Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

Institute for
Genomic
Statistics and
Bioinformatics

Master's Thesis

Quantitative Comparison of Deep Learning Classifiers and Human Attention in Assessing Rare Disorders

Anna Rose Johny

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.-Ing. Sebastian Houben
Prof. Dr. med. Dipl. Phys. Peter Krawitz
Dr. rer. nat. Tzung-Chien Hsieh

November 2023

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	Anna Rose Johny

Abstract

Around 6% of the overall population is affected by genetic conditions. The recent advances in deep learning (DL) based approaches have shown rare-disorder identification and classification more accessible to clinicians for identifying the disorders even though some are extremely rare.

The GestaltMatcher-Arc model performed better than the ResNet-50 for the identification of ultra-rare disorders. The model can help clinicians in identifying some rare disorders when the disorders are extremely rare. However, there exists a black box in the predictions made by the convolutional neural network (CNN) model. This research work aims to eliminate the black box in predictions of the GestaltMatcher-Arc and ResNet-50 models trained on frontal facial images using image saliency methods. The SIDU image saliency method approach performed better than the GRAD-CAM in indicating the regions that influenced the predictions of the models.

As an evaluation of the XAI methods, an eye-tracking study is conducted. The visual explanations generated by the XAI methods are compared to the visual explanations generated from the eye tracker using the intersection over union (IoU) score. An IoU score of 0.62 was obtained from the GestaltMatcher-Arc model with SIDU explanations, which is the highest among other models. Based on the IOU score, a conclusion is made that the model and human might be looking at different features in an image while observing the facial image. In future, we could consider datasets with normal faces for training the network and also an extension to the eye-tracking study.

Acknowledgements

I would like to take this opportunity to thank the efforts of the people who have supported, guided and helped me in doing this master's thesis. I would like to thank my supervisor Prof. Dr.-Ing. Sebastian Houben, Prof. Dr. med. Dipl. Phys. Peter Krawitz, Dr. rer. nat. Tzung-Chien Hsieh for their support and timely input. I would also like to thank different members of the IGSB department for their valuable input and suggestions. I would like to thank Behnam Javanmardi, Alexander Hustinx, and Aron Kirchoff for their valuable support and suggestions for doing this work. They have given me timely suggestions regarding improving my research work. I would also like to thank Duong Dat, and Ben Solomon from the National Institutes of Health for providing ideas for reaching into this project. Thanks for their valuable suggestions during the project and also for evaluating the outputs of this project. I would like to thank my parents and family for their continuous support during my master's studies. Thank you all for your guidance and support throughout this journey.

Contents

List of Abbreviations

AGDev	Arbeitsgemeinschaft für Gen-Diagnostik e.V.
AI	Artificial Intelligence
AOI	Area Of Interest
CAM	Class Activation Mapping
CDL	Cornelia de Lange
CFPS	Clinical Face Phenotype Space
CNN	Convolutional Neural Network
Deep SHAP	Deep SHapley Additive exPlanations
DeepLIFT	Deep Learning Important FeaTures
DL	Deep learning
GMDB	GestaltMatcher DataBase
Grad-CAM	Gradient-weighted Class Activation Mapping
IoU	Intersection Over Union
LIME	Local Interpretable Model-agnostic Explanations
ML	Machine Learning
NIH	National Institute of Health
ReLU	Rectified linear unit

RISE	Randomized Input Sampling for Explanation
SHAP	SHapley Additive exPlanations
SIDU	Similarity Difference and Uniquness
SVM	Support Vector Machine
TS	Turner Syndrome
XAI	Explainable AI

List of Figures

List of Tables

1

Introduction

Around 6% of the overall population is affected by genetic conditions. Rarity and diversity make detecting these genetic conditions even more difficult [24]. The facial features of a genetic condition can be similar to some other existing genetic conditions. However, the diagnosis of rare diseases by clinicians can be based on their experience with those disorders. The rarity of some genetic conditions makes some diagnoses be missed if a clinician is not experienced enough with them.

The current advances in DL approaches made these predictions much easier. The deep learning classifiers can predict the syndromes depending on the learned facial features. One of the DL classifiers such as GestaltMatcher-Arc can predict the ultra-rare genetic disorder that is trained on GMDB [24] and learns the facial features. We can also use the DL approach that is trained on various facial features to predict rare genetic disorders based on the facial features.

These DL classifiers are not able to tell which facial features are responsible for giving such a prediction. For that, we can use XAI methods to highlight those regions in a face that a model would have chosen to give the prediction. XAI techniques are those set of process that allows humans to interpret what is happening inside a Machine Learning (ML) algorithm to arrive at its prediction. XAI methods are used to describe the outcomes of the Artificial Intelligence (AI) models. By using the XAI based approaches, we will be able to specify why the model results from such a prediction.

