Region-based Saliency Explanations on the Recognition of Facial Genetic Syndromes

Ömer Sümer [1]   OEMER.SUEMER@INFORMATIK.UNI-AUGSBURG.DE
Rebekah L. Waikel [2]   REBEKAH.WAIKEL@NIH.GOV
Suzanna E. Ledgister Hanchard [2]   SUZANNA.LEDGISTERHANCHARD@NIH.GOV
Dat Duong [2]   DAT.DUONG2@NIH.GOV
Peter Krawitz [3]   PKRAWITZ2@UNI-BONN.DE
Cristina Conati [4]   CONATI@CS.UBC.CA
Benjamin D. Solomon [2]   SOLOMONB@MAIL.NIH.GOV
Elisabeth André [1]   ANDRE@INFORMATIK.UNI-AUGSBURG.DE

[1] *Human-Centered Artificial Intelligence, University of Augsburg, Augsburg, Germany*

[2] *Medical Genetics Branch, National Human Genome Research Institute, Bethesda, MD, USA*

[3] *Institute for Genomic Statistics and Bioinformatics, University of Bonn, Bonn, Germany*

[4] *Department of Computer Science, University of British Columbia, BC, Canada*

Abstract

Deep neural networks in computer vision have shown remarkable progress in recognizing facial genetic syndromes. Many genetic syndromes are difficult to detect, even for experienced clinicians, and computer-aided phenotyping can accelerate clinical diagnosis. High-stakes clinical tasks using deep learning, as in clinical genetics, require human understandable explanations for model decisions. Saliency methods are used to explain DNN predictions in various image analysis domains but have yet to be studied in facial genetics. The syndromic features of most genetic conditions are often localized to areas like the eyes, nose, and mouth. In this paper, to summarize the contribution of key facial regions to a specific disease prediction, we propose a face region relevance score that can be applied to any saliency method. We also investigate how prior knowledge, namely human phenotype ontology and DNN model explanations, align. Quantitative experiments are performed on a new database containing over 3,500 images of 11 rare facial syndromes, a healthy control group, and an additional test set of 171 facial images, whose respective facial phenotypes are labeled by clinicians. Current saliency methods are good at capturing dysmorphism in particular regions (parts of the face), but they may not completely capture all the relevant features in a given person or condition. Our study indicates which saliency explanations and face regions are more consistent with the phenotypes of specific genetic syndromes and could be used in large-scale clinical evaluations.

1. Introduction

Genetic disorders refer to medical conditions caused by abnormalities in or affecting a person's DNA. Precise diagnosis is important for optimal care. As genetic conditions often affect the face, clinicians rely on careful physical examination of the face to identify which condition a patient may have. Traditional computer vision based on feature engineering and, more recently, deep learning approaches have been shown to predict genetic

syndromes from facial images (Gurovich et al., 2019). However, such high-stake decisions require model explanations and, more specifically, a human-AI collaboration that supports clinicians' decision-making and diagnosis.

In recent years, machine learning applications, primarily based on deep and representation learning, have gained importance in digital healthcare and medical image processing tasks. Deep learning models showed remarkable performance in many tasks, which also caused much enthusiasm in the medical domain. However, human experts need explanations for high-stake decisions, as in human medicine and healthcare. Deep neural networks (DNN) are not inherently interpretable, so they must be explained. In computer vision, saliency maps (also called visual feature attribution methods) are well established, and they quantify the importance of image pixels for a network's decision.

Saliency methods either use a model's internal mechanisms and parameters (model specific) or are independent of the DNN model (model agnostic). In medical image analysis, there are many domains, such as histopathology, radiology, or ophthalmology, where saliency methods are used to explain DNN models (see for an overview, Van der Velden et al. (2022)). Even though the evaluation of saliency maps is not yet standardized, computational tasks such as object detection or recognition in various medical imaging applications can be evaluated relatively easily. However, in face images, multiple or all regions can play a role in the decision with varying importance. For instance, an image with Williams syndrome can have syndromic features affecting the eyes, nose, and mouth at the same time, and measuring localization performance by the intersection of regions is not straightforward. Furthermore, the specific underlying syndromes may be subtle or difficult to identify even for a highly trained clinician (Figure 1). Recognition of facial genetic conditions is more challenging than other popular face analysis tasks, such as identity or facial expression recognition, and requires human genetics expertise and clinical experience.

Explainable machine learning approaches in facial genetics are important in clinical applications, where phenotypic features help inform the differential diagnosis and decision-making, such as the genetic testing strategy. The Human Phenotype Ontology (HPO) is the most comprehensive and standardized collection of human diseases and phenotypes and a highly-used source for clinical genetics creating a computational bridge between genome biology and clinical medicine (Köhler et al., 2020). Previous computer vision work in facial genetics did not investigate the relationship between clinicians' and DNNs' decisions. Relating DNN decisions and HPO is an open research direction to evaluate DNN decisions and investigate their strengths and drawbacks.

In this study, we aim to generate more structured, robust, and understandable explanations based on saliency maps that relate the network's decision to facial regions and

---

1. All images used in the study are from publicly available sources, and we make the versions of the publicly available images included in our analyses available (via CC0 license) for the purpose of reproducibility and research, with the assumption that these would only be used for purposes that would be considered fair use. These data were compiled to produce a new, derivative work, which we offer as a whole. All images are publicly available (all are used with appropriate permission), and sources are listed in the appendix.

2. 22q11.2DS: 22q11.2 deletion syndrome; Angelman: Angelman syndrome; BWS: Beckwith-Wiedemann syndrome; CdLS: Cornelia de Lange syndrome; Down: Down syndrome; KS: Kabuki syndrome; NS: Noonan syndrome; PWS: Prader-Willi syndrome; RSTS: Rubinstein-Taybi syndrome; Unaffected: Unaffected individual; WHS: Wolf-Hirschhorn syndrome; WS: Williams syndrome
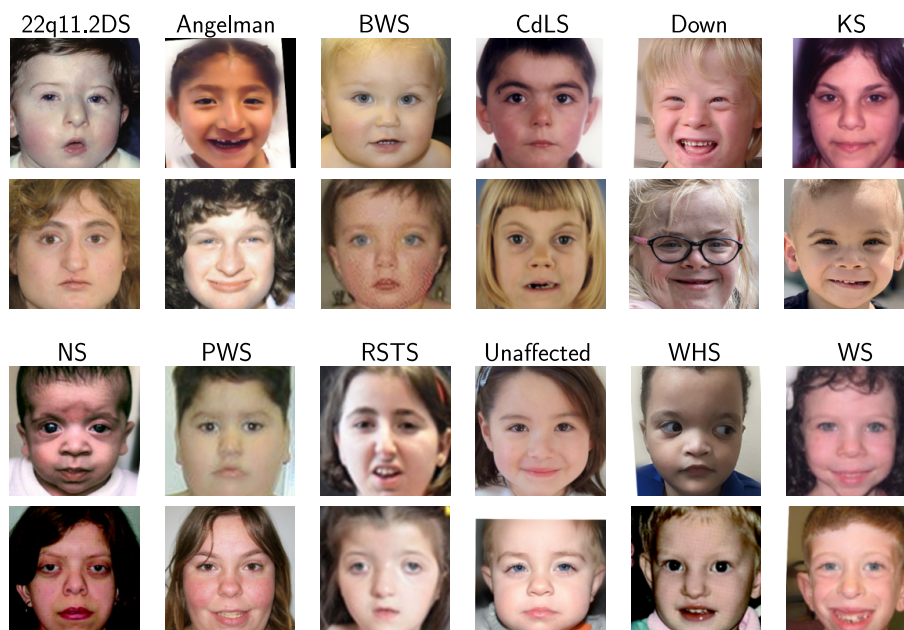
Figure 1: Representative images of individuals with specific genetic conditions, as well as unaffected individuals[1,2].

phenotypes that are known to be highly relevant based on the HPO taxonomy. We do this by proposing a region-based explanation approach that uses any saliency method and face parsing and generates coefficients representing the importance of face regions for DNN's prediction.

To validate our approach, we create a new facial image database of genetic conditions and an unaffected control group, carefully collected and confirmed by geneticists from previously published and other publicly available web-based resources.

Our paper aims to fill this gap. Our main contributions are as follows:

1. We propose a region-based saliency explanation approach to validate and compare DNN decisions in recognition of facial genetic syndromes. Our approach highlights the importance of face regions for DNN decisions. In this way, model explanations, together with predictions, provide more structured evidence about model predictions and potentially support clinicians' examination.

2. We create a new face database containing over 3,500 images of 11 rare facial syndromes and a healthy group to validate the proposed approach and further machine learning research in facial genetics. Clinicians validated the disease labels and carefully labeled the most prominent HPO terms in a test set for fine-grained analysis [3].

3. The evaluation of the region-based saliency explanation also contains twofold contributions: one is statistical testing of how DNN's decisions are faithful to show the most

---

3. The links and directions to recreate the data, phenotype annotations (HPO terms) are publicly available under https://doi.org/10.5281/zenodo.8209022. Source code is also available at https://github.com/sumeromer/facial-gestalt-xai.

influential facial regions to distinguish unaffected and syndromic groups, and the second is a correlation analysis of region-based explanations of DNN's predictions and the physicians' decisions based on HPO taxonomy. Our preliminary evaluation shows the strengths and drawbacks of saliency map methods and a promising direction toward human understandable explanations in recognizing rare facial genetic syndromes.

Generalizable Insights about Machine Learning in the Context of Healthcare

Computer vision and machine learning bring powerful solutions to the recognition of facial genetic syndromes; however, current DNN-based approaches are opaque models and lack the necessary evidence needed for clinical applications. To the best of our knowledge, none of the previous studies deploying machine learning have used model explanations yet.

