GestaltMatch: breaking the limits of rare disease matching using facial phenotypic descriptors

Hsieh, Tzung-Chien[1]; Bar-Haim, Aviram[2]; Nadav, Guy[2]; Pantel, Jean Tori[1,3]; Fleischer, Nicole[2]; Krawitz, Peter[1,*]

1. Institute of Genomic Statistics and Bioinformatics, University of Bonn, Bonn, Germany
2. FDNA Inc., Boston Massachusetts, United States
3. Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Institute of Medical Genetics and Human Genetics, Berlin, Germany

* corresponding author, pkrawitz@uni-bonn.de

Abstract

Introduction:

Recent advances in next-generation phenotyping (NGP) for syndromology, such as DeepGestalt, have learned phenotype representations of multiple disorders by training on thousands of patient photos. However, many Mendelian syndromes are still not represented by existing NGP tools, as only a handful of patients were diagnosed. Moreover, the current architecture for syndrome classification, e.g., in DeepGestalt, is trained "end-to-end," that is photos of molecularly confirmed cases are presented to the network and a node in the output layer, that will correspond to this syndrome, is maximized in its activity during training. This approach will not be applicable to any syndrome that was not part of the training set, and it cannot explain similarities among patients. Therefore, we propose "GestaltMatch" as an extension of DeepGestalt that utilizes the similarities among patients to identify syndromic patients by their facial gestalt to extend the coverage of NGP tools.

Methods:

We compiled a dataset consisting of 21,400 patients with 1,451 different rare disorders. For each individual, a frontal photo and the molecularly confirmed diagnosis were available. We considered the deep convolutional neural network (DCNN) in DeepGestalt as a composition of a feature encoder and a classifier. The last fully-connected layer in the feature encoder was taken as Facial Phenotypic Descriptor (FPD). We trained the DCNN on the patients' frontal photos to optimize the FPD and

to define a Clinical Face Phenotype Space (CFPS). The similarities among each patient were quantified by cosine distance in CFPS.

Results:
Patients with similar syndromic phenotypes were located in close proximity in the CFPS. Ranking syndromes by distance in CFPS, we first showed that GestaltMatch provides a better generalization of syndromic features than a face recognition model that was only trained on healthy individuals. Moreover, we achieved 87% top-10 accuracy in identifying rare Mendelian diseases that were excluded from the training set. We further proved that the distinguishability of syndromic disorders does not correlate with its prevalence.

Conclusions:
GestaltMatch enables matching novel phenotypes and thus complements related molecular approaches.

Introduction

Worldwide, rare genetic disorders affect more than 8% of the population. The rarity and diversity of genetic disorders make it time-consuming and challenging for a clinician to achieve a correct diagnosis, which is the so-called “diagnostic odyssey.”[1] Craniofacial abnormalities present in 30-40% genetic disorders.[2] The patients with these syndromic disorders usually show recognizable faces such as Down syndrome and Fragile X syndrome. Hence, the facial manifestation provides a crucial visual hint for a clinician to identify related disorders, which speeds up the diagnostic workup with gene panel or exome sequencing because it helps reduce the search space of candidate genes. However, the ability to recognize these syndromic disorders highly relies on the human expert’s experience. It will be very challenging to make a diagnosis if the clinician has not seen the ultra-rare disorder or novel disease on the patient. Therefore, there is an urgent need for the next-generation phenotyping (NGP) tool to analyze the facial phenotypic information by the aid of a computer.

With the rapid development of machine learning and computer vision, a considerable number of NGP tools has emerged for analyzing facial dysmorphology with patients’ 2D portrait images.[3–10] Clinical Face Phenotype Space (CFPS), formed by the facial features extracted from facial images, was proposed to perform syndrome classification on the scale of training on more than 1000 patient photos with eight different syndromes.[3] Moreover, face recognition technologies were improved significantly in recent years and were at the core of the deep learning revolution in

computer vision. DeepFace[11] demonstrated, for the first time, human-level performance in identity verification on the Labeled Faces in the Wild dataset.[12] As a result, the face recognition system trained on CCTV images was utilized to match the patients with one of ten syndromic disorders with intellectual disability.[6] In addition, the facial recognition model from healthy individuals can also be integrated with the CFPS as a hybrid model, and it was proved to discriminate the facial gestalt on three novel disease-genes.[10] The current state-of-the-art syndrome classification framework DeepGestalt showed record-breaking results for syndrome classification using facial phenotypic cues, achieving 91% top-10 accuracy in identifying the correct syndrome in a test set of 502 images spanning more than 200 syndromes.[9] DeepGestalt also demonstrated strong separation ability for specific syndromes and subtypes, surpassing human experts' performance. These results demonstrated the power of a community-driven platform to gather patients and collect phenotypes.

Although NGP tools have shown the discriminative ability for syndrome classification, they still suffer from the limited data for rare genetic disorders and limited scalability of the model. In Figure 1, the two most well-known studies[3,9] for multi-syndromes classification focused mainly on the disease-genes with around 50 up to 500 pathogenic submissions in ClinVar such as *UBE3A*, *SMC1A*, and *HDAC8* which can be considered as common amongst the rare. The discriminative facial gestalt was identified in *PACS1*, *PPM1D*, and *PHIP*, which moved the border towards the genes with around ten submissions.[10] In addition, two unrelated patients with the same disease-causing mutation in *LEMD2* successfully matched by DeepGestalt syndrome similarity scores.[13] However, it is still challenging to push the limit to the ultra-rare disease-genes fall in the right tail of the distribution because the NGP approaches require a certain amount of images to learn the facial representation of syndromic disorder.

