Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

Master's Thesis

"On the Explainability of Neural Network Models to Classify Rare Genetic Syndromes from Frontal Facial Images"

Aswinkumar Vijayananth

Submitted to Hochschule Bonn-Rhein-Sieg,
Department of Computer Science
in partial fullfilment of the requirements for the degree
of Master of Science in Autonomous Systems

Supervised by

Prof. Dr Paul G. Plöger
Prof. Dr Ralf Thiele
Prof. Dr. med. Dipl. Phys. Peter Krawitz

November 2022

© 2020 Aswinkumar Vijayananth. All Rights Reserved

I, the undersigned below, declare that this work has not previously been submitted to this or any other university and that it is, unless otherwise stated, entirely my own work.

_____ _____

Date Aswinkumar Vijayananth

Abstract

Your abstract

Acknowledgements

Thanks to ....

Contents

List of Figures

List of Tables

1

Introduction

"Artificial Neural Networks (ANNs) are increasingly applied for medical image diagnostics" [6]. Medical image data such as scans produced from imaging devices like x-rays [7], Magnetic Resonance Imaging (MRI) [8], Computed Tomography (CT) [9], and ultra sound [10], waveforms produced from procedures like ElectroCardioGraphy (ECG) [11] and ElectroEncephaloGraphy (EEG), histological images [12], images of body parts, and admission notes [13] are fed into ANNs to perform tasks such as segmentation [14], classification [15] and abnormality detection [12].

Predictions made by neural network models are typically intended to be used in a clinical setting, to aid medical practitioners in diagnosing their patients. As a result, it can help in an overall reduction of "misdiagnosis", which is one of the most severe problems in health care [13]. Besides, the adoption of ANN based Artificial Intelligence (AI) systems facilitates early screening and identification of life-threatening conditions such as cancer in population with limited or no access to sub-specialty trained clinicians.

ANN models such as the ones used in AI systems for intracranial haemorrhage classification [9] and breast cancer prediction [16] are claimed to outperform clinicians in their respective tasks. In spite of their success in terms of predictive performance, the black box nature of ANNs restrains them from getting deployed in a clinical setting. Inherently, ANNs lack transparency and therefore are considered to be less dependable for high stake applications like medical diagnosis and autonomous driving. Besides, regulatory bodies such as "US FDA (United States Food and Drug Administration) require any clinical decision support software to explain the rationale or support for its decisions to enable the users to independently review the basis of their recommendations" [9]. Such contradictions further restrict the deployment of ANN based AI systems for performing medical diagnoses in a clinical setting.

"Explainable Artificial Intelligence (XAI) is a set of processes and methods that allows human users to comprehend and trust the results and output created by machine learning algorithms" [17]. Application of XAI techniques to Machine Learning (ML) models enables its human users to better understand their behavior and rationale behind their predictions. As a result, they improve transparency of models like ANNs, in turn making them more trustworthy for applications like medical diagnosis.

In the recent times, there has been a steep rise in the adoption of XAI methods to explain ANN models for medical diagnoses. A considerable number existing works focus on explaining neural network models that use radiological image data to perform tasks such as COVID classification [18], breast cancer risk assessment [10] and haemorrhage detection [9] respectively. A few other leverage XAI techniques to

explain neural network models used for detecting cancer from histopathological [12] and skin images [19].

The objective of this thesis is to use XAI to explain one such neural network model called GestaltMatcher [1], which surpasses the performance of clinical practitioners in the identification of certain rare genetic syndromes, from frontal facial images of patients. The findings of this work shall make the GestaltMatcher model more transparent and dependable, thereby taking it a step closer to be deployed in a clinical setting.

1.1 Motivation

"Rare genetic disorders affect more than 6.2% of global population" [1]. A significant fraction of the population with certain such disorders are characterized by facial abnormalities, which make up their respective facial phenotypes. The facial phenotypic information is used by clinical geneticists along with results from other laboratory tests such as molecular, to reach a diagnosis. However, the rarity in occurrence of such disorders combined with the lack of distinctive traits for a subset of them, makes their diagnosis a challenging task even for experienced medical practitioners.

ANN models such as the ones presented in DeepGestalt [15] and GestaltMatcher [1] comprise a promising step forward in using AI for the task of recognizing rare genetic disorders from facial phenotypes. Such works rely on databases like Face2Gene [20] and GestaltMatcher database (GMDB) [21] which offer a valuable collection of frontal facial images and other medical data of the patients with such rare disorders. The GestaltMatcher [1] model surpasses human expert's performance in the task of recognizing certain syndromes and their sub types. Therefore, deployment of such models in a clinical setting has the potential to significantly improve the speed and accuracy of diagnoses.

The constraints on deploying ANN models for medical diagnoses in a clinical setting, as discussed in the previous section, apply to the genetic disorder classifiers as well. Therefore, there is an obligation for such models to provide bases for their predictions in order to make them dependable. However, none of the existing works on genetic disorder classification from frontal facial images focus on the explainability of their models.

In general, ANN models are capable of learning novel discriminative features or regions from the training data, that are relevant to the task at hand. Associating such learned features or regions with the real world knowledge (with respect to the dataset) can provide new insights about the data and task at hand. In the case of genetic disorder classification, associating attention regions of SOTA models like DeepGestalt [15] and GestaltMatcher [1] in their frontal facial input images can enhance human kind's understanding about facial phenotypes of genetic disorders. Such a study to associate the attention regions of genetic disorder classifiers with the facial phenotypic information known to the medical community, is yet to be conducted.

1.2 Importance

Making a genetic disorder classifier explainable by determining its attention regions, offers a means for its users to check whether the model focuses on features relevant to a given disorder, or something irrelevant like background, for example. In the former case, a clinician could use the model's prediction to reinforce their diagnoses. In the latter, they could simply ignore its decisions. Thus an explainable genetic disorder classifier provides a verifiable second opinion to a clinical geneticist.

The knowledge about facial phenotypes of rare genetic disorders are contained in resources like Human Phenotype Ontology (HPO) [22] and relevant medical literature. However, due to rare occurrences of such disorders, not all variations in their phenotypes are known to the medical community. Analyzing the attention regions of genetic disorder classifier models offers a way to discover facial regions that contain novel phenotypic traits, there by enhancing human kind's understanding of such rare medical conditions.

Datasets like GMDB [21] only contain a fractional number of instances per class, when compared with sizes of general purpose image classification datasets like ImageNet [23]. Most of the disorder classes contains samples in the order of tens and a few in the order of hundreds. This in turn demands use of highly effective data pre-processing and model training techniques to learn the most from a dataset. Analyzing attention regions of a classifier model trained on such datasets, enables an ML practitioner to identify any possible biases that the model could have learned, and consecutively enables him to adopt suitable techniques to remove them. This eventually makes the model more generalized and possibly increases its predictive performance.

1.3 Challenges and Difficulties

This section lists and discusses some of the key challenges associated with this research work.

- **Dataset size and imbalance:** As briefly mentioned in the previous section, datasets like GMDB [1] are small-sized, and also are characterized by problems such as dataset imbalance and low image resolution. Figure **??** depicts the class-wise distribution of samples in GMDB. Such problems are caused by various factors like inherent rarity in occurrence of genetic syndromes, data-privacy constraints and lack of openly available datasets. The above listed issues of genetic syndrome datasets often have consequential effects on the performance and explainability of machine learning models they are trained with.

- **Low predictive performance of the classifier:** Although, SOTA genetic syndrome classifier models such as DeepGestalt [15] and GestaltMatcher [1] surpass human-level performance in diagnosing rare genetic disorders, their predictive perfomances are exceptionally low when compared with that of top classification models trained on large general purpose datasets. Besides, over-fitting is a common problem experienced by these models. In most of the existing works on XAI methods for neural networks, the research community has benchmarked them on high performance models. This raises doubts about the quality of explanations generated by XAI methods, when applied to low performance models.

- **Lack of ground truth explanations:** Evaluating the performance and effectiveness of XAI methods remains a challenge till date. In most of the cases, this is due to the lack of any ground truth and/or metrics to evaluate their explanations. Besides, the working principle of every XAI method is different, with each focusing on explaining a particular aspect of model and its predictions. Due to this reason, often XAI methods are evaluated by subjecting their corresponding artifacts to be assessed by humans. In the case of this work, such an evaluation needs to be performed by clinicians and dysmorphologists who specialize in the diagnosis of rare genetic syndromes. Certain

practical difficulties in conducting such an evaluation like the willingness of clinicians to participate in the process pose a challenge. In addition, new findings and discoveries about phenotypic features and new variants of genetic syndromes change the medical community's understanding of them time to time, questioning the correctness of clinical evaluation conducted at a given point in time.

1.4 Problem Statement

This research work systematically approaches the problem in hand. Firstly, a literature review is conducted to identify SOTA XAI methods, which when applied to the GestaltMatcher model explains the rationale behind its predictions, in the form of post-hoc attention maps. In a classification setup, post-hoc attention heat maps such as the ones in Figure **??**, signal regions in the input image that were relevant for a classifier model to produce a certain class label. A handful of XAI methods are shortlisted based on their advantages and drawbacks, recommendations from the research community, and their suitability to the task at hand.

The selected set of techniques are applied to GestaltMatcher [1], in order to generate explanations associated with the model's predictions for every patient image in the GMDB [1] dataset. The generated explanations are then analyzed with intents to understand the behaviours of both the model and the chosen set of XAI techniques with different input categories. Besides analyzing the model's regions of interest in individual patient images, this work also studies its characteristic attention regions on a class level for specific syndromes. Such characteristic representations are produced by combining patient-wise attention maps of individual classes.

As mentioned in 1.3, dataset imbalance is one of the key challenges in the application of machine learning technqiues to medical data. This can have consequences on the quality of attention maps generated for classes of different sizes. This work conducts experiments to analyze the effects of class imbalance on the quality of explanation artifacts. This is achieved by comparing attention maps, produced from classifier models that are trained with different numbers and choices for the syndrome classes.

Inorder to know the association between GestaltMatcher's attention regions and the medical community's knowledge on facial phenotypic features of genetic syndromes, the model's attention maps need to be evaluated by clinicians. It is a time-consuming process which involves the participation of atleast a few dysmorphologists and sub-specialty trained clinicians. Such an evaluation is not included in this work, due to the time constraints, however, a questionnaire to facilitate the process in future is formulated and presented. Besides, this thesis also suggests and proposes ways to quantitatively evaluate the attention maps.

Finally, a use case scenario is presented, depicting the application of techniques presented in this work to a real-world genetic syndrome diagnosis situation.

Research Questions

Concisely put, this thesis intends to address the following research questions:

RQ1. What are the XAI methods to determine important regions in an input image for a CNN based classifier model to make its predictions?

RQ2. What are the key regions in frontal facial images fed to CNN models trained for the task of classifying rare genetic disorders?

RQ3. How do the regions obtained from findings for RQ2 compare with the knowledge known to the medical community on facial phenotypes of the corresponding rare genetic disorders?

1.5 Structure

This report consists of seven chapters (including this chapter) with each discussing different aspects of the conducted research work. The first chapter introduced the reader to the research topic, and discusses the scope and significance of this work. Chapter 2 gives the necessary background knowledge on diagnosis of rare genetic condiitions using GestaltMatcher [1], which is necessary to appreciate this work. The chapter also briefly introduces the reader to the field of XAI. In Chapter 3, a literature review of SOTA explanation methods considered for this research work is provided. Chapter 4 describes the systematic approach taken to address the problem at hand. The chapter gives the rationale behind the design of experiments and other choices made. The fifth chapter discusses implementation details and provides specifications of datasets and models used for experimentation. The results obtained from the conducted experiments are listed and analyzed in Chapter 6. Besides, it also presents and describes the formulated evaluation questionnaire. Finally, Chapter 7 concludes the report by summarizing the contributions of this project, and also discusses the possible future research directions.

2

Background

Last chapter introduced the reader to this research work by discussing its objectives and formulating them into research questions. This chapter provides the necessary background information for the reader to understand this research work. It begins with a description of GestaltMatcher's working and how it surpasses human performance in diagnosis of rare genetic syndromes from facial images. Subsequently, an introduction to the topic of Explainable Artificial Intelligence (XAI) is given. Information on the genetic syndromes considered for this research work is provided alongside their respective result discussions in Chapter 6 for ease of reading.

2.1 Genetic Syndrome Diagnosis using GestaltMatcher

GestaltMatcher is a state-of-the-art (SOTA) neural network based tool developed to diagnose rare genetic syndromes from frontal facial images of people. It surpasses the ability of clinicians to recognize images of certain rare genetic conditions. GestaltMatcher stands out from the rest of all similar works for its ability to diagnose a patient even if the disorder he/she suffers was not a part of the training set.
In the beginning, we describe the construction of GestaltMatcher. Subsequently, we explain the two applications of the tool. Finally, we provide details about GestaltMatcher's performance on the GestaltMatcherDataBase (GMDB) dataset.

2.1.1 Construction

Under the hood GestaltMatcher consists of a Convolutional Neural Network (CNN) based encoder which acts as a feature extractor to transform inputted images into embeddings in a highly discriminative embedding space. Patient photos from datasets such as GMDB are pre-processed before feeding them into the encoder. The pipeline shown in Figure 2.1 illustrates the sequence of steps involved in processing a patient photo using GestaltMatcher.

Pre-processing Pipeline

The pre-processing pipeline is responsible for detecting facial landmarks and aligning the orientation of a given patient photo, and finally cropping the face from it. Authors of GestaltMatcher use RetinaFace [24]

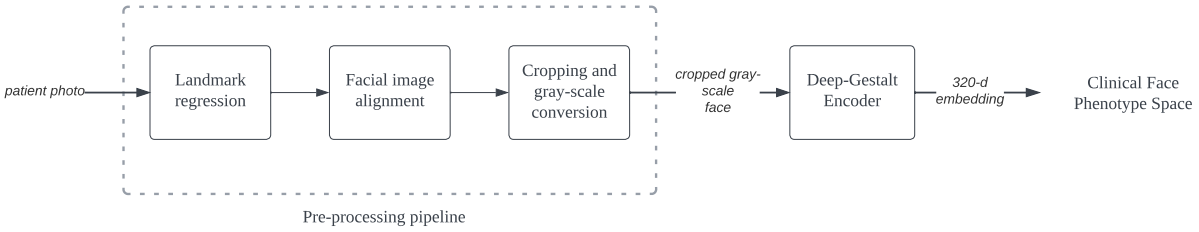


Figure 2.1: Illustration of the pipleline to process a patient photo using GestaltMatcher

to fetch detect landmark points from a patient's facial image. In a nutshell, RetinaFace is a single-stage, multi-scale face localization method which employs multi-task learning to perform five different tasks: face detection, landmark position and score regression, 3D position and pixel correspondence prediction. It uses a Resnet-50 [25] backbone. The facial landmark coordinates outputted by RetinaFace is used to crop and align faces from patient photos. The facial images are resized to 100 x 100, and converted to gray-scale before feeding them into the encoder module.

Deep Gestalt Encoder

The encoder in GestaltMatcher is called as "Deep Gestalt" encoder, named after the work [15] in which its architecture was first proposed. The CNN based encoder consists of ten convolutional layers and uses Rectified Linear Unit (ReLU) activation function. Please refer Figure 5.2 in Chapter 5 an enumeration of the network architecture. Deep Gestalt encoder extracts features from input facial images and represents them in 320-dimensional embedding space called the Clinical Face Phenotype Space (CFPS). The embeddings in CFPS are called as Facial Phenotype Descriptors (FPDs). Authors of GestaltMatcher use the encoder differently for two of its applications: syndrome classification and patient matching.

2.1.2 Applications

An intuitive application of GestaltMatcher is genetic syndrome recognition. In this case, a dense layer which acts as a classifier is appended to the encoder. Top-k predictions within the scope of trained classes can be obtained for any patient's facial image. The GestaltMatcher classifier is useful when a clinical practitioner suspects his patient to have a particular genetic condition and seeks to validate his diagnosis.

The primary application of GestaltMatcher is patient matching. CFPS acts as a discriminative embedding space to measure similarities between FPDs of patients with known or unknown disorders. The cosine similarity metric is used to quantify similarities between cases. The measure can computed using the following formula:

$$\cos(\mathbf{p}, \mathbf{q}) = \frac{\mathbf{pq}}{\|\mathbf{p}\|\|\mathbf{q}\|} = \frac{\sum_{i=1}^{n} \mathbf{p}_i \mathbf{q}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{p}_i)^2} \sqrt{\sum_{i=1}^{n} (\mathbf{q}_i)^2}} \tag{2.1}$$

where p and q represent vectors each of dimensions n. GestaltMatcher helps to perform an objective comparison of patients with same or different disorders, with shared phenotypic features. Authors of GestaltMatcher provide an example for such an application scenario, and describe how the tool was used to match patients from different families, who suffered from the same disorder.

Besides diagnosing known disorders, GestaltMatcher can be used to identify novel genetic conditions. For example, when a syndromologist is unable to find the molecular cause for a patient's phenotype, he/she could use the matching tool, to match the case's FPD with the existing instances in the CFPS. Position of the FPD can be used to determine whether the patient suffers from an unidentified disorder. The CFPS neighborhood of such a case can be used in the identification of genes associated with the unknown disorder.

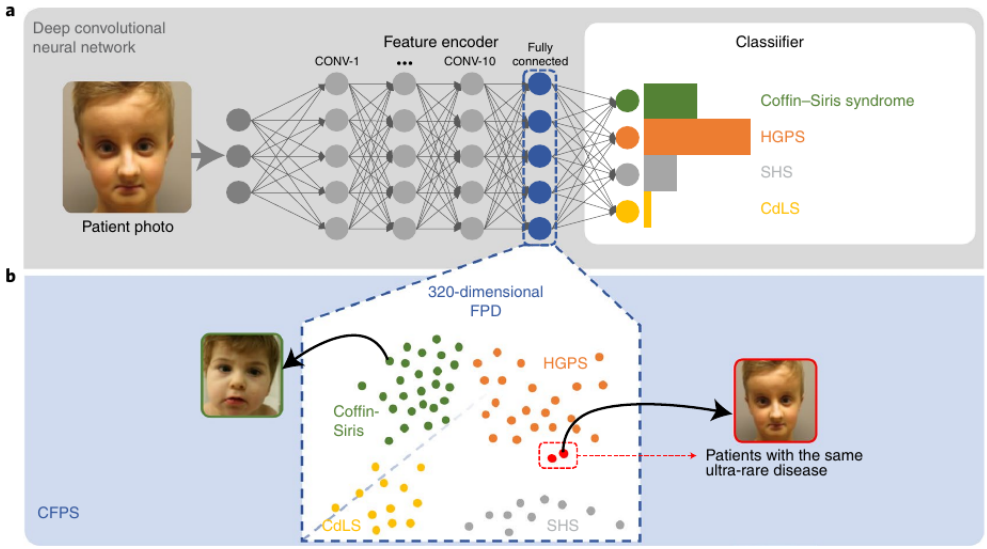


Figure 2.2: Use cases of GestaltMatcher: a. syndrome classification b. patient matching. Image source: [1] (p.3)

2.1.3 Performance

GestaltMatcher achieves state-of-the-art performance in the task of syndrome recognition, measured using the top-k accuracy metric. Its authors report the model's performance on two datasets: Face2Gene, a propreitary one, and GMDB, which consists of images obtained from 902 scientific publications. The model's performance on GMDB is presented in the table below, as the same dataset was used for our research work.

2.2 Introduction to XAI

"XAI is artificial intelligence (AI) in which humans can understand the decisions or predictions made by the AI" [26]. It also refers to a branch of study which deals with the development of processes and

methods to explain decisions of AI systems. The term "interpretability" is often used interchangeably with "explainability". However, it is important to note that interpretability has to do with describing the internals of an AI system, in a way that is understandable to humans. In this section, we give an overview of different types of methods developed to visualize internals of neural network models, especially CNNs such as the one used in GestaltMatcher, and explain their predictions.

Methods for Neural Networks

Figure 2.3 depicts the taxonomy of explanation methods for neural network methods as presented in [27].

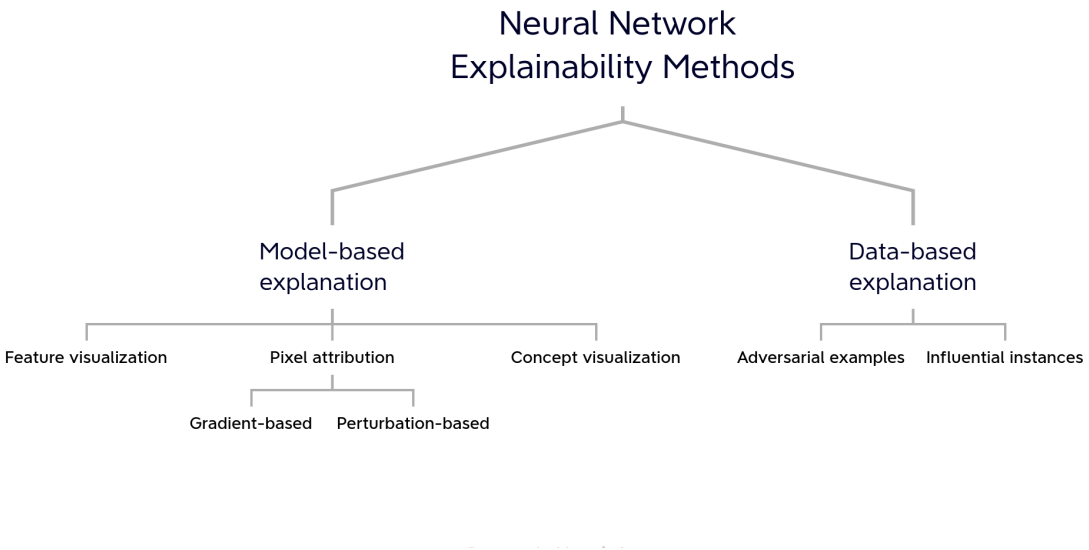


Figure 2.3: A categorization of neural network explanation methods

Feature Visualization

This class of methods focus on visualizing features learned in hidden layers of a CNN. Entities like edges, textures, patterns, parts and objects can be visualized by maximizing activations of convolutional layers. Last layers of neural network model learn features of higher complexity than the earlier ones. Feature visualization can be achieved through optimization, or by using approaches such as network dissection [28], in which highly activated channels of a given layer are associated with human concepts such as color. Although feature visualization seems to be a simple tool for understanding internals of a neural network, often its visualization artifacts are not interpretable. Besides, another challenge lies in choosing the right set of layers and channels for visualization, and summarizing information represented in them.

Attribution Methods

Attribution methods provide the rationale behind a CNN model's output by producing an attribution map representation, in which every pixel or feature of the input image is assigned a score based on how much it impacts a given prediction. Artifacts generated by attribution methods are called with different names like saliency maps, attribution maps and sensitivity maps, based on the approach used to generate them.

Attribution methods can be categorized in two types, based on how they compute attribution scores: perturbation/occlusion-based and gradient-based. Perturbation methods manipulate sections of the input to compute their importance scores. On the other hand, gradient-based methods make use of gradient information obtained by computing derivatives of output score with respect to feature values to calculate attributions. This research work uses attribution methods to understand GestaltMatcher's regions of attention in syndromic faces and to explain its predictions.

Concept Visualizations

Concept based approaches are built with an intent to detect and visualize user-defined concepts such as a patterns or any abstraction in the latent space learned by a neural network model. Testing with Concept Activation Vectors (TCAV) [29] can be taken as an example for a concept visualization method. It quantifies any given concept's influence on a neural network model's prediction for a given class.

