problem: distinguishing a specific syndrome from a set of other syndromes. Many studies on genetic syndrome classification deal with binary problems, differentiating

[1]FDNA Inc., Boston, MA, USA. [2]Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel. [3]Recanati Genetic Institute, Rabin Medical Center & Schneider Children's Medical Center, Petah Tikva, Israel. [4]Institute for Genomic Statistic and Bioinformatics, University Hospital Bonn, Rheinische-Friedrich-Wilhelms University, Bonn, Germany. [5]Institute of Human Genetics, University Hospital Magdeburg, Magdeburg, Germany. [6]Department of Pediatrics, University of California San Diego, San Diego, CA, USA. [7]Division of Genetics/Dysmorphology, Rady Children's Hospital San Diego, San Diego, CA, USA. [8]Division of Medical Genetics, A. I. du Pont Hospital for Children/Nemours, Wilmington, DE, USA. *e-mail: yaron@fdna.com

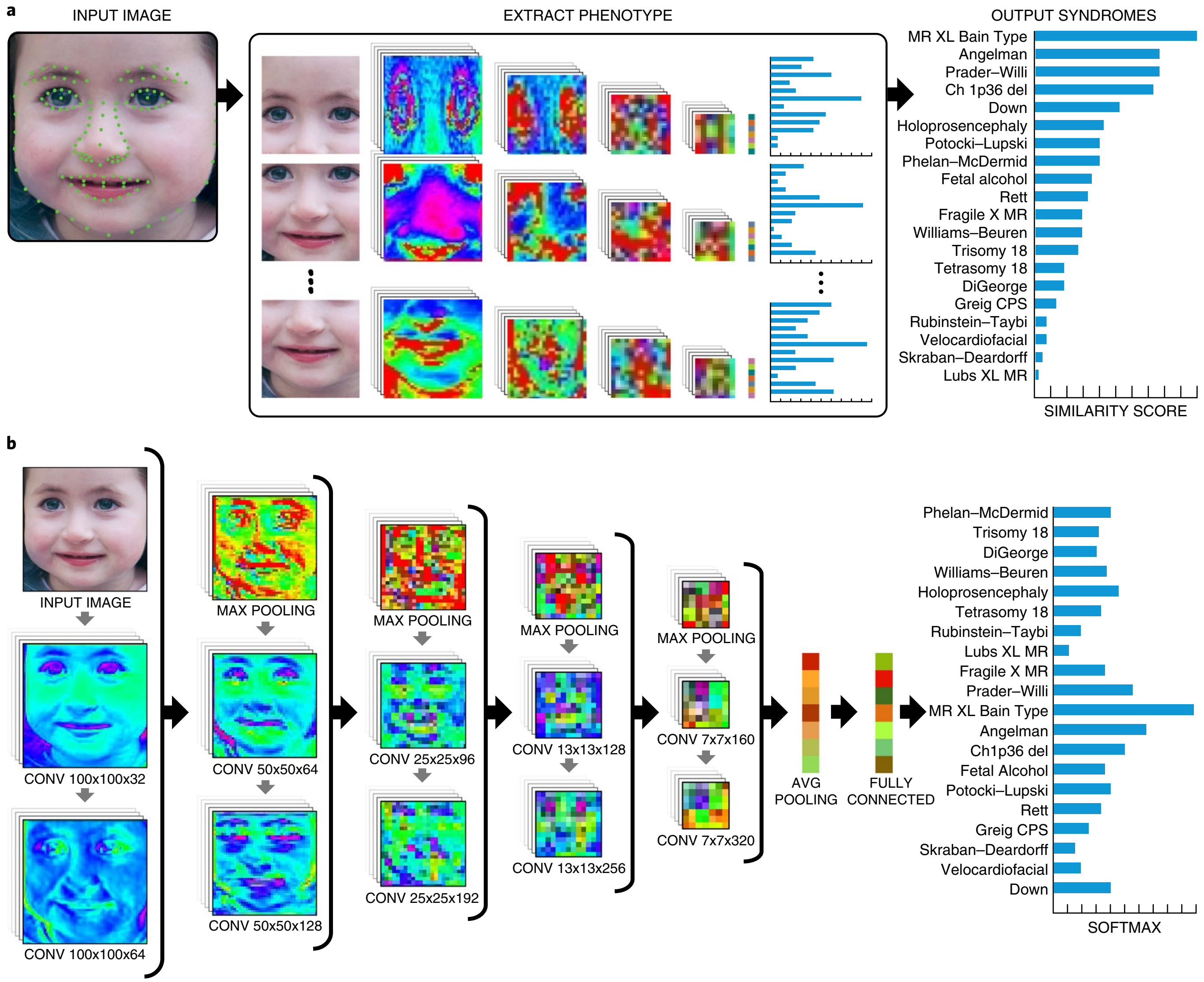

Fig. 1 | DeepGestalt: high-level flow and network architecture. a, A new input image is first preprocessed to achieve face detection, landmarks detection and alignment. After preprocessing, the input image is cropped into facial regions. Each region is fed into a DCNN to obtain a softmax vector indicating its correspondence to each syndrome in the model. The output vectors of all regional DCNNs are then aggregated and sorted to obtain the final ranked list of genetic syndromes. The histogram on the right-hand side represents DeepGestalt's output syndromes, sorted by the aggregated similarity score. **b**, The DCNN architecture of DeepGestalt. A snapshot of an image passing through the network. The network consists of ten convolutional layers, and all but the last are followed by batch normalization and a rectified linear unit (ReLU). After each pair of convolutional (CONV) layers, a pooling layer is applied (maximum pooling after the first four pairs and average pooling after the fifth pair). This is then followed by a fully connected layer with dropout (0.5) and a softmax layer. A sample feature map is shown after each pooling layer. It is interesting to compare the low-level features of the first layers with respect to the high-level features of the final layers; the latter identify more complex features in the input image, and distinctive facial traits tend to emerge while identity-related features disappear. The photograph is published with parental consent.

unaffected from affected individuals or distinguishing one specific syndrome from several others. We performed two binary experiments of the latter type.

The model was trained using 614 Cornelia de Lange syndrome (CdLS) images as positive cohort, and 1079 other images as negative cohort. The test sets contained 23 images of CdLS and nine of non-CdLS patients[4] (Supplementary Table 3). DeepGestalt achieved an accuracy of 96.88% (95% CI, 90.62–100%), sensitivity of 95.67% (95% CI, 87–100%) and specificity of 100% (95% CI, 100–100%) (for all binary experiments, accuracy is top-1 accuracy). We compared this result with previous studies on the same test set (Table 1). Basel-Vanagaite et al.[4] reported an accuracy of 87% and compared their method's performance with that of Rohatgi et al.[18], where the same images were assessed by 65 experts, achieving 75% accuracy. We measured statistical significance using the population proportions test and calculated P values of 0.01 and 0.22 for the results of DeepGestalt and Basel-Vanagaite et al.[4], respectively, versus the baseline of Rohatgi et al.[18].