Moreover, the end-to-end offline trained architecture is suboptimal for scaling the model to support new syndromes, to keep the model updated, or to change its original goals. In order to support a new syndrome in DeepGestalt's model, the developer has to go through the six steps described in Supplementary Figure 1. In addition, the model for multi-syndrome classification cannot be used to quantify the similarities among patients that is crucial for clinicians to interpret the patient's phenotype. Therefore, the main limitations of the current approaches are: network architectures that do not scale and that do not allow comparison of single patients.

In the nosology of genetic diseases, there has been a discussion about splitters and lumpers for decades.[14] Deep learning approaches cannot only contribute to this

dialogue by quantifying distinguishability.[15] The architecture of a well-performing artificial neural network that serves as a classifier for syndromic disorders might reveal something about the complexity of the problem itself. In this work, we consider DeepGestalt as a composition of an image encoder which converts images to a vector of numbers, and a classification head which classifies the encoded vector to soft syndrome probabilities. While the last layer in DeepGestalt consisted of all the syndromes that the network learned to distinguish, we can refer to the layer preceding this last one, as the feature layer.

We hypothesize that the new framework, called GestaltMatch, is suitable to overcome these limitations:

1. Support new syndromes on the fly (without extra training);
2. Support multiple tasks (e.g., matching patients/syndromes etc.);
3. Support new explainability approaches (e.g., showing clusters separability in the dataset for different categories);
4. Be easily customized and allow low maintenance.

We show that the features vector created by DeepGestalt encoder can be used as a Facial Phenotypic Descriptor (FPD), which can be further used for syndrome classification and patient clustering. The concept of GestaltMatch is shown in Figure 2. Moreover, we show that features created using DeepGestalt encoder are better for matching cases with similar syndromic features, than features extracted from modern models used for face verification and no syndromic phenotype context. Interestingly, we show that our new FPD based framework, named GestaltMatch, has improved scalability for long-tailed syndromes distribution in Figure 1, without the need for retraining. Furthermore, it provides built-in support for patient matching. We show that given a facial image, one can use our system to search patients and syndromes with similar visual phenotypes. Moreover, the similarity between multiple FPDs spans a metric space between syndromes and can be used for finding new phenotypic series or discriminate between affected and non-affected subjects. Our new system is a natural extension to DeepGestalt and can help to develop new visual phenotype matching applications.

Method

Datasets

We collected the images of subjects with clinically or molecularly confirmed diagnoses from Face2Gene database. The images with poor quality or duplicated images were

removed from the dataset. After removing the problematic images, the dataset consisted of 33,434 images and 21,400 subjects of 1451 syndromes in total.

GestaltMatch aims to evaluate syndromes with different properties. We separated the syndromes by the number of subjects in each syndrome and whether they were learned by the DeepGestalt model. The overview of how the dataset was divided is shown in Supplementary Figure 2. The current DeepGestalt approach can only learn the syndromes which have more than six subjects. Hence, based on this threshold, we first separated the syndromes into frequent and rare syndromes. We denoted rare syndromes as target syndromes because these are the syndromes on which this study targets. However, not all frequent syndromes can be modeled by DeepGestalt. Some of them might have no dysmorphic features, so DeepGestalt cannot learn their facial representation. We denoted these syndromes as non-distinct, whereas the syndromes supported by DeepGestalt as distinct. The distinct syndromes were used for validating syndrome prediction and the separation ability of subtypes of a phenotypic series because these syndromes were known to have facial dysmorphic features, and the facial features were well recognized by DeepGestalt encoder. For target syndromes, we aim to prove that GestaltMatch is able to predict the syndrome even if only a few subjects are publicly available. It is noteworthy that currently, for more than half of all known disease-genes, less than ten cases with pathogenic mutations have been submitted to ClinVar (Figure 1). By the type of syndromes, we split the entire dataset into three datasets: distinct, non-distinct, and target syndromes, and they contained 301, 265, and 885 syndromes, respectively. Non-distinct and target syndromes are not yet applicable to DeepGestalt.

We further sampled each dataset into a gallery and test set. The gallery is a set of subjects we intend to match, given a subject from the test set. First of all, 1422 subjects in distinct and non-distinct datasets were kept out of the training set as a blind set for validating the DeepGestalt training. The subjects in the blind set were assigned to either distinct test set or non-distinct test set based on the type of syndromes, and the subjects not in the blind set were assigned to the gallery of the corresponding dataset. For the target dataset, we performed 10-fold cross-validation. 90% and 10% of subjects were assigned to the gallery and test set, respectively.

However, if we only performed the matching within the same dataset, it will not be the real-world scenario. The galleries of three datasets were later combined as a unified gallery, and we try to find the matched patients in the unified gallery. We called the gallery of each dataset as a partial gallery. It is used for the performance comparison between the DeepGestalt model and GestaltMatch on distinct syndromes because

DeepGestalt only predicts distinct syndromes, so we should only use the partial gallery of the distinct set as the gallery.