Influential Instances

A training data sample is considered "influential" when its deletion from the train set considerably affects the predictions of a model. Identification of influential instances offers a way to debug machine learning models and explain their predictions.

Counterfactual Explanations

A counterfactual explanation of a model prediction explains the least change to the input feature values which changes the prediction to a predefined output. Adversarial examples [] make a great example for counterfactual explanations. However, they are synthesized with an intent to deceive a model rather than explaining its predictions.

3

State of the Art

The previous chapter gave necessary background knowledge for the reader to understand this research work. This chapter contains a literature review of various state-of-the-art feature attribution methods considered for our project. Kokhlikyan *et al.* [30] categorize attribution methods into three groups, based on the entity to which a model's predictions are attributed: input/primary, layer and neuron. The same dimension was used to group the set of methods considered for this work, and inturn organize this chapter.

3.1 Input Attribution Methods

Input or primary attribution methods evaluate contribution of each input pixel to the output of a neural network model. This section discusses six input attribution methods considered for this research work.

3.1.1 Occlusion Sensitivity Maps

Occlusion sensitivity mapping [2] is a model-agnostic perturbation based method, which generates explanations by manipulating parts of the input image. The approach is computationally expensive, O(#simultaneous occlusions * #features * #ablations_per_eval * 1/#strides), and is included in this work to verify if Gestalt Matcher model is focusing on key facial features, or simply using the surrounding context to produce predictions. This is achieved by systematically occluding different portions of the input image with a black square or rectangular mask, and computing the difference in outputs (logit scores of the target class). In this work, we use a black square mask of dimensions 10x10. Important portions of the input when occluded, result in relatively larger logit score differences, than the trivial ones. The differences are plotted on the image, yielding the occlusion sensitivity maps.

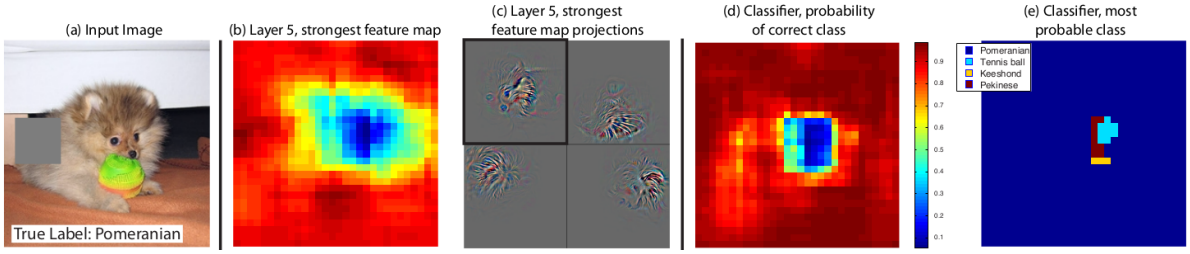


Figure 3.1: An illustration of occlusion sensitivity mapping. Image source: [2]

3.1.2 Deconvolution

Zeiler and Fergus proposed the "Deconvolution" [2] approach to visualize and provide insights into the functions learned by intermediate layers of a CNN. It is one of the earliest attribution techniques, which produces visualizations based on computing gradient of loss function with respect to a given input. The work acts a baseline till date for development and evaluation of new pixel attribution techniques.

The method uses a deconvolution counterpart for every building block of a CNN, to obtain reverse mapping from features to input pixels. The idea of deconvolution was first introduced by Zeiler et al. [31], as a way to perform unsupervised learning. In order to obtain attribution maps using the Deconvolution approach, the first step is to attach each block of convnet with its deconvolution counterpart as shown in the figure 1. Every activation except the ones belonging to the class of interest is set to zero. The activation value is then backpropagated through the deconvolution blocks such as unpooling, rectification and transposed convolution, all the way to the input layer. Deconvolution blocks act according to a pre-defined set of rules. The transposed convolution block performs the inverse of convolution operation by using transposed versions of the same filters. This is equivalent to flipping a given filter both in vertical and horizontal directions. In order to backtrack activations through max-pooling layer (.i.e. using the unpooling layer), indices corresponding to maximum activations in every layer, are first stored during the forward pass and later retrieved during the back propagation phase. However, the use of indices or switches from the forward pass, constrains the visualization on the input image [32].

Authors test their method on an AlexNet [33] trained on the ImageNet [33], Caltech-101 [34] and Caltech-256 [35] and PASCAL2012 [36] datasets. As a first step, they visualize the top 9 feature maps of the each of the first five layers, to show the proportional increase of complexity in features with respect to their receptive fields. The visualizations are obtained by backtracking the strongest activation of a feature map for most of the data samples, all the way until a given input, using the deconvolution rules.The paper also discusses about the proportionality between the time taken for a given layer to learn features and its corresponding depth. Further, it shows that features learned by top layers are more invariant to transformations like translations, rotations and scale changes.

The work evaluates itself by qualitatively comparing its resulting attribution maps with occlusion sensitivity maps. Occlusion sensitivity maps are obtained by systematically occluding portion of an image and analyzing the given classifier's output, to determine the most discriminative regions as shown in the first

image of Figure 3.2.

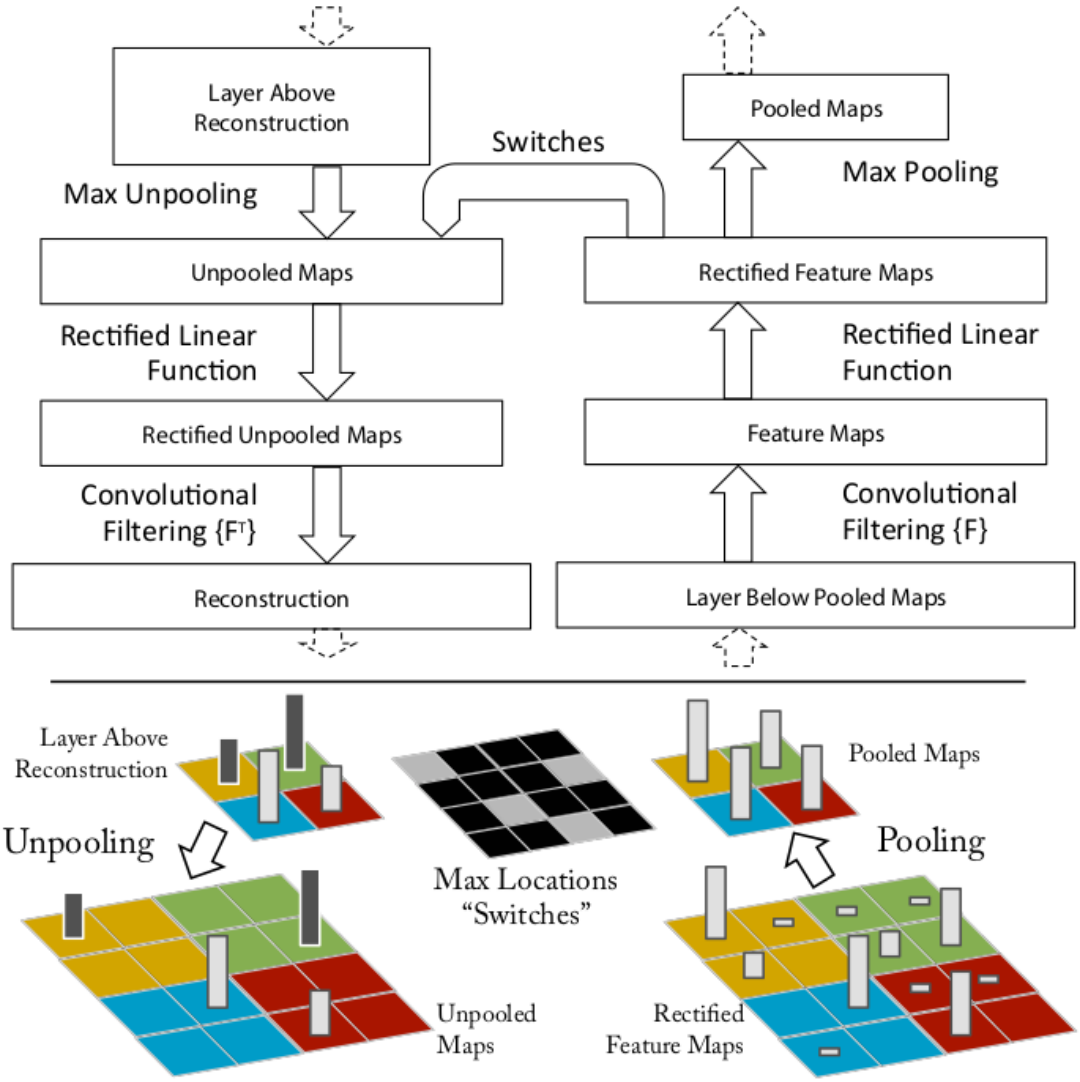


Figure 3.2: Top: A convnet layer (right) attached to its deconv counterpart(left). Bottom: An illustration of the unpooling process. Image source: [2]

3.1.3 Saliency Maps

Simonyan et al. [3] propose two visualization techniques with intents to generate an image which maximizes the class score, and to compute a class-specific saliency map for a given input. The first technique numerically optimizes the input image while the other computes the spatial support of a given class in an input. This work is one of the earliest to leverage saliency maps for the task of weakly-supervised

object segmentation. Authors demonstrate the proposed techniques by applying to a deep convnet trained on the ILSVRC-2013 dataset [37].

Class Model Visualization

The intention of this technique is to numerically generate an image which is representative of the target category with respect to the convolutional net's class scoring model. This is achieved by finding a L2-regularised image such that the logit S_c of a given class c is maximized:

where λ refers to the regularization parameter and I is a local optimum, which can be found with help of back propagation. The optimization process uses the mean image of the data set as the initial value. The work also mentions about the prominence of visualizations produced by using logit scores over soft-max/unnormalized class scores.

Image-Specific Class Saliency

The objective of this technique is to rank pixels of an input image, based on their impact on class scores (S_c). Authors provide a couple of interpretations for the class score values/ logits with respect to which saliency maps are created:

1. A linear approximation of the function learned by neural network in a local neighbourhood of the input image.

2. Higher the saliency associated with a pixel, lesser it needs to be altered in order to increase it respective class's score.

The derivative of class score with respect to input image is found using back propagation as described by the equation below:

In order to obtain the saliency map for a multi-channel image, the maximum magnitude of gradient for a given position across channels is used. Class saliency maps thus produced are used as initial points to compute object segmentation mask using the GraphCut algorithm [38]. Foreground and background portions are considered as Gaussian Mixture Models and the former is estimated from pixels with saliency value higher than the 95% quantile of the image's saliency distribution.On the other hand, the latter is estimated from pixels with saliency smaller than 30% quantile.
The work evaluates its outputs on test split of the data set for the localization task in the ILSVRC-2013 [37] challenge, where it achieves 46.4% top-5 error in spite of its simplicity. In hindsight, apart from the strategy used to reverse the ReLU layer this approach is equivalent to Deconvnet [2].

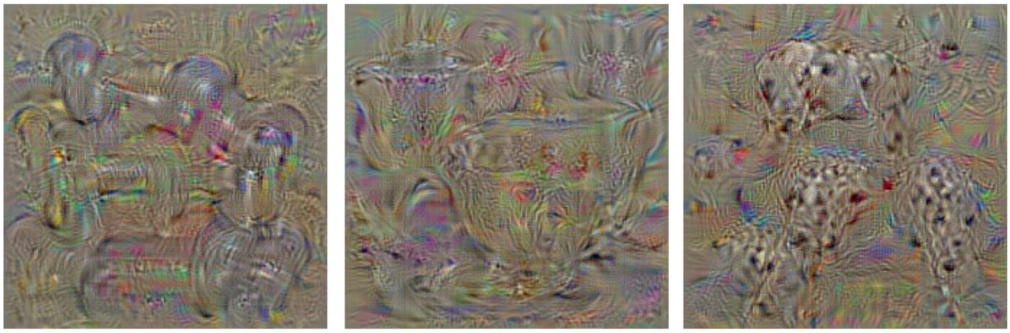


Figure 3.3: An example for class model visualization. Image source: [3]

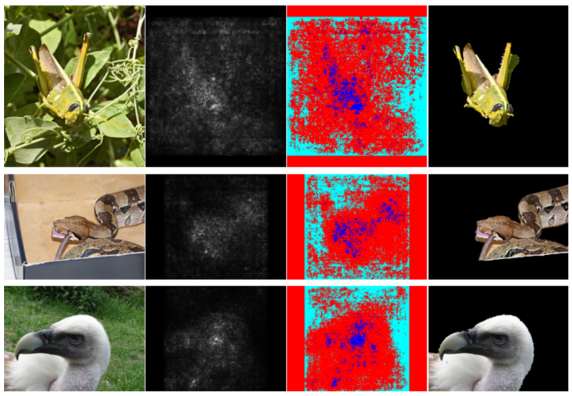


Figure 3.4: Examples of saliency maps. Image source: [3]

3.1.4 Guided Backpropagation

Springenberg et al. proposed a new variant of the deconvolution approach in their work, as a means to analyze their "All Convolutional Net" architecture, which replaces max-pooling layers by convolutional layers with increased stride. The first objective of this work was to empirically prove the equivalence (in terms of predictive performance) between a max-pooling layer and a convolutional layer with an increased stride. This was achieved by evaluating a custom cnn model with max-pooling layers against its convolutional counterpart on three datasets: CIFAR-10, CIFAR-100 and ILSVRC 2012. In all cases, performances of the all convolutional models were on par with their max pooling counterparts. The second objective was to determine the quality of representations learned by the intermediate layers of the all convolutional neural network models. In order to achieve this, authors proposed a visualization approach, which can be considered as a slight modification of the Deconvnet [2] technique.

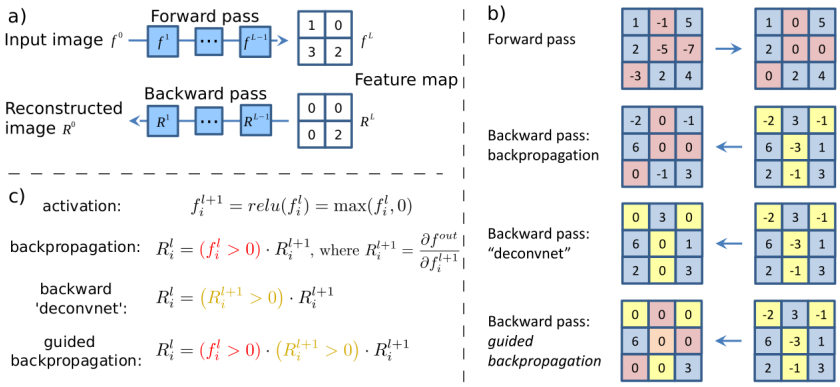


Figure 3.5: A graphical representation of the saturation problem. Image source: [4]

Back propagation through ReLU

One of the most significant difference between Saliency Maps [3], Deconvnet [2] and Guided backprop [32] approaches is the strategy used by these methods to backpropagate gradients through the ReLU layer.

- Saliency maps approach backpropagates gradients of positions with respective to non-negative activations.
- On the other hand, the deconvnet approach allows only positive gradients to flow in reverse direction.
- The guided-backprop approach combines the above mentioned methods and masks out values for which at least one of activation or gradient values is negative. This is performed with an intention to avoid the reverse of negative gradients of neurons which reduce the activation of the target neuronal unit.

The term "guided backpropagation" comes from the use of the additional navigation signal, to selectively back propagate only the positive gradients of the positively activated neuronal units. Though guided backprop was devised to show the learning capability of the all convolutional network architectures, authors show the effectiveness of the technique on the ones with max-pooling units. Guided backprop produced significantly more accurate representation, especially for higher layers, when compared to Deconvnet and Saliency maps.

3.1.5 Deep LIFT

Shrikumar *et al.* [4] propose an attribution technique, which assigns scores to input pixels based on their contribution to change in activation of each neuronal unit with respect to a reference value, which is chosen based on the problem in hand. Authors claim the technique to be computationally inexpensive, and yield meaningful representations in comparison with other methods. Besides, the technique is devised

in a way that it is suitable for neural net variants apart from CNN like Recurrent Neural Networks (RNN). Intuitively, Deep LIFT seeks to explain the deviation in output from some reference output in terms of deviation in input from its respective reference. The motivation to use a reference value, arises from the need to handle the "saturation problem".

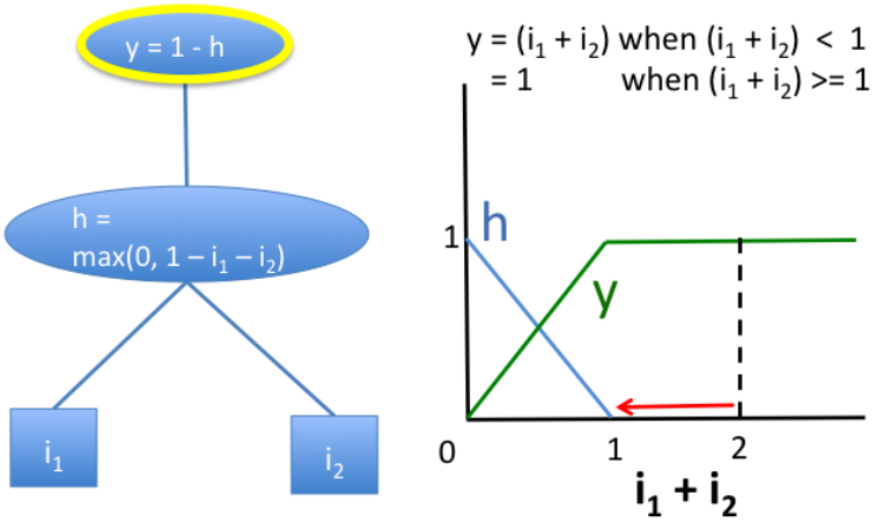


Figure 3.6: A graphical representation of the saturation problem. Image source: [4]

Saturation problem

The saturation problem can be explained intuitively with an example. The figure illustrates a simple neural network whose outputs saturates, when the sum of its inputs (i_1 and i_2) exceeds 1. Application of any perturbation or gradient based attribution method, to this scenario would lead to creation of undesirable artifacts. For the methods of earlier type, perturbing inputs will not cause any changes in the output. On the other hand, gradient based methods will suffer from lack of gradients in the region where $i_1 + i_2 > 1$.

Working

Deep LIFT handles this problem by using a reference value, enabling it to backpropagate the importance signal even in zero and discontinuous gradient situations.
The contributions computed by DeepLIFT is analogous to the idea of partial derivatives, except for the fact that change in input is computed with respect to a finite value (activation difference) in the earlier, in contrast to an infinitesimal value in the latter. The concept of "multipliers" is used to achieve the same. x is the input neuron with a difference from reference Δx, and t is the target neuron with a difference from reference Δt. C is then the contribution of Δx to Δt. Multipliers obey chain rule Along with the

custom chain rule, the paper also defines a set of rules for neurons of every kind of neural network layer, to assign contribution scores their inputs:

- Linear rule for linear layers such as the fully-connected and convolutions

- Rescale rule for non-linear transformations such as ReLU, tanh or sigmoid

- Reveal-cancel rule which enables the measurement of marginal effect of having an input over all possible orderings, similar to Shapely values [39].

The technique also takes in to account that a zero contribution could be due to cancellation of positive and negative contributions from a pair of entities. The application of DeepLIFT also depends on the choice of target neuron's layer, logit or softmax layer in a classifier network, for example. Authors demonstrate the approach by applying it to a CNN trained on MNIST dataset [40] and an RNN trained on simulated genomic data.
The benchmarking on MNIST classifier for various methods was performed on basis of the change in log-odds score, when a selected pixels of a given image belonging to class c0 were erased to convert it to an image of some target class ct. The upper section of figure illustrates the result of masking pixels ordered as the most significant for converting to target classes 3 and 6. As graphically represented in bottom part of Figure 3.8, DeepLIFT outperformed other backpropagation based methods.

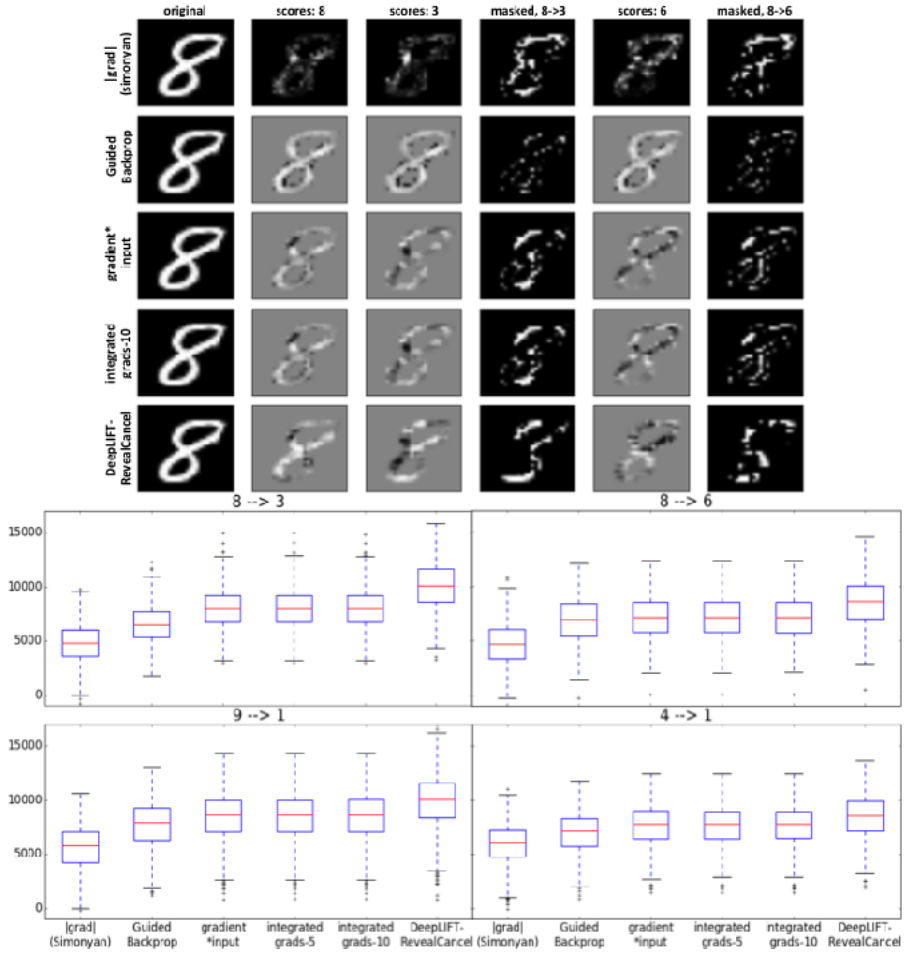


Figure 3.7: A graphical representation of the saturation problem. Image source: [4]

3.1.6 Layer-wise Relevance Propagation

Layer-wise Relevance Propagation (LRP) [41] is a complete path attribution method, whose propagation procedure obeys the conservation principle. This ensures that during back propagation, every neuronal unit receives its share of a given network's output, and continues to equally redistribute it to its predecessor units, until the input layer is reached. Authors of the work claims that the method suffers less from the "Clever Hans" effect, which occurs when a model produces correct predictions based on wrong features. Besides they show the equivalence between their approach and Taylor decomposition of the network output/activations into contribution scores of individual neurons in lower layers. Further more, they stress upon the fact that their method is computationally less expensive when compared to other approaches like occlusion sensitivity maps and Local Interpretable Model-agnostic Explanations (LIME) [42].