For a binary experiment on distinguishing patients with Angelman syndrome from other syndromes, the model was trained on 766 Angelman syndrome images as positive cohort and 2,669 images as negative cohort. In a survey by Bird et al.[19], 20 dysmorphologists examined 25 patient images for Angelman syndrome. The test set included 10 patients with Angelman syndrome and 15 with other syndromes (Supplementary Table 4). Bird et al.[19] reported an accuracy of 71% (range, 56–92%), sensitivity of

Table 1 | Results comparison for the two binary experiments

Experiment	Method	Accuracy (%) (95% CI)	Sensitivity (%) (95% CI)	Specificity (%) (95% CI)	*P* value
CdLS	Rohatgi et al.[18]	75 (NA)	–	–	–
CdLS	Basel-Vanagaite et al.[4]	87 (NA)	–	–	0.22
CdLS	DeepGestalt	96.88 (90.1–100)	95.67 (87–100)	100 (100–100)	0.01
Angelman syndrome	Bird et al.[19]	71 (NA)	60 (NA)	78 (NA)	–
Angelman syndrome	DeepGestalt	92 (80–100)	80 (50–100)	100 (100–100)	0.05

The results of detecting Cornelia de Lange syndrome (CdLS) patients using a sample size of $n = 32$ independent images are reported on the top three rows. The results of detecting Angelman syndrome patients using a sample size of $n = 25$ independent images are reported on the bottom two rows. To produce the CI values, we used the percentile bootstrap method with 10,000 independent experiments. We measured statistical significance using a two-sided population proportions test and calculated a *P* value. For CdLS the *P* value is a result for DeepGestalt and Basel-Vanagaite et al.[4] versus the baseline of Rohatgi et al.[18]. For Angelman syndrome, the *P* value is a result for DeepGestalt versus the baseline accuracy of Bird et al.[19]. NA indicates not available where CI calculation was not possible.

60% (range, 30–100%) and specificity of 78% (range, 47–100%). On the same test set, DeepGestalt achieved an accuracy of 92% (95% CI, 80–100%), sensitivity of 80% (95% CI, 50–100%) and specificity of 100% (95% CI, 100–100%) (Table 1). The *P* value is 0.05, calculated with the population proportions test, versus the baseline of Bird et al.[19].

Specialized Gestalt model: classifying different genotypes of the same syndrome. DeepGestalt may be used for small-scale problems, with only a few images per cohort. Here, the goal is to distinguish between molecular subtypes of a heterogeneous syndrome resulting from different mutations affecting the same pathway. Allanson et al.[20] explored whether dysmorphologists can predict the correct Noonan syndrome–related genotype from the facial phenotype. They presented 81 images of patients with Noonan syndrome with mutations in *PTPN11*, *SOS1*, *RAF1* or *KRAS* to two dysmorphologists and concluded that facial phenotype alone was insufficient to predict the genotype[20].