DeepGestalt encoder

The preprocessing pipeline of DeepGestalt includes points detection, facial alignment (frontalization), and facial regions cropping. During inference, every facial region crop is forward passed through a deep convolutional network (CNN), and finally, the results for all of the image regions are aggregated to the final prediction for the input face image. DeepGestalt network consists of ten convolutional layers with batch normalization (BN) and ReLU for embedding the input features. After every Conv-BN-ReLU layer, a max pooling layer is applied for reducing the spatial size while increasing the semantic representation. The classifier part of the network consists of a fully connected linear layer with dropout (0.5). In this work, we considered DeepGestalt architecture as an encoder-classification composition, pipelined during inference. We chose the last fully connected layer before the softmax classification as the facial feature representation, resulting in a vector of size 320. Our first hypothesis is that images with the same molecularly diagnosed syndromes or phenotypic series, which also share similar phenotypes, can be encoded to similar feature vectors, under some set of metrics.

Moreover, we claim that the specific design choice of DeepGestalt of using a predefined, offline trained, linear classifier, can be replaced by other classification "heads," for example, *k*-Nearest Neighbors using cosine distance or a Random Forest. Interestingly, we found that the data used during the FPD encoder training is essential to generalize unseen syndromes, subjects, and the space represented by the FPD encoder.

Descriptor projection - Clinical Face Phenotype Space

Each image was encoded by the DeepGestalt encoder and resulted in a 320-dimensional facial phenotypic descriptor. These facial phenotypic descriptors were further used to form a 320-dimensional space which is called Clinical Face Phenotype Space (CFPS), and each image is a point located in CFPS, as shown in Figure 2. The similarity between the two images is quantified by the cosine distance between them in CFPS. The smaller the distance is, the higher similarity between two images is. Therefore, the clusters of subjects in CFPS can represent the similarities among the different disorders or show the substructure under a phenotypic series.

Evaluation

To evaluate GestaltMatch, we take the images in the test set as input and position them in the CFPS that is defined by the images of the gallery. We calculated the cosine distance between each of the test set images to all the gallery images, and benchmark the performance by top-*k* accuracy. For each test image, if an image from another subject with the same disorder in the gallery is among the top-*k* nearest neighbors, we call it a top-*k* match. We further compare the accuracy of each syndrome in distinct, non-distinct, and target syndrome subsets to investigate whether GestaltMatch can extend DeepGestalt to support more syndromes.

Results

Comparing DeepGestalt and face recognition encoders

We first investigated the importance of using a syndromic features encoder rather than a normal facial features encoder. We compared FPDs created by DeepGestalt encoder to another encoder with the same architecture, trained on the CASIA-WebFace[16] recognition task. We then trained these two encoders and encoded all images by these two encoders separately.

- Enc-DeepGestalt encoder, trained on the gallery of 301 distinct syndromes.
- Enc-CASIA encoder, trained on the CASIA-WebFace dataset, with the same architecture of DeepGestalt.

We evaluated the performance by testing distinct and target test sets on the unified gallery. Table 1 shows the superiority of the features created by DeepGestalt in the matching performance, which emphasizes the importance of training the encoder on data with phenotypic cues. The features created by DeepGestalt improves the top-10 accuracy by 30% for the distinct category. Further, the top-10 accuracy was improved by 43% for the target syndromes, which contains a different, mutually exclusive list of syndromes. These results suggest that the features encoded by DeepGestalt are a better fit for the task of syndromes classification than the features encoded by the modern face recognition model. Moreover, DeepGestalt's FPD provides a better generalization than the FPD encoded by the modern face recognition model for unseen target syndromes.

Comparing distinct and non-distinct FPDs

In order to demonstrate the separability of syndromes with facial dysmorphism, we applied *t*-SNE[17] to project 4353 images of ten distinct syndromes with the largest number of subjects and 872 images of ten non-distinct syndromes to two-dimensional space, and we further calculated Silhouette index[18] for both of two datasets. Autism

syndrome has 1171 images which is the largest non-distinct syndrome. We did not take Autism into this analysis because it leads to an extreme imbalance of the number of subjects to the other non-distinct syndromes. As shown in Supplementary Figure 3, the FPDs of distinct syndrome show ten clear clusters of subjects. However, when applying *t*-SNE projection on subjects of non-distinct syndromes, no clear clusters are created. Besides, the Silhouette index of distinct syndromes is 0.07, which is higher than the index of non-distinct syndromes, which is -0.01. The negative Silhouette index of non-distinct syndromes indicates the poor separation of different syndromes. The results show the evidence for the phenotypic information encoded in the FPDs created by DeepGestalt.

GestaltMatch on unseen dysmorphic syndromes

For the purpose of proving GestaltMatch can match the patients with a novel syndrome unseen to the encoder and to better understand the important characteristics of the training dataset, we trained four different encoders for comparison. We sampled 21228 images of 13872 subjects with 279 known dysmorphic syndromes. The four encoder variants Enc-1 to Enc-4 are:

- Enc-1, trained on 90% of the 279 syndromes' subjects;
- Enc-2, trained on 90% of the 239 smallest syndromes' subjects;
- Enc-3, trained on 90% of the 239 largest syndromes' subjects;
- Enc-4, trained on 90% of 239 random syndromes' subjects.