Our study can be an exemplar for further research in explainable machine learning to evaluate and attribute model decisions to medical ontologies as we did by using HPO. These insights are not limited to facial genetics, but also applicable to different modalities such as medical imaging or electronic health records to align decisions of machine learning classifiers and ontologies in support of clinical decision-making.

The rest of the paper is organized as follows: In Section 2, we review the literature on both computer vision approaches in recognition of facial genetic syndromes and explainable AI methods in deep learning. Section 3 presents our region-based saliency explanation approach for facial genetic syndromes. Section 4 describes data collection, HPO annotation, and region of the interest selection that will be used in our approach as well as classification model and experimental settings. Subsequently, in Section 5, we present our experimental results to distinguish unaffected and syndromic groups (how faithful DNN's prediction is) and the relationship between our region-based saliency explanations and HPO terms labeled by clinicians (understandability). Finally, in Sections 6 and 7, we conclude this paper with a discussion including a summary of the results and limitations and discuss future implications.

2. Background and Related Work

In this section, we first review the automated methods to recognize rare facial genetic syndromes. As our work aims to integrate more explainable approaches to facial genetics and produce region-based relevance information for human experts, afterward, we review explainability methods in deep neural networks.

2.1. Recognition of Facial Genetic Syndromes

Since the early days of computer vision, several studies have aimed at automatically recognizing dysmorphic faces. Loos et al. (2003) gathered manually labeled 48 facial nodes from a limited number of images (only 55 images of mucopolysaccharidosis type III, Cornelia de Lange, fragile X, Prader–Willi, and Williams–Beuren syndromes), extracted texture information around these nodes using Gabor wavelets, and used elastic bunch graph matching to classify dysmorphic faces. Later, computer vision methodologies like Active Appearance Models (AAM) led to automated facial landmark estimation, and Ferry et al. (2014) classified genetic conditions and healthy controls in a larger dataset (2878 images) based

on facial landmarks and the pixel values around them as appearance features using Large Margin Nearest Neighbor (LMNN) metric learning. In understanding the subtle relationships between facial phenotypes and genetic conditions (beyond significant malformations such as a cleft palate or heart defect, i.e., minor variants and combinations, such as a slight difference in the angulation of the eyes or the shape of the tip of the nose), 3D facial models have also been very helpful Ham (2007). In contrast to previous works that require 3D face scanners to acquire face meshes, 3D Morphable Face Models (3DMM) have shown remarkable progress in estimating 3D facial geometry from a single-view image (Blanz and Vetter, 1999; Egger et al., 2020).

Traditional computer vision methods used handcrafted feature extractors and separate classifiers (Zhao et al., 2003); however, the progress in deep learning methods offered an opportunity to learn feature representations from data, notably, the performance of face recognition (Masi et al., 2018). Thanks to the availability of large-scale face databases with 4-5 million facial images, self-supervised learning further aims to learn proxy tasks (that align with different downstream tasks) on a large amount of unlabeled data. For instance, the use of contrastive learning on face image-text pairs and image inpainting that learns to retrieve tokens of masked facial image regions showed state-of-the-art results in face parsing and face alignment (Zheng et al., 2022).

Genetic syndromes can involve sequelae affecting different facial regions, and recognition of facial genetic conditions is similar to facial biometrics, with a significant difference: most syndromes appear very rarely in large populations, and finding training databases is challenging. Recent studies used pre-trained face recognition models on large databases and applied transfer learning to recognition of facial genetic syndromes (Gurovich et al., 2019; Hsieh et al., 2022). The power of learned representation showed promising results on unseen syndromes by retrieving the most similar samples compared to several gallery images, even in a population where the training set lacks enough samples from concordant genetic characteristics Mishima et al. (2019).

In addition to learned face representations, generative face modeling can also tackle data scarcity, for instance, as Duong et al. (2022) used StyleGAN2 with adaptive discriminator augmentation for age progression and showed improved recognition performance using the same persons' age-manipulated images.

Despite computer vision and machine learning literature in recognition of genetic syndromes in the last two decades, to the best of our knowledge, none of them incorporated explainable artificial intelligence approaches to relate facial images with particular facial regions or phenotype ontology and produce human-understandable interpretations.

2.2. Explainable AI Methods in Deep Neural Networks

Deep learning models perform far beyond other machine learning approaches in facial phenotyping; however, they are known not to be transparent. Saliency maps are often used in computer vision applications. Here, we mostly focus on feature attribution methods that assign an importance score for particular label or layer activations.

We consider a multiclass classifier $f_\theta : \mathbb{R}^{H \times W \times 3} \to \mathbb{R}^K$, where H and W are the height and width of input images, and K is the number of categories (in our case, the number of syndromes and healthy groups). An explanation method creates a relevance mapping

$h_{f_\theta} : \mathbb{R}^{H \times W \times 3} \to \mathbb{R}^{H \times W \times 3}$ that associates each input sample to a relevance score for classifier's decision.

2.2.1. GRADIENT-BASED METHODS

Simonyan et al. (2013) (Gradients) proposed visualization methods, one is to generate an image (in the input size) that maximize the class score. The second is computing a class saliency map for a given image and class using backpropagation. For a given image, x, they used $h_{f_\theta}(x) = \frac{\partial f_\theta}{\partial x}$. However, these raw gradients were typically noisy, and Smooht-Grad (Smilkov et al., 2017) sharpened gradient-based saliency maps by sampling many small perturbations to the input and averaging. A different approach for this problem was Integrated Gradients Sundararajan et al. (2017):

$$h_{f_\theta}(x) = (x - x') \times \int_{\alpha=0}^{1} \frac{\partial f(x' + \alpha \times (x - x'))}{\partial x} \partial \alpha \tag{1}$$

Here, they assigned importance scores for each feature by approximating the integral of gradients from a given baseline, x'.

Guided Backprop (Springenberg et al., 2015) is another gradient-based method; it also calculated the gradient of the target output with respect to the input image but masked out the negative values by changing the gradients of ReLU functions.

Zhang et al. (2018a) proposed a probabilistic Winner-Take-All (WTA) formulation for modeling top-down attention, inspired by Selective Tuning model (Tsotsos et al., 1995) that is a deterministic scheme to decide the most relevant neurons. In current DNN architectures, activation neurons are also being used, and their Excitation Backprop approach propagates in activation neurons. This approach passes top-down signals only through excitatory (non-negative neurons).

DeconvNet (Zeiler and Fergus, 2014) is another method that is based on Gradients except for the backpropagation through the ReLU nonlinearity. In pooling, activation, and convolutional layers, they reversed the filters by using transposed operations.

Another gradient-based approach is Layer-wise Relevance Propagation (LRP) (Bach et al., 2015; Montavon et al., 2019). The main difference, LRP is based on the decomposition of the decision and produces a relevance score between activations of each neuron and their inputs. Epsilon rule (LRP-ϵ) distributes the contributions of activations as follows:

$$R_j = \sum_{K} \frac{a_j w_{jk}}{\epsilon + \sum_{0,j} a_j w_{jk}} R_k \tag{2}$$

where j and k are neurons at consecutive layers, w are the weights of the classifier $f_\theta(.)$, a_j's are the activations at layer k. ϵ is a term to filter out weak and unrelated contributions. This rule at $\epsilon = 0$ is also called as the basic rule (LRP-0). Another rule, LRP-$\alpha\beta$ weights the positive and negative contributions as follows:

$$R_j = \sum_{K} \left(\alpha \frac{(a_j w_{jk})^+}{\sum_{0,j} (a_j w_{jk})^+} - \beta \frac{(a_j w_{jk})^-}{\sum_{0,j} (a_j w_{jk})^-} \right) R_k \tag{3}$$

Flat rule applies uniform weights only in the first layer, whereas w^2-rule similarly use squares of the weights in lower layers.

Following LRP, Shrikumar et al. (2017); Ancona et al. (2018)(DeepLIFT) calculated the relevance of neurons between consecutive layers in backpropagation. However, DeepLIFT compared the input x w.r.t. some reference input $\bar{x}$, and the selection of baseline depended on the domain knowledge (i.e., blurred version of input images). It also brought additional backpropagation rules to basic LRP, such as rescale and reveal cancel rules.

There are attribution methods that associate the feature importance of higher convolutional layers with the gradients of each class probability. For instance, Selvaraju et al. (2017) (GradCAM) used weighted activations of feature maps or Guided GradCAM when scores of Guided Backpropagation were used.

2.2.2. Occlusion Methods

Gradient-based methods in the previous section were all using the internal structure of the classifiers. Another way to investigate the contributions of particular regions of the input is occlusion analysis (Zeiler and Fergus, 2014; Zintgraf et al., 2017). In occlusion analysis, $h_{(i,j)} = f(x) - f(x \cdot m_{(i,j)})$. Here, a rectangular mask, m, is applied to the image, and the importance of the occluded part is decided according to the drop in classifier performance. Similarly, SHAP and Kernel SHAP approaches (Lundberg and Lee, 2017) can also be regarded as occlusion methods; however, they occlude not only a patch but a set of occlusion patterns.

Localization of Saliency Maps. There are different ways to evaluate the localization performance of saliency maps in the literature. Pointing game (Zhang et al., 2018b) evaluates whether the highest attribution of a saliency map lies inside the target object's bounding box. Kohlbrenner et al. (2020) used the ratio of positive attributions within the object bounding box to the total image area. Arras et al. (2020) evaluated localization performance using the ratio of highly attributed pixels in a bounding box to the size of the area. All these approaches are most suitable for generic object detection and recognition; however, in face analysis, all input faces are aligned. It is not straightforward to determine a particular bounding box as several dysmorphic features can be present in the same image and possibly be overlapping between different diseases. In different face analysis tasks such as the face, gender, expression recognition (John et al., 2021) or depression recognition (Zhou et al., 2018), saliency maps were deployed to explain decisions or detect biases in the model. However, none of the previous works reported a structured approach to relating the localization performance of DNN and human explanations.