We examined whether DeepGestalt performs better at a similar task using images of patients with Noonan syndrome due to a mutation in *PTPN11*, *SOS1*, *RAF1*, *RIT1* or *KRAS*. To train this model, we used 278 Noonan syndrome images curated from articles and clinical data. To test the performance, we composed a set of 25 images, 5 images per gene (class), excluded from the training set and curated from published articles[20–25] (Supplementary Table 5). Figure 2a shows composite photos created by averaging the training images, illustrating the general appearance of each cohort.

The Specialized Gestalt Model is a truncated version of DeepGestalt, predicting only the five desired classes with a top-1 accuracy of 64% (95% CI, 44–84%) (Fig. 2b), superior to the random chance of 20%. A permutation test yields a *P* value lower than 1×10^{-5}.

DeepGestalt performs facial Gestalt analysis at scale. A multiclass Gestalt model trained on a large database of 17,106 images of diagnosed cases spanning 216 distinct syndromes (Supplementary Table 2) was evaluated on two test sets: (1) a clinical test set of 502 patient images of cases submitted and solved over time by clinical experts; and (2) a publications test set of 329 patient images from the London Medical Databases[26], a resource of photos and information about syndromes, genes and clinical phenotypes that is accessible through Face2Gene Library.

DeepGestalt uses an aggregation of facial regions to improve performance and robustness. To examine how each region contributes to the final model, we evaluated the performance on both test sets for each region separately and in comparison to the aggregated model. The aggregated model performed better than each separate component (Table 2).

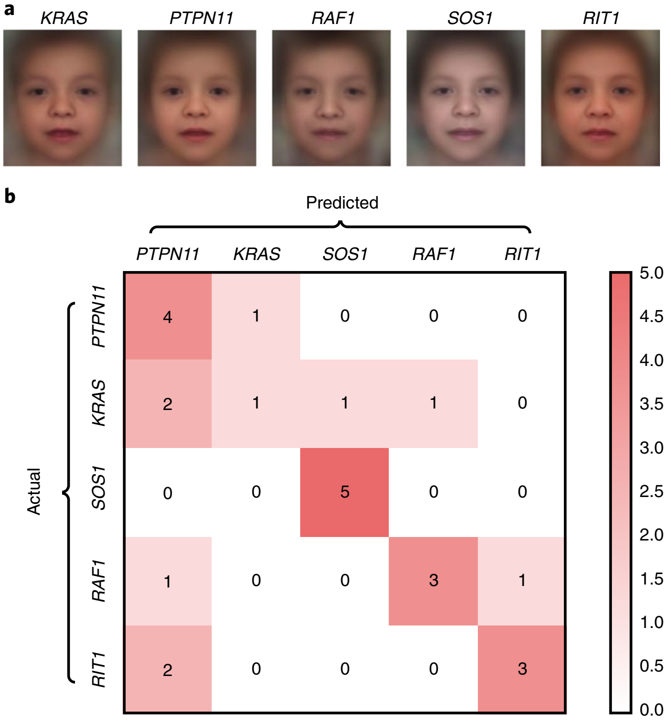

Fig. 2 | Composite photos and test set results of the Specialized Gestalt Model. a, Composite photos of patients with Noonan syndrome with different genotypes show subtle differences, such as less prominent eyebrows in individuals with a *SOS1* mutation, which might reflect the previously recognized sparse eyebrows as an expression of the more notable ectodermal findings associated with mutations in this gene. The numbers of images used to create the composite photo for *KRAS*, *PTPN11*, *RAF1*, *SOS1* and *RIT1* are 34, 123, 21, 54 and 46, respectively. **b**, Test set confusion matrix for the Specialized Gestalt Model. Rows indicate the diagnosed gene, while columns indicate the model's predicted gene. The value in each cell is the number of images with the same gene and prediction. The diagonal represents the true positive predictions.

DeepGestalt achieved a top-10 accuracy of 90.6% (95% CI, 88–93%) on the clinical test set and 89.4% (95% CI, 86–92.7%) on the publications test set. For patients with more than one frontal image, random selection of one image per patient led to similar results with very small variance. The top-5 and top-1 accuracy for the clinical test set was 85.4% (95% CI, 82.3–88.4%) and 61.3% (95% CI, 57.2–65.5%), respectively, and for the publications test set 83.2% (95% CI, 79–87.2%) and 68.7% (95% CI, 63.52–73.55%), respectively.

The permutation test for all experiments yielded a *P* value lower than 1×10^{-6}.