For each model, we used the remaining 10% of the subjects, sampled across the syndromes in the training set, as a validation set. Moreover, we used the remaining 40 syndromes, of encoders 2-4 (the eliminated syndromes for each encoder) as an external test set of unseen distinct syndromes, denoted by Test-Large, Test-Small, and Test-Random, respectively. For example, the 40 syndromes in Test-Large are the largest 40 syndromes in 279 distinct syndromes, complementing the 239 syndromes trained in Enc-2.

To evaluate FPDs generalization ability of each encoder on unseen syndromes, we compared each of three encoders (Enc-2, Enc-3, and Enc-4), trained on a subset of 239 syndromes, to Enc-1. We used GestaltMatch to estimate the similarity between the test images to the gallery images, with cosine distance. The results are shown in

Table **2**. Enc-1 outperformed Enc-2 when testing on Test-Large, the top-10 accuracy dropped from 85.28% to 78.49%. The poor performance of Enc-2 could be due to losing too much training data because Test-Large contained the largest 40 distinct syndromes consisting of 12429 images, which is more than half of the total images.

However, Enc-3 and Enc-4 showed comparable results to Enc-1 on Test-Small and Test-Random, respectively. The top-10 accuracy of Enc-4 was even slightly higher than Enc-1 when testing on Test-Random. It means that the encoder without the 40 smallest or random distinct syndromes, leads to comparable performance on these 40 syndromes with an encoder trained with these 40 syndromes. Therefore, the results proved that GestaltMatch could generalize the facial dysmorphic features well on unseen syndromes, which means we are able to support new syndrome without retraining the model.

Target syndromes matching accuracy

We defined a syndrome as a target syndrome if it has less than seven subjects in our dataset. To understand the potential of matching target syndromes, we trained an encoder on 2215 images of 526 target syndromes, which have more than three subjects, and less than seven subjects, denoted by Enc-Target. We then compared Enc-Target to Enc-DeepGestalt trained on the 301 distinct syndromes from the previous section. Results in

Table **3** show that Enc-Target with a softmax classifier provides the best results, which means it learned important phenotypic features. However, in GestaltMatch only, Enc-DeepGestalt, which trained on distinct syndromes and did not see any of the target syndromes during its training, showed very similar results compared to Enc-Target. Although the results showed that cosine distance is inferior to a trained softmax classifier, encoders trained on distinct syndromes provide a similar accuracy on the unseen subject of target syndromes, compared to encoders trained on these target syndromes. Therefore, GestaltMatch is a more suitable choice for target syndromes because it achieved comparable performance to the encoder train on target syndrome, that means we could save resources for retraining the encoder. Moreover, training the model on both distinct and target syndromes, which have very few high-quality photos, might lead to poor performance due to the extremely imbalanced training dataset.

Correlation between prevalence and accuracy

Training an end-to-end network for classifying faces to syndromes such as DeepGestalt requires many subjects for each of the supported syndromes. Since this minimum subject requirement is no longer a must for GestaltMatch, we were interested in whether the matching accuracy of a syndrome correlates with its prevalence. We used Enc-2 from the previous section, trained on the 239 syndromes out of the full list of 279 syndromes. To remove the confounding effect from prevalence, we randomly

down-sampled each of the 40 avoided syndromes by selecting five subjects to the gallery and one subject to the test set. This experiment is repeated 1000 times. Figure 3 shows the average top-10 accuracy and the prevalence by Orphanet of each syndrome. We can see that the top-10 accuracy does not correlate with the prevalence. Several syndromes with low prevalence still perform very well. We can further consider the accuracy as the distinguishability of the syndrome. Therefore, as distinguishability does not depend on the prevalence, GestaltMatch can extend DeepGestalt to cover the ultra-rare disorders with a high distinguishability.

Hierarchical clustering

Phenotypic series is defined as a heterogeneous set of genetic disorders sharing similar phenotypes. We were interested in testing the visual clusters created with a two-dimensional projection of FPDs. We sampled subjects from subtypes of four large phenotypic series in our database: Noonan syndrome, Cornelia De Lange Syndrome (CDLS), Kabuki syndrome, and Mucopolysaccharidosis (MPS). As demonstrated in Supplementary Figure 4, using *t*-SNE projection on the FPDs of 743 subjects, sampled from the four phenotypic series, resulted in highly separable four clusters composed of the different subtypes of each phenotypic series. This result is a piece of evidence for the phenotypic features encoded in the FPDs.

Dysmorphism estimation

We were interested in testing GestaltMatch separation ability between FPDs of affected subjects with a dysmorphic genetic disorder and non-affected subjects (without a known genetic disorder). We used *t*-SNE projections in two different formats:

1. We sampled 1000 faces of healthy individuals and added to the ten largest app-valid syndromes (4346 subjects) projection;
2. We sampled 1000 faces of healthy individuals and 1000 faces of non-healthy individuals, evenly across the ten largest app-valid syndromes.

In the first experiment, we projected into eleven classes, while in the second experiment, we used a binary classification into two clusters. Results in Supplementary Figure 5 and Supplementary Figure 6 show that in both cases, non-affected subjects create reasonably separatable clusters, emphasizing the syndromic context encoded in the FPDs by GestaltMatch.