Similarly, Saporta et al. (2022) reported that several saliency maps performed worse than the human benchmark and described several oracles; for instance, a model with perfect AUROC but saliency maps (1) can pick up confounders on the image, (2) do not localize well, or (3) do not correctly reflect the model's attention. However, the evaluation metrics they used were mean intersection over union (mIoU) and hit rate metrics, not a structured evaluation based on an ontology as we adopted by using the HPO taxonomy.

IoU and pointing game in Saporta et al. (2022) and the previously mentioned metrics define some regions of interest to judge how well saliency maps perform against ground truth (and human experts). However, these metrics are not suitable for facial genetics. The same facial regions can have varying severity (for example, at the mouth region, a Williams Syndrome image may be annotated with more HPO terms than an Angelman

Syndrome image, and vice versa). We often have varying numbers of phenotypes covering the eyes, nose, and mouth regions. Thus, the comparison of binarized segmentation maps and drawn regions by clinicians cannot correctly capture the alignment of saliency maps and the syndromic facial features.

Human understanding of object detection or recognition is more based on meaningful properties such as texture, color, object parts, or scene composition. To give more concrete examples, we perceive a person in an image as having certain characteristics (i.e., having two legs and arms and a head, walking or standing on a surface). Similarly, a pathologist decides about a cellular structure as a tumor based on a grading scale. Providing only bounding boxes does not tell us too much about human explanations, and we need more structured evidence. Also, in the general domain of computer vision, there are recent works to explain DNN decisions based on a set of samples representing the concept of human interest (Kim et al., 2018; Ghorbani et al., 2019). In the facial phenotyping of genetic syndromes, concepts of interest are well defined, often key regions like eyes, nose, and mouth, and more precisely as in HPO. We aim to relate this ontology and DNN explanations. This is the main motivation of this work, and we describe our approach, region-based saliency explanations, in the following section.

3. Region-Based Saliency Explanations

Genetic conditions that affect facial morphology, which are our primary interest here, are rare in populations. The identification of each genetic condition can depend on a number of visible dysmorphic features and phenotypes in faces. The primary motivation for using DNNs is to assist clinicians in practice, and the explanation of decisions plays an important role. However, the benefit of showing raw saliency maps is limited, and there is a need to associate decisions with human-interpretable regions. Our motivation is to transform saliency maps that contain the information of "where" into a summary of "what/which region".

Figure 2 describes our workflow to extract face region-based saliency explanations. We initially train a classifier network, f_θ, on labeled images to classify syndromic and unaffected groups. In the same image input provided to the classifier, we use a segmentation model for face parsing and acquire segmentation maps, for instance, eyes, eyebrows, nose, forehead, mouth, and so on. These fine-scale segmentation maps can be directly used, or different regions of interest can be created according to analyzed facial syndromes. In each of these regions, we will acquire region relevance values for DNN's decision.

In facial images, raw saliency explanations, h_{f_θ}, tell about the importance of each feature dimension. Using different saliency maps (as described in Section 2.2), we create a pixel-wise relevance map maximizing the most likely predicted category. Saliency maps usually contain both excitatory and inhibitory information. In a clinical use case, we should provide evidence for the predicted category. Furthermore, we compare excitatory information with human experts' evaluation. Thus, we applied feature normalization, $\frac{h_{f_\theta} - \min(h_{f_\theta})}{\max(h_{f_\theta}) - \min(h_{f_\theta})}$ to scale attributions to the range of $[0, 1]$. Subsequently, we calculate the mean activation of
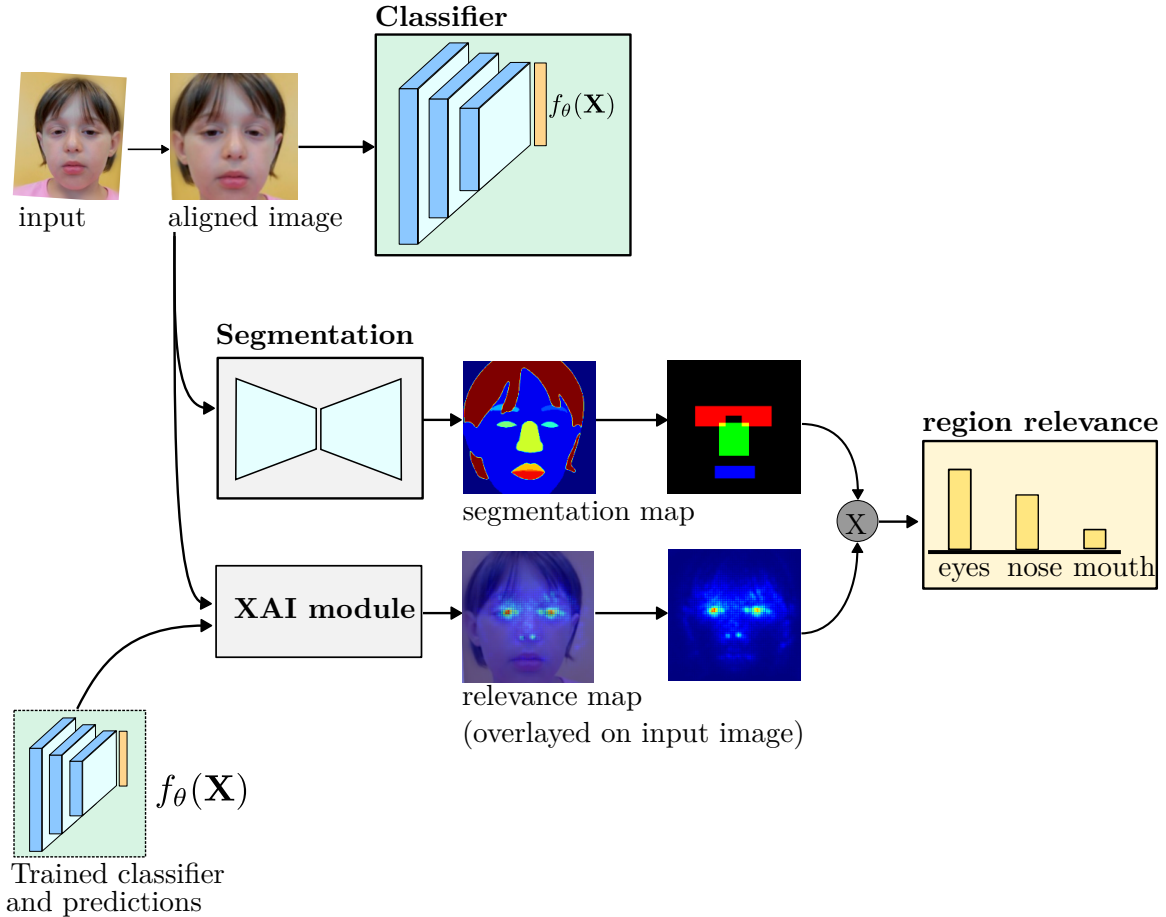
Figure 2: Region-based saliency explanations. Our approach takes the weights of a trained DNN, region map acquired from a pre-trained face parser network, and an XAI module. the outputs are single floating values for each region.

the relevance map per region as follows:

$$C_i = \frac{\sum h_{f_\theta(x) \,\odot\, m_i(x)}}{\sum m_i(x)} \tag{4}$$

where x is the aligned face image, $h(.)$ is the saliency map calculated with any available approach, m_i is the binary mask for any facial region whose importance needs to be calculated, and $\odot$ is the element-wise product sign. In this way, we acquire the total amount of activation, C_i for region i, in the raw saliency map normalized by the size of the region of interest. Therefore, C_i is a variable that summarizes the importance of a particular region behind DNN's prediction. In other words, instead of using pixel-based information (*where?*), we explain by "which part/what" plays a role in DNN's particular decision.

To summarize, our approach generates region-based relevance scores, which form the building blocks of our explanations by highlighting the importance of major facial parts relevant to the diseases investigated. Figure 3 depicts a sample image, three regions of interest, eyes, nose, and mouth, and observed HPO terms that will be used in further
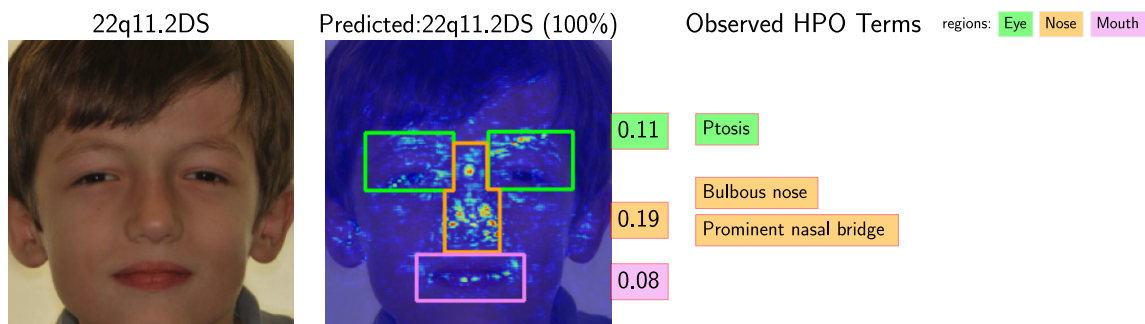
Figure 3: A test sample with 22q11.2 deletion syndrome. The original image, region-based explanation (using SmoothGrad), and observed HPO terms are shown.

evaluation. Region relevance coefficients represent the respective regions' importance in DNN's prediction based on a given saliency map [4].