Table 2 | Performance comparison between facial regions and the aggregated DeepGestalt model (as an ensemble of regional predictors)

Facial area	Clinical test	Publications test
	Top-10 accuracy (%)	Top-10 accuracy (%)
Face, upper half	82	82.4
Middle face (ear to ear)	81	80.2
Face, lower half	76.8	77.2
Full face	88.2	87.5
Aggregated model	**90.6**	**89.4**

Results are reported for both test sets: clinical test ($n = 502$ images of 92 syndromes from 375 patients); publications test ($n = 329$ images of 93 syndromes from 320 patients).

Discussion

We present a facial analysis framework for genetic syndrome classification called DeepGestalt. This framework leverages deep-learning technology and learns facial representation from a large-scale face-recognition dataset, followed by knowledge transfer to the genetic syndrome domain through fine-tuning.

DeepGestalt is able to generalize for different problems, as demonstrated on binary models for CdLS and Angelman syndrome, for which its performance surpassed that of human experts. It can be optimized for specific phenotypic subsets, as shown on a Specialized Gestalt Model focused on identifying the correct facial phenotype of five genes related to Noonan syndrome, allowing geneticists to investigate phenotype–genotype correlations. DeepGestalt's performance on hundreds of genetic syndromes characterized by unbalanced class distributions, as evaluated on two external test sets wherein 90% of cases the correct syndrome appeared in the top 10, suggests that this technology can highlight possible diagnostic direction in clinical practice. The common clinical practice is to describe the patient's phenotype in discrete clinical terms[27] and to use semantic similarity search engines for syndrome suggestions[28]. This approach is subjective and depends greatly on the clinician's phenotyping experience. Adding an automated facial analysis framework to the clinical workflow could achieve better syndrome prioritization and diagnosis.

DeepGestalt, like many artificial intelligence systems, cannot explicitly explain its predictions and provides no information about which facial features drove the classification. To address this, a heatmap visualization shows the goodness-of-fit between areas of the individual image and each suggested syndrome, achieved by backpropagating the information through the DCNN to the input image (Supplementary Fig. 2). While it is possible to calculate ratios from the 130 detected landmarks, such as that between inner and outer canthal distance defining hypertelorism, this is not an intrinsic part of DeepGestalt.

Given the assumption underlying the clinical use of DeepGestalt that the patient has some syndrome, one scientific question not included here is the ability to determine whether a subject has a genetic syndrome. Such comparisons have been previously conducted[429,30]. The results in this report are limited to patients with certain syndromes and, therefore, are not transferable to a test set including unaffected individuals.

A limitation of this study is the lack of comparison to other methods or human experts in some experiments. Previous work in this field lacks large datasets to allow fair comparison. We had access to small benchmarks in the two binary experiments and the specialized Gestalt experiment, where 25–32 images were used. To enable comparison in future studies, the publications test set is available for research (Supplementary Table 6).

DeepGestalt, a form of next-generation phenotyping technology[31], assists with syndrome classification. Similar to genotypic data, phenotypic data are sensitive patient information, and discrimination based thereon is prevented by the Genetic Information Nondiscrimination Act. Unlike genomic data, facial images are easily accessible. Payers or employers could potentially analyze facial images and discriminate based on the probability of individuals having pre-existing conditions or developing medical complications. Effective monitoring strategies mitigating abuse may include the addition of a digital footprint through blockchain technologies to applications using DeepGestalt.

The increased ability to describe phenotype in a standardized manner enables identification of new genetic syndromes by matching undiagnosed patients sharing a similar phenotype. We believe that coupling of automated phenotype analysis with genome sequencing data will enable improved prioritization and interpretation of gene variant results, and may become a key factor in precision medicine.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41591-018-0279-0.

Received: 18 December 2017; Accepted: 29 October 2018;
Published online: 7 January 2019