Discussion

Syndrome matching

As described in the evaluation section, the GestaltMatch framework can be used to match syndromes with an input image. The difference from DeepGestalt, lies in the ability to match unseen syndromes (no patient with these syndromes included in the encoder's training set). GestaltMatch framework also allows us to abstract away the encoding of a dataset from the classification task, and thus support multiple targets within the same evaluation. For example, one can evaluate both phenotypic series and subtypes levels within a single inference, or get the most similar patients for each of the matched syndromes with a minor computational cost which is only a few seconds.

GestaltMatch framework computes the similarity between each of the test set images to the entire dataset of images. The similarity can be computed using different metrics, for example, cosine or euclidean distance. Then the results are aggregated according to the chosen configuration. For example, image similarity can be aggregated at the patient level or in the syndrome level. Furthermore, filtering according to different dataset parameters (such as ethnicity, number of affected genes, and age), can be done to customize the evaluation further.

Patient matching

Matching patients with high similarities of facial dysmorphic features is one of the most important applications of GestaltMatch. Finding the second patient is always a challenging problem for most of the physicians when analyzing the novel or extremely rare Mendelian disorders. There are several online platforms such as Gene Matcher,[19] MyGene2, and Exchange Maker,[20] which allow physicians to look for similar patients by uploading phenotypic data, such as HPO terms or genomic information. These platforms have already matched thousands of patients in the past few years. However, the automated facial matching technology was not included in any of these platforms yet, although the facial phenotypes are crucial information for physicians to determine whether two patients have similar disorders or not. Therefore, there is an urgent need to support the patient matching approach by analyzing the facial images to facilitate the matching procedure.

As a proof of concept, we have matched two unrelated patients from different countries, with the same novel disease successfully by gestalt matching approach.[13] They both shared similar progeria-like features, and later the same *de novo* disease-causing mutation in the *LEMD2* gene was identified by the diagnostic workup. Although further analysis with more unrelated patients is needed to be done, the GestaltMatch approach could be a promising patient match application. Moreover, this approach can be integrated with the other matching platforms to enhance the matching ability to

reduce the amount of time further when looking for the second patient with the same rare disorder.

Ethnic bias

Ethnic bias could influence the performance of GestaltMatch dramatically, especially in target syndromes, because some patients with the same target syndromes were from the same paper. Moreover, most of those patients with extremely rare diseases in the same publication are usually from the same family. It is hard to tell whether they were matched by the dysmorphic features or the ethnic similarity. Therefore, testing the matching of subjects of distinct syndromes with different ethnic backgrounds by a statistical setting to assess the influence of ethnic bias is needed in the future experiment.

The future of phenotype representation

Using semantic descriptors for similarity estimation is common in many areas of artificial intelligence, such as face recognition, text understanding, visual tracking, speech recognition, and more. In recent years, converting structured data into semantic vectors is becoming common for non-visual phenotype matching as well, for example, in HPO2VEC[21] and NODE2VEC.[22]

Moreover, to improve the matching accuracy, input signals can be sourced from different modalities. For example, DeepGestalt uses different regions of a patient face and aggregates the classification result of each region. Moreover, in PEDIA,[23] semantic and visual phenotypic cues are aggregated to improve the prioritization of variant analysis.

Since semantic descriptors share the same format, one can aggregate these descriptors from different sources, to allow multimodal signals contribution to the final accuracy. Due to the generic structure of the GestaltMatch framework and the abstractions used for encoding datasets, future work can extend the GestaltMatch framework to support different input types such as text, speech, video, or other sources of medical imaging, to improve classification accuracy.

Designing a unified classification approach

One of the main challenges of productionizing the GestaltMatch technology lies in the ability to aggregate different categories. As shown in unseen syndromes analysis, an internal bias in the encoder's dataset can deteriorate the matching performance for both seen and unseen syndromes. Moreover, training a softmax classifier (as in DeepGestalt) provides better accuracy than a naive cosine distance over FPDs. The

question raised from these insights is - how to use GestaltMatch for supporting all types of syndromes? Accurately, future work will test whether it is better to combine GestaltMatch classification (for unseen or target syndromes) and DeepGestalt (for distinct syndromes) in a hybrid manner or use a single model to directly classify an image to all syndromes (using GestaltMatch or DeepGestalt).

Conclusion

GestaltMatch can match syndromes with facial dysmorphism in the CFPS and can be treated as an extension of DeepGestalt to cover the syndromes which are not supported in DeepGestalt model. Moreover, the sub-structure under a phenotypic series or novel diseases can be explored by the clustering of subjects in CFPS. Eventually, matching patients is one of the most important applications of GestaltMatch. It could be integrated into other online matching platforms such as MatchMaker Exchange or MyGene2 further to accelerate the matching process of unknown diagnosed patients and explore novel phenotype-genotype correlation.

Reference

1. Baird, P. A., Anderson, T. W., Newcombe, H. B. &Lowry, R. B. 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. 1–22 (2014). doi:10.7554/eLife.02020

4. Cerrolaza, J. J. *et al.* Identification of dysmorphic syndromes using landmark-specific local texture descriptors. *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)* 1080–1083 (2016). doi:10.1109/ISBI.2016.7493453

5. Wang, K. &Luo, J. Detecting Visually Observable Disease Symptoms from Faces. *EURASIP J. Bioinform. Syst. Biol.* **2016**, 13 (2016).