Similar to Deep LIFT, LRP operates based on a set of propagation rules, which are framed based on the conservation principle. The rule set defines the logic to propagate relevance scores through different layers,

and also provides parameters that can be set based on the application domain, to produce meaningful attribution maps. Although, there are a few variants of LRP each with its own rule set available, only three basic variants were considered for this work and discussed here.

LRP-0

In general, the relationship between the relevance score Rk which is propagated from layer k, onto a neuron in a subsequent lower layer j as Rj can be expressed as follows:

<eqn1>

where R_j is represented as a linear combination of relevance scores from layer k, weighted by the contribution Z_{jk} of neuron j to k, normalized with respect to sum of all contribution from the layer to R_k.

The contribution Z_{jk} is measured as the activation times its gradient with respect to R_k. Substituting the same in eqn 1 gives the following:

<eqn2>

The rule also satisfies the fact that

<Eqn3>

which means to say that a neuron with a zero activation or zero weight association has zero relevance to the output score.

LRP-ϵ

LRP-ϵ is obtained by slightly modifying the LRP-0 rule. The rule incorporates an extra parameter ϵ in the expression's denominator, which is responsible for absorbing some relevance from Rk, when it is contradictory. With an increase in ϵ, only the most important explanation factors survive, leading to less noisy explanations. Though, a very high value of ϵ can result in sparse attribution maps as depicted in Figure **??**.

<eqn>

LRP-γ

LRP-γ is another enhancement to LRP-0, and it introduces the hyper-parameter γ, which indicates the weightage given to the positive contributions. With an increase in the value of γ, negative contributions begin to disappear. There are other rules of choice to adopt such LRP-$\alpha\beta$ enabling the treatment of positive and negative contributions in an asymmetric manner.

<eqn>

Hyperparameters such as ϵ and γ play a vital role in deciding the meaningfulness of attribution maps and are to be set based on repeated experimentation and domain knowledge. Often a composite version of LRP, using all three rules is recommended to obtain human understandable attributions. Authors recommend to use LRP-0 for the last layers, as the propagation rule closely approximates the underlying neural network function. This is desirable in handling those layers where concepts from different classes are closely entangled. LRP-ϵ is recommended for the middle layers, which are relatively less entangled

when compared to the last ones. They claim that LRP-ϵ effectively filters out spurious variations resulting from weight sharing, and retains the most salient explanation factors. LRP-γ is preferred for the lower layers as this rule spreads the propagated relevance score uniformly to features in an uniform manner.

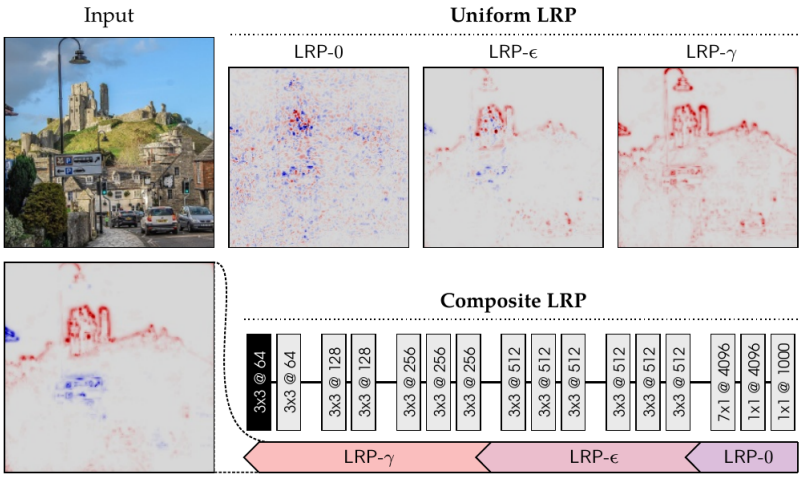


Fig. 10.4. Input image and pixel-wise explanations of the output neuron 'castle' obtained with various LRP procedures. Parameters are $\epsilon = 0.25\,\mathrm{std}$ and $\gamma = 0.25$.

Figure 3.8: A graphical representation of the saturation problem. Image source: [4]

3.2 Layer Attribution Methods

Layer attribution methods evaluate contribution of each neuronal unit in a given convolutional layer to a model's output. This section discusses three SOTA methods of the category, considered for our work.

3.2.1 GradCAM

Selvaraju et al. propose GradCAM [5], a pixel-attribution technique leveraging the gradients of a given target class with respect to a convolution layer. This approach can be considered as a generalization of CAM for CNNs with fully connected layers. Besides, the technique is applicable to neural network architectures for tasks such as image captioning and visual-question answering. The work also comes up with a means to convert class-agnostic fine grained visualization approaches like Guided-Backprop and Deconvolution to be class-discriminative in nature.
Grad-CAM is one of the most widely used approaches in obtaining attribution map based explanations, for neural network models used with medical data. Concisely put, Grad-CAM creates a visualization of which parts of an input image a convolutional layer "looks" for a certain class prediction. The working of Grad-CAM can be described in the following steps:

1. Perform a forward pass of the input image through CNN to obtain logit scores for all classes.

2. Except for the logit activation of the class of interest, set other activations to zero.

3. Backpropagate the gradient of the class of interest all the way to the chosen convolutional layer containing k feature maps Ak. $\partial y_c / \partial A_k$

4. Global pooling: For every feature map, weigh their pixels based on the gradient value, and obtain their weighted mean value scaled by constant Z where y_c refers to logit score for class c, A^k refers to the k^{th} activation map of dimensions ixj and α refers to the computed weightage.

5. Obtain the α weighted average of all pooled feature maps and apply a ReLU to filter out negative values, if any present. where Lc refers to the localization map produced by the Grad-CAM for the class of interest c and Ak refers to the kth feature map.

6. An element-wise multiplication operation is applied on the scaled version (input image dimensions) of Grad-CAM's maps, and outputs of fine-grained visualization approaches like Guided-Backpropagation and Deconvolution, to obtain fine-grained yet class discriminative visualizations. These combinations are termed as Guided-GradCAM and Guided-Deconvolution approaches respectively. Guided Grad-CAM visualizations are both high-resolution and class-discriminative.

Typically, last convolutional layers of a neural network model are chosen for Grad-CAM as they contain high-level semantics and detailed spatial information [5]. This is because the attribution maps produced by the methods becomes progressively worse qualitatively, as we use earlier convolutional layers which have relatively lesser receptive fields. In a classification network, logit scores of the target class are used for gradient computations. However, any differentiable activation value can be backpropagated. The embedding based Grad-CAM discussed in the section, draws motivation from this fact. In the original implementation of the attribution method, a Rectified Linear Unit (ReLU) function is applied on heat maps, to obtain only regions that positively affect the given prediction. However, to get better insights into prediction decisions, this work uses unfiltered heat maps, which consists of negatively correlated regions as well.
Authors evaluate the method on three different tasks: weakly-supervised localization, weakly-supervised segmentation and pointing game experiment [43]. The approach is also evaluated and benchmarked on its ability to generate discriminative, trust-worthy, faithful and interpretable attribution maps. A couple of neural network models with different architectures (AlexNet [33] and VGG-16) are used for evaluation, to determine the method's performance consistency across architectures. Finally, the work also demonstrates the association of a given concept with a neuron, similar to the ones presented in [2] and [44].
Authors also demonstrate it use in analyzing failure modes in neural network models and identifying bias in dataset. The approach is considered to be computationally in-expensive when compared to perturbation based methods such as LIME or SHapley Additive exPlanations (SHAP) [45], yet producing interpretable and discriminative attribution maps.

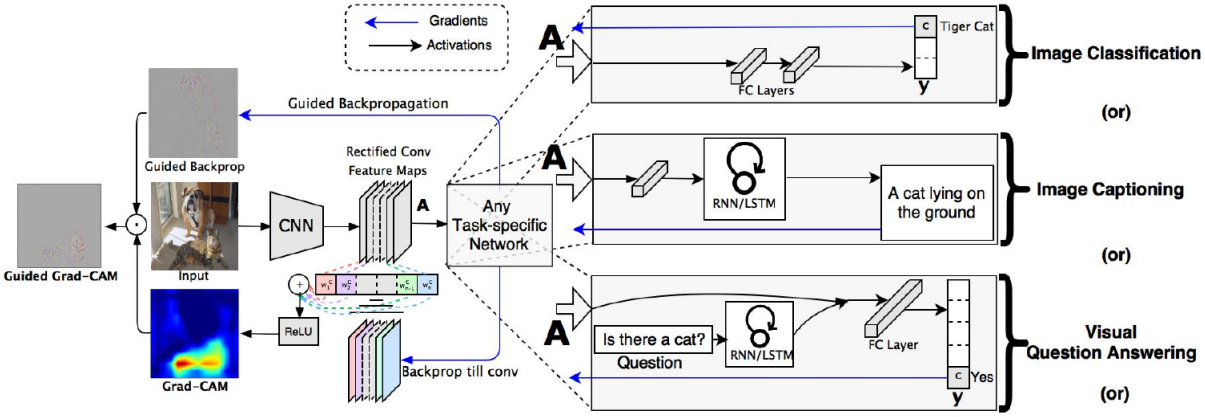


Figure 3.9: An illustration depicting application of GradCAM for different tasks. Image source: [5]

3.2.2 HiResolution Class Activation Mapping (HiResCAM)

Draelos et. al proposed HiResCAM [46] as a generalization of CAM, and the method aims to use feature maps of a layer directly for visualization without averaging unlike GradCAM. The authors demonstrate why the gradient averaging step in attribution methods like GradCAM limit them from being reliable to highlight a neural network model's regions of attention in an image. Besides, they mathematically show the conceptual relationship between HiResCAM and other CAM approaches. The faithfulness of explanations produced by HiResCAM is shown, by conducting certain experiments on natural images whose results were verified using crowd-sourced assessment.

HiResCAM can be seen as a modification of Grad-CAM, as it is primarily designed to address the later's limitation caused by its averaging step. As described in 3.2.1 and illustrated in the figure below, importance weight α^f for a feature map f, is computed by computing the mean of gradients over spatial dimensions. However, the process of averaging may cause the final attribution map not highlight locations within the image which the model used to make predictions.

The above figure illustrates the working of GradCAM and HiResCAM methods. Let s_m be the logit score of a CNN model for a given image input, with respect to class m. GradCAM computes the derivative $\partial s_m / \partial A$, which yields the gradient for every position of a feature map. The gradients are then averaged to yield ${\alpha_m}^f$ for every map. Following which, the gradient values are multiplied with feature activation values ($\alpha_m^f A^f$). While HiResCAM also computes $\partial s_m / \partial A$, they are multiplied directly with feature map activation values leaving out the averaging procedure. Feature map values across channels are then combined as sum to yield the chosen layer's attribution map. By skipping the averaging procedure, HiResCAM preserves the effect of the gradients across every feature map.

Similar to GradCAM, authors recommend using the last convolutional layers for visualization. Apart from proposing the method, the authors also argue that explaining a model doesn't correspond to weakly-supervised segmentation, as the objective of the former to reveal the working of the model while the latter is to localize an object of interest. The work reiterates the fact that every explanation method

aims to describe different aspects of model. They also suggest the usage of metrics like IOU to evaluate the localization capability of neural network model than leveraging them to evaluate the correctness of an explanation method. HiResCAM aims to produce faithful explanations even if a model's regions of interest lie outside an object of interest.
The work compares its method with GradCAM by conducting experiments on PASCAL VOC 2012 [36] and a custom made dataset of 20 classes with their ground-truth segmentation maps. Results of the benchmarking experiments reveal the following:

- HiResCAM better reflects computations of the model than GradCAM
- Humans perceive explanations produced by both the methods differently

3.2.3 Full-Grad

Full-grad [47] is a gradient-based attribution method, which produces saliency map explanations by providing attributions to both the input image and the neuronal units of intermediate layers of a given neural network. The method is layer-agnostic and produces a single attribution map for an input image passed through a network by model, considering different scales of features, starting from pixel-level to high-level features. This makes the method to produce sharper saliency maps than other techniques.
Throughout the work, the authors emphasize the importance of a saliency map based explainability method to satisfy two key properties:

1. Completeness: The ability of a saliency map $S(x)$ representation for an input x, to fully encode the computation performed by a function $f(x)$, such that it can be recovered using $S(x)$ and x.
2. Weak dependence: In a piece-wise linear model, $S(x)$ is dependent only on local neighborhood of x in the data space.

Authors claim that other saliency mapping methods are able to satisfy only either of these properties. This inability arises from the exclusion of gradient contributions from the bias components of a neural network model. Besides, they also stress on the need for a saliency mapping method to attribute importance to individual (local attribution) and regions of pixels (global attribution) simultaneously. The novelty of this work lies in its ability to simultaneously satisfy both the completeness and weak dependence properties, by considering gradients associated with bias components for producing saliency maps.
The following equation expresses the idea that a ReLU neural network with bias b can be approximated as the sum of input gradients and bias gradients. The bias term b here considers both explicit bias component and implicit bias such as running mean of batch normalization layers.

Figure shows the ability of the Full-grad method to produce meaningful saliency maps by accumulating both bias and input gradients across all hidden layers. The work also provides the rationale behind its ability to be sensitive to saturation scenarios like zero input attribution, and to produce in corrrect input-output mappings in the case of parameter randomization. The method first obtains spatial maps (R^D) for every convolutional filter called neuron-wise maps. Such maps are then aggregated to form

layer-wise maps. The layer-wise maps are further processed by rescaling supplemented with an absolute value operation. The following equation mathematically expresses the process:

The authors evaluate the effectiveness of the method by conducting two quantitative experiments:

1. Pixel perturbation: Replace the least salient input pixels for classification with black pixels, and observe the output variation. The lower the variation in salient regions of the attribution maps, the better is a method.

2. RemOve And Retrain (ROAR) experiment [48]: In this experiment, the top-k pixels highlighted by an attribution method for an entire data set are removed, and the considered classifier is retrained on this modified data set. The attribution method corresponding to the least accurate classifier model is considered to be the best, as it indicates that the method correctly identified the salient pixels.

The work reports that the FullGrad method outperformed the other considered methods (Input-gradients [4], Integrated-gradients [49], Smooth Grad [50] and GradCAM [5]) in both the experiments. Apart from the quantitative evaluation, the authors also conducted a visual inspection in which they found the method to produce meaningful attribution maps which highlight both a given object's boundaries and interiors clearly. Figure

4

Methodology

Last chapter gave a literature review of different attribution methods considered for our research work. This chapter explains the methodical approach taken to address the research problem at hand, by discussing the design of experiments and rationale behind different choices made to conduct them.

4.1 Datasets

This research work uses two datasets for conducting experiments. Frontal facial images of patients with rare genetic syndrome are taken from the GestaltMatcher DataBase (GMDB) dataset. The dataset contains 12849 facial images from 7640 patients with 742 different syndromes. Besides patient images, the dataset also contains other details such as patient history. GMDB acts a valuable resource for the research community.
The UTKFace [51] is used for the dataset imbalance - explanation quality experiment 4.3 that requires faces from a healthy cohort. The dataset has over 20,000 face images with labels for age, gender and ethnicity. For the experiment, we handpick 244 samples from the dataset, while ensuring the diversity of the dataset.

4.2 Selection of Methods

In this section, we analyze the neural network explainability methods dicussed in Chapter 2 based on the following dimensions and identify the ones to be considered for further experimentation and evaluation:

- **Saturation problem:** As briefly described in 3.1.5, saturation problem occurs when the output of a neural network model gets saturated and its gradients with respect to inputs become insensitive, thereby affecting the quality of generated attribution maps. Shrikumar et al. [4] reported that occlusion sensitivity maps, layer-wise relevance propagation, saliency maps, guided-backpropation and guided-GradCAM methods suffer from this problem. Besides, they also report the failure of deconvolution approach in the presence of discontinuos input gradients.

- **Sensitivity to changes in model and data:** Adebayo et al. [52] investigated faithfulness of explanations produced by popular attribution methods. The chosen methods were evaluated based on their sensitivity to randomization induced in the considered neural network model's weights and labels of the training instances. Authors report that guided-backpropagation and guided-GradCAM

remained insensitive to the changes in model and data, having functioned merely as edge detectors. GradCAM passed this sanity check.

- **Robustness of explanations:** An interpretability method is robust when it generate similar explanations for similar inputs. David Alvarez et al. [53] and Yi-han Sheu [54] reported that DeepLIFT experiences robustness issues and produces inaccurate results in the presence of multiplicative interactions between features. Besides the lack of robustness, the success of DeepLIFT depends on factors such the choice of reference/baseline image, which is difficult to determine.

- **Relevance to the problem at hand:** Along with the above list dimensions, this work shortlists the methods based on their ability to produce meaningful explanations in the context of genetic syndrome recognition. Figures 4.1 and 4.2 shows few sample attribution maps generated from applying both the considered sets of input and layer attribution maps to the GestaltMatcher model. It can be observed that contours of regions in attribution maps produced by layer attribution methods closely match the scales of phenotypic features of genetic syndromes and parts of human face, in general. This effect is possibly due to the inherent working principles of the two classes of attribution methods.
  Layer attribution maps are obtained by scaling and superimposing feature map activations of the last convolutional layers of a CNN, which are responsible for extracting high-level features from a given input image. This nature makes them more suitable for the problem at hand than input attribution methods, which represent pixel-wise attribution methods.

4.2.1 More Reasons to Consider Layer Attribution Methods

In addition to the above discussed reasons, it is also observed that layer attribution approaches like GradCAM are widely used to explain neural network models for medical diagnosis [] [] []. However, the downside of GradCAM is it approaches attribution as weakly supervised segmenation problem and "sometimes higlight locations the model didnot actually use" [46]. HiResCAM [46], a recent successor of GradCAM overcomes this issue and produces more faithfulness explanations. Therefore, the method is included in the scope of this work.
GradCAM and HiResCAM are layer-specific and therefore their explanations are limited by the choice of the convolutional layers made. FullGrad [47] overcomes this deficit by considering attributions across all neuronal units of a model and thus becomes a candidate for experimentation. Along with these CAM techniques, occlusion sensitivity mapping is included as a reference explanation method, to better understand the classifier model's regions of attention.

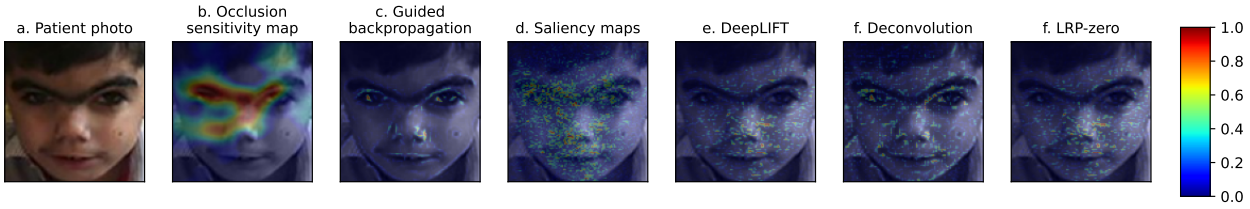


Figure 4.1: An example showing input attribution maps of discussed methods, generated for a patient image

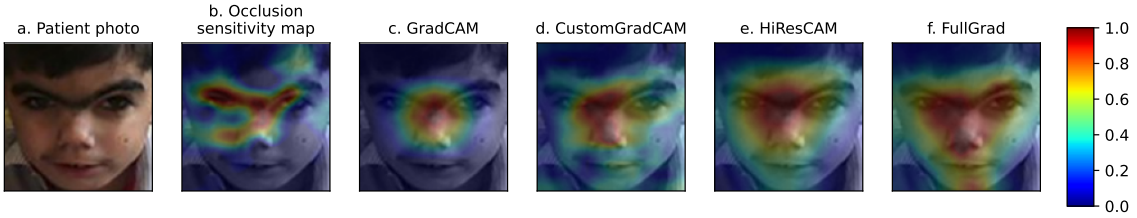


Figure 4.2: An example for layer attribution maps of discussed methods, generated for a patient image

4.3 Design of Experiments

Figure 4.3 illustrates the methodical approach taken to realize the objectives of this research work. The previous chapter and Section 4.2 discussed the XAI methods considered for this work and explained the rationale to select three of them for further experimentation. The next step is to integrate the selected set of methods with the GestaltMatcher model, on which the planned experiments are conducted.

A. Patient-wise Attribution Maps Generation

This experiment focuses on one of our research objectives, to use XAI methods to generate explanations for individual predictions produced by the GestaltMatcher model. Patient-wise attribution maps help us in knowing GestaltMatcher's attention regions in every patient face from GMDB dataset. In turn, this helps us to study the patterns (if there is any) behind the model's varying regions of interests for different categories of inputs.

B. Composite Face Generation

Experiments B and C are conducted with an intent to obtain a characteristic representation of GestaltMatcher's regions of interest in a syndrome(class)-wise manner. Composite face representation of a given class is obtained by applying certain image transformations and averaging patient faces belonging to it (see **??** Chapter 5 for the detailed description). Thus generated composite faces are then processed by the XAI integrated GestaltMatcher model to produce its corresponding attribution maps using all three selected methods.

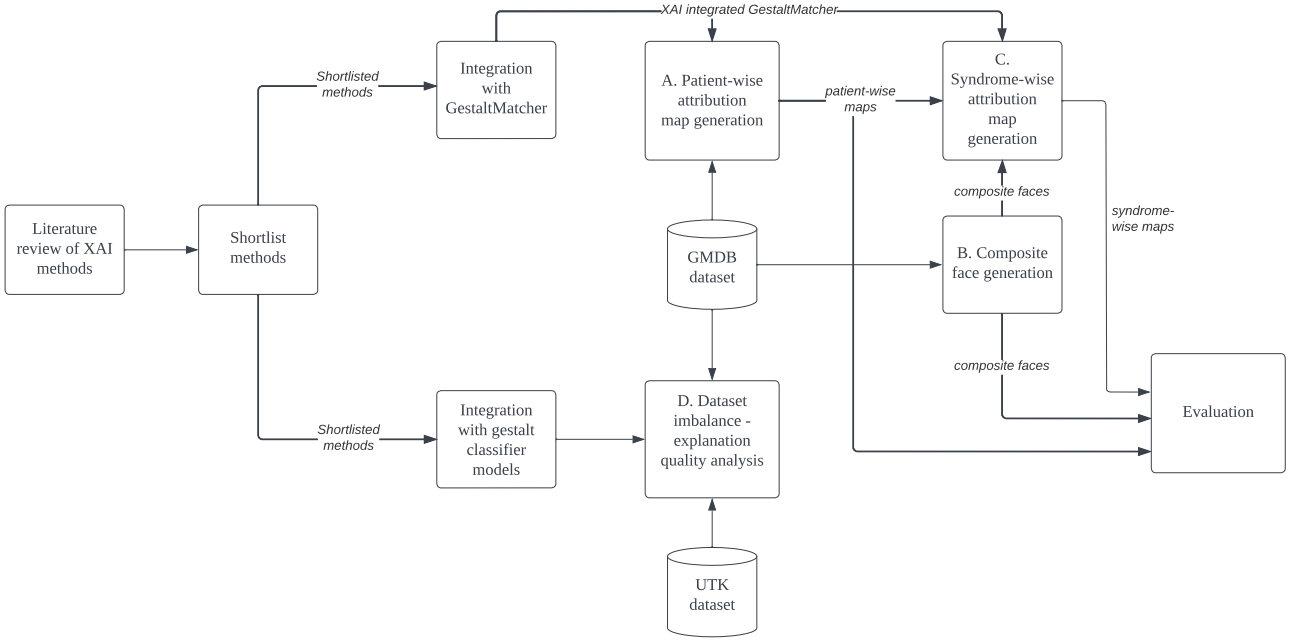


Figure 4.3: Design of experiments

C. Syndrome-wise Attribution Maps Generation

Although, attribution maps of composite faces are generated with an intent to give a generic representation, they suffer from a couple of drawbacks. The process of applying image transformations to facial images of patients to generate composite faces, may alter the key facial abnormalities and dysmorphic features associated with corresponding genetic syndromes. Besides, there is a possibility of the GestaltMatcher model to misclassify them, raising concerns about the correctness of generated attribution maps.