In order to evaluate the potential of this approach to support the explainability of DNN predictions, we focus on two questions that cover different aspects of the problem:

1. Do our region relevance coefficients enable us to distinguish syndromic and unaffected samples? An initial step to answering this question is to verify that the distribution of our region-based saliency coefficients is statistically different between syndromic and unaffected images. This statistical difference is a necessary condition for region-based explanations to provide suitable visual support to humans to verify that a given prediction is correct. Furthermore, this analysis can also be regarded as a measure of faithfulness. We cannot consider these explanations faithful if our region relevance coefficients give comparable importance values on these two groups.

2. Do region-based explanations align with HPO terms labeled by clinicians? Asking this question is important because HPO terms represent the features that clinicians rely on to diagnose syndromic faces. On the other hand, DNN-based classifiers of syndromic vs. unaffected faces are entirely unaware of HPO terms. Finding an alignment of region-based explanations with HPO terms would validate the usefulness of explanations as they focus on features known to be discriminative for humans. Furthermore, this alignment would suggest a great potential for the region-based explanations to be understandable by clinicians since they reflect the elements that the clinicians are familiar with, where human understandability is the final step of evaluation for any explainable machine learning approach.

---

4. In our study, we used eyes, nose, and mouth regions, however, depending on the syndromes analyzed and their relation to facial regions, our approach can be applied to any parts (i.e., hair, upper forehead or chin).

4. Experiments

4.1. Dataset

We build the dataset to evaluate our proposed approach by using Google and PubMed to select publicly available images depicting individuals with one of the following 11 genetic conditions: 22q11.2 deletion, Angelman, Beckwith-Wiedemann (BWS), Cornelia de Lange (CdLS), Down, Kabuki (KS), Noonan (NS), Prader-Willi (PWS), Rubinstein-Taybi (RSTS), Wolf-Hirschhorn (WHS), and Williams (WS) syndromes. We selected these particular conditions because they often involve subtle but recognizable facial phenotypes and are common enough that there are sufficient publicly available images for the purposes of this study.

From websites and academic papers in human genetics and medicine, we initially collected more than 4,000 images. While we did collect multiple images from some of the papers we used, we usually reviewed three to five times as many papers to find a paper with usable images. Subsequently, 2 to 3 clinicians reviewed the images to ensure the patient's phenotype was consistent with the disorder, particularly if the genetic result was not given with the image. Together with an unaffected control group, this resulted in 3547 images. The number of samples per category is as follows: 22q11.2DS (591), Angelman (456), BWS (308), CdLS (120), Down (352), KS (246), NS (327), PWS (104), RSTS (105), Unaffected (228), WHS (178), and WS (529).

Unlike other image datasets, understanding facial phenotypes requires human genetics expertise and more careful data curation. Some images collected from the internet might contain an inaccurate diagnosis, which may introduce noise in the created database. Thus, the syndrome labels of each image in the database were verified by clinicians to prevent such situations. We also included facial images from public resources to get samples of individuals without genetic conditions. However, individuals we retrieved as "unaffected" could theoretically have an undiagnosed condition. To mitigate this, clinicians carefully inspected the unaffected images for a constellation of dysmorphic features (phenotypes) that would suggest a specific genetic condition. Figure 1 shows representative images from our dataset.

4.2. HPO Labeling and Defining Region of Interests

From over 13,000 HPO terms that are listed in the HPO taxonomy, the clinicians identified the set of relevant phenotypes (HPO terms) that are known to be visible in facial images from the Clinical Synopsis section of OMIM (https://omim.org), as described in Köhler et al. (2020). Among them, only HPO terms whose occurrences in a syndrome are frequent or very frequent are considered in our analysis. This resulted in 50 HPO terms represented in Table 1[5].

Labeling HPO terms is a time-consuming task requiring careful facial features inspection. For this reason, we considered 171 images containing varying phenotypes of 22q11.2 deletion, Angelman, Kabuki, Noonan, and Williams syndromes. Three clinicians viewed these images and independently annotated the absence (0) or presence (1) of the 50 HPO terms in Table 1.

---

5. Short descriptions, synonyms, PubMed references, syndromes, and gene associations of each HPO term can be found at https://hpo.jax.org/app

Table 1: Selected phenotypes (HPO terms) that are visible on facial images of the selected conditions in this study.

Overall		
Abnormal facial shape	Narrow face	Elfin facies
Long face	Coarse facial features	Webbed neck
Microcephaly	Triangular face	Midface retrusion
Hypopigmentation of the skin		

Ears		
Low-set ears	Small earlobe	Protruding ear
Overfolded helix	Macrotia	Thickened helices
Low-set posteriorly rotated ears		

Eyes		
Epicanthus	Strabismus	Proptosis
Upslanted palpebral fissure	Iris hypopigmentation	Highly arched eyebrow
Abnormal eyelid morphology	Blepharophimosis	Sparse lateral eyebrow
Downslanted palpebral fissures	Ptosis	Long eyelashes
Eversion of lateral third of lower eyelids	Hypertelorism	Telecanthus

Nose		
Prominent nasal bridge	Bulbous nose	Short columella
Wide nasal bridge	Short nose	

Mouth		
Long philtrum	Widely spaced teeth	Microdontia
Open bite	Protruding tongue	Wide mouth
Abnormality of the dentition	Thick lower lip vermilion	Everted lower lip vermilion

Forehead	Hair	Chin
High forehead	Fair hair	Pointed chin
Broad forehead		

After having HPO terms, we need to identify which of the regions in the face are conspicuous enough to be included in our region-based explanations and define a ground truth measure providing their importance. First, we aggregated clinicians' ratings and created 50-dimensional weight vectors that encode each HPO value varying from 0 to 3. Using these weight vectors, we plotted the t-SNE distribution of 171 test samples colored by their syndrome labels in Figure 4(a). As expected, five syndromic conditions can be clearly distinguished by using these phenotype annotations. So, this validates relevant HPO terms as a standard to evaluate explanations of DNN predictions.

We also investigated these 50 HPO terms' ability to distinguish five genetic conditions and further to see which terms are most useful in this context. For this purpose, we used the χ^2 statistic that measures the expected and observed frequencies of two events, in our case, between the dimensions of the weight vector and syndrome labels. The χ^2 statistic of HPO terms can be seen in Figure 4(b). There we see that, for instance, very relevant terms are eversion of the lateral third of lower eyelids, downslanted palpebral

Figure 4: **(a)** t-SNE embedding of 171 images represented via 50-dimensional weight vector, where each syndrome label is colored. **(b)** χ^2 statistic between HPO terms and syndrome categories (all p-values are below 0.05, and the most relevant features are highlighted in red).

fissures, and a short columella. The phenotypes under the "overall" group are related to the morphology of overall faces; thus, they cannot be explained (or localized) easily by saliency explanations. Ears are not visible in most images. Similarly, foreheads might be occluded by hair, particularly in toddlers and children. Furthermore, the number of prominent phenotypes in the hair and chin area is only one per area. Considering these conditions, Table 1 and feature importances in Figure 4(b), we decided on three regions of interest in the face: the eyes, nose, and mouth.

4.3. Classification Model

We train a convolutional neural network from frontal facial images to classify 11 genetic syndromes (22q11.2 deletion, Angelman, Beckwith-Wiedemann, Cornelia de Lange, Down, Kabuki, Noonan, Prader-Willi, Rubinstein-Taybi, Wolf-Hirschhorn, Williams) and healthy groups. We allocated HPO-annotated images (N = 171) for the test set. Considering the limited size of the database, the rest is used in person-independent 5-fold cross-validation to train the classifier, where we kept a small number of samples with kinship relations, either training or test sets at a time. We applied the RetinaFace (Deng et al., 2020) face detector, and 5-point similarity transform on eyes, nose, and mount points.

Due to the nature of the problem, we cannot create large image datasets as in face recognition. On the other hand, the information needed to analyze facial dysmorphism and biometric tasks is similar, and most prior works leveraged face recognition models. Thus, we apply transfer learning.

Most DNN architectures trained on face recognition tasks are in a lower resolution, for instance, in the size of 112×112 pixels. As we aim to create human-understandable saliency explanations, having fine-grained saliency maps would be a reasonable option. Thus, we chose a backbone trained on larger input images (of 224×224 pixels). The majority of face recognition studies focus on loss formulation, using comparable DNN architectures, and ResNet50 He et al. (2016) backbone is one of the most widely used ones. This is another reason to use a ResNet-50 trained on a large-scale face recognition database, VG-GFace2 (Cao et al., 2018).

All the networks in 5-folds were trained for 35 epochs using an SGD optimizer with an initial learning rate of 0.001 and a momentum of 0.9 and a Cosine annealing warm restarts scheduler.

Table 2: Validation performance evaluation for recognition of facial genetic syndromes on NIH-Faces. All metrics are reported mean/std. of 5-fold cross-validation.