6. Dudding-Byth, T. *et al.* Computer face-matching technology using two-dimensional photographs accurately matches the facial gestalt of unrelated individuals with the same syndromic form of intellectual disability. *BMC Biotechnol.* **17**, 1–9 (2017).

7. Shukla, P., Gupta, T., Saini, A., Singh, P. &Balasubramanian, R. A Deep Learning Frame-Work for Recognizing Developmental Disorders. *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)* 705–714

(2017). doi:10.1109/WACV.2017.84

8. 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 (2018).

9. Gurovich, Y. *et al.* Identifying facial phenotypes of genetic disorders using deep learning. *Nature Medicine* **25**, 60–64 (2019).

10. van derDonk, R. *et al.* Next-generation phenotyping using computer vision algorithms in rare genomic neurodevelopmental disorders. *Genet. Med.* **21**, 1719–1725 (2019).

11. Taigman, Y., Yang, M., Ranzato, M. &Wolf, L. DeepFace: Closing the gap to human-level performance in face verification. in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 1701–1708 (IEEE Computer Society, 2014). doi:10.1109/CVPR.2014.220

12. Learned-Miller, E., Huang, G. B., RoyChowdhury, A., Li, H. &Hua, G. Labeled faces in the wild: A survey. *Adv. Face Detect. Facial Image Anal.* 189–248 (2016). doi:10.1007/978-3-319-25958-1_8

13. Marbach, F. *et al.* The Discovery of a LEMD2-Associated Nuclear Envelopathy with Early Progeroid Appearance Suggests Advanced Applications for AI-Driven Facial Phenotyping. *Am. J. Hum. Genet.* **104**, 749–757 (2019).

14. McKusick, V. A. On lumpers and splitters, or the nosology of genetic disease. *Perspect. Biol. Med.* **12**, 298–312 (1969).

15. Knaus, A. *et al.* Characterization of glycosylphosphatidylinositol biosynthesis defects by clinical features, flow cytometry, and automated image analysis. *Genome Med.* **10**, 3 (2018).

16. Yi, D., Lei, Z., Liao, S. &Li, S. Z. Learning Face Representation from Scratch. (2014).

17. Van DerMaaten, L. &Hinton, G. *Visualizing Data using t-SNE*. **9**, (2008).

18. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).

19. Sobreira, N., Schiettecatte, F., Valle, D. &Hamosh, A. GeneMatcher: A Matching Tool for Connecting Investigators with an Interest in the Same Gene. *Hum. Mutat.* **36**, 928–930 (2015).

20. Philippakis, A. A. *et al.* The Matchmaker Exchange: A Platform for Rare Disease Gene Discovery. *Hum. Mutat.* **36**, 915–921 (2015).

21. Shen, F. *et al.* HPO2Vec+: Leveraging heterogeneous knowledge resources to enrich node embeddings for the Human Phenotype Ontology. *J. Biomed. Inform.* **96**, 103246 (2019).

22. Grover, A. &Leskovec, J. Node2vec: Scalable feature learning for networks. in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **13**-**17**-**Augu**, 855–864 (Association for Computing Machinery, 2016).