References

1. Baird, P. A., Anderson, T., Newcombe, H. & Lowry, R. Genetic disorders in children and young adults: a population study. *Am. J. Hum. Genet.* **42**, 677–693 (1988).
2. Hart, T. & Hart, P. Genetic studies of craniofacial anomalies: clinical implications and applications. *Orthod. Craniofac. Res.* **12**, 212–220 (2009).
3. Ferry, Q. et al. Diagnostically relevant facial gestalt information from ordinary photos. *eLife* **3**, e02020 (2014).
4. Basel-Vanagaite, L. et al. Recognition of the Cornelia de Lange syndrome phenotype with facial dysmorphology novel analysis. *Clin. Genet.* **89**, 557–563 (2016).
5. Rai, M. C. E., Werghi, N., Al Muhairi, H. & Alsafar, H. Using facial images for the diagnosis of genetic syndromes: a survey. In *2015 International Conference on Communications, Signal Processing, and their Applications (ICCSPA)* (2015).
6. Shukla, P., Gupta, T., Saini, A., Singh, P. & Balasubramanian, R. A deep learning frame-work for recognizing developmental disorders. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)* (IEEE, 2017).
7. Hadj-Rabia, S. et al. Automatic recognition of the XLHED phenotype from facial images. *Am. J. Med. Genet. A.* **173**, 2408–2414 (2017).
8. Valentine, M. et al. Computer-aided recognition of facial attributes for Fetal Alcohol Spectrum disorders. *Pediatrics* **140**, e20162028 (2017).
9. Gripp, K. W., Baker, L., Telegrafi, A. & Monaghan, K. G. The role of objective facial analysis using FDNA in making diagnoses following whole exome analysis. Report of two patients with mutations in the BAF complex genes. *Am. J. Med. Genet. A.* **170**, 1754–1762 (2016).
10. Delgadillo, V., Maria del Mar, O., Gort, L., Coll, M. J. & Pineda, M. Natural history of Sanfilippo syndrome in Spain. *Orphanet J. Rare Dis.* **8**, 189 (2013).
11. Kole, A. et al. The Voice of 12,000 Patients: experiences and expectations of rare disease patients on diagnosis and care in Europe. *Eurordis* http://www.eurordis.org/IMG/pdf/voice_12000_patients/EURORDISCARE_FULLBOOKr.pdf (2009).
12. Taigman, Y., Yang, M., Ranzato, M. & Wolf, L. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2014* 1701–1708 (IEEE, 2014).
13. Huang, G. B., Ramesh, M., Berg, T. & Learned-Miller, E. Labeled faces in the Wild: a database for studying face recognition in unconstrained environments. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition* (2008).
14. Yi, D., Lei, Z., Liao, S. & Li, S. Z. Learning face representation from scratch. Preprint at https://arxiv.org/abs/1411.7923 (2014).
15. Schroff, F., Kalenichenko, D. & Philbin, J. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015 815–823* (IEEE,2015).

16. Li, H., Lin, Z., Shen, X., Brandt, J. & Hua, G. A convolutional neural network cascade for face detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015* 5325–5334 (IEEE, 2015).
17. Karlinsky, L. & Ullman, S. Using linking features in learning non-parametric part models. *Computer Vision–ECCV* **2012**, 326–339 (2012).
18. Rohatgi, S. et al. Facial diagnosis of mild and variant CdLS: Insights from a dysmorphologist survey. *Am. J. Med. Genet. A.* **152**, 1641–1653 (2010).
19. Bird, L. M., Tan, W. H. & Wolf, L. The role of computer-aided facial recognition technology in accelerating the identification of Angelman syndrome. In *35th Annual David W Smith Workshop* (2014).
20. Allanson, J. E. et al. The face of Noonan syndrome: does phenotype predict genotype. *Am. J. Med. Genet. A.* **152**, 1960–1966 (2010).
21. Gulec, E. Y., Ocak, Z., Candan, S., Ataman, E. & Yarar, C. Novel mutations in PTPN11 gene in two girls with Noonan syndrome phenotype. *Int. J. Cardiol.* **186**, 13–15 (2015).
22. Zenker, M. et al. SOS1 is the second most common Noonan gene but plays no major role in cardio-facio-cutaneous syndrome. *J. Med. Genet.* **44**, 651–656 (2007).
23. Rusu, C., Idriceanu, J., Bodescu, I., Anton, M. & Vulpoi, C. Genotype-phenotype correlations in Noonan Syndrome. *Acta Endocrinologica* **10**, 463–476 (2014).
24. Cavé, H. et al. Mutations in RIT1 cause Noonan syndrome with possible juvenile myelomonocytic leukemia but are not involved in acute lymphoblastic leukemia. *Eur. J. Hum. Genet.* **24**, 1124–1131 (2016).
25. Kouz, K. et al. Genotype and phenotype in patients with Noonan syndrome and a RIT1 mutation. *Genet. Med.* **18**, 1226–1234 (2016).
26. Winter, R. M. & Baraitser The London Dysmorphology Database. *J. Med. Genet.* **24**, 509–510 (1987).
27. Robinson, P. N. & Mundlos, S. The human phenotype ontology. *Clin. Genet.* **77**, 525–534 (2010).
28. Köhler, S. et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *A. J. Hum. Genet.* **85**, 457–464 (2009).
29. Zarate, Y. A. et al. Natural history and genotype-phenotype correlations in 72 individuals with SATB2-associated syndrome. *Am. J. Med. Genet. A.* **176**, 925–935 (2018).
30. Liehr, T. et al. Next generation phenotyping in Emanuel and Pallister Killian Syndrome using computer-aided facial dysmorphology analysis of 2D photos. *Clin. Genet.* **93**, 378–381 (2017).
31. Hennekam, R. & Biesecker, L. G. Next-generation sequencing demands next-generation phenotyping. *Hum. Mutat.* **33**, 884–886 (2012).