	N	Precision	Recall	F1-score
22q11DS	557	0.816 (0.02)	0.813 (0.04)	0.814 (0.02)
Angelman	420	0.817 (0.03)	0.872 (0.03)	0.843 (0.02)
BWS	308	0.732 (0.01)	0.729 (0.05)	0.730 (0.03)
CdLS	120	0.858 (0.06)	0.791 (0.09)	0.818 (0.03)
Down	352	0.914 (0.04)	0.926 (0.03)	0.919 (0.03)
KS	212	0.805 (0.02)	0.802 (0.08)	0.802 (0.05)
NS	293	0.828 (0.05)	0.778 (0.07)	0.799 (0.03)
PWS	104	0.634 (0.10)	0.509 (0.09)	0.563 (0.09)
RSTS	105	0.813 (0.11)	0.801 (0.09)	0.806 (0.10)
Unaffected	228	0.703 (0.04)	0.726 (0.06)	0.710 (0.01)
WHS	178	0.775 (0.06)	0.767 (0.10)	0.765 (0.04)
WS	496	0.916 (0.03)	0.917 (0.04)	0.916 (0.03)
Overall	3373	0.801 (0.015)	0.786 (0.007)	0.790 (0.009)

The average Top-1 accuracy over 5-folds is 81.8%, and precision, recall, and F1-scores are presented in Table 2. The difficulty of each facial image to be recognized by clinicians and also DNN may depend on many factors, including the distinctiveness of facial features or the size of the region. Among the 11 syndromes, particularly, both the precision and recalls of Prader-Willi and Beckwith-Wiedemann seem below average. Interestingly, the unaffected group also has lower performance, with an F1 score of 0.710. This can be because the dataset is limited, and the initial facial representation is very powerful. Even after transfer learning on our database, we retain the knowledge of unaffected facial features. Furthermore, only a part of the face is affected in syndromic images. The best-performing categories are Down and Williams syndromes. The performance on 171 held-out test images of 22Q11.2DS, Angelman, Kabuki, Noonan, and Williams syndromes is much better, with an average accuracy of 90.2%(0.008).

4.4. Experimental Settings

The salience methods that we experiment with as the basis for our region-based approach are Gradient, SmothGrad, IntegratedGradients GuidedBackprob, ExcitationBackprop, DeconvNet, LRP, DeepLIFT and GuidedGradCam (they are described in Section 2.2.) In LRP, we used $\epsilon - z^+$, $\epsilon - z^+ - flat$, $\epsilon - \alpha_2 - \beta_1$, and $\epsilon - \alpha_2 - \beta_1 - flat$ rules. For DeepLIFT and GradCam, we also used the layer attribution of ResNet50's layer4. Furthermore, we used the Occlusion maps with a stride of 8 pixels and a sliding window of 15. In order to standardize saliency map computation, we used the open-source libraries Captum (Kokhlikyan

et al., 2020) and Zennit (Anders et al., 2021). An example of these saliency maps are given in Figure 5.
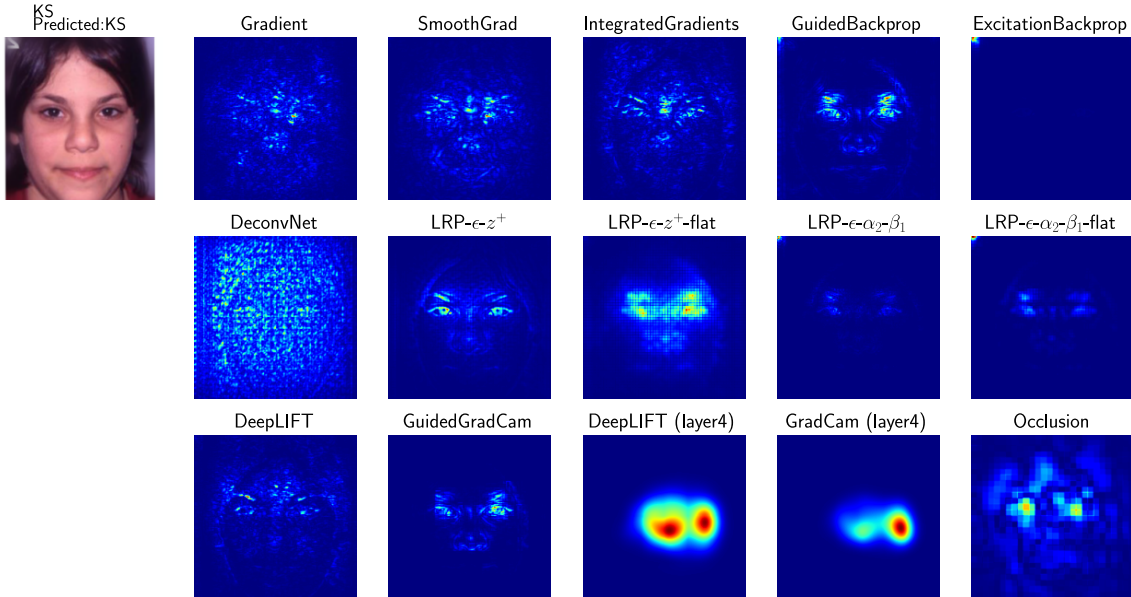

Figure 5: A test sample with Kabuki syndrome and all saliency methods used in this study (this particular subject has more prominent phenotypes in the eyes and nose regions).

When creating region coefficients, we initially ran FaRL face parsing model (Zheng et al., 2022) which is the state-of-the-art face parsing method in benchmarks such as LaPa (Liu et al., 2020) and CelebAMask-HQ (Lee et al., 2020), to acquire face segments. However, our region-based saliency explanation is rather generic, and it works on the manually drawn region of interest using another face parsing method or areas acquired by facial landmarks. After having face parsing, we created the non-overlapping eye, nose, and mouth regions as depicted in Figures 3 and 5 (on the input image above). As phenotypes related to the nasal bridge are considered in the nose regions, the area between the eyes is included in the nose. The reason why we picked these three regions was described in Section 4.2, and our experimental evaluation can be extended into different regions of interest.

We investigated two research questions described in Section 3. First is whether region-based saliency explanations distinguish unaffected and syndromic images or not. Second is region-based saliency explanations' relationship with HPO terms labeled by clinicians. Whereas the first one does not require any labeling, we used the respective test partitions of the entire dataset in a cross-validation setting. The second question is a comparison between region-based saliency explanations and human evaluation. As it requires HPO labels, we used only 171 held-out test images.

Do Region-based saliency explanations distinguish unaffected and syndromic images? To answer this question, we derive the distributions of region relevance coefficients, C_i's in unaffected and a particular syndromic group based on different state-of-the-art saliency-based techniques and test if there is a statistical difference between these two groups

or not. Here, our null hypothesis is there is no difference between the average region-based relevance score of unaffected and a particular syndromic category in a face region. In each comparison, we initially assess normality (Shapiro-Wilk) and equality of variances (Levene) tests and apply two-tailed t-test or Mann-Whitney-U tests, respectively. As we deployed many pairwise comparisons (number of regions $\times$ number of syndromes, 15 hypotheses in total), we used one-step Bonferroni correction at an alpha level of 0.001 among the tests used to evaluate a particular approach.

Do region-based explanations align with HPO terms labeled by clinicians? This question requires a ground truth importance measure of HPO terms. As the affected regions of respective HPO terms are known (see Table 1), we aggregated the number of annotations for each region in images to create a region importance measure. The distributions did not fulfill the normality conditions; thus, we used the Spearman rank correlation between these groups to compare DNN and human explanations. This correlation analysis does not tell about a specific HPO term, but it is a good indicator of how region-based explanations mimic the visual cues behind physicians' analysis.

5. Results

This section describes our experimental results in region-based saliency explanations of unaffected vs. syndromic groups and the relationship between model explanations and HPO terms.

5.1. Results on Region-based Saliency Explanations of Unaffected and Syndromic Images

We first look into the first question is whether our region-based explanations using different saliency maps helped to distinguish these five syndromes from unaffected groups. Table 3 summarizes the number of statistically significant outcomes using 15 different saliency maps (hence, 15 statistical tests).

These results depict that the importance of each region varies according to the syndrome that we compare with the unaffected control group. For instance, in our dataset, samples with 22q11.2DS differ from the unaffected group at the eyes and nose region because there are 11/15 and 9/15 significant tests for these two areas as compared to 3/15 significant tests for the mouth region. This result is in line with the phenotypic characteristics of 22q11.2DS, as it has more prominent phenotypes in the eyes and nose regions (i.e., epicanthus, upslanted palpebral fissure, abnormal eyelid morphology, ptosis, and telecanthus in eyes, and prominent nasal bridge, wide nasal bridge, and bulbous nose in nose). However, in our experiment, this condition only has one phenotype, long philtrum, annotated in the mouth region. There may be other useful phenotypes in the mouth regions, but the ones labeled by clinicians as prominent features in their decision-making suggest they are the most notable or important.

All the Angelman images in the test set have a wide mouth phenotype, and more than half of the images have protruding tongues. In contrast, strabismus and iris hypopigmentation in the eye region is not as common. Having more statistically significant outcomes of region-based coefficients in Williams syndrome has a similar condition. Williams syndrome

Table 3: The number of significant tests: genetic condition vs. unaffected comparisons using different region explanations and saliency methods (the total number of tests is 15).

Region/Syndrome	22q11.2DS	Angelman	Kabuki	Noonan	Williams
eye	11	2	2	9	9
nose	9	4	4	4	9
mouth	3	8	8	6	11

has wide mouth, thick lower lip vermilion, long philtrum, everted lower lip vermilion, and open bite. These phenotypes are observed in most of the images by clinicians.

In the case of Kabuki syndrome, only two tests are significant, in spite of having more prominent phenotypes in the eye regions, and they were manually observed in most of the images. Having a lower significance in the eye region's relevance (particularly in Angelman and Kabuki) can be due to major phenotypes such as strabismus and iris hypopigmentation, which require high-resolution, detailed imaging in the eye regions or might otherwise not be readily apparent and might be missed in typical quality facial images.

We can consider the ability to distinguish explanations of unaffected and syndromic groups as a necessary condition. There are different factors of variation, and it is not easy to tell about region-based coefficients' performance using a particular saliency map by only looking at this comparison.