23. Hsieh, T. C. *et al.* PEDIA: prioritization of exome data by image analysis. *Genet. Med.* **21**, 2807–2814 (2019).

Figures and tables

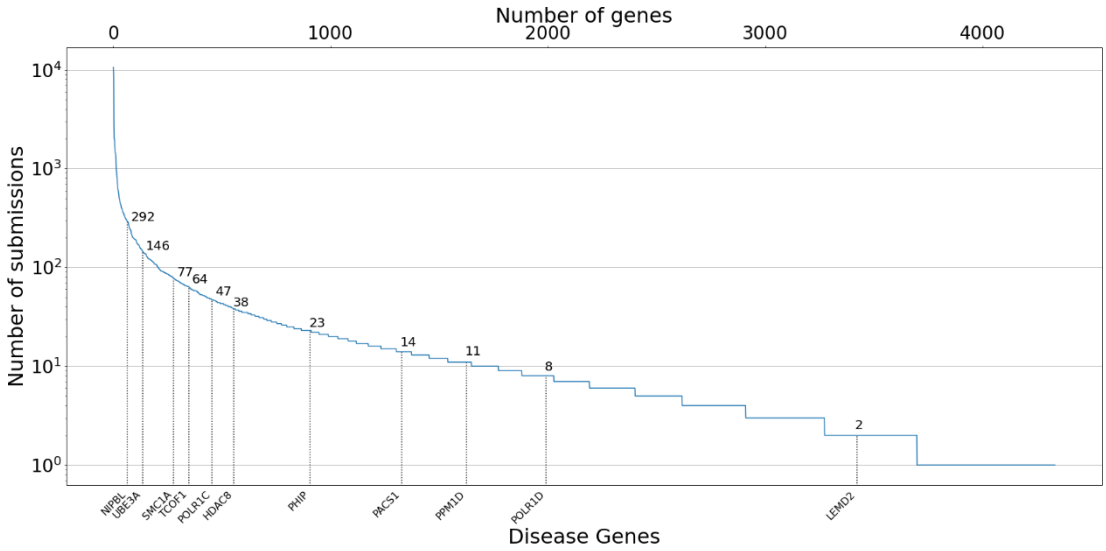
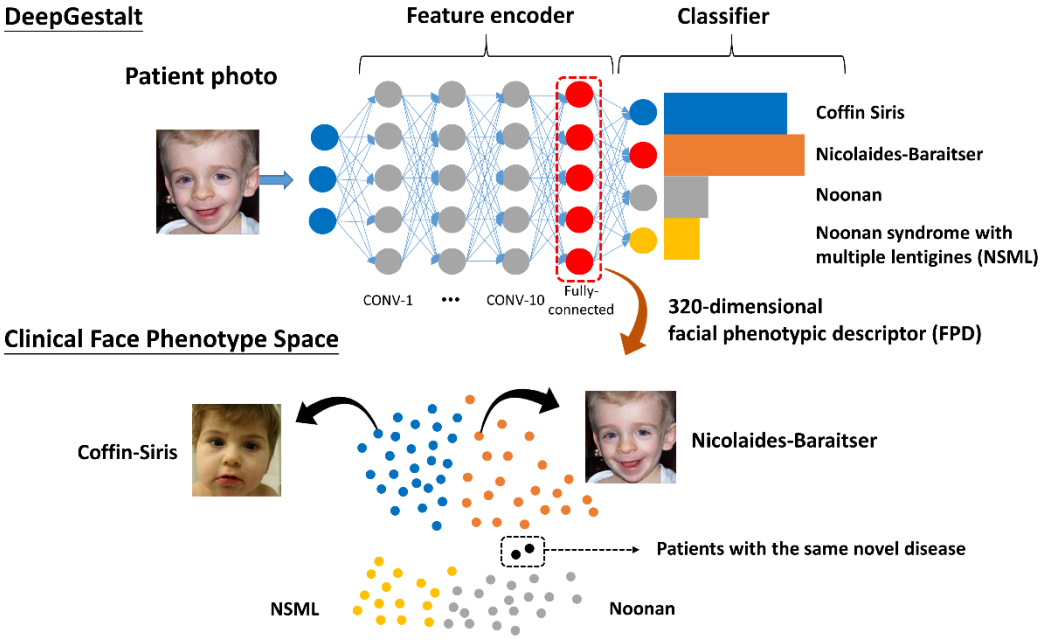
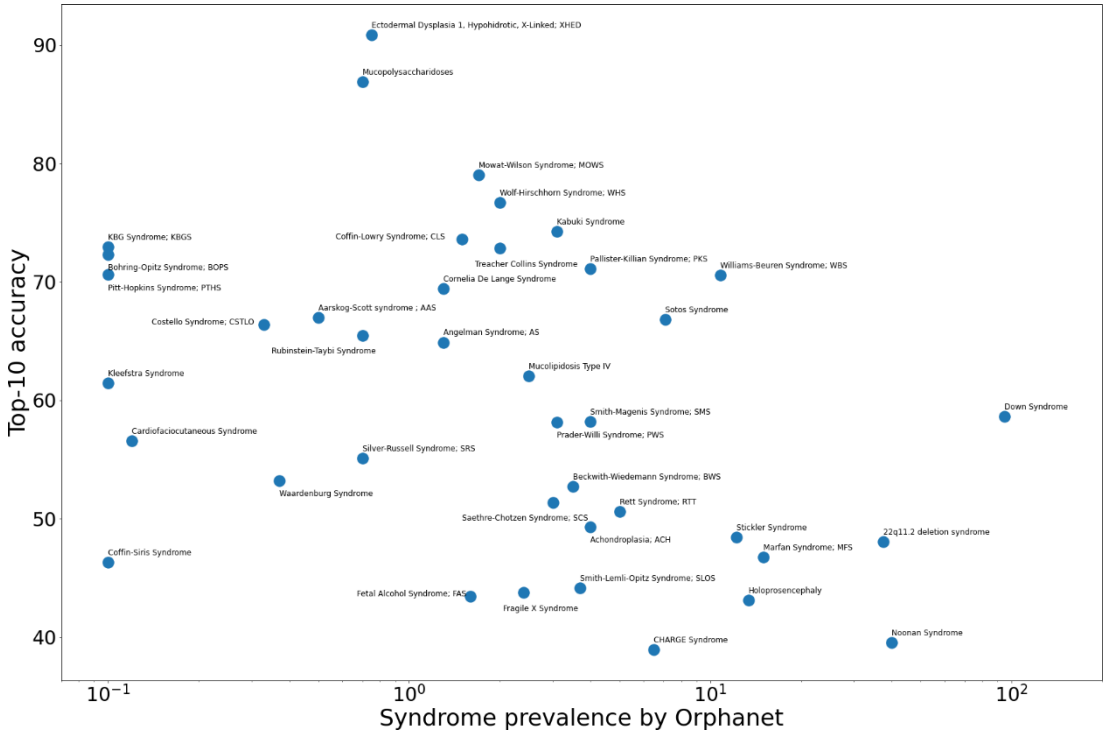