Acknowledgements

The authors thank the patients and their families, as well as Face2Gene users worldwide who contribute with their knowledge and dedication to the improvement of this and other tools for the ultimate benefit of better healthcare.

Author contributions

Y.G., Y.H. and O.B. initiated the project. Y.G., Y.H. and D.G. developed the DeepGestalt framework. N.F., L.B.-S., P.M.K., S.B.K., M.Z., L.M.B. and K.W.G. designed and conducted the clinical experiments. O.B. and G.N. finalized the dataset, computed statistics and contributed to the software engineering. Y.G., Y.H., O.B., P.M.K. and K.W.G. contributed to the writing of the manuscript.

Competing interests

Y.G., Y.H., O.B., G.N., N.F. and D.G. are employees of FDNA; L.B.-S., P.M.K. and K.W.G. are advisors of FDNA; L.B.-S., P.M.K., M.Z., L.M.B. and K.W.G. are members of the scientific advisory board of FDNA.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-018-0279-0.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to Y.G.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2019

Methods

Ethics statement. The authors affirm that human research participants provided informed consent for publication of the images in Fig. 1 and Supplementary Figs. 1 and 2.

Study approval. This paper describes studies governed by the following Institutional Review Board (IRB) approval: Nemours Children's Health System, DE, USA (IRB no. 2005-051); Charité–Universitätsmedizin Berlin, Germany (EA2/190/16); Rady Children's Hospital, San Diego, CA, USA (31542 and 091451); Beilinson Rabin Medical Center, Israel (0114-17); and UKB Universitätsklinikum Bonn, Germany (Lfd.Nr.386/17). The authors have obtained patient consent, where applicable, per the respective IRB.

The building blocks of the technology behind DeepGestalt. We detail our image-preprocessing pipeline, phenotype extraction and syndrome classification methods, datasets used, training details, evaluation protocol and statistical analysis. Typically, facial images were captured by clinicians during patient visits using consumer cameras, usually smartphone cameras. There are no specific hardware requirements. Following upload, image quality is assessed by whether a frontal face can be detected or not.

From an end-to-end perspective, our goal is to achieve a function $F(x)$, which maps an input image x into a list of genetic syndromes with a similarity score per syndrome. When sorted by this Gestalt score, the top listed syndromes represent those with the most similar phenotype (Fig. 1a).

Image preprocessing. Our model is designed for real-world uncontrolled 2-D images. The first step is to detect a patient's face in an input image. Since real clinical images have a large variance due to face size, pose, expression, background, occlusions and lighting, a robust face detector is needed in order to identify a valid frontal face. We adopt a deep-learning method, based on a DCNN cascade, proposed in ref. [16] for face detection in an uncontrolled environment. We adjust this method to fit our needs and operate optimally on images of children with genetic syndromes, in order to identify a frontal face from the image background.

We then detect 130 facial landmarks on the patient's face (Fig. 1a). This landmarks detection algorithm works in a chain of multiple steps, starting from a coarse step of identifying a small number of landmarks up to a more subtle detection of all landmarks of interest[17].

The resulting face and landmarks detected are first used to geometrically normalize the patient's face. The alignment of images reduces the pose variation among patients and shows improved performance on recognition tasks such as face verification[32]. An example of these steps is presented in Supplementary Fig 1.

The aligned image and its corresponding facial landmarks are then processed through a regions generator, which creates multiple predefined regions of interest from the patient's face. As illustrated in Fig. 1a, the different facial crops contain holistic face crops and several distinct regional crops which contain the main features of the human face, including the eyes, nose and mouth. The final step in the preprocessing stage is to scale each facial cropped region to a fixed size of 100 × 100 pixels and convert it to grayscale.

Phenotype extraction and syndromes classification. DeepGestalt uses DCNNs, which belong to a type of machine-learning techniques that are composed of interconnected data units, known as artificial neurons. Each of these neurons has its own specialized knowledge and shares information with other neurons. Neurons are organized in stacked layers from input to output, where each layer's output is the following layer's input. Each layer is typically also followed by a nonlinear step (a sigmoid function, for example). The layers closer to the input extract low-level information, such as edges and corners from images, whereas layers closer to the output usually aggregate information from previous layers into more complex features. This structure allows the network to extract information from the input for a specific objective function (classification or other). Each layer's parameters (weights) are initialized as random and updated incrementally while using training data samples, where the true class or value is known. This process repeats until convergence (typically using the backpropagation algorithm). Given a large and sufficiently variable training set, these networks learn a generalizable and powerful model to use for test images, where the label is unknown. In a DCNN, some layers perform a convolution kernel operation on their input layer, which was shown to be an effective way to extract information from images.