Figure 6 depicts the boxplot distribution of these five syndromes and unaffected groups using selected saliency maps: SmoothGrad, LRP($\epsilon + flat$ rule), GradCam (using the last convolutional group, the layer4 of ResNet50), and Occlusion maps with a stride of 8 pixels and a sliding window of 15. Particular trends are easily noticeable; for instance, Angelman and Williams's relevance in the mouth region is significantly above the level of the unaffected group. These findings help us find the most prominent features to explain DNN's decisions, but we still need more structured HPO information.

5.2. Results on the Relationship between Region-based Saliency Explanations and HPO Terms

As clinicians labeled these images based on the most prominent HPO terms, they represent clinicians' visual attention based on prior phenotype ontology. We conducted a Spearman rank correlation analysis between the number of HPO terms labeled per region and region-based relevance coefficients in 171 held-out test images. Table 4 depicts the results of this analysis for region and each saliency map. In order to make a more reliable comparison, we used all five models trained in a cross-validation setting and reported average and standard deviations of correlation coefficients.

Independent of the saliency maps used, correlation analysis gave the best results in the mouth region. The highest correlation is acquired from GradCam (layer4) with 0.530. DeepLIFT, Occlusion maps, and SmoothGrad follow in the mouth region. Among LRP rules, there is a large variation; however, the ones with z^+-rule, LRP-ϵ-z^+-flat, and LRP-ϵ-α_2-β_1-flat with correlations of 0.422 and 0.429 are among the best. Interestingly, the DeconvNet approach consistently showed a negative correlation of -0.415 with HPO-based region coefficients, particularly in the mouth region. DeconvNet visually did not give relevant outcomes
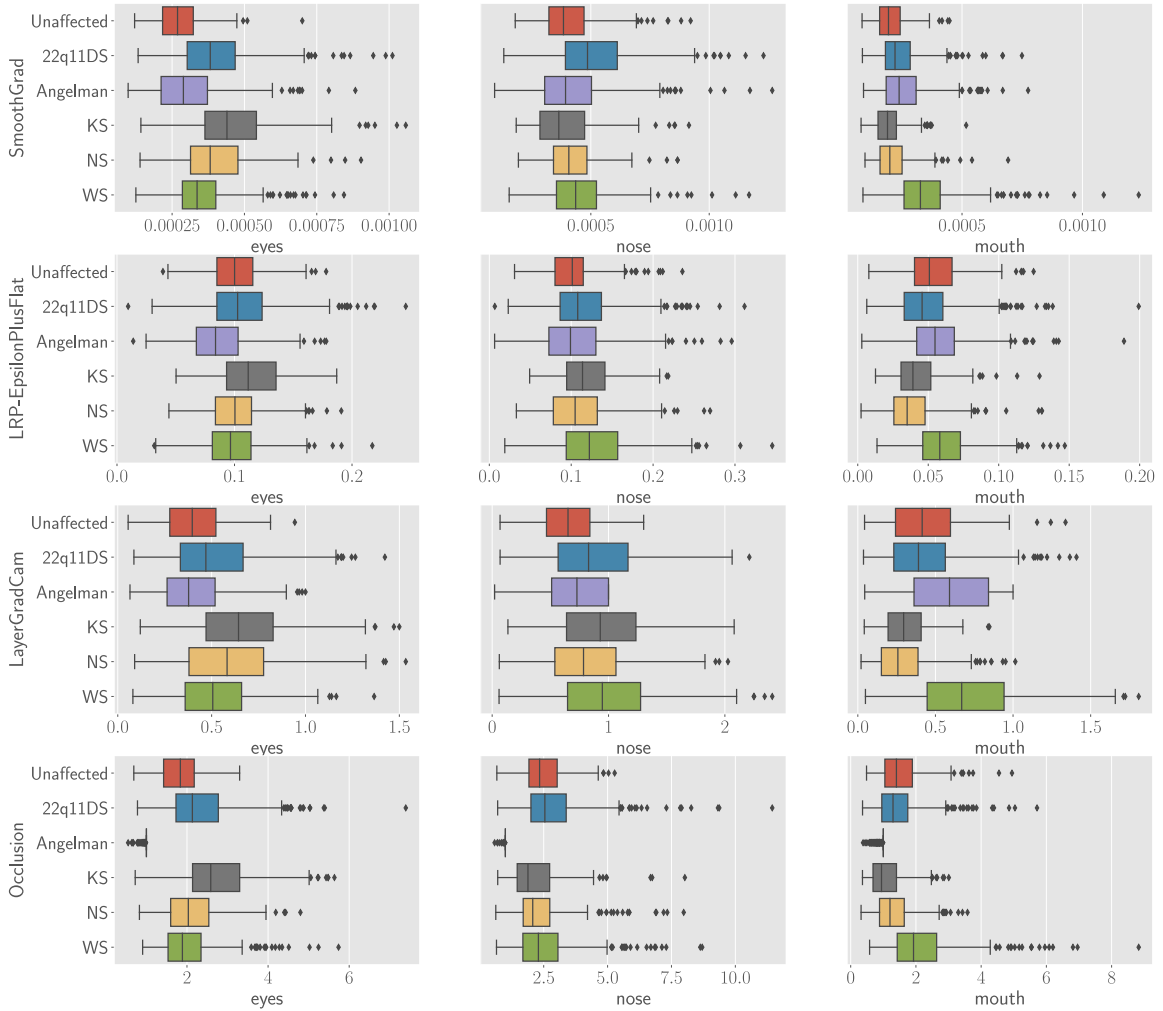
Figure 6: Region-based explanation coefficients using selected attribution methods. Here in the entire database (using respective trained models in cross-validation), the distribution of coefficients across healthy vs. five syndromes is depicted.

(as in Figure 5), and this can be regarded as an artifact of this saliency map on our data domain.

In the eye region, the average performance is far below that of mouth; however, several saliency methods performed well, for instance, GradCam (layer4) and GuidedGradCam, with correlations of 0.441 and 0.332. When we looked at the number of HPO terms per region, 15 HPO terms out of 51 are from eye regions. However, these phenotypes could not be explained by most of the saliency maps compared.

The worst-performing region is the nose. In contrast to the eye and mouth regions, we would expect having fewer HPO terms in the nose (prominent or wide nasal bridge, bulbous nose, short nose, and short columella) to facilitate the explanation of DNN decisions. Understanding these HPO terms in the nose region requires a different view of the same image. This is not the case in the eyes or mouth. In the eye region, the issue can be due to

Table 4: Correlation analysis between our region-based explanations and amount of labeled HPO terms for each regions (eyes, nose, and mouth), and various attribution methods.

	eye	nose	mouth
Gradient	0.145 (0.042)	0.078 (0.061)	0.335 (0.193)
SmoothGrad	0.043 (0.149)	0.034 (0.070)	0.486 (0.136)
IntegratedGradients	-0.039 (0.145)	0.023 (0.040)	0.425 (0.161)
GuidedBackprop	0.055 (0.091)	-0.046 (0.065)	0.331 (0.112)
ExcitationBackprop	0.022 (0.203)	-0.006 (0.072)	0.345 (0.078)
DeconvNet	-0.101 (0.065)	-0.033 (0.090)	-0.414 (0.096)
LRP-ϵ-z^+	-0.007 (0.055)	-0.015 (0.061)	0.403 (0.076)
LRP-ϵ-z^+-flat	0.050 (0.079)	-0.011 (0.088)	0.422 (0.091)
LRP-ϵ-α_2-β_1	0.060 (0.151)	0.017 (0.148)	0.380 (0.091)
LRP-ϵ-α_2-β_1-flat	-0.014 (0.124)	-0.008 (0.153)	0.429 (0.086)
DeepLIFT	0.162 (0.119)	0.049 (0.063)	0.512 (0.097)
GuidedGradCam	0.332 (0.129)	0.074 (0.121)	0.449 (0.123)
DeepLIFT (layer4)	0.087 (0.120)	0.016 (0.136)	0.410 (0.213)
GradCam (layer4)	0.441 (0.085)	0.259 (0.215)	0.530 (0.137)
Occlusion	0.111 (0.125)	-0.145 (0.117)	0.512 (0.090)

a number of factors (i.e., having small affected areas). However, having better HPO correlations also in the eyes and nose area can potentially improve the human understandability of region-based explanations.

6. Discussion

In this study, we trained a classifier with decent performance in recognizing genetic syndromes and investigated the saliency explanations of these DNN models. Our region-based explanation approach depicted the strengths and weaknesses of different saliency maps. Furthermore, we proposed a quantitative method to compare DNNs' explanations and clinicians' decisions based on phenotype ontology. In contrast to previous works on facial genetics that relied on machine learning models, we carefully investigated the explainability of deep learning models in facial genetic syndromes.

Our work also has certain limitations. Due to the scarcity of available image data, we have to transfer the pre-trained representations from other face analysis domains. The most relevant task is face recognition, which learns morphology information from large-scale image databases. Different forms of bias in face recognition have been known in recent years (Leslie, 2020), and the computer vision community aims to mitigate this. Thus, as well as explainable machine learning, the fairness of algorithmic solutions is critical. For instance, there is limited diversity of race and ancestry in most face recognition databases. The performance of a trained deep learning model varies across ethnicities, genders, and age groups. Thus, ensuring the fairness criteria of classifiers is another aspect of the problem that needs to be carefully addressed.

We created an extensive, clinically validated face database that can be used in further research to improve automated methods to recognize syndromes and explain classifiers. However, most diseases are rare, and it is more challenging to scale up the size of the database than in face recognition databases. Furthermore, the average viewer cannot easily perceive the syndrome from faces; field experts and clinicians' support is needed to verify labeling. There are commercial tools (i.e., web-based and mobile applications such as Face2Gene[6]), or more recently, publicly available web services and databases like Gestalt-Matcher (Hsieh et al., 2022). These tools have a great potential to create awareness among clinicians and contribute to larger databases of genetic syndromes. The use of few-shot learning on recognition of facial genetic syndromes (Sümer et al., 2022) is a promising direction to address data scarcity, and there is a line of work in the explainability of few-shot learning models (Wang et al., 2022). We are also limited by the images being taken in 2D, like passport photos. Statistical shape modeling is another alternative; for instance, creating 3D databases of faces with dysmorphism (Matthews et al., 2021) and the explainability of deep learning models in shape analysis is another open research direction.