In order to overcome these challenges, we propose and deploy two other techniques to produce syndrome-wise attribution maps: averaging and Singular Value Decomposition (SVD). These techniques are applied to only attribution maps associated with correctly predicted inputs. However, syndrome-wise attribution maps generated using all the three techniques are considered for evaluation, provided the composite face of a given syndrome is correctly predicted.

D. Dataset Imbalance - Explanation Quality Analysis

As discussed in Section 1.3, GMDB dataset is imbalanced. We hypothesize that the problem could affect the quality of attribution maps generated from the GestaltMatcher model. This experiment verifies the hypothesis.

We train three different classifier models: two trained with balanced sets of classes, and one with an

imbalanced set. We observe and analyze the changes in attribution maps produced from these models. Besides this investigation, we study the variations in GestaltMatcher's regions of attention in syndromic and healthy faces. This is realized by analyzing attribution maps produced from a binary classifier, which is trained to differentiate faces of syndromic and healthy individuals. Facial images of healthy people are obtained from the UTKFace dataset.

4.4 Evaluation

Evaluating XAI methods based on their explanation correctness remains a challenge till date. Yeh *et al.* [55] classifies explanation evaluation measures into two categories: objective or computational measures and subjective measures. Objective measures evaluate explanations based on whether a given XAI method satisfies certain properties or axioms, such as completeness [49]. On the other hand, subjective measures focus on aspects such as explanation usefulness, and the evaluation procedure involves humans. This research work uses subjective measures to evaluate the explanation artifacts produced by the chosen set of XAI methods.
In the beginning, this section briefly explains the two predominantly used objective measures to evaluate saliency mapping methods: "infidelity and sensitivity" [55], to give the reader an idea about what computational metrics typically look for. Subsequently, the rationale behind using subjective measures over objective ones for this work is provided. Finally, the proposed evaluation procedure is described in detail.

4.4.1 Objective Measures

Yeh *et al.* [55] proposed infidelity and sensitivity metrics in their work, to benchmark various saliency mapping methods. These measures are computed based on the changes observed in model explanations, when its inputs are subject to perturbations.

Infidelity

$$content... \tag{4.1}$$

Infidelity computes the expected mean squared difference between an explanation scaled by the magnitude of an input perturbation, and differences between the predictor function of the given model at its input and perturbed input. Intuitively, this refers to the correctness of an XAI method in generating true explanations for model predictions.

Sensitivity

$$content... \tag{4.2}$$

Sensitivity can explained as the change in model explanation resulting from a small perturbation in the input.

4.4.2 Rationale to Adopt Subjective Evaluation

Although, computational measures such as the ones presented offer a quantitative means to benchmark XAI methods, they cannot be used for evaluating them on abstract objectives such as meaningfulness of explanations. Such an evaluation can only be performed in subjective manner, by end users of the AI system.
Mohseni *et al.* [56] list various possible goals with which XAI methods can be used, and also recommend measures to evaluate them, based on the context of their application. The goals listed in their work are as follows: algorithmic transparency, user trust and reliance, bias mitigation, privacy awareness, model visualization and inspection, model tuning and selection, model interpretability and debugging.
The primary goals of this research work can be related to three of the above mentioned: "algorithmic transparency", "user-trust and reliance" and "model interpretability". For such cases, authors recommend to evaluate XAI methods by conduct human-subject studies using tools such as Likert-scale questionnaires [57]. This research work follows the recommendation, and formulates a questionnaire for clinicians (end-users) to evaluate the generated explanations.

4.4.3 Proposed Evaluation Procedure

The three types of artifacts (patient-wise map, composite face and syndrome-wise map) generated from the above mentioned experiments are meant to be evaluated by clinicians, who are familiar with the syndromes in our scope. This research work formulates evaluation procedures for each artifact. Clinical practitioners were consulted to develop the procedures. The proposed evaluation strategies are realized in the form of a questionnaire, which is presented in **??** Chapter 5.

i. Patient-wise Attribution Maps

Steps to evaluate patient-wise maps are presented in Figure 4.4 as a flowchart. Besides illustrating the process flow, the chart enumerates possible responses and what can be inferred from them. Scenarios B and C are desirable, as they indicate the usefulness of attribution maps. The final step ranks attribution maps generated by the four chosen methods based on their usefulness.

It is important to note that this evaluation procedure is performed on attribution maps of correctly predicted test samples from GMDB dataset. The idea to evaluate attribution maps of test instances is based on the recommendation from Molnar in his book titled "Interpretable Machine Learning" [27].

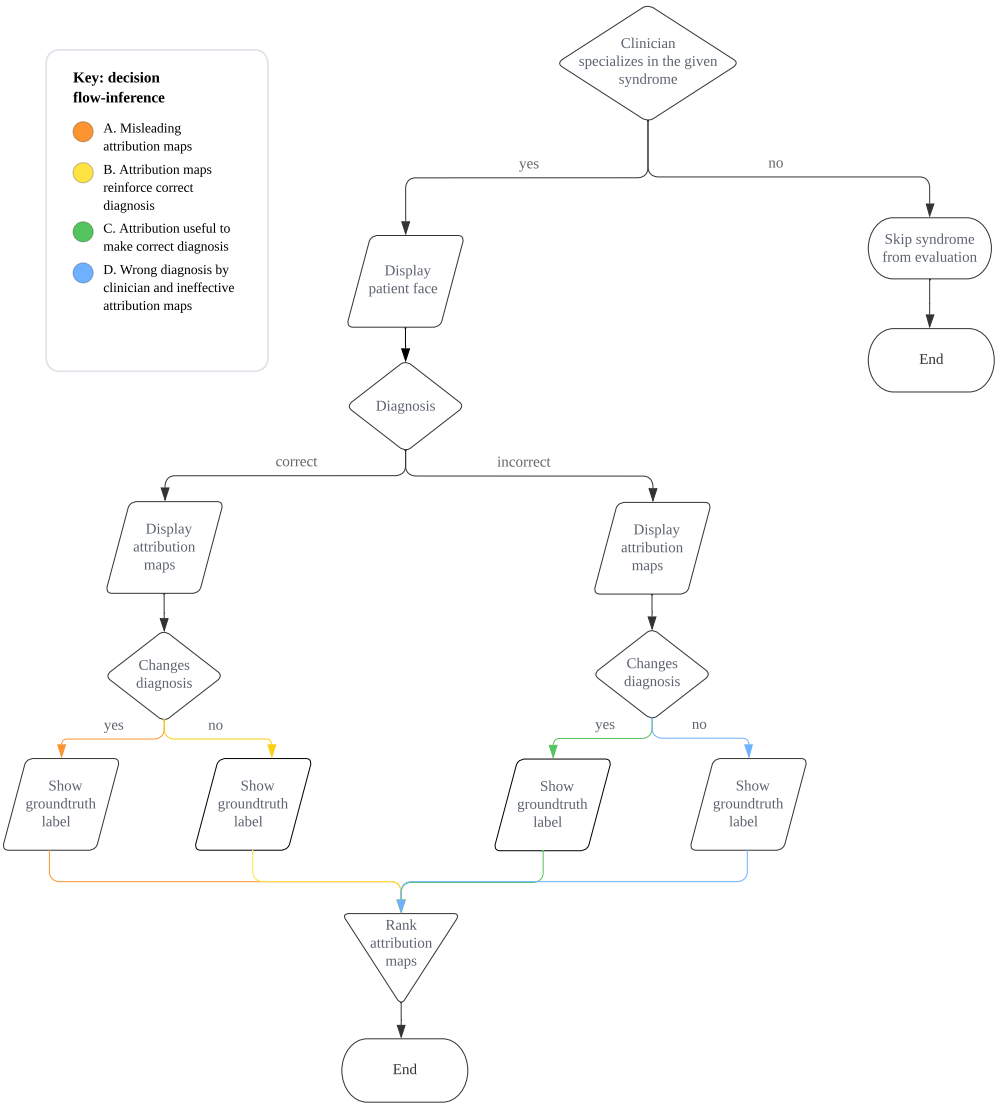


Figure 4.4: Proposed procedure to evaluate patient-wise attribution maps

ii. Composite Faces

The implicit goal of evaluating composite faces is to know if they contain all or most of characteristic facial features of a genetic syndrome. Unlike the case of individual patient images, ground truth labels associated with composite faces are not known before hand, as they are synthesized using facial features from multiple images of a class. Therefore, at first it is important to know whether a given composite face atleast resembles a patient face of a particular syndrome, before looking for its prominent features. The same idea is illustrated as a flowchart in Figure 4.6.

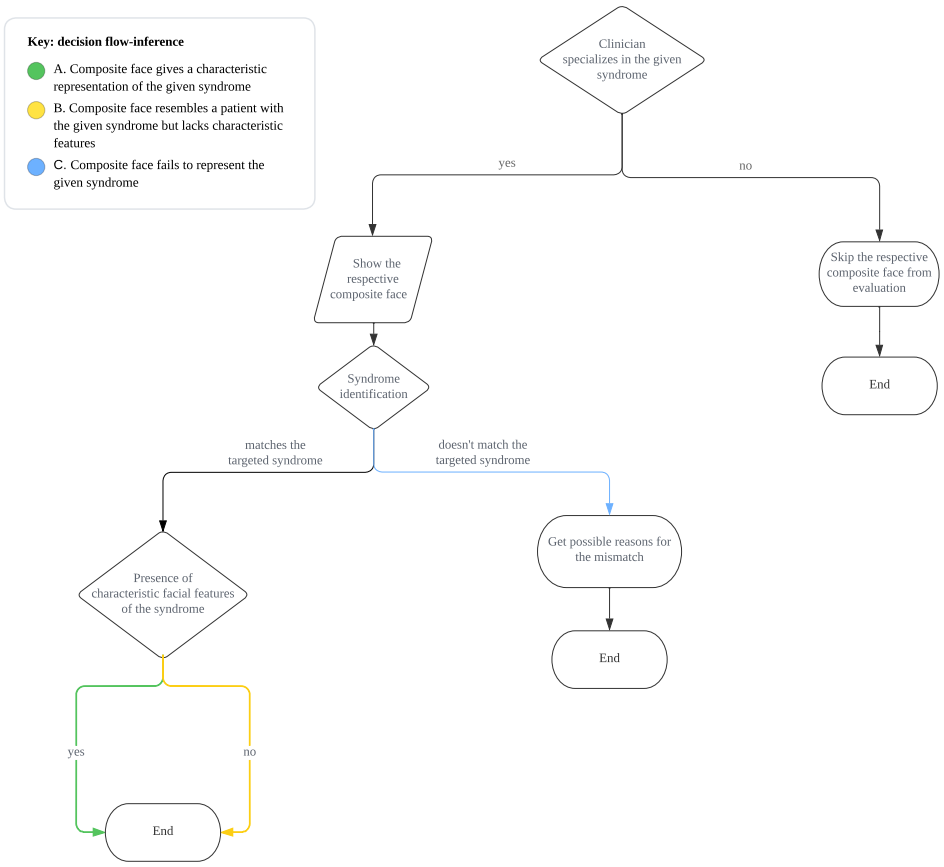


Figure 4.5: Proposed procedure to evaluate composite faces

iii. Syndrome-wise Attribution Maps

Syndrome-wise attribution maps are evaluated in a single step process. The target syndrome is revealed, and the evaluator is asked to pick top three attribution maps, based on how well they highlight the facial regions containing all or most of the disorder's features.

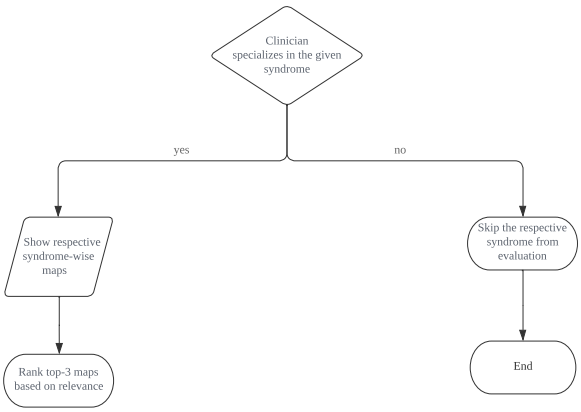


Figure 4.6: Proposed procedure to evaluate composite faces

5

Implementation

This chapter describes gives implementation details of the experiments conducted during this thesis work. In the beginning training details of the GestaltMatcher model are provided. Subsequently, we describe the procedure to integrate explanation methods with the model. Furthermore, implementation details of the experiments proposed in the previous chapter are discussed. Finally, the evaluation questionnaire is presented.

5.1 GestaltMatcher Model Training

The architecture and working of GestaltMatcher was explained in Chapter 2. This work uses the classifier variant of GestaltMatcher for its experiments (refer Figure 5.2 for architectural details). The neural network is first trained on facial images from the CASIA-WebFace [58] dataset. Subsequently, it is trained on 139 frequent-syndrome classes listed by the authors of GestaltMatcher. The training details are presented in Table 5.1.

Hyperparameter	**Value**
Loss function	weighted cross entropy loss
Input image size	100 x 100
Batch size	280
Epochs	200
Learning rate	0.05
Momentum	0.9

Table 5.1: GestaltMatcher classifier training details

5.2 Explanation Methods - GestaltMatcher Integration

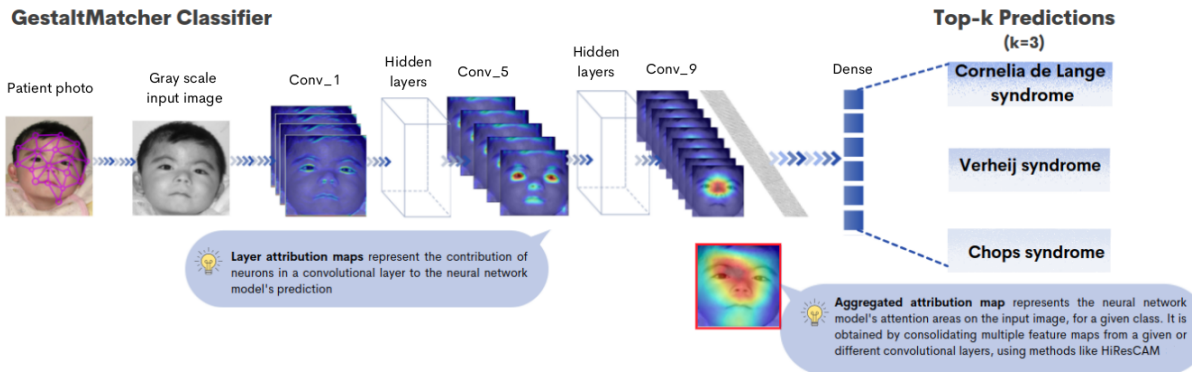


Figure 5.1: An illustration of attribution map computation from the GestaltMatcher model

The Class Activation Mapping (CAM) methods used for this work typically generate class-specific attribution maps, by evaluating contributions of neurons in a given convolutional layer to the model output. Therefore, the choice of target convolutional layer plays a vital role. Selvaraju *et al.* [5] recommend to use the last convolutional layers, as they contain more high-level semantic and spatial information, when compared to earlier layers, which have smaller receptive fields. The same is reinforced by Muhammad *et al.* [59]. We validated this recommendation by computing and comparing attribution maps generated from all ten convolutional layers, using both GradCAM and HiResCAM techniques. FullGrad was excluded from this analysis, as the method is layer-agnostic in nature.

Based on the recommendation and our observations, we chose layer 14 (refer conv_9 in Figure 5.3), the last convolutional layer to compute attribution maps.

CustomGradCAM

Apart from the last convolution layer, we also observed that applying GradCAM to layers 10.a and 11.a produced visualizations in which certain key features of human face are highlighted (refer conv_6 and conv_7 in Figure 5.3). Therefore, in order to include the layers in our scope, we averaged their attribution maps along with the ones produced by layer 14. The resulting map is presented under the name of "CustomGradCAM". We used the Application Programming Interfaces (APIs) offered by PyTorch [60], Captum [30] and an open-source library [61] to integrate the explanation methods to the GestaltMatcher classifier model.

Layer (type)	Input Shape [channels, rows, columns]	Output Shape [channels, rows, columns]	Kernel Shape [count, channels, rows, columns]
1. Conv_Norm_Act	[1, 100, 100]	[32, 100, 100]	--
a. Conv2d	[1, 100, 100]	[32, 100, 100]	[1, 32, 3, 3]
b. BatchNorm2d	[32, 100, 100]	[32, 100, 100]	[32]
c. ReLU	[32, 100, 100]	[32, 100, 100]	--
2. Conv_Norm_Act	[32, 100, 100]	[64, 100, 100]	--
a. Conv2d	[32, 100, 100]	[64, 100, 100]	[32, 64, 3, 3]
b. BatchNorm2d	[64, 100, 100]	[64, 100, 100]	[64]
c. ReLU	[64, 100, 100]	[64, 100, 100]	--
3. MaxPool2d	[64, 100, 100]	[64, 50, 50]	--
4. Conv_Norm_Act	[64, 50, 50]	[64, 50, 50]	--
a. Conv2d	[64, 50, 50]	[64, 50, 50]	[64, 64, 3, 3]
b. BatchNorm2d	[64, 50, 50]	[64, 50, 50]	[64]
c. ReLU	[64, 50, 50]	[64, 50, 50]	--
5. Conv_Norm_Act	[64, 50, 50]	[128, 50, 50]	--
a. Conv2d	[64, 50, 50]	[128, 50, 50]	[64, 128, 3, 3]
b. BatchNorm2d	[128, 50, 50]	[128, 50, 50]	[128]
c. ReLU	[128, 50, 50]	[128, 50, 50]	--
6. MaxPool2d	[128, 50, 50]	[128, 25, 25]	--
7. Conv_Norm_Act	[128, 25, 25]	[96, 25, 25]	--
a. Conv2d	[128, 25, 25]	[96, 25, 25]	[128, 96, 3, 3]
b. BatchNorm2d	[96, 25, 25]	[96, 25, 25]	[96]
c. ReLU	[96, 25, 25]	[96, 25, 25]	--
8. Conv_Norm_Act	[96, 25, 25]	[192, 25, 25]	--
a. Conv2d	[96, 25, 25]	[192, 25, 25]	[96, 192, 3, 3]
b. BatchNorm2d	[192, 25, 25]	[192, 25, 25]	[192]
c. ReLU	[192, 25, 25]	[192, 25, 25]	--
9. MaxPool2d	[192, 25, 25]	[192, 13, 13]	--
10. Conv_Norm_Act	[192, 13, 13]	[128, 13, 13]	--
a. Conv2d	[192, 13, 13]	[128, 13, 13]	[192, 128, 3, 3]
b. BatchNorm2d	[128, 13, 13]	[128, 13, 13]	[128]
c. ReLU	[128, 13, 13]	[128, 13, 13]	--
11. Conv_Norm_Act	[128, 13, 13]	[256, 13, 13]	--
a. Conv2d	[128, 13, 13]	[256, 13, 13]	[128, 256, 3, 3]
b. BatchNorm2d	[256, 13, 13]	[256, 13, 13]	[256]
c. ReLU	[256, 13, 13]	[256, 13, 13]	--
12. MaxPool2d	[256, 13, 13]	[256, 7, 7]	--
13. Conv_Norm_Act	[256, 7, 7]	[160, 7, 7]	--
a. Conv2d	[256, 7, 7]	[160, 7, 7]	[256, 160, 3, 3]
b. BatchNorm2d	[160, 7, 7]	[160, 7, 7]	[160]
c. ReLU	[160, 7, 7]	[160, 7, 7]	--
14. Conv2d	[160, 7, 7]	[320, 7, 7]	[160, 320, 3, 3]
15. AvgPool2d	[320, 7, 7]	[320, 1, 1]	--
16. Dropout	[320]	[320]	--
17. Linear	[320]	[139]	[320, 139]

Figure 5.2: Architecture of GestaltMatcher classifier. The convolution layer 14 used to generate attribution maps using GradCAM and HiResCAM methods is highlighted using a blue box. Layers (10.a, 11.a and 14) that are used by the CustomGradCAM method are highlighted using cyan boxes.

5.2.1 Visualization Details

Captum's visualizer tool was used to generate color coded attribution maps. We post-processed attribution maps to remove noise and visual artifacts. This section provides details on visualization parameters and the post-processing procedure.

Interpolation

The attribution maps generated by layer-specific methods such as GradCAM and HiResCAM are shaped identical to the feature maps outputted by the target layers. For example, layer 14 of the gestaltmatcher model spits out 320 7x7 (height x width) feature maps, and any attribution map generated from the

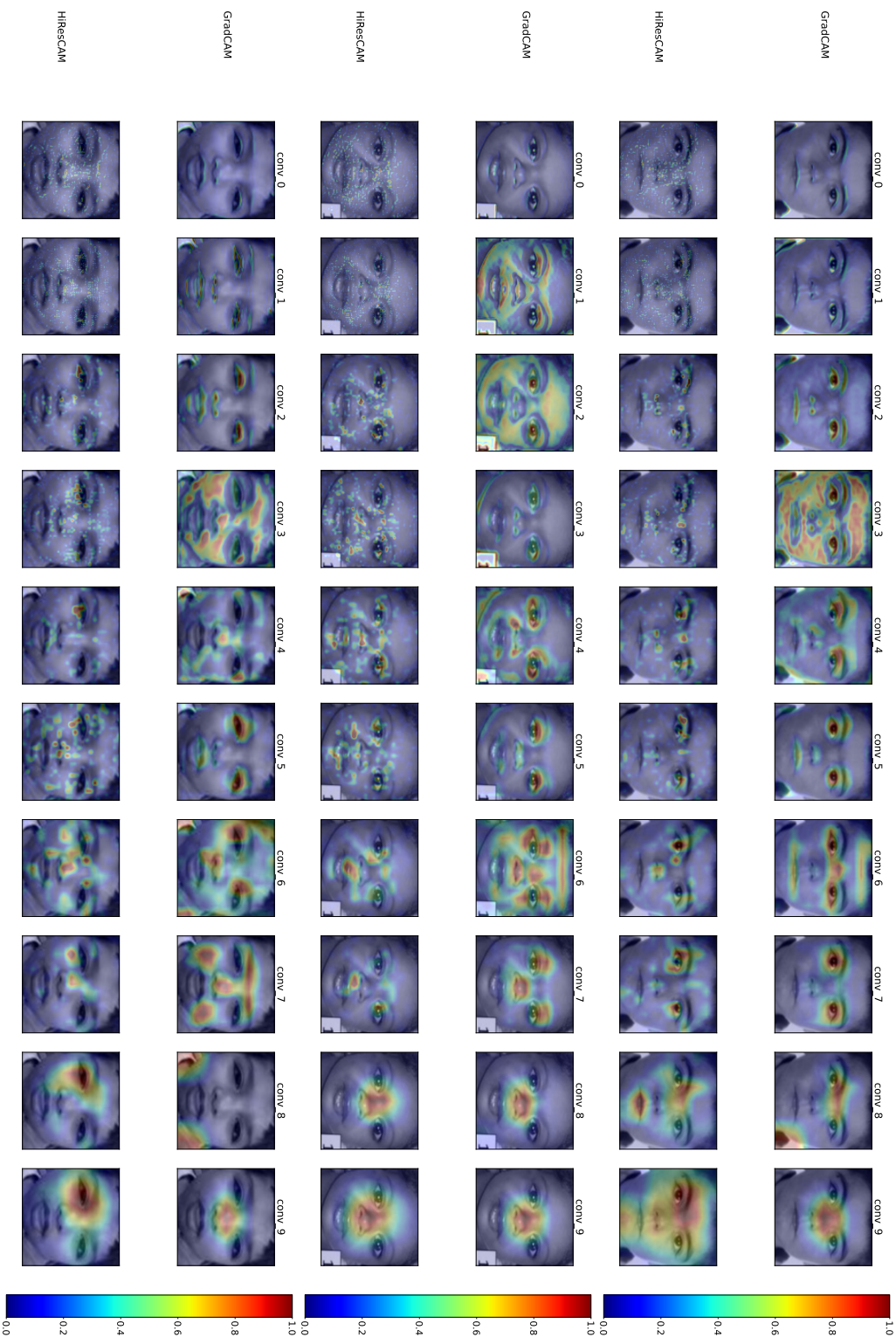


Figure 5.3: Layer-wise visualization of attribution maps generated by GradCAM and HiResCAM methods.

layer is of dimensions 7x7 (height x width). The bilinear interpolation technique [62] was used to rescale attribution maps to the input image size of 100x100 (height x width), so that the map could be overlaid on the input image.