In order to mitigate the main challenge of our specific problem, a small training database with unbalanced classes, we train the DeepGestalt model in two steps. First, we learn a general face representation and then fine-tune it into the genetic syndromes classification task.

To learn the baseline facial representation, we train a DCNN on a large-scale face identity database. Our backbone architecture is based on that suggested by Yi et al.[14] and is illustrated in Fig. 1b. We train separately for each facial crop, and combine the trained models to form a robust facial representation.

Once the general face representation model is obtained, we fine-tune the DCNN for each region with a smaller-scale phenotype dataset for the task of syndrome classification. In practice, this step acts as a transfer learning step between a source domain (face recognition) and a target domain (genetic syndromes classification)[33,34]. Effectively, we use the powerful face recognition model for face representation (which performs comparably to the state-of-the-art results on the Labeled Faces in the Wild benchmark[13]), and train the model to classify different genetic syndromes rather than classifying identities.

We use the different facial regions, both as expert classifiers and as an ensemble of classifiers[35,36]. Each region's specific DCNN separately makes a prediction, and these are combined by averaging the results and producing a robust Gestalt model for a multiclass problem (Fig. 1a).

At the time of real clinical use, an image of a patient that has not been used during training is processed through the described pipeline. The output vector is a sorted vector of similarity scores, indicating the correlation of the patient's photo to each syndrome supported in the model.

In order to better understand the predictions made by DeepGestalt, we create a heatmap describing the spatial correlation between the input image and any chosen syndrome. This is done by backpropagating the information from the output of one of the specialized DCNNs and visualizing the most correlative areas in the face with respect to a specific syndrome, as done in ref. [37] (Supplementary Fig. 2).

Datasets. In order to train the model for face recognition, the publicly available CASIA WebFace dataset[14], which contains 494,414 images from 10,575 different subjects, is aligned, scaled and cropped, as described above. In order to fine-tune the networks to capture phenotypic information, we used clinical data, including facial images, uploaded to Face2Gene.

In this dataset, the diagnosis of cases is based on users' annotation, and further validation of these diagnoses is not possible due to strict privacy rules. For training we use a snapshot of the dataset, supporting 216 different syndromes and using 17,106 images of 10,953 subjects (mean and s.d. of 1.56 ± 1.70 images per subject, median value of 1) derived from the full set of images in the current database (see Supplementary Table 2 for demographic and clinical information about the dataset).

We use only cases that have been either clinically or molecularly diagnosed by relevant healthcare professionals, and automatically exclude images of low resolution and those where no frontal face was detected. This database is exposed to annotation errors. However, we believe that the DeepGestalt framework is able to generalize well even when errors in training exist. We assume that the presence of such mistakes is small and is not creating a large bias in the learned model. Other publications in deep learning also support a similar bias assumption[38]. For system evaluation, we built two test sets:

1. Clinical test set. Within a certain period of time, we sampled all 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 were in either the training set or the test set. We ended up with 502 images covering 92 different syndromes. The test set is skewed towards ultrarare 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 (Supplementary Table 2 for demographic and clinical information about the dataset).
2. Publications test set. We composed a new test set of 329 images covering 93 syndromes, published with the appropriate consent in the London Medical Databases (https://www.face2gene.com/lmd-history/)[26]. A complete list of links to images and relevant annotations is provided in Supplementary Table 6.

In order to create a high-quality test set, we applied a set of data-pruning rules on the full London Medical Databases dataset of thousands of images. We excluded images with no frontal face, images of bad quality or where the subject was under 1 or over 18 years old, and images where the subject was occluded (wearing glasses for example). In this test set, 80% of the syndromes presented in only 1 to 5 images and 20% in more than 6, with a median of 2 and mean of 3.54 images per syndrome (see Supplementary Table 2 for demographic and clinical information about the dataset).

To comply with high standards of security and privacy, a fully automated processing system is used. Images are automatically processed within the same environment as they were uploaded by users, maintaining the privacy and security of those images. In order to evaluate performance, only final results are reported.

Training. For each facial region, we train a face recognition DCNN using the large-scale face recognition dataset previously described. The training dataset is randomly split into training (90%) and validation (10%). The region's networks are then fine-tuned for the genetic syndromes classification task. The DCNN architecture is similar to that described[14] but with several modifications, including the addition of batch normalization[39] layers after each convolutional layer (Fig. 1b).

The training is done using Keras[40] with TensorFlow[41] as the backend. Baseline model training uses He Normal Initializer[42] weight initialization, which produced superior results compared to other known initializations. The optimization process uses Adam[43], with an initial learning rate of 1×10^{-3}, using a cross-entropy loss function. After 40 epochs (an epoch is one pass of training on the full dataset), we continue training the network for an additional 10 epochs using Stochastic Gradient Descent (SGD) with a learning rate of 1×10^{-4} and a momentum of 0.9.