Figure 1: The distribution of the number of pathogenic submissions of each gene in ClinVar (April, 2020). The lower x-axis shows the disease genes, and the upper x-axis is the number of genes cumulative from zero on the left. Y-axis is the number of pathogenic submissions in ClinVar for the respective gene. The most two well-known multi-syndromes classification studies[3,9] mainly focused on the syndromes with relative common in the rare disorders such as Angelman syndrome (*UBE3A*), Cornelia de Lange syndrome (*NIPBL*, *SMC1A*, *HDAC8*), and Treacher Collins syndrome (*TCOF1*, *POLR1D*). The three novel disease-genes (*PACS1*, *PPM1D*, and *PHIP*,) which is proved to show discriminative facial gestalt,[10] were relatively rare compared to the previous studies. Later, the new disease related to *LEMD2* was found by two matching patients with similar facial phenotype.[13] *LEMD2* even only had two submissions so far. It shows that NGP approaches keep pushing the limit to more ultra-rare diseases on the right tail. However, 59% (2562 out of 4335) of disease genes with less than ten pathogenic submissions in ClinVar. The limited patients of rare disorders are a challenge to the current NGP approach since it requires a certain number of images to learn the facial representation of a disorder.



509

510 **Figure 2: Concept of GestaltMatch.** The DeepGestalt was trained on 301 distinct syndromes
511 to learn the facial dysmorphic features. The last fully-connected layer in the feature encoder is
512 taken as Facial Phenotypic Descriptor (FPD) and can be used to form a Clinical Face
513 Phenotype Space (CFPS). In this space, the distance between each patient can be considered
514 as the similarities of facial phenotypic features, which can be further used for syndrome
515 classification or clustering patients with unknown diagnosis.



516

517 **Figure 3: Correlation between syndrome prevalence and average top-10 accuracy.** X-axis

is the birth prevalence by Orphanet, and the unit is 1 in 100,000. For each syndrome, we randomly selected five subjects to the gallery and one subject to the test set to remove the confounding effect from prevalence. We further performed the classification on these 40 syndromes for 1000 times. Y-axis is the average top-10 accuracy of the experiments with 1000 times. From this figure, we can see that the top-10 accuracy does not correlate with disease prevalence.

Table 1: Performance comparison of DeepGestalt and CASIA encoder on distinct, non-distinct and target test set. Enc-DeepGestalt and Enc-CASIA have the same architecture. Enc-DeepGestalt was initiated with the CASIA-WebFace and further fine-tuned on patients' photos. Enc-DeepGestalt outperformed Enc-CASIA on distinct and target syndromes. It shows the importance of fine-tuning on patients' photos for learning facial dysmorphic features.

Test set	Model	Syndromes		Top 1	Top 5	Top 10	Top 30
		Gallery	Test				
Distinct	Enc-DeepGestalt	1451	168	**33.46%**	**56.11%**	**66.73%**	**80.54%**
Distinct	Enc-CASIA	1451	168	20.10%	41.49%	51.33%	70.23%
Non-distinct	Enc-DeepGestalt	1451	75	6.44%	**11.85%**	**15.79%**	**27.99%**
Non-distinct	Enc-CASIA	1451	75	**7.30%**	11.16%	14.25%	20.26%
Target	Enc-DeepGestalt	1451	885	**6.84%**	**12.97%**	**16.03%**	**21.69%**
Target	Enc-CASIA	1451	885	4.67%	8.49%	11.21%	15.41%

Table 2: Results of unseen syndromes classification with four encoders. Each pair of results below shows the comparison between training without the syndromes in the test set and with them. For example, Enc-1 was trained on 279 distinct syndromes, and Enc-2 was trained on the 239 distinct syndromes. The 40 syndromes in Test-Large are the unseen syndromes to Enc-2. When testing on Test-Random, Enc-4 shows the comparable results to Enc-1.

Test set	Model	Images		Syndromes	Top 1	Top 5	Top 10	Top 30
		Gallery	Test					
Test-Large	Enc-2	12429	1311	40	32.57%	65.45%	78.49%	97.79%
Test-Large	Enc-1	12429	1311	40	**44.85%**	**74.07%**	**85.28%**	**98.32%**
Test-Small	Enc-3	532	87	40	37.93%	**73.56%**	82.76%	96.55%
Test-Small	Enc-1	532	87	40	**44.83%**	67.82%	**85.06%**	**97.70%**
Test-Random	Enc-4	4025	430	40	47.44%	77.44%	**87.44%**	**99.07%**
Test-Random	Enc-1	4025	430	40	**53.02%**	**77.67%**	86.74%	**99.07%**

Table 3: Comparison of different models for matching target syndromes. Enc-DeepGestalt is the encoder trained on 301 distinct syndromes, and Enc-Target is the encoder trained on 526 target syndromes. The last row used DeepGestalt method, which is the softmax in DeepGestalt model to predict the syndrome, so it did not use the gallery.

Model	Method	Images		Syndromes	Top 1	Top 5	Top 10	Top 30
		Gallery	Test					
Enc-DeepGestalt	GestaltMatch	2215	749	526	14.81%	23.98%	29.57%	41.84%
Enc-Target	GestaltMatch	2215	749	526	14.55%	24.70%	30.04%	42.59%
Enc-Target	DeepGestalt	-	749	526	**17.35%**	**26.56%**	**32.84%**	**44.30%**