Attribution Sign and Representation

We considered only positive attribution values for visualization and used the "jet" colormap scheme offered by Captum's [30] visualization tool.

Smoothing Attribution Maps

Test Time Augmentation (TTA) is a commonly used technique to boost the performance of image classification models [63]. During the inference process for a given input image, instead of passing the actual sample, its augmented counterparts are fed to obtain predictions from the model. Thus obtained predictions are averaged to obtain the final output. Augmented copies are obtained by applying transformations such as rotations and flips to the original input image.
In our work, we used TTA to enhance the quality of generated attribution maps. Augmented copies of every test input image were obtained by individually applying horizontal flipping and multiplications (with factors of 0.9, 1 and 1.1) operations. Attribution maps were generated for each augmented copy, and finally averaged to obtain the smoothened map. Attribution maps produced from HiResCAM were exempted from TTA based on the guidelines provided by Jacob Gildendblat [61].

5.3 Experiments

Implementation details of the conducted set of experiments are presented in this section.

5.3.1 A. Patient-wise Attribution Map Generation

Patient-wise attribution maps are generated by applying the chosen set of methods to individual patient images. The attributions are computed with respect to the logit scores of the samples' ground truth classes. This is because the attribution "methods work only when the classification result is correct" [59]. However, an experiment was also conducted to observe the differences between explanations of misclassified samples, obtained by computing attributions with respect to their ground truth and predicted labels. Its results are discussed in the next chapter.

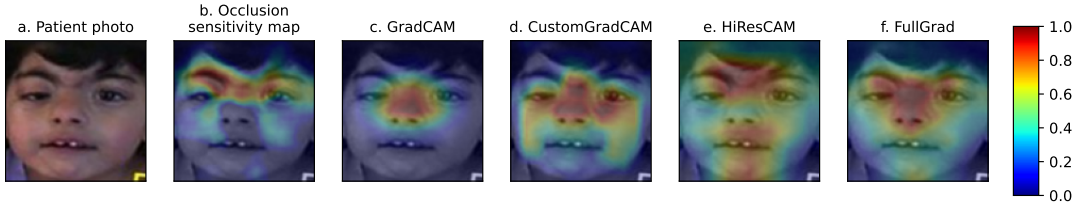


Figure 5.4: An example for patient-wise attribution maps

5.3.2 B. Composite Face Generation

Composite faces are generated with an intent to synthesize a characteristic face for each genetic syndrome class under our scope. Concisely put, they are obtained from aligning and averaging facial images of each category separately. This section provides a detailed description of the process involved.

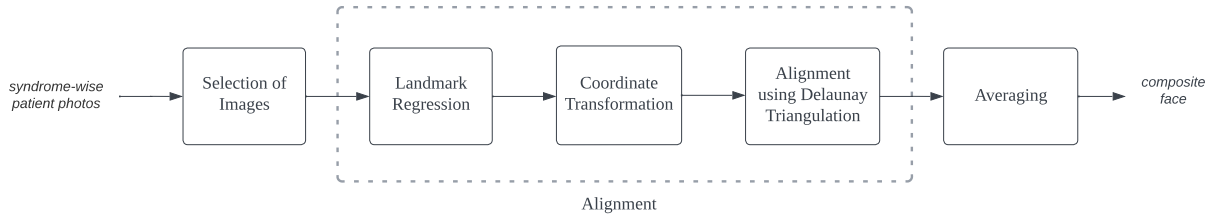


Figure 5.5: Block diagram representing composite face generation process

a. Selection of Images

The first step involved in generating composite faces is to select the right consituent images. A considerable number of images in GestaltMatcherDataBase (GMDB) are characterized by poor visual quality and/or presence of artifacts. A few such as the ones shown in Figure 5.6 also contain objects like nasal catheter tubes, spectacles, and black strips which were introduced to conceal patient identities. Therefore, we hand picked images from each of 139 syndrome classes to remove the unsuitable candidates for generating composite faces.

Figure 5.6: Examples of images rejected for composite face generation

b. Facial Landmark Regression

Landmark points act as anchors to align the constituent images of composite faces. 68 landmark points are calculated for each input facial image using the dlib [1]library's face detector functionality. Internally, dlib uses a CNN based max-margin object detector to perform the task.

[1]dlib library - `http://dlib.net/`.

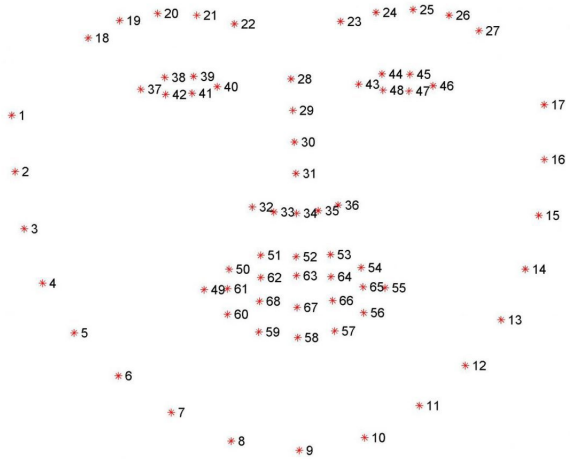


Figure 5.7: An illustration showing positions of facial landmark points considered by dlib face detector. Image source: Pyimagesearch [2].

c. Coordinate Transformation

The next step warps the consituent facial images in a such a way that their landmark points corresponding to the right corner of the right eye and the left corner of the left eye are shifted to locations P_1 (0.3 x image width, 0.3 x image height) and P_2 (0.7 x image width, 0.3 x image height) respectively. In practice this is achieved by computing and applying a similarity transformation matrix using OpenCV library's "estimateRigidTransform" method.

Similarity transformation [64] is a specialization of projective transformations, composed of an euclidean transformation (rotation and translation), and an isotropic scaling. Planar similarity transformation has four Degrees Of Freedom (DOFs) can be expressed in the following form: <equation>

Planar affine transformation has two more DOFs offered by non-isotropic scaling and shear. A planar affine transformation can represented in the form of following matrix:

d. Facial Alignment using Delaunay Triangulation

Facial images are now anchored at points P_1 and P_2. However, their regions are yet to be aligned to a common reference. Such a reference is obtained by computing the mean of landmark points of all constituent faces. Subsequently, every facial image is split into regions using Delaunay triangulation technique. "A Delaunay triangulation of a vertex set is a triangulation of the vertex set with the property that no vertex in the vertex set falls in the interior of the circumcircle (circle that passes through all three vertices) of any triangle in the triangulation" [65]. An affine transformation is computed for every triangular region using its vertices, to warp and fit it in the corresponding region formed by the mean landmark points.

[2]Pyimagesearch - `https://pyimagesearch.com/2017/04/03/facial-landmarks-dlib-opencv-python/`

e. Averaging

The aligned sets of images are averaged to form their corresponding composite faces.

5.3.3 C. Syndrome-wise Attribution Map Generation

Syndrome-wise attribution maps are generated using three different methods and their implementation details are discussed below.

i. Attribution Maps of Composite Faces

The first type of syndrome-wise maps are obtained by treating composite faces as individual patient faces and applying the patient-wise attribution map generation procedure.

ii. Average Attribution Maps

Average attribution map of a given syndrome is obtained by computing the mean across attribution maps of the correctly predicted instances from the class's test split.

iii. Singular Value Decomposition (SVD) of Attribution Maps

Eigen-CAM [59] is a layer attribution technique, which computes saliency from the first principal component of feature activations of channels, across the target convolutional layer. For this experiment, we apply the same technique to the attribution maps of correct predicted samples of a given class, for obtaining its characteristic saliency map. The procedure is described in a step-wise manner below:

1. Vectorize the attribution maps of correctly predicted samples of a given syndrome class c into a matrix A_c.

2. Mean subtraction <>

3. Factorize A_c using SVD to compute principal components of A_c

$$A_c = U\Sigma V^T \tag{5.1}$$

   where U is an MXM orthogonal matrix and V is an NXN orthogonal matrix. The columns of U and V form the left and right singular vectors of A_c respectively. The resultant attribution map R_c is obtained by projecting A on the first right eigen vector V_1:

$$R_c = A_c V_1 \tag{5.2}$$

   The procedure is applied individually on attribution maps of all the three GradCAM, HiResCAM and FullGrad methods, for all syndromes under our scope. In the cases of average and SVD of

attribution map visualizations, a given syndrome's composite face is used as the background image to project the maps, although there is no relationship between them.

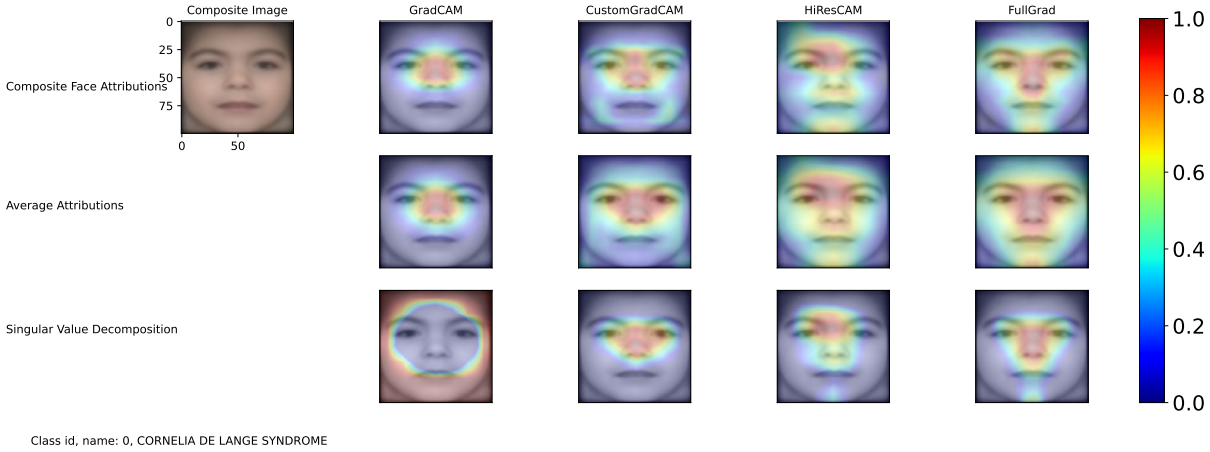


Figure 5.8: An example for syndrome-wise attribution maps

5.3.4 D. Dataset Imbalance - Explanation Quality Analysis

The aim of this experiment is to investigate the effects of class imbalance in GMDB on contents and quality of generated attribution maps. We trained four classifier models each with different numbers and choices for syndrome classes, using the same network architecture and hyperparameters as that of the 139 class GestaltMatcher classifier model. Table **??** summarizes the experimental setup.

5.4 Evaluation Questionnaire

We considered six genetic syndrome classes from GMDB for evaluation: Cornelia de Lange syndrome (CDLS), Williams Beuren syndrome (WBS), Coffin-Siris syndrome (CSS), Nicolaides-Baraitser syndrome (NBS), Hyperphosphatasia-intellectual disability syndrome or Hyperphosphatasia mental retardation syndrome (HPMRS) and Smith-Lemli-Opitz-Syndrome (SMOS). The syndromes were chosen as they make up majority classes of GMDB. Their respective composite faces, patient-wise attribution maps of correctly predicted test samples and syndrome-wise attribution maps, were compiled into a questionnaire which can be accessed at `https://forms.gle/aRru9aeCyZaqpURP9`. The survey was designed to be self-explanatory in nature, containing all necessary information for the evaluator.

Following is an outline of the questionnaire:

- Section 1. Professional Details and Introduction
- Section 2. Composite Face Evaluation
  - Introduction

  - Questions

- Section 3. Patient-wise Attribution Map Evaluation
  - Introduction
  - Questions
- Section 4. Syndrome-wise Attribution Map Evaluation
  - Introduction
  - Questions

Section 1. Professional Details and Introduction

The first section asks the evaluator (clinician) for his professional details, and gives a brief introduction to the questionnaire. Details including his professional experience, specialization and his familiarity with the syndromes presented for evaluation are collected.

Section 2. Composite Face Evaluation

This section contains an introduction followed by a couple of questions. The first question asks the evaluator to identify the syndrome represented by the given composite face. Subsequently, the targeted syndrome name is revealed, and the clinician is asked whether the presented face contains its characteristic phenotypic features.

Section 3. Patient-wise Attribution Map Evaluation

This part of the questionnaire contains an introduction which briefly explains the objective of our research work to the evaluator, especially regarding the integration of XAI methods to the GestaltMatcher model. Subsequently, it presents the clinician a set of five questions for every patient-wise attribution map in the questionnaire.

Section 4. Syndrome-wise Attribution Map Evaluation

The final section of the questionnaire contains a background subsection followed by a question asking the evaluator to pick upto four of the twelve presented syndrome-wise maps and rank them.

Tables 5.2, 5.3, 5.4 contain the list of questions included in sections 2 till 4 respectively. Each row in the tabular columns represent an individual page in the evaluation questionnaire. The first section and introductory parts of all the other three sections are provided in the appendix**??**.

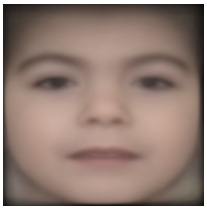
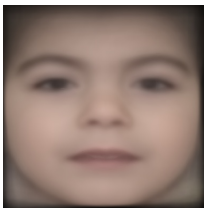
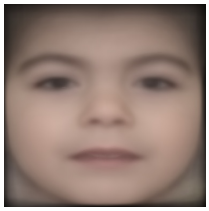
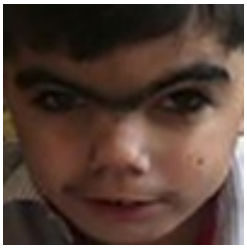
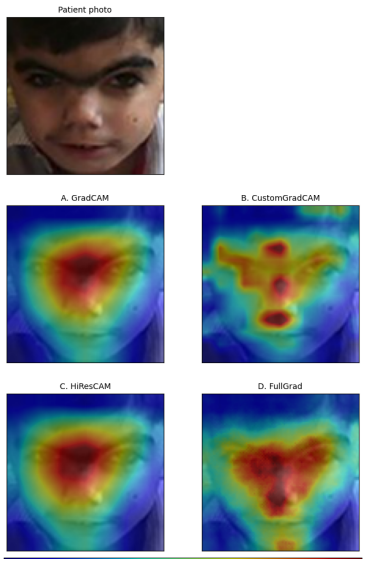
Section 2	
Displayed Images	**Questions/Displayed Text**
	1a. What genetic syndrome do you think the given image most likely represents? (Select one) • Coffin-Siris syndrome • Cornelia de Lange syndrome • Hyperphosphatasia - intellectual disability syndrome • Nicolaides Baraitser syndrome • Williams syndrome (Williams Beuren syndrome) • Smith Lemli Opitz syndrome • None
	Show the true label: "The composite face represents **Cornelia de Lange syndrome**"
	1b. Do you feel that the composite face contains characteristic facial features of Cornelia de Lange syndrome? • Yes • No

Table 5.2: Questions in the composite face section of the evaluation questionnaire

Section 3	
Displayed Images	**Questions/Displayed Text**
	2a. What genetic syndrome do you think the given most likely represents? (Select one) • Coffin-Siris syndrome • Cornelia de Lange syndrome • Hyperphosphatasia - intellectual disability syndrome • Nicolaides Baraitser syndrome • Williams syndrome (Williams Beuren syndrome) • Smith Lemli Opitz syndrome • None 2b. What phenotypic features were considered most important for your choice?
	2c. Do you change your diagnosis after seeing the attribution maps? • Yes • No

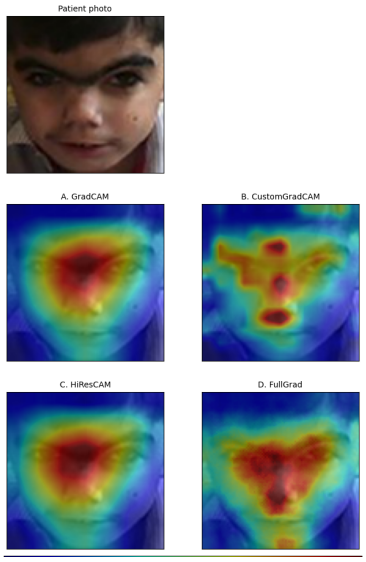
	Show the true label: "The patient has **Cornelia de Lange syndrome**" 2d. Which of the attribution maps shown best highlights the important phenotypic features (those used for diagnosis) of the above mentioned syndrome in the patient ? (Rank them from 1 to 4 (with 1 as the top choice)). If you feel that none of them highlight the key features select None in all rows) <Options A till D to assigned ranks> 2e. Do you find attribution maps to be helpful in evaluating the image?

Table 5.3: Questions in the patient-wise maps section of the evaluation questionnaire

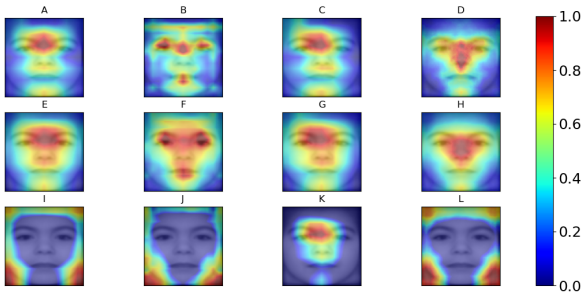
Section 4	
Displayed Images	**Questions/Displayed Text**
	3. Which of the attribution maps shown below most highlight the key characteristic facial phenotypic features of Cornelia de Lange syndrome? (Select top four. If none of them represent select None in all rows)) <Top four from A till L to be selected>

Table 5.4: Questions in the syndrome-wise maps section of the evaluation questionnaire

6

Evaluation and Results

This chapter presents the results of experiments whose implementation details were provided in the last chapter. Besides, the outcomes are systematically analyzed and evaluated.

6.1 Background

We conducted the first three experiments (A. patient-wise attribution map generation, B. composite face generation and C. syndrome-wise attribution map generation) on each of the 139 frequent syndrome classes in the GestaltMatcherDataBase (GMDB) dataset. However, as mentioned earlier, experimental artifacts related to six of the syndromes (Cornelia de Lange syndrome (CDLS), Williams Beuren syndrome (WBS), Coffin-Siris syndrome (CSS), Nicolaides-Baraitser syndrome (NBS), Hyperphosphatasia-intellectual disability syndrome or Hyperphosphatasia mental retardation syndrome (HPMRS) and Smith-Lemli-Opitz-Syndrome (SMOS)) were evaluated by an experienced clinical geneticist. The clinician specialized in the diagnosis of three of the syndromes (CDLS, WBS, HPMRS), and had less familiarity with others. Therefore, analyses and discussions on experimental results mostly revolve around the three syndrome categories. Besides, some important findings from other syndromes are also provided. In this section, some background information regarding facial phenotypic features of the syndromes discussed in this chapter is given.

Cornelia De Lange Syndrome

Cornelia de Lange syndrome (CDLS) is a rare genetic condition present at birth. It is characterized by multiple physical and intellectual abnormalities. There are different variants of the syndrome each identified with its own set of facial features. However, the profile of eyebrows is predominantly considered for diagnosing the syndrome. Figures 6.1 and 6.2 show samples contain other facial features related to the genetic condition.

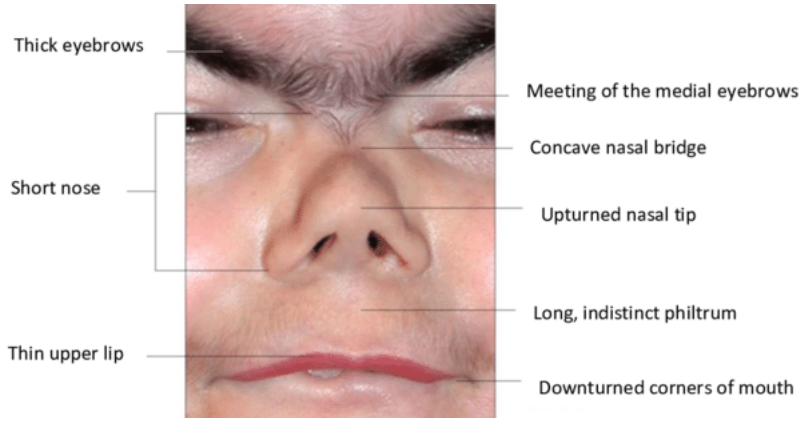
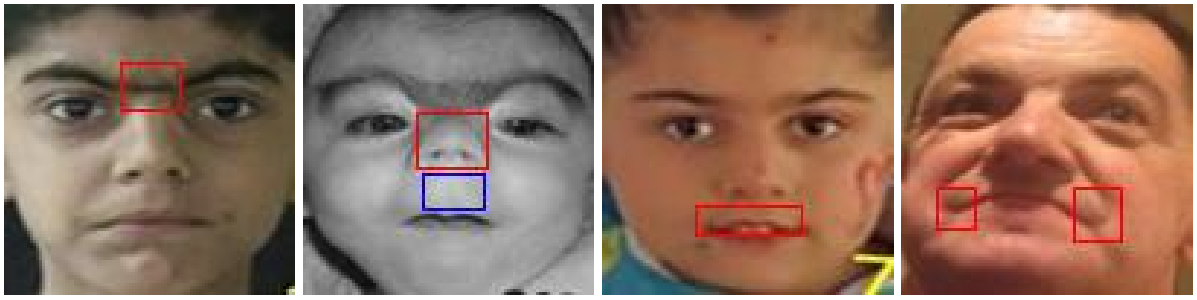


Figure 6.1: Cardinal facial features of CDLS. Image source: [66]



(a) Synophrys (b) Depressed nasal bridge (red), long philtrum (blue) (c) Thin upper lip (d) Down-turned corners of mouth

Figure 6.2: Instances of CDLS from GMDB dataset annotated with phenotypic facial features

Williams Beuren Syndrome

Williams Beuren Syndrome (WBS) or Williams syndrome is a rare genetic condition which affects many parts of the human body. The disorder is marked by both prenatal (before birth) and postnatal growth retardation. Flat midface, long philtrum and thick lips are some of the characteristic facial features associated with the syndrome. Refer figures **??** and 6.4 contain an animated representative face and examples from GMDB dataset respectively.