In the fine-tuning phase, we replace the final layer output to match the number of syndromes in training. We found that the initialization for the fine-tuned layer is very important, and the best results are achieved when using a modified version of Xavier Normal Initializer[44]. We experimented with different scales of Xavier Normal Initializer and found that the best result was with a scale of 0.3.

The fine-tuning optimizer is SGD with a learning rate of 5×10^{-3} and a momentum of 0.9. No weight decay or kernel regularization is used, since we found that the addition of batch normalization[39] to the original architecture[14], which also includes dropout (we set the rate to 50%), performed better.

Augmentation was shown to be significantly important. Each region is randomly augmented by rotation with a range of 5 degrees, small vertical and horizontal shifts (shift range of 0.05), shear transformation (shear range of 5π / 180) and random zoom (zoom range of 0.05) horizontal flip. Without augmentation, training quickly overfitted, especially on the non-full-face regions.

In conclusion, each region DCNN is independently trained with 50 epochs for the face recognition task and an additional 500 epochs for the fine-tune step.

Evaluation. In the binary case, we measure the model's performance using top-1 accuracy (the percentage of cases where the model predicted the correct syndrome as the first result). We also measure the sensitivity (percentage of correctly predicted positive cohort cases from all positive cohort cases) and specificity (percentage of correctly predicted negative cohort cases from all negative cohort cases) of the model. The statistical significance of the comparison to human predictions is measured with the P value, calculated using the population proportions test.

In the multi-class case we measure top-K accuracy, where $K=1$, 5 or 10 (the percentage of images where the model predicted the correct syndrome within the top 1, 5 or 10 results out of 216 possible syndromes). In order to measure the statistical significance of our results for an unbalanced multiclass problem, we use a permutation test.

Statistical analysis. All values are reported with their 95% CI, calculated using the percentile bootstrap method[45].

For the binary experiments (CdLS and Angelman syndrome), when comparing our experiments to experts' performance or previous studies, we measured statistical significance using the P value with the two-sided population proportions test. This test measures the difference between two proportions on a single binary characteristic. The test's result is a Z-score and the associated P value, which is subjected to a null hypothesis significance test.

For the multiclass experiment, we derive the statistical significance using a permutation test, by measuring the distribution of the test set accuracy statistic under the null hypothesis. We randomly permute the test set labels 1×10^{6} times over the test data images, and calculate the top-K accuracy for each of the permutations. This allows us to sample the accuracy distribution and to calculate its P value.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Code availability. DeepGestalt is a proprietary framework. While its source code cannot be shared, the framework is accessible for use by healthcare professionals free of charge in Face2Gene (www.face2gene.com).

Data availability

The data that support the findings of this study are divided into two groups, published data and restricted data. Published data are available from the reported references and also in Supplementary Table 6. Restricted data are curated from Face2Gene users under a license and cannot be published, to protect patient privacy.

References

32. Huang, G., Mattar, M., Lee, H. & Learned-Miller, E. G. Learning to align from scratch. In *Advances in Neural Information Processing Systems 2012* 764–772 (2012).
33. Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems 2014* 3320–3328 (2014).
34. Taigman, Y., Yang, M., Ranzato, M. & Wolf, L. Web-scale training for face identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015* 2746–2754 (IEEE, 2015).
35. Zhou, E., Cao, Z. & Yin, Q. Naive-deep face recognition: touching the limit of LFW benchmark or not? Preprint at https://arxiv.org/abs/1501.04690 (2015).
36. Liu, J., Deng, Y., Bai, T., Wei, Z. & Huang, C. Targeting ultimate accuracy: face recognition via deep embedding. Preprint at https://arxiv.org/abs/1506.07310 (2015).
37. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: visualising image classification models and saliency maps. Preprint at https://arxiv.org/abs/1312.6034 (2013).
38. Parkhi, O. M., Vedaldi, A. & Zisserman, A. Deep face recognition. In *Proceedings of the British Machine Vision Conference* **1**, 6 (2015).
39. Ioffe, S. & Szegedy, C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *Proc. International Conference on Machine Learning 2015* 448–456 (2015).
40. Chollet, F. et al. Keras. http://keras.io (2015).
41. Abadi, M. et al. Tensorflow: large-scale machine learning on heterogeneous distributed systems. Preprint at https://arxiv.org/abs/1603.04467 (2016).
42. He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In *Proc. IEEE International Conference on Computer Vision* 1026–1034 (IEEE, 2015).
43. Kingma, D. and Ba, J. Adam: a method for stochastic optimization. Preprint at https://arxiv.org/abs/1412.6980 (2014).
44. Glorot, X. & Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international Conference on Artificial Intelligence and Statistics* 249–256 (2010).
45. Efron, B. Bootstrap methods: another look at the jackknife. In *Breakthroughs in Statistics* 569–593 (Springer, New York, 1992).

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.