Our approach is an initial step towards generating human understandable explanations by using region relevance for DNN's decisions. In other words, explanations should be in the following form: "This model predicts this syndrome because the following phenotypes are present in this image." Such an explanation requires more time consuming labeling work at a larger scale and, in return, improves explanations.

As well as learning problems and datasets, explainable machine learning methods depend on the DNN architectures, too. We used a well-performing Resnet architecture and presented a proof of concept to systematically evaluate region-based explanations. In future work, the effects of different DNN architectures can also be investigated.

Facial phenotyping provides a structured and objective way to measure the shift between DNNs and humans' attention. However, a large-scale clinical validation of our region-based explanations is needed. How does explanation-based visual support help clinicians' diagnoses? This question is essential to deploying these methods at a larger scale for clinical use. We left this for future work.

7. Conclusion

This study proposes a region-based saliency explanation approach for the deep neural networks trained to recognize facial genetic syndromes. We evaluated different saliency maps based on the importance of the eyes, nose, and mouth regions in face images. We trained a classification model that performed with an acerate accuracy of 81.8% and investigated how the region-based saliency explanations differ between unaffected and syndromic images. Depending on the syndrome investigated, all three regions are comparatively informative in distinguishing syndromic images' explanations from the unaffected set. Williams and Noonan syndromes showed more distinctive explanations that separated them from the unaffected group. The importance of regions varies; for instance, 22Q11.2DS has a better separation in the eye regions, whereas it is the mouth region in Angelman syndrome. Overall, our findings aligned with the prominent affected areas of particular syndromes.

---

6. https://www.face2gene.com/

Furthermore, we acquired HPO annotations of an image set and compared region-based explanations of different saliency maps and HPO annotations. Explanations were more in line with clinicians' HPO annotations in the mouth region; however, eyes and nose were less precise. Among the saliency methods compared, GradCam, DeepLIFT, and Occlusion maps were the best-performing approaches. Our study provided a quantitative approach to comparing saliency explanations and pointed out the weaknesses to guide future research in this direction.

Acknowledgments

References

The use of 3d face shape modelling in dysmorphology. *Archives of Disease in Childhood*, 92(12):1120–1126, 2007. ISSN 0003-9888. doi: 10.1136/adc.2006.103507. URL https://adc.bmj.com/content/92/12/1120.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Sy21R9JAW.

Christopher J Anders, David Neumann, Wojciech Samek, Klaus-Robert Müller, and Sebastian Lapuschkin. Software for dataset-wide xai: from local explanations to global insights with zennit, corelay, and virelay. *arXiv preprint arXiv:2106.13200*, 2021.

Leila Arras, Ahmed Osman, and Wojciech Samek. Ground truth evaluation of neural network explanations with clevr-xai. *arXiv preprint arXiv:2003.07258*, 2020.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999.

Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the*

IEEE/CVF conference on computer vision and pattern recognition, pages 5203–5212, 2020.

Dat Duong, Ping Hu, Cedrik Tekendo-Ngongang, Suzanna E Ledgister Hanchard, Simon Liu, Benjamin D Solomon, and Rebekah L Waikel. Neural networks for classification and image generation of aging in genetic syndromes. *Frontiers in Genetics*, 13, 2022.

Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (TOG)*, 39(5):1–38, 2020.

Quentin Ferry, Julia Steinberg, Caleb Webber, David R FitzPatrick, Chris P Ponting, Andrew Zisserman, and Christoffer Nellåker. Diagnostically relevant facial gestalt information from ordinary photos. *elife*, 3:e02020, 2014.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems*, 32, 2019.

Yaron Gurovich, Yair Hanani, Omri Bar, Guy Nadav, Nicole Fleischer, Dekel Gelbman, Lina Basel-Salmon, Peter M Krawitz, Susanne B Kamphausen, Martin Zenker, et al. Identifying facial phenotypes of genetic disorders using deep learning. *Nature medicine*, 25(1):60–64, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Tzung-Chien Hsieh, Aviram Bar-Haim, Shahida Moosa, Nadja Ehmke, Karen W Gripp, Jean Tori Pantel, Magdalena Danyel, Martin Atta Mensah, Denise Horn, Stanislav Rosnev, et al. Gestaltmatcher facilitates rare disease matching using facial phenotype descriptors. *Nature genetics*, 54(3):349–357, 2022.

Thrupthi Ann John, Vineeth N Balasubramanian, and CV Jawahar. Canonical saliency maps: Decoding deep face models. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(4):561–572, 2021.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

Maximilian Kohlbrenner, Alexander Bauer, Shinichi Nakajima, Alexander Binder, Wojciech Samek, and Sebastian Lapuschkin. Towards best practice in explaining neural network decisions with lrp. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020.
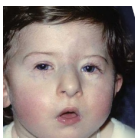
Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*, 2020.

Sebastian Köhler, Michael Gargano, Nicolas Matentzoglu, Leigh C Carmody, David Lewis-Smith, Nicole A Vasilevsky, Daniel Danis, Ganna Balagura, Gareth Baynam, Amy M Brower, Tiffany J Callahan, Christopher G Chute, Johanna L Est, Peter D Galer, Shiva Ganesan, Matthias Griese, Matthias Haimel, Julia Pazmandi, Marc Hanauer, Nomi L Harris, Michael J Hartnett, Maximilian Hastreiter, Fabian Hauck, Yongqun He, Tim Jeske, Hugh Kearney, Gerhard Kindle, Christoph Klein, Katrin Knoflach, Roland Krause, David Lagorce, Julie A McMurry, Jillian A Miller, Monica C Munoz-Torres, Rebecca L Peters, Christina K Rapp, Ana M Rath, Shahmir A Rind, Avi Z Rosenberg, Michael M Segal, Markus G Seidel, Damian Smedley, Tomer Talmy, Yarlalu Thomas, Samuel A Wiafe, Julie Xian, Zafer Yüksel, Ingo Helbig, Christopher J Mungall, Melissa A Haendel, and Peter N Robinson. The Human Phenotype Ontology in 2021. *Nucleic Acids Research*, 49(D1):D1207–D1217, 12 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa1043. URL https://doi.org/10.1093/nar/gkaa1043.

Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5558, 2020.

David Leslie. Understanding bias in facial recognition technologies. *arXiv preprint arXiv:2010.07023*, 2020.

Yinglu Liu, Hailin Shi, Hao Shen, Yue Si, Xiaobo Wang, and Tao Mei. A new dataset and boundary-attention semantic segmentation for face parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11637–11644, 2020.

Hartmut S Loos, Dagmar Wieczorek, Rolf P Würtz, Christoph von der Malsburg, and Bernhard Horsthemke. Computer-based recognition of dysmorphic faces. *European Journal of Human Genetics*, 11(8):555–560, 2003.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Iacopo Masi, Yue Wu, Tal Hassner, and Prem Natarajan. Deep face recognition: A survey. In *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 471–478, 2018. doi: 10.1109/SIBGRAPI.2018.00067.

Harold S Matthews, Richard L Palmer, Gareth S Baynam, Oliver W Quarrell, Ophir D Klein, Richard A Spritz, Raoul C Hennekam, Susan Walsh, Mark Shriver, Seth M Weinberg, et al. Large-scale open-source three-dimensional growth curves for clinical facial assessment and objective description of facial dysmorphism. *Scientific reports*, 11(1): 12175, 2021.

Hiroyuki Mishima, Hisato Suzuki, Michiko Doi, Mutsuko Miyazaki, Satoshi Watanabe, Tadashi Matsumoto, Kanako Morifuji, Hiroyuki Moriuchi, Koh-ichiro Yoshiura, Tatsuro
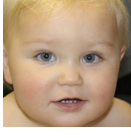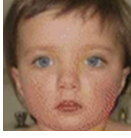
Kondoh, et al. Evaluation of face2gene using facial images of patients with congenital dysmorphic syndromes recruited in japan. *Journal of human genetics*, 64(8):789–794, 2019.

Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 193–209, 2019.

Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, Steven QH Truong, Chanh DT Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G Blankenberg, Andrew Y Ng, et al. Benchmarking saliency methods for chest x-ray interpretation. *Nature Machine Intelligence*, 4(10):867–878, 2022.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR, 2017.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

J Springenberg, Alexey Dosovitskiy, Thomas Brox, and M Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR (workshop track)*, 2015.

Ömer Sümer, Fabio Hellmann, Alexander Hustinx, Tzung-Chien Hsieh, Elisabeth André, and Peter Krawitz. Few-shot meta learning for recognizing facial phenotypes of genetic disorders. *arXiv preprint arXiv:2210.12705*, 2022.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

John K Tsotsos, Scan M Culhane, Winky Yan Kei Wai, Yuzhong Lai, Neal Davis, and Fernando Nuflo. Modeling visual attention via selective tuning. *Artificial intelligence*, 78 (1-2):507–545, 1995.

Bas HM Van der Velden, Hugo J Kuijf, Kenneth GA Gilhuijs, and Max A Viergever. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, page 102470, 2022.

Bowen Wang, Liangzhi Li, Manisha Verma, Yuta Nakashima, Ryo Kawasaki, and Hajime Nagahara. Match them up: visually explainable few-shot image classification. *Applied Intelligence*, pages 1–22, 2022.
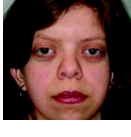
Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014.

Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018a.

Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018b.

W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Comput. Surv.*, 35(4):399–458, dec 2003. ISSN 0360-0300. doi: 10.1145/954339.954342. URL https://doi.org/10.1145/954339.954342.

Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18676–18688, 2022. doi: 10.1109/CVPR52688.2022.01814.

Xiuzhuang Zhou, Kai Jin, Yuanyuan Shang, and Guodong Guo. Visually interpretable representation learning for depression recognition from facial images. *IEEE Transactions on Affective Computing*, 11(3):542–552, 2018.

Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. In *International Conference on Learning Representations*, 2017.
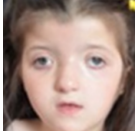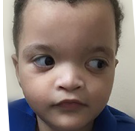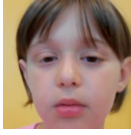
Appendix A. Web Resources

Table 5: All the images used throughout the paper were acquired from publicly available resources, and references to publications or websites are listed below.

Image	Reference
22q11.2DS	Maria Cristina Digilio, Bonnie Marino, Rossella Capolino, and B Dallapiccola. Clinical manifestations of deletion 22q11. 2 syndrome (digeorge/velo-cardio-facial syndrome). *Images in paediatric cardiology*, 7(2):23, 2005.
22q11.2DS	Elena Michaelovsky, Amos Frisch, Miri Carmel, Miriam Patya, Omer Zarchi, Tamar Green, Lina Basel-Vanagaite, Abraham Weizman, and Doron Gothelf. Genotype-phenotype correlation in 22q11. 2 deletion syndrome. *BMC medical genetics*, 13(1):1–11, 2012.

Angelman	Emiy Yokoyama-Rebollar, Adriana Ruiz-Herrera, Esther Lieberman-Hernandez, Del Castillo-Ruiz, Silvia Sanchez-Sandoval, Silvia M Avila-Flores, Jose Luis Castrillo, et al. Angelman syndrome due to familial translocation: unexpected additional results characterized by microarray-based comparative genomic hybridization. *Molecular Cytogenetics*, 8(1):1–8, 2015.
Angelman	Thomas Liehr. Uniparental disomy is a chromosomic disorder in the first place. *Molecular Cytogenetics*, 15(1):1–12, 2022.
BWS	Kathleen H Wang, Jonida Kupa, Kelly A Duffy, and Jennifer M Kalish. Diagnosis and management of beckwith-wiedemann syndrome. *Frontiers in pediatrics*, 7:562, 2020.
BWS	Silvia Russo, Luciano Calzari, Alessandro Mussa, Ester Mainini, Matteo Cassina, Stefania Di Candia, Maurizio Clementi, Sara Guzzetti, Silvia Tabano, Monica Miozzo, et al. A multi-method approach to the molecular diagnosis of overt and borderline 11p15. 5 defects underlying silver–russell and beckwith–wiedemann syndromes. *Clinical epigenetics*, 8(1): 1–15, 2016.
CdLS	Antonie D Kline, Joanna F Moss, Angelo Selicorni, Anne-Marie Bisgaard, Matthew A Deardorff, Peter M Gillett, Stacey L Ishman, Lynne M Kerr, Alex V Levin, Paul A Mulder, et al. Diagnosis and management of cornelia de lange syndrome: first international consensus statement. *Nature Reviews Genetics*, 19(10):649–666, 2018b.
CdLS	Antonie D Kline, Joanna F Moss, Angelo Selicorni, Anne-Marie Bisgaard, Matthew A Deardorff, Peter M Gillett, Stacey L Ishman, Lynne M Kerr, Alex V Levin, Paul A Mulder, et al. Diagnosis and management of cornelia de lange syndrome: first international consensus statement. *Nature Reviews Genetics*, 19(10):649–666, 2018a.
Down	Wikipedia contributors. Down syndrome, 2023. URL https://en.wikipedia.org/wiki/Down_syndrome. [Online; accessed 20-April-2023].

Down	Wikipedia contributors. List of people with down syndrome, 2023. URL https://en.wikipedia.org/wiki/List_of_people_with_Down_syndrome. [Online; accessed 20-April-2023].
KS	Victor Faundes, Stephanie Goh, Rhoda Akilapa, Heidre Bezuidenhout, Hans T Bjornsson, Lisa Bradley, Angela F Brady, Elise Brischoux-Boucher, Han Brunner, Saskia Bulk, et al. Clinical delineation, sex differences, and genotype–phenotype correlation in pathogenic kdm6a variants causing x-linked kabuki syndrome type 2. *Genetics in medicine*, 23(7): 1202–1210, 2021.
KS	Wikipedia contributors. Category:kabuki syndrome, 2023. URL https://commons.wikimedia.org/wiki/Category:Kabuki_syndrome?uselang=de. [Online; accessed 20-April-2023].
NS	M Digilio and Bonnie Marino. Clinical manifestations of noonan syndrome. *Images Paediatr Cardiol*, 3(2):19–30, 2001.
NS	Ellen Denayer, Th de Ravel, and Eric Legius. Clinical and molecular aspects of ras related disorders. *Journal of medical genetics*, 45(11):695–703, 2008.
PWS	MA Angulo, MG Butler, and ME Cataletto. Prader-willi syndrome: a review of clinical, genetic, and endocrine findings. *Journal of endocrinological investigation*, 38:1249–1263, 2015.
PWS	Merlin G Butler, Jennifer L Miller, and Janice L Forster. Prader-willi syndrome-clinical genetics, diagnosis and treatment approaches: an update. *Current pediatric reviews*, 15 (4):207–244, 2019.
RSTS	Virginia Perez-Grijalba, Alberto Garcia-Oguiza, Maria Lopez, Judith Armstrong, Sixto Garcia-Minaur, Jose Maria Mesa-Latorre, Mar O'Callaghan, Merce Pineda Marfa, Maria Antonia Ramos-Arroyo, Fernando Santos-Simarro, et al. New insights into genetic variant spectrum and genotype–phenotype correlations of rubinstein-taybi syndrome in 39 crebbp-positive patients. *Molecular Genetics & Genomic Medicine*, 7(11):e972, 2019a.

RSTS	Virginia Perez-Grijalba, Alberto Garcia-Oguiza, Maria Lopez, Judith Armstrong, Sixto Garcia-Minaur, Jose Maria Mesa-Latorre, Mar O'Callaghan, Merce Pineda Marfa, Maria Antonia Ramos-Arroyo, Fernando Santos-Simarro, et al. New insights into genetic variant spectrum and genotype–phenotype correlations of rubinstein-taybi syndrome in 39 crebbp-positive patients. *Molecular Genetics & Genomic Medicine*, 7(11):e972, 2019a.
Unaffected	Virginia Perez-Grijalba, Alberto Garcia-Oguiza, Maria Lopez, Judith Armstrong, Sixto Garcia-Minaur, Jose Maria Mesa-Latorre, Mar O'Callaghan, Merce Pineda Marfa, Maria Antonia Ramos-Arroyo, Fernando Santos-Simarro, et al. New insights into genetic variant spectrum and genotype–phenotype correlations of rubinstein-taybi syndrome in 39 crebbp-positive patients. *Molecular Genetics & Genomic Medicine*, 7(11):e972, 2019a.
Unaffected	Unaffected subject, Apr 2023. URL https://pxhere.com/en/photo/908228. [Online; accessed 20-April-2023]
WHS	Mona K Mekkawy, Alaa K Kamel, Manal M Thomas, Engy A Ashaat, Maha S Zaki, Ola M Eid, Samira Ismail, Saida A Hammad, Hisham Megahed, Heba ElAwady, et al. Clinical and genetic characterization of ten egyptian patients with wolf–hirschhorn syndrome and review of literature. *Molecular Genetics & Genomic Medicine*, 9(2):e1546, 2021a.
WHS	Mona K Mekkawy, Alaa K Kamel, Manal M Thomas, Engy A Ashaat, Maha S Zaki, Ola M Eid, Samira Ismail, Saida A Hammad, Hisham Megahed, Heba ElAwady, et al. Clinical and genetic characterization of ten egyptian patients with wolf–hirschhorn syndrome and review of literature. *Molecular Genetics & Genomic Medicine*, 9(2):e1546, 2021a.
WS	Griet Van Buggenhout, C Melotte, Binita Dutta, Guido Froyen, Paul Van Hummelen, Peter Marynen, Gert Matthijs, Thomy de Ravel, Koenraad Devriendt, Jean-Pierre Fryns, et al. Mild wolf-hirschhorn syndrome: micro-array cgh analysis of atypical 4p16. 3 deletions enables refinement of the genotype-phenotype map. *Journal of Medical Genetics*, 41(9): 691–698, 2004.

WS	Colleen A Morris and Carolyn B Mervis. Williams syndrome. *Cassidy and Allanson's Management of Genetic Syndromes*, pages 1021–1038, 2021.
KS	Francesca Romana Lepri, Dario Cocciadiferro, Bartolomeo Augello, Paolo Alfieri, Valentina Pes, Alessandra Vancini, Cristina Caciolo, Gabriella Maria Squeo, Natascia Malerba, Iolanda Adipietro, et al. Clinical and neurobehavioral features of three novel kabuki syndrome patients with mosaic kmt2d mutations and a review of literature. *International Journal of Molecular Sciences*, 19(1):82, 2017.
22q11.2DS	Paul Kruszka, Yonit A Addissie, Daniel E McGinn, Antonio R Porras, Elijah Biggs, Matthew Share, T Blaine Crowley, Brian HY Chung, Gary TK Mok, Christopher CY Mak, et al. 22q11. 2 deletion syndrome in diverse populations. *American Journal of Medical Genetics Part A*, 173(4):879–888, 2017.