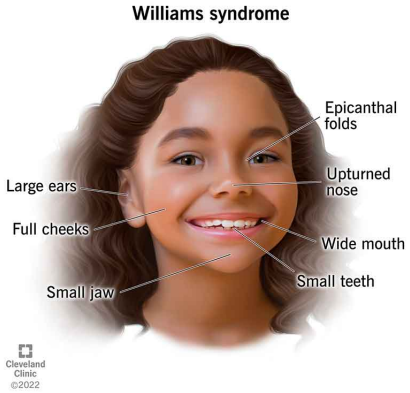
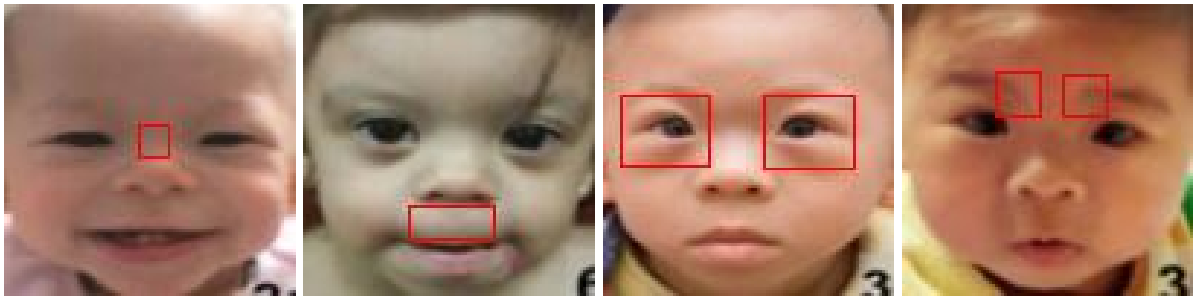


Figure 6.3: An animated characteristic face of WBS. Image source: Cleveland clinic [1].

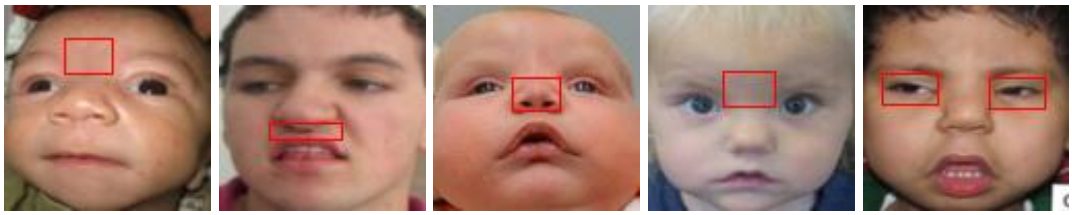


(a) Depressed nasal bridge (b) Long philtrum (c) Puffy eyes (d) Medial eyebrow flares

Figure 6.4: Instances of WBS from GMDB dataset annotated with phenotypic facial features

Hyperphosphatasia with Mental Retardation Syndrome

Hyperphosphatasia with intellectual disability or Mental Retardation Syndrome (HPMRS)



(a) Hypertelorism (b) Short philtrum (c) Short nose (d) Broad nasal bridge (e) Upslanting palpebral fissures

Figure 6.5: Instances of HPMRS from GMDB dataset annotated with phenotypic facial features

[1] Cleveland clinic - `https://my.clevelandclinic.org/health/diseases/15174-williams-syndrome`

Coffin Siris Syndrome

(a) Short philtrum (b) Thin upper-lip vermilion and thich lower-lip vermilion (c) Small chin (d) Downslanting palpebral fissures (e) Broad nasal tip

Figure 6.6: Instances of CSS from GMDB dataset annotated with phenotypic facial features

6.2 Experiment A. Patient-wise Attribution Map Generation

The clinician was presented with attribution maps of 23 instances from six different syndromes classes in GMDB. However, as mentioned earlier, his specialization was limited to three syndromes which were represented by 15 samples in the questionnaire. Table 6.1 shows the sample distribution, and diagnostic performance of the clinician for each syndrome class, without the aid of attribution maps. As described in Chapter 4, it is important to note that only samples from the test split of GMDB that were correctly classified by the GestaltMatcher classifier model, were chosen for evaluation.

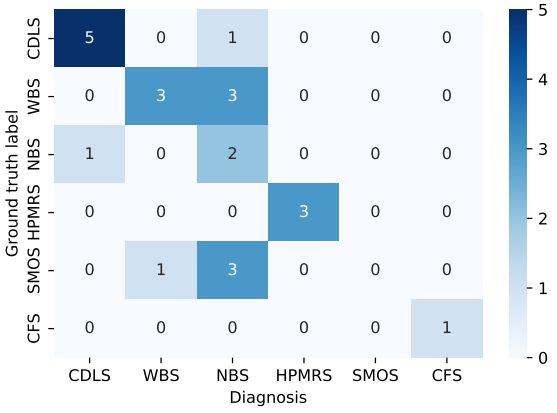


Figure 6.7: Confusion matrix representing the clinician's diagnostic performance

Syndrome	Sample count	Clinician specializes in syndrome	Accuracy
CDLS	6	yes	0.83
WBS	6	yes	0.50
NBS	3	no	0.67
HPMRS	3	yes	1.00
SMOS	4	no	0.00
CFS	4	no	1.00

Table 6.1: Diagnostic performance of the clinician on samples in the questionnaire

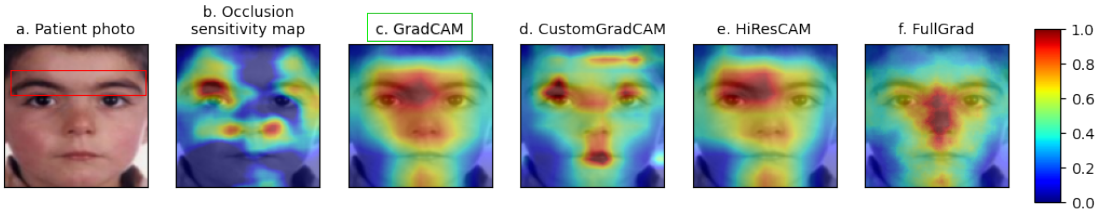
Usefulness of Attribution Maps in Diagnoses

Fig 4.4 in Chapter 4 enumerated the following possible inferences, which can obtained from responses in the patient-wise attribution map evaluation section of the questionnaire:

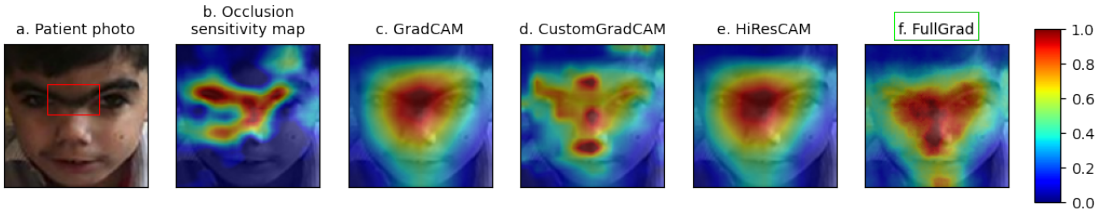
- A. Attribution maps misleads the clinician to make incorrect diagnosis
- B. Attribution maps reinforces the clinician's correct diagnosis
- C. Attribution maps helps the clinician to correct his diagnosis
- D. Attribution maps fail to help the clinician to correct his diagnosis

Here, we describe the usefulness of attribution maps and performance of methods used to generate them, by binning clinician responses into one of the above mentioned, and analyzing them.

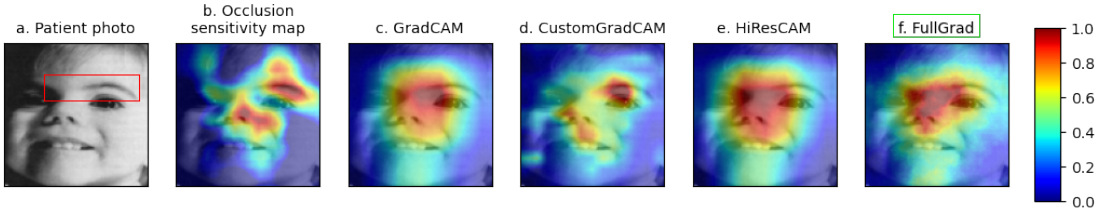
As shown in Figure 6.7, the clinician correctly diagnosed 11 out of the total 15 patient images, presented to him from the syndromes of his specialty, without the aid of attribution maps. His responses to the subsequent questions (Refer questions 2b - 2e in Table 5.4) asked after the 11 correctly diagnosed cases, indicate that the attribution maps reinforced correct diagnoses (scenario B) in eight cases, and didnot prove helpful (scenario A) in three. Within the 8 occurrences of scenario B, attribution maps using FullGrad were marked to best highlight the associated facial features in five, followed by the maps of other methods in one each. Among the four misdiagnosed instances, none of the attribution maps proved effective in correcting three of his predictions (scenario D), and a FullGrad attribution map helped him correct one (scenario C). To summarize, the clinician found attribution maps in general to be helpful for diagnosis, in 9 out of 15 instances, and also chose FullGrad maps to highlight relevant facial features in most of the cases.



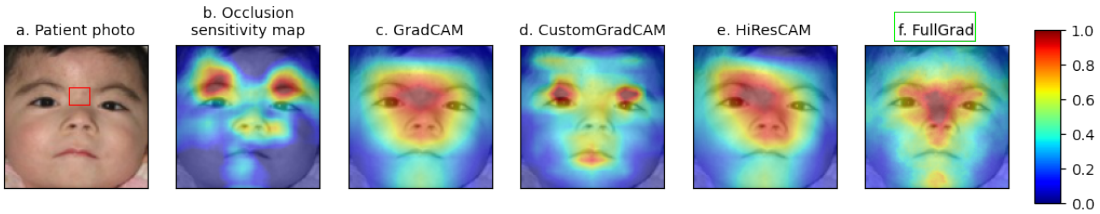
(a) Patient with CDLS and eyebrows specified as the key feature



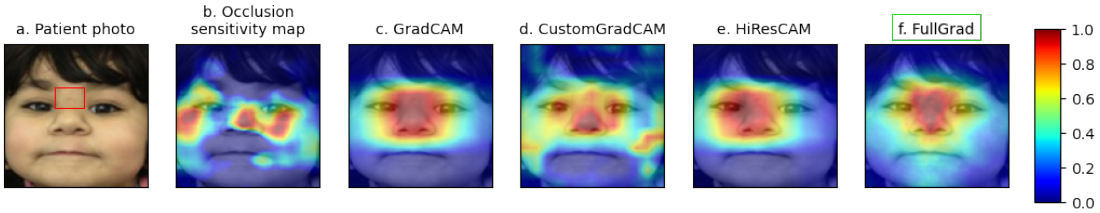
(b) Patient with CDLS and synophrys feature



(c) Patient with CDLS and eyebrows specified as the key feature



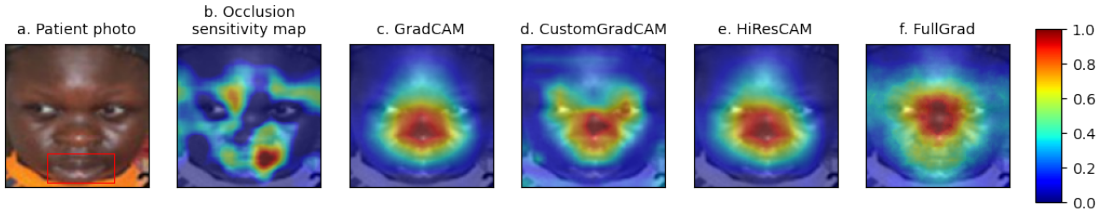
(d) Patient with CDLS and nose, depressed nasal bridge reported as the key features



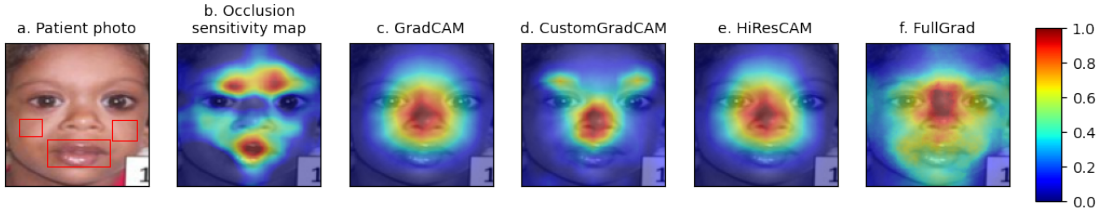
(e) Patient with HPMRS and nasal bridge reported as the key feature

Figure 6.8: Attribution maps of instances in which clinician's attention regions matched that of the classifier model. Key features specified and methods chosen by the clinician are boxed in red and green colors respectively

(a) Patient with HPMRS, and nose and mouth specified as key features



(b) Patient with WBS, and lips specified as the key feature



(c) Patient with WBS, and lips and cheeks specified as key features

Figure 6.9: Attribution maps of instances in which clinician's attention regions differed from that of the classifier model. Key features specified by the clinician are boxed in red. Method chosen for the first instance is boxed in green color.

Comparing Attention Regions

Besides identifying syndromes and rating attribution maps, the clinician was asked to specify the key facial features which he used for diagnosis (Refer question 2b in 5.4). From his answers, let us compare his regions of attention with that of the GestaltMatcher classifier model, represented in form of attribution maps. Figure 6.8 and Figure 6.9 contain examples of attribution map sets of correctly diagnosed samples, for which features considered crucial by the clinician, matched and differed with that of the classifer respectively.

He reported the eyebrow region to contain key features for diagnosing CDLS, in most of the cases, especially the occurrence of synophrys such as found in figures 6.8a and 6.8b. Besides, features in the nose region such as the depressed nasal bridge (refer Figure 6.8e) were also considered important for his identification. Attribution maps generated using FullGrad were reported to correctly highlight key features of CDLS in three out of six cases, followed by GradCAM in two. The clincial geneticist found none of the maps to be helpful in diagnosing one instance. On the whole, attribution maps were found to be helpful in diagnosing CDLS.

In the case of HPMRS samples, nose and mouth regions were pointed out to contain the syndrome's characteristic features (refer figures 6.8e and 6.9a). FullGrad maps are reported to highlight the characteristic nose region in most of the cases. However, none of the maps highlighted the mouth feature, as it can observed from Figure 6.9a.
Samples presented from WBS were reported to contain key facial features in lip and cheek regions of the face (refer figures 6.9b and 6.13c). Analyzing clinician responses pertaining the syndrome reveal that none of attribution maps proved helpful for him, both in cases of reinforcing correct diagnoses and correcting misdiagnoses. More importantly, it can be observed that the maps highlighted the nose region, a feature that is not characteristic of the syndrome. However, regions highlighted in occlusion sensitivity maps contain features that were considered by the clinician. This finding raised doubts on faithfulness [1] of the considered attribution methods. Therefore, we attempted to understand the cause of the issue by visualizing and analyzing layer-wise activation maps of different samples.

Layer-wise Activation Visualization and Analysis

We generated attribution maps for all ten convolutional layers of the classifier model using GradCAM and HiResCAM methods. This was done to check whether the missing features get highlighted in attributions of layers other than what were chosen for the experiment. We chose conv_9 for GradCAM and HiResCAM, and conv_6, conv_7 and conv_9 to generate maps using CustomGradCAM. FullGrad is a layer agnostic approach, and therefore was excluded.
Figure 6.10 contains layer-wise visualizations for samples whose attribution maps failed to highlight the key facial features. The first pair of visualizations correspond to the image in Figure 6.9a whose maps failed to highlight the mouth region. The area gets highlighted in the conv_7 layer visualization using GradCAM. Likewise, incases of visualizations of WBS samples (refer second and third pair of rows in Figure 6.10), the key features in lip and cheek regions get highlighted in other convolutional layers.
We extended this analysis to samples in WBS which were not included in the questionnaire. This was performed to find whether visualizations of any particular layer, or set of layers highlighted key features in all instances. Figure 6.11 contain visualizations of three of such samples, whose key features are in the lip region. The region was found to be highlighted in visualizations of different layers in different samples. This inconsistent behavior of attribution methods poses a challenge in using them to explain GestaltMatcher predictions in a clinical setting.

[1] Accurate representation of the reasoning process behind a model's prediction

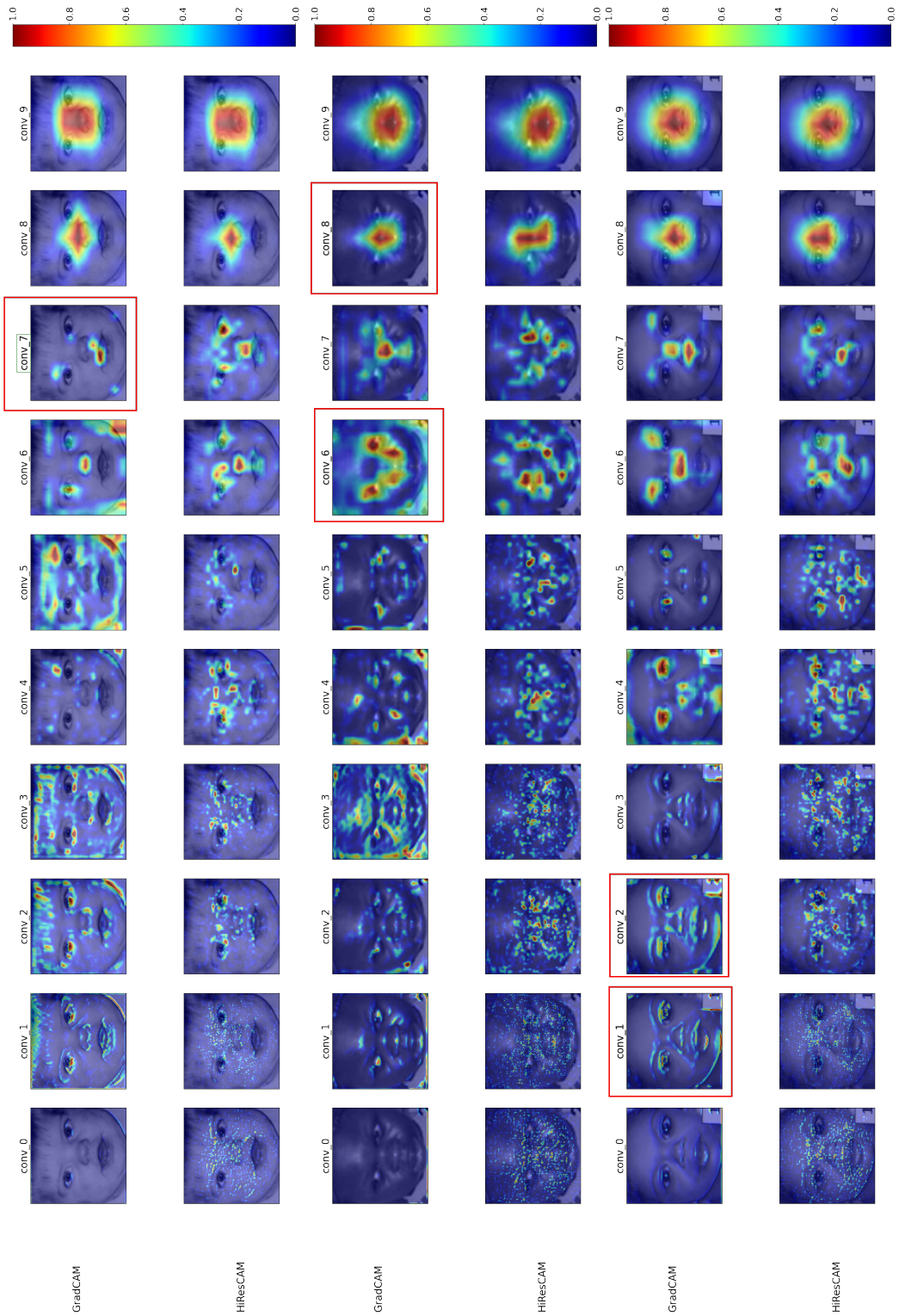


Figure 6.10: Example layer-wise activation map visualizations for instances presented in the questionnaire. Layers highlighting syndromic features are boxed in red.

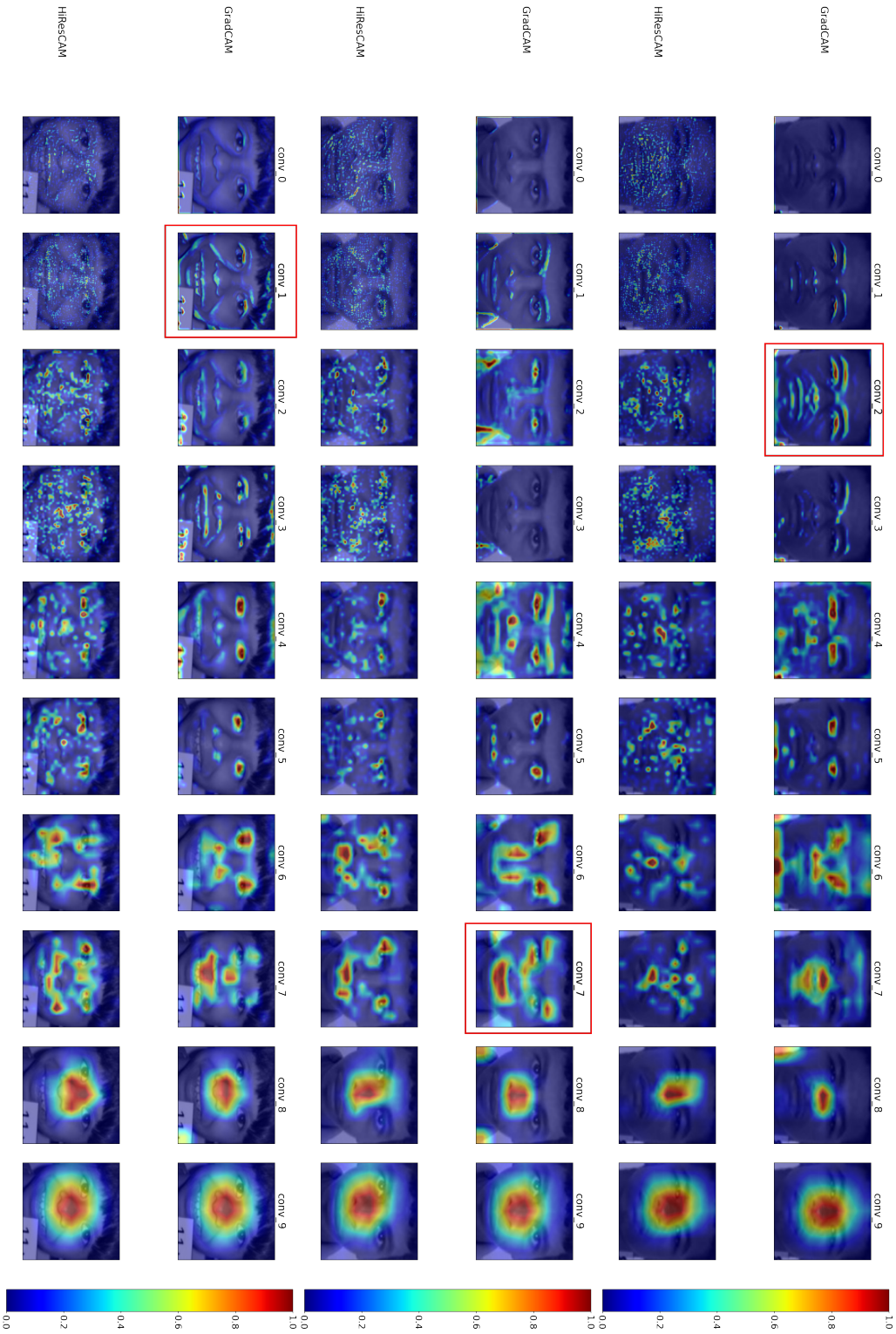
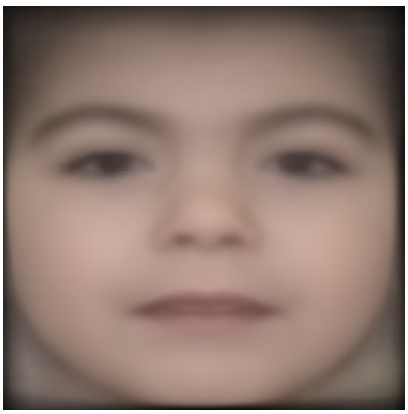


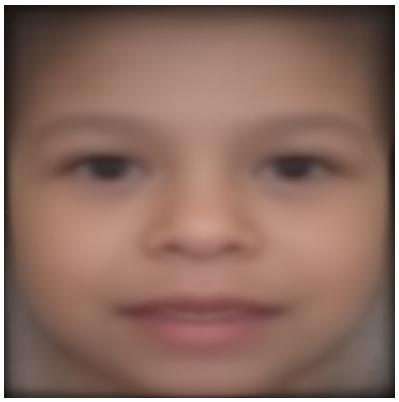
Figure 6.11: Example layer-wise activation map visualizations for instances not present in the questionnaire but in GMDB dataset. Layers highlighting syndromic features are boxed in red.

6.3 Experiment B. Composite Face Generation

A composite face provides a characteristic representation of the facial phenotype of a given genetic syndrome. Figure 6.12 contains composite faces of the twelve largest classes of the GMDB dataset.

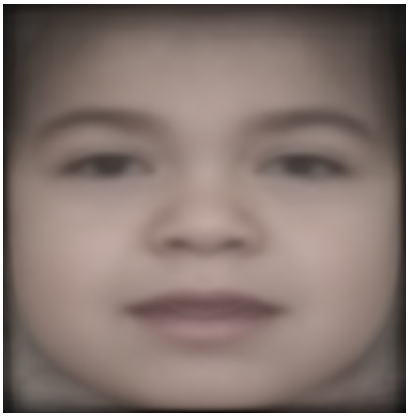


(a) Cornelia de Lange syndrome

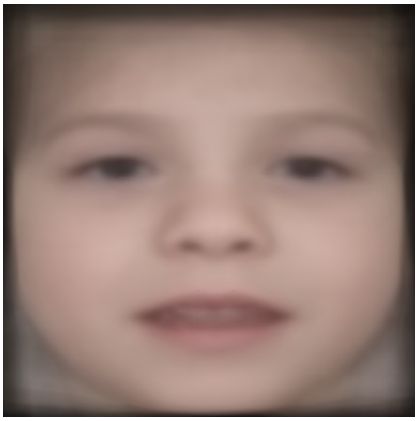


(b) Williams syndrome

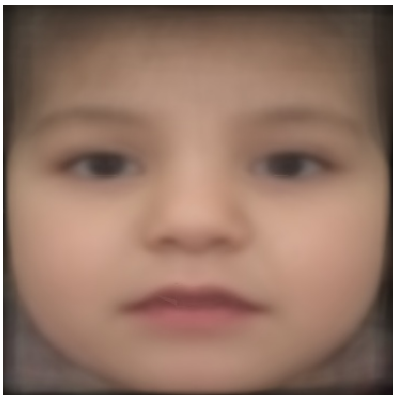
(c) Wiedemann-Steiner syndrome



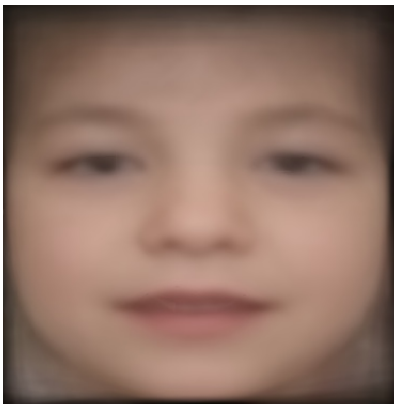
(d) Mucopolysaccharidoses



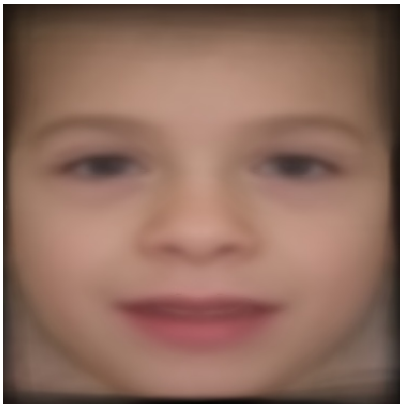
(e) Nicolaides-Baraitser syndrome



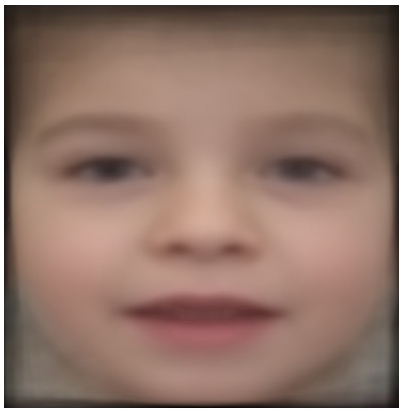
(f) Hyperphosphatasia with mental retardation syndrome



(g) Baraitser-Winter syndrome

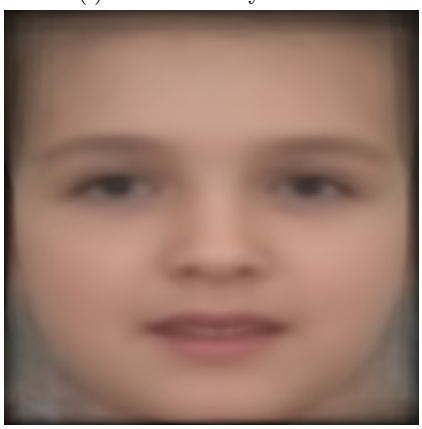


(h) Smith-Lemli-Opitz syndrome



(i) Coffin-Siris syndrome

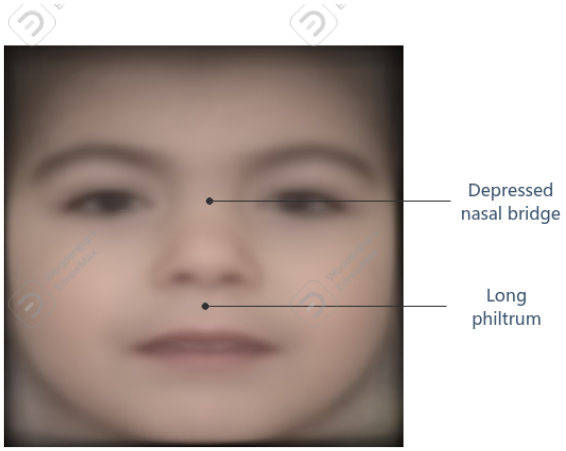
(j) Treacher Collins syndrome



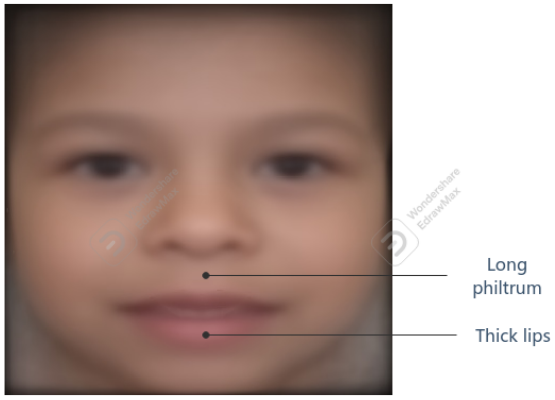
(k) Sotos syndrome

(l) Kabuki syndrome

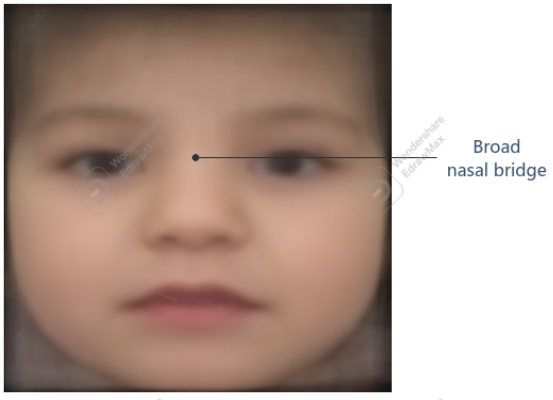
Figure 6.12: Composite faces of the twelve largest syndrome classes in GMDB dataset



(a) Composite face of CDLS labeled with phenotypic features



(b) Composite face of WBS labeled with phenotypic features



(c) Composite face of HPMRS labeled with feature

Figure 6.13: Composite faces of the syndromes evaluated by the clinician labeled with phenotypic features

Similar to other experimental artifacts, the clinician evaluated composite faces representing six syndromes. He found all three faces pertaining syndromes he specialized (CDLS, WBS, HPMRS), to

contain their respective characteristic features. The same was reported for SMOS, although he was not experienced in diagnosing the syndrome. We observed the presence of a few though not all distinctive facial features in composite faces of CDLS, WBS and HPMRS (Refer Figure 6.13). The absence of other characteristic features can be attributed to the fact that composite faces were generated by averaging variations in their constituent images. Besides acting as an input artifact for syndrome-wise attribution map generation, composite faces can serve as reference images for clinicians.

6.4 Experiment C. Syndrome-wise Attribution Map Generation

The clinician evaluated four out of six syndrome-wise attribution maps presented to him representing CDLS, WBS, HPMRS and CSS (refer figures 6.14 and 6.15). The following could be deduced from his responses, and by analyzing the maps:

- The composite face-GradCAM and average attribution-GradCAM combinations were chosen to best highlight characteristic feature of CDLS. It can be observed from Figure 6.14 that the occurrence of synophrys is featured in both the maps.

- In the case of WBS, representations generated using composite face-FullGrad and average attribution-FullGrad pairs were preferred over other options. Both the attribution maps marked the depressed nasal bridge feature of the syndrome (refer Figure 6.14).

- The clinician considered composite face-GradCAM and average attribution-GradCAM combinations to closely represent the broad nasal bridge and hypertelorism features associated with HPMRS (refer Figure 6.15).

- Unlike the above discussed syndrome-wise maps, down-slanting palpebral fissures, a feature in the eye region was highlighted in the composite face-CustomGradCAM and average attribution-CustomGradCAM maps of CSS (refer Figure 6.15). This led the clinician to pick them over others.

- Maps generated using the SVD approach were not found to be useful in any of the evaluations. In addition, we found the approach to produce maps in which image background was highlighted (refer Figure 6.14).

In general, the syndrome wise attribution maps produced from average attribution - GradCAM and average attribution - FullGrad combinations, successfully highlighted the most characteristic feature of respective syndromes. However, similar to the case of patient-wise attribution maps most of the maps highlighted the nasal region, except for CSS.

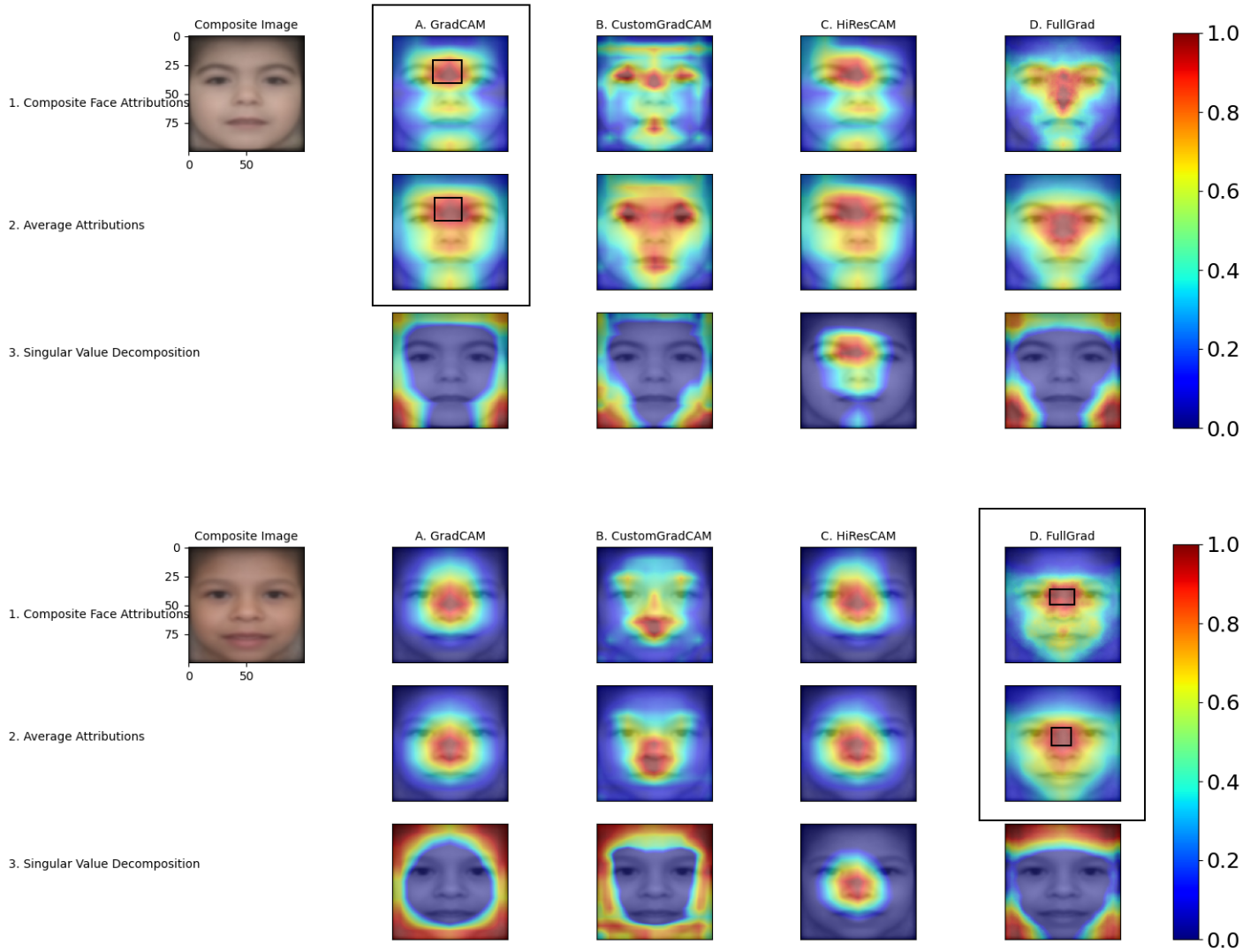


Figure 6.14: Syndrome-wise attribution maps of CDLS (top three rows) and WBS (bottom three rows). Options chosen by the clinician are boxed in black.

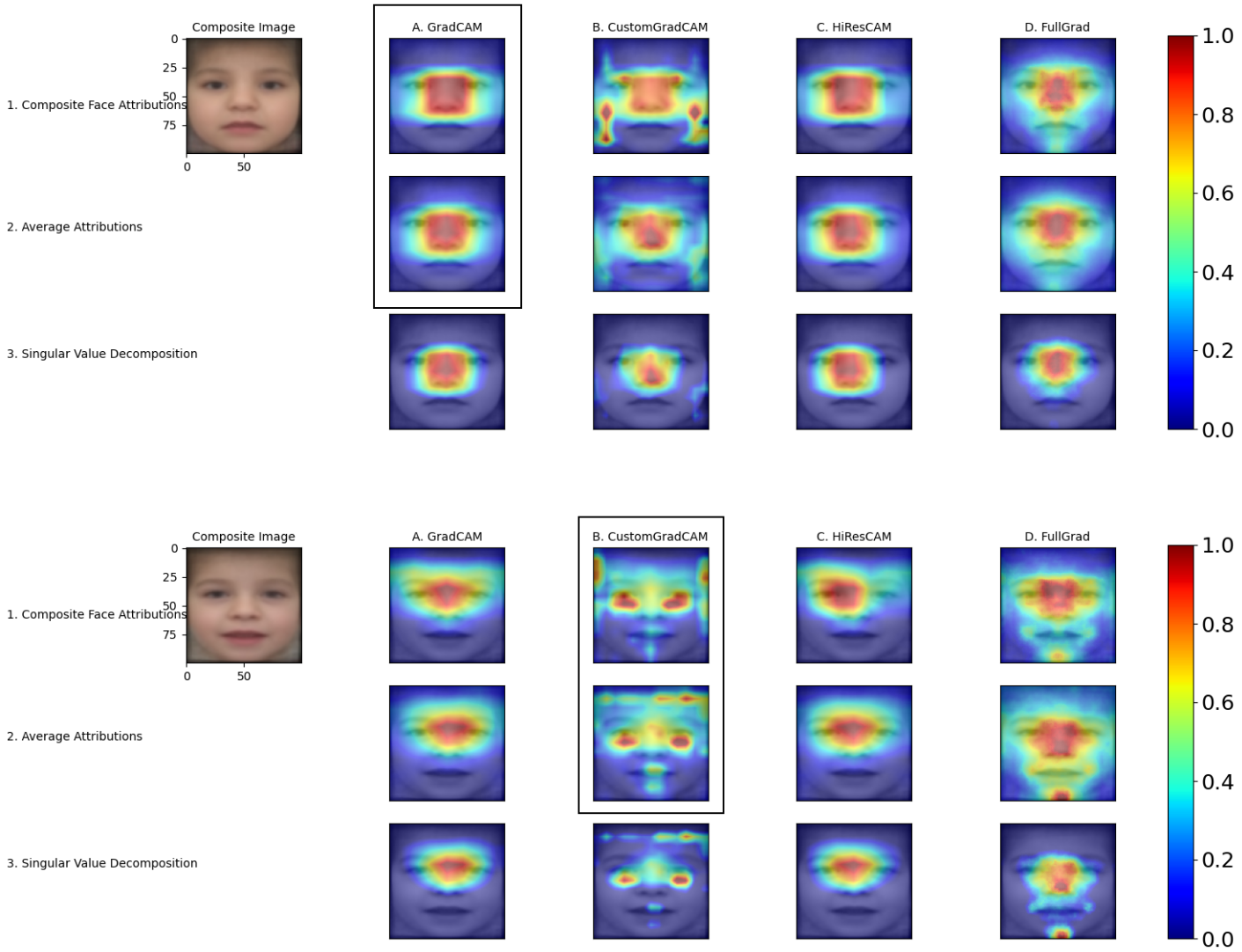


Figure 6.15: Syndrome-wise attribution maps of HPMRS (top three rows) and CSS (bottom three rows). Options chosen by the clinician are boxed in black.

6.5 Experiment D. Dataset Imbalance - Explanation Quality Analysis

This experiment was conducted to study the impact of dataset imbalance on attribution maps. The following four classifier models, each with different choice and number of target classes were trained for this experiment:

1. GC1: CDLS vs healthy
2. GC2: CDLS vs WBS
3. GC3: CDLS vs WBS vs mucopolysaccharidoses (MPS) vs HPMRS vs CSS, with class-wise imbalance
4. GC4: CDLS vs WBS vs MPS vs HPMRS vs CSS, with equal number of samples

Besides, we considered attribution maps of the 139-class GestaltMatcher classifier model (GM) as the baseline for comparison.

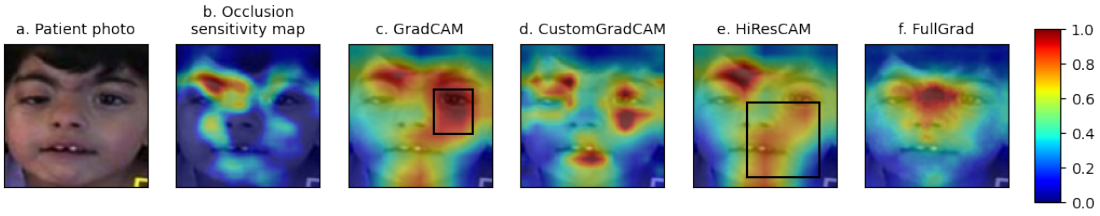
Unlike other experiments, analysis of this experiment's artifacts was done without the help of a clinician. Most of the changes observed in the attribution maps of the above listed classifiers, couldn't be interpreted in a meaningful way. Therefore, in this section, a qualitative analysis of some important observations are provided.

- Figure 6.16 contains attribution maps of a CDLS patient image generated from the set of classifier models considered for this experiment (a-d), and the GM model. It can be observed that the regions highlighted in FullGrad maps remain unchanged in all cases. However, attention regions of GradCAM and HiResCAM maps differ for every model.

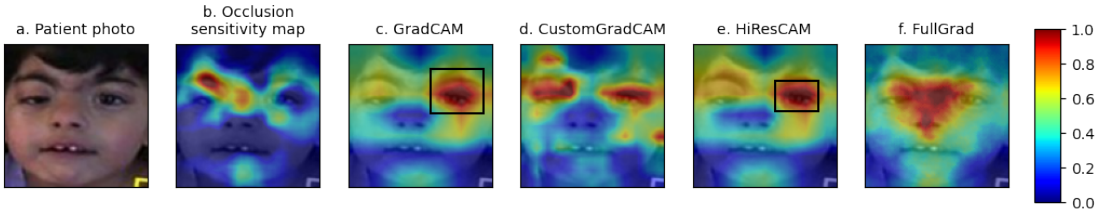
- It can be observed that the attention region of GC1, the syndrome vs healthy classifier has a wider region of attention than the GM model. Also, their regions of attention differ from each other. GC1 focused more on the right cheek region while GM focuses on labella (the meeting point of eyebrows), which is characteristic to CDLS. Attention regions of GC2 were observed to be narrower than that of GC1. "No significant changes were observed between attribution maps GC3 and GC4, the classifiers trained on balanced and imbalanced sets of classes respectively".

- Attribution maps shown in Figure 6.17 reveal that except GM all other classifier model get affected by the presence of spectacles on the patient's face. The GM model focuses on the characteristic eyebrow region (refer Figure 6.17e).

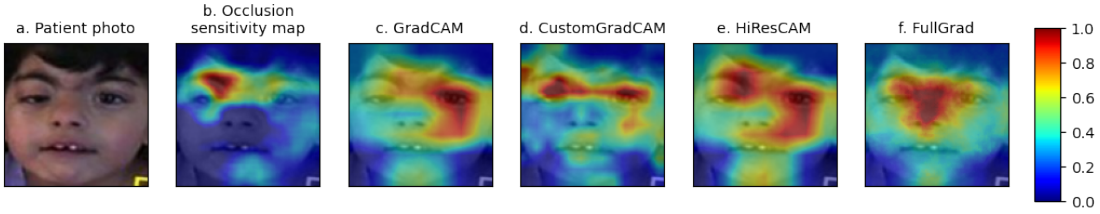
- Attribution maps of healthy faces produced using GC1 reveal that the model uses same set of features to recognize images of both the classes. This indicates that the neural network model looks for the presence or absence of certain discriminative features in an image, to classify the same. It can be observed from visualizations in Figure 6.18 that the nasal region was used to recognize both the healthy and syndromic images.



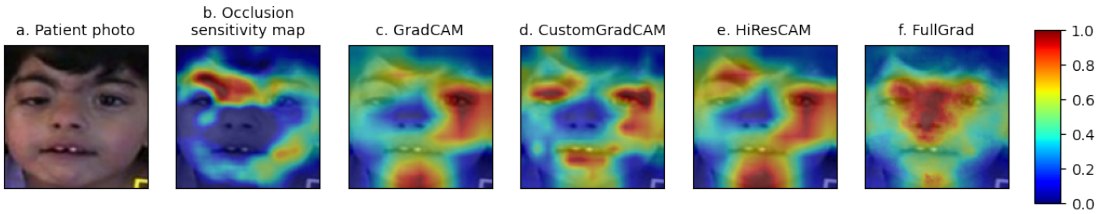
(a) GC1



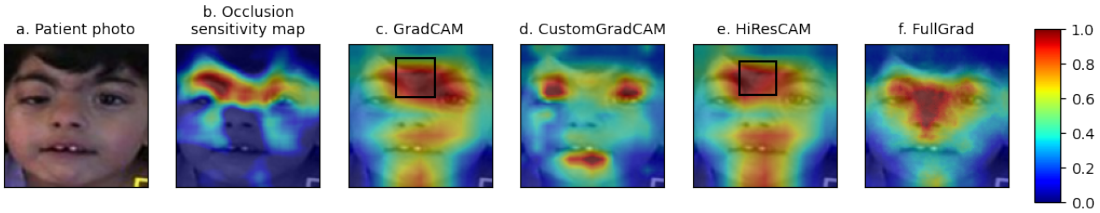
(b) GC2



(c) GC3

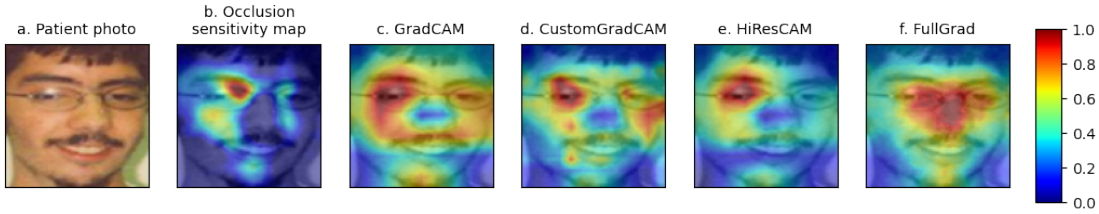


(d) GC4

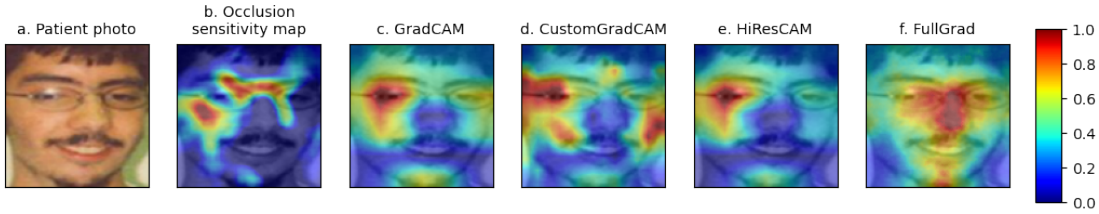


(e) GM

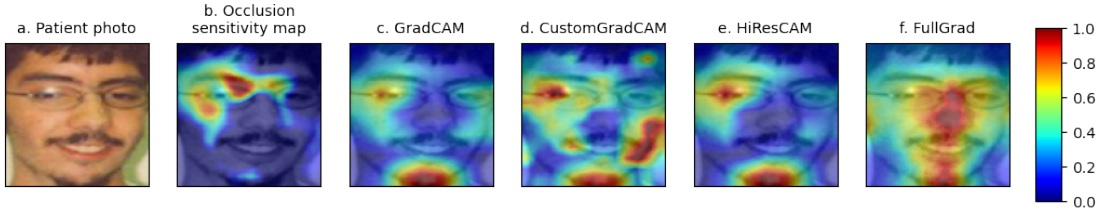
Figure 6.16: Attribution maps of a CDLS patient image generated using different classifier models. Meaningful changes in regions of attention are marked with a black bounding box.



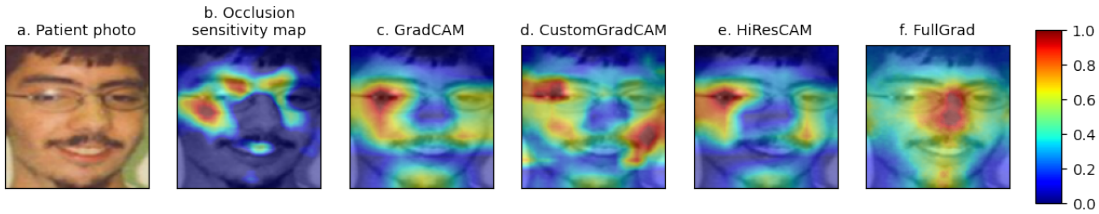
(a) GC1



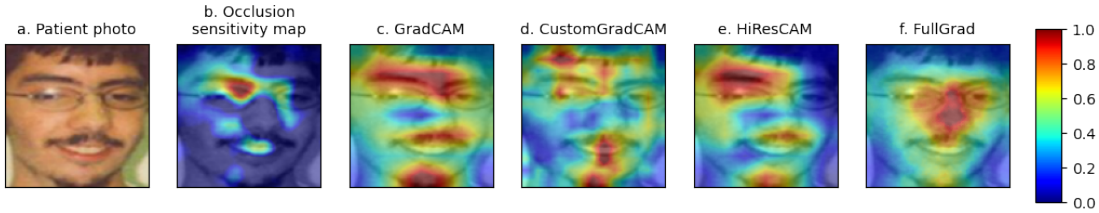
(b) GC2



(c) GC3



(d) GC4

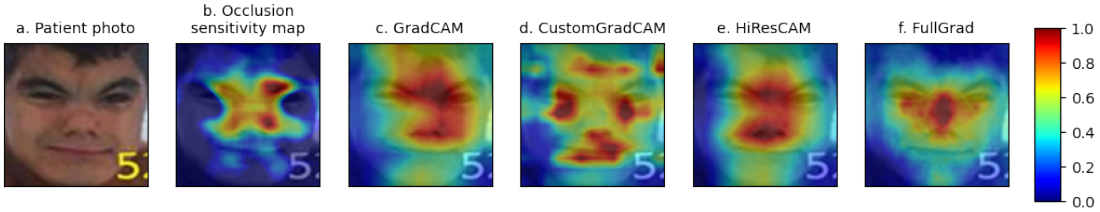


(e) GM

Figure 6.17: Attribution maps depicting the effect of wearables on attention of different classifier models

(a) An example of correctly classified healthy face

(b) An example of correctly classified healthy face



(c) An example of correctly classified CDLS face

Figure 6.18: Attribution maps of healthy and syndromic facial images

6.6 Summary of Results

This section summarizes results presented earlier in this chapter:

- Experiment A: Patient-wise attribution maps representing six genetic syndromes was evaluated by an experienced clincal geneticist. We considered his responses related to three out of the six syndromes he was familiar, for our analyses (CDLS, WBS and HPMRS). Attribution maps generated using the FullGrad method were voted to highlight relevant facial features in majority of samples, representing CDLS and HPMRS classes. None of the presented maps highlighted "thick lips", the characteristic feature of WBS. Layer-wise activation visualization revealed that the missing feature gets highlighted in maps of convolutional layers different from the ones considered earlier. Extending the analysis to other samples showed that maps of different layers contained highlighted key feature regions for different instances.

- Experiment B: Results from the clinician's evaluation of composite faces were presented. The expert was convinced that composite faces of CDLS, WBS, HPMRS and SMOS represented their respective syndromes. Besides, it was reported that they contained one or two, but not all characteristic features of their corresponding syndromes. Faces of CDLS, WBS and HPMRS labeled with the

identified features were presented, along with the ones representing other nine syndromes in GMDB dataset.

- Experiment C: The clinician evaluated four syndrome-wise attribution maps representing CDLS, WBS, HPMRS and CSS. The composite face - average attribution combination was picked to better highlight key regions of CDLS and HPMRS. The composite face - FullGrad and average attribution - FullGrad maps closely matched the clinician's region of attention for WBS. Downslanting palpebral fissures, a characteristic feature of CSS was contained in the regions highlighted by composite face - customGradCAM and average attribution-custom GradCAM maps. Maps generated using the SVD approach were not found to be useful.

- Experiment D: No significant effects of dataset imbalance were observed in attribution maps. Results of experiment show that the classifier model trained to differentiate CDLS faces from healthy faces, relies on features of the same facial region, to classify an input into either of the two. In addition, it was found that the GestaltMatcher classifier model's attention on key features, does not get affected by the presence of wearables like spectacles in the facial image.

6.7 Inferences

In this section, inferences drawn from the results of conducted experiments are presented.

- **1. Nasal profile, a discriminative feature for syndrome recognition:** We observed the nose region to be highlighted in attribution maps generated for samples of many other classes, which were not represented in the questionnaire. Figure **??** contain some examples of such instances. The recurrent behavior of the attribution methods to highlight the nose region, may indicate the facial region's distinctiveness and importance in the diagnosis of rare genetic conditions. In order to verify this hypothesis, we looked into Online Mendelian Inheritance in Man (OMIM), an online compendium of human gene and phenotypes. We searched for facial features linked to syndromes of the ten largest classes of GMDB (Refer Table 6.2). We found eight out of the ten syndrome to be linked to atleast feature in the nose region. This finding supports our hypothesis, and possibly could help the scientific community in discovering new gene-phenotype associations.

[1] `https://www.omim.org`

Syndrome	Features in the nose region
Cornelia de Lange syndrome I	Anteverted nostrils, depressed nasal bridge
Williams syndrome	Anteverted nares, depressed nasal bridge
Wiedemann-Steiner syndrome	Broad nose, wide nasal bridge, depressed nasal tip
Mucopolysaccharidoses	None
Nicolaides-Baraitser syndrome	Narrow nasal bridge, Broad nasal base, upturned nasal tip, thick alae nasi, anteverted nares
Hyperphosphatasia with mental retardation syndrome I	Broad nasal bridge, broad nasal tip, short nose
Baraitser-Winter syndrome	Broad nasal bridge, broad nasal tip, short nose, large, squared nose tip, prominent nasal root on profile
Smith-Lemli-Opitz syndrome	Anterverted nares, broad and flat nasal bridge
Coffin-Siris syndrome I	Broad nasal tip
Treacher Collins syndrome I	None

Table 6.2: Features in the nose region associated with the ten largest classes of GMDB dataset. Source: OMIM

- **2. Attribution methods are limited by their expressiveness**: Attribution methods highlight the regions that were relevant for a certain prediction by a neural network. Although attribution maps are helpful in localizing a network's attention regions, they lack the capability to provide the context of the explanation. In the case of this research work, the nose region gets highlighted in most of the syndromes' attribution maps. However, we are unable to deduce what particular feature attributes are used by GestaltMatcher for its predictions, for example, symmetry or shape profile of the nose. This could be determined using other XAI techniques such as feature and concept visualizations (Refer 2.2).

- **3. All explanations do not carry a real-world meaning:** Each XAI method considered in this work approaches the problem of attribution in its own way. GradCAM converts the problem into "weakly supervised localization", and aims to localize the target class in form of an object. This proves helpful when classes represent real-world objects, which have a well defined physical form. In our case, the target categories represent genetic syndromes which are entities characterized by distinct features, but dont exist as individual objects.
  HiResCAM, a more faithful approach avoids the gradient-averaging step to reveal all attention regions of a neural network model. It can be inferred from the clinician's evaluation that unfiltered explanations produced by the method was not chosen to highlight the facial features, that are considered important for diagnosis by humans. This is possibly because the GestaltMatcher model focuses on multiple sparse sections of a facial image, whose composition does not correspond to a feature of a given syndrome. The success of FullGrad approach can be attributed to its layer-agnostic

nature and ability to consolidate both input and neuronal attributions into a single representation.

- **4. XAI methods are not ready for high-stake applications**

7

Conclusions

7.1 Contributions

7.2 Lessons learned

7.3 Future work

A

Design Details

Your first appendix

B

Parameters

Your second chapter appendix

References

[1] T.-C. Hsieh, A. Bar-Haim, S. Moosa, N. Ehmke, K. W. Gripp, J. T. Pantel, M. Danyel, M. A. Mensah, D. Horn, S. Rosnev *et al.*, "Gestaltmatcher facilitates rare disease matching using facial phenotype descriptors," Nature Publishing Group, Tech. Rep., 2022.

[2] D. Matthew Zeiler and F. Rob, "Visualizing and understanding convolutional neural networks." ECCV, 2014.

[3] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[4] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *International conference on machine learning*. PMLR, 2017, pp. 3145–3153.

[5] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[6] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan, "Medical image analysis using convolutional neural networks: a review," *Journal of medical systems*, vol. 42, no. 11, pp. 1–13, 2018.

[7] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. R. Acharya, "Automated detection of covid-19 cases using deep neural networks with x-ray images," *Computers in biology and medicine*, vol. 121, p. 103792, 2020.

[8] Z. Xue, S. Antani, L. R. Long, D. Demner-Fushman, and G. R. Thoma, "Window classification of brain ct images in biomedical articles," in *AMIA Annual Symposium Proceedings*, vol. 2012. American Medical Informatics Association, 2012, p. 1023.

[9] H. Lee, S. Yune, M. Mansouri, M. Kim, S. H. Tajmir, C. E. Guerrier, S. A. Ebert, S. R. Pomerantz, J. M. Romero, S. Kamalian *et al.*, "An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets," *Nature biomedical engineering*, vol. 3, no. 3, pp. 173–182, 2019.

[10] X. Qian, J. Pei, H. Zheng, X. Xie, L. Yan, H. Zhang, C. Han, X. Gao, H. Zhang, W. Zheng *et al.*, "Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning," *Nature biomedical engineering*, vol. 5, no. 6, pp. 522–532, 2021.

[11] S. A. Hicks, J. L. Isaksen, V. Thambawita, J. Ghouse, G. Ahlberg, A. Linneberg, N. Grarup, I. Strümke, C. Ellervik, M. S. Olesen *et al.*, "Explaining deep neural networks for knowledge discovery in electrocardiogram analysis," *Scientific reports*, vol. 11, no. 1, pp. 1–11, 2021.

[12] A. Binder, M. Bockmayr, M. Hägele, S. Wienert, D. Heim, K. Hellweg, M. Ishii, A. Stenzinger, A. Hocke, C. Denkert *et al.*, "Morphological and molecular breast cancer profiling through explainable machine learning," *Nature Machine Intelligence*, vol. 3, no. 4, pp. 355–366, 2021.

[13] C. Li, D. Konomis, G. Neubig, P. Xie, C. Cheng, and E. Xing, "Convolutional neural networks for medical diagnosis from admission notes," *arXiv preprint arXiv:1712.02768*, 2017.

[14] F. Schwendicke, T. Golla, M. Dreher, and J. Krois, "Convolutional neural networks for dental image diagnostics: A scoping review," *Journal of Dentistry*, vol. 91, p. 103226, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0300571219302283

[15] Y. Gurovich, Y. Hanani, O. Bar, G. Nadav, N. Fleischer, D. Gelbman, L. Basel-Salmon, P. M. Krawitz, S. B. Kamphausen, M. Zenker *et al.*, "Identifying facial phenotypes of genetic disorders using deep learning," *Nature medicine*, vol. 25, no. 1, pp. 60–64, 2019.

[16] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. S. Corrado, A. Darzi *et al.*, "International evaluation of an ai system for breast cancer screening," *Nature*, vol. 577, no. 7788, pp. 89–94, 2020.

[17] "Explainable AI," Mar. 2021, [Online; accessed 4. May 2022]. [Online]. Available: https://www.ibm.com/watson/explainable-ai

[18] B. Ghoshal and A. Tucker, "Estimating uncertainty and interpretability in deep learning for coronavirus (covid-19) detection," *arXiv preprint arXiv:2003.10769*, 2020.

[19] F. Nunnari, M. A. Kadir, and D. Sonntag, "On the overlap between grad-cam saliency maps and explainable visual features in skin cancer images," in *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, 2021, pp. 241–253.

[20] "Home - Face2Gene," Mar. 2022, [Online; accessed 5. May 2022]. [Online]. Available: https://www.face2gene.com

[21] "GestaltMatcher Database," May 2022, [Online; accessed 5. May 2022]. [Online]. Available: https://db.gestaltmatcher.org/publications

[22] "Human Phenotype Ontology," Apr. 2022, [Online; accessed 6. May 2022]. [Online]. Available: https://hpo.jax.org/app

[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[24] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5203–5212.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[26] G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Information Fusion*, vol. 76, pp. 89–106, 2021.

[27] C. Molnar, *Interpretable Machine Learning*, 2019.

[28] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6541–6549.

[29] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas *et al.*, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," in *International conference on machine learning*. PMLR, 2018, pp. 2668–2677.

[30] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan *et al.*, "Captum: A unified and generic model interpretability library for pytorch," *arXiv preprint arXiv:2009.07896*, 2020.

[31] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *2011 international conference on computer vision*. IEEE, 2011, pp. 2018–2025.

[32] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[34] Li, Andreeto, Ranzato, and Perona, "Caltech 101," Apr 2022.

[35] Griffin, Holub, and Perona, "Caltech 256," Apr 2022.

[36] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[37] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[38] Y. Boykov and G. Funka-Lea, "Graph cuts and efficient nd image segmentation," *International journal of computer vision*, vol. 70, no. 2, pp. 109–131, 2006.

[39] E. Štrumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Knowledge and information systems*, vol. 41, no. 3, pp. 647–665, 2014.

[40] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[41] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, "Layer-wise relevance propagation: an overview," *Explainable AI: interpreting, explaining and visualizing deep learning*, pp. 193–209, 2019.

[42] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[43] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," *International Journal of Computer Vision*, vol. 126, no. 10, pp. 1084–1102, 2018.

[44] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene cnns," *arXiv preprint arXiv:1412.6856*, 2014.

[45] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[46] R. L. Draelos and L. Carin, "Hirescam: Faithful location representation in visual attention for explainable 3d medical image classification," *arXiv preprint arXiv:2011.08891*, 2020.

[47] S. Srinivas and F. Fleuret, "Full-gradient representation for neural network visualization," *Advances in neural information processing systems*, vol. 32, 2019.

[48] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, "Evaluating feature importance estimates," 2018.

[49] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.

[50] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.

[51] S. Y. Zhang, Zhifei and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[52] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," *Advances in neural information processing systems*, vol. 31, 2018.

[53] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," *arXiv preprint arXiv:1806.08049*, 2018.

[54] Y.-h. Sheu, "Illuminating the black box: interpreting deep neural network models for psychiatric research," *Frontiers in Psychiatry*, vol. 11, p. 551299, 2020.

[55] C.-K. Yeh, C.-Y. Hsieh, A. Suggala, D. I. Inouye, and P. K. Ravikumar, "On the (in) fidelity and sensitivity of explanations," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[56] S. Mohseni, N. Zarei, and E. D. Ragan, "A multidisciplinary survey and framework for design and evaluation of explainable ai systems," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 11, no. 3-4, pp. 1–45, 2021.

[57] Wikipedia contributors, "Likert scale — Wikipedia, the free encyclopedia," 2022, [Online; accessed 6-November-2022]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Likert_scale&oldid=1109849563

[58] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[59] M. B. Muhammad and M. Yeasin, "Eigen-cam: Class activation map using principal components," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–7.

[60] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[61] J. Gildenblat and contributors, "Pytorch library for cam methods," https://github.com/jacobgil/pytorch-grad-cam, 2021.

[62] A. C. Bovik, *The essential guide to image processing*. Academic Press, 2009.

[63] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[64] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[65] "Triangle: Definitions." [Online]. Available: https://www.cs.cmu.edu/~quake/triangle.defs.html

[66] A. D. Kline, J. F. Moss, A. Selicorni, A.-M. Bisgaard, M. A. Deardorff, P. M. Gillett, S. L. Ishman, L. M. Kerr, A. V. Levin, P. A. Mulder *et al.*, "Diagnosis and management of cornelia de lange syndrome: first international consensus statement," *Nature Reviews Genetics*, vol. 19, no. 10, pp. 649–666, 2018.

[67] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, p. e0130140, Jul. 2015.

[68] D. Castelvecchi, "Can we open the black box of ai?" *Nature News*, vol. 538, no. 7623, p. 20, 2016.

[69] S. Jetley, N. A. Lord, N. Lee, and P. H. S. Torr, "Learn to pay attention," in *International Conference on Learning Representations*, 2018.

[70] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, 2021.

[71] A. Santoro, D. Raposo, D. G. T. Barrett, M. Malinowski, R. Pascanu, P. W. Battaglia, and T. P. Lillicrap, "A simple neural network module for relational reasoning," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 4967–4976.

[72] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, "What clinicians want: contextualizing explainable machine learning for clinical end use," in *Machine learning for healthcare conference*. PMLR, 2019, pp. 359–380.

[73] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.

[74] B. Ait Skourt, A. El Hassani, and A. Majda, "Lung ct image segmentation using deep neural networks," *Procedia Computer Science*, vol. 127, pp. 109–113, 2018, pROCEEDINGS OF THE FIRST INTERNATIONAL CONFERENCE ON INTELLIGENT COMPUTING IN DATA SCIENCES, ICDS2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050918301157

[75] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in mri images," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1240–1251, 2016.

[76] C. Olah, A. Mordvintsev, and L. Schubert, "Feature visualization," *Distill*, 2017, https://distill.pub/2017/feature-visualization.

[77] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[78] Q. Ferry, J. Steinberg, C. Webber, D. R. FitzPatrick, C. P. Ponting, A. Zisserman, and C. Nellåker, "Diagnostically relevant facial gestalt information from ordinary photos," *elife*, vol. 3, p. e02020, 2014.

[79] I. E. Nielsen, D. Dera, G. Rasool, N. Bouaynaya, and R. P. Ramachandran, "Robust explainability: A tutorial on gradient-based attribution methods for deep neural networks," *arXiv preprint arXiv:2107.11400*, 2021.

[80] L. Chen, J. Chen, H. Hajimirsadeghi, and G. Mori, "Adapting grad-cam for embedding networks," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 2794–2803.

[81] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan *et al.*, "A unified and generic model interpretability library for pytorch, 2020," 2009.

[82] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5810–5818.

[83] A. Ghorbani, A. Abid, and J. Zou, "Interpretation of neural networks is fragile," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 3681–3688.

[84] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, aug 2018. [Online]. Available: https://doi.org/10.1145/3236009

[85] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4813, 2021.

[86] V. Miglani, N. Kokhlikyan, B. Alsallakh, M. Martin, and O. Reblitz-Richardson, "Investigating saturation effects in integrated gradients," *arXiv preprint arXiv:2010.12697*, 2020.

[87] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[88] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable ai: Challenges and prospects," *arXiv preprint arXiv:1812.04608*, 2018.

[89] A. Rosenfeld, "Better metrics for evaluating explainable artificial intelligence," in *Proceedings of the 20th international conference on autonomous agents and multiagent systems*, 2021, pp. 45–50.

[90] Y.-S. Lin, W.-C. Lee, and Z. B. Celik, "What do you see? evaluation of explainable artificial intelligence (xai) interpretability through neural backdoors," *arXiv preprint arXiv:2009.10639*, 2020.

[91] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas *et al.*, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," in *International conference on machine learning*. PMLR, 2018, pp. 2668–2677.