To compare the predictions of XAI, we can use heat maps obtained from an eye-tracker that is used with similar images. An eye-tracker is fed with a set of images of a person with or without a genetic condition. These images were provided to clinicians and non-clinicians to identify the regions which they feel are important.

The heat maps generated by this method are then compared to the XAI method to have a human vs AI comparison.

In this thesis, we will be addressing the deep learning approaches that give a prediction based on an input image and then compare those predictions to the outputs obtained from an eye-tracker.

1.1 Motivation

"Rare genetic disorders affect around 6% *of the overall population" [24].* Most of these rare disorders have some facial abnormalities. Most of these genetic disorders are diagnosed based on the facial features and the clinical tests performed. However, the rarity of some genetic conditions and their similarity to some other genetic conditions make this diagnosis difficult even for clinicians experienced in such fields.

Some of the DL models such as DeepGestalt[61] and GestaltMatcher-Arc [24] models were able to identify some of these rare-genetic disorders based on the facial phenotypes. These models mainly used GMDB for training the network, which mainly contains frontal facial images of patients with genetic disorders. The GestaltMatcher-Arc model has shown improvement in recognising patient faces, which are not even recognisable by clinicians. The development of such models for the clinical settings helped the diagnosis of such rare diseases much more easy. These models are able to identify and classify rare genetic disorders, but these models lack an explanation of how these predictions are made. There exists a black box between the model predictions.

Explainability of the model refers to explaining the portions in an input to give the responsible regions in an input, the model used in giving predictions. The existing architecture of the GestaltMatcher-Arc model is able to classify frontal facial images into rare genetic disorders but lacks explainability. The explainability of these models is yet to be conducted.

Another experiment is to analyse how clinicians or non-clinicians analyse the frontal facial images of rare genetic disorder individuals. For that, an eye-tracking experiment has to be conducted to analyse what are the features clinicians look at the most while analysing images. A study has yet to be conducted to compare these eye-tracking data to those generated by the explainable methods, to compare if AI and humans are comparable. No comparison has been provided for XAI to

eye-tracking heat maps for healthy and non-healthy individuals. No comparison was provided based on clinicians and non-clinicians analysing images with or without genetic conditions.

1.2 Challenges and difficulties

This section discusses the challenges and difficulties of this research work.

- **Lower performance of the classifier**

  Even though the higher performance of the existing GestaltMatcher-Arc model in recognising the rare-disorder frontal facial images, their prediction scores are less compared to the other existing classifiers that use XAI methods in their prediction. The existing XAI methods use the predictions from the classifiers that have higher prediction scores. In this case, there is a doubt that the classifier will be able to generate more reasonable explanations for the images.

- **Smaller dataset and class imbalance**

  The existing database GMDB contains frontal facial images of different patients with some rare genetic disorders [29]. But the dataset has imbalances between the classes and some images have lower resolution. This dataset imbalance and low-resolution images can have a significant effect on the prediction and explanation generated.

- **Lack of methods to validating outputs**

  The predicted outputs or the explanations generated are not able to be validated due to the lack of methods available. Due to the lack of such approaches, these methods are often validated by humans or in this case, clinicians experienced in such fields [29]. In our case, we will be comparing the explanations to those from an eye-tracker results. A proper evaluation approach cannot be given in such cases due to the lack of methods available. Existing approaches on XAI methods do not give a comparison of Area Of Interest (AOI) in the facial features of people with or without genetic conditions.

1.3 Problem statement

The proposed work aims to analyse the eye-tracking data obtained from an eye-tracking device for rare-disorder identification and classification. Firstly, a literature review will be conducted to identify the image saliency method to generate attention heat maps. The selected image saliency method will be then used to integrate with the rare-genetic disorder classifier to generate predictions based on the input images. Finally, these generated attribution maps will be compared to the heat maps obtained from an eye-tracker. This master thesis intends to address the following research questions:

RQ1. What are the image saliency methods available to identify the AOI in a given image to make predictions?

RQ2. What are the important AOI in a face to make predictions related to genetic disorders?

RQ3. How can we compare the attribution maps obtained from image saliency methods to heat maps from an eye-tracker? Do the saliency maps of clinicians and eye-tracker saliency maps align?

RQ4. Do the clinicians and non-clinicians look at the same facial features?

1.4 Structure

This report consists of seven chapters including the current chapter. Chapter 1 introduced the topic, motivation, challenges and difficulties of this work, the problem statement, and research questions relevant to this research work, and the outline of this overall report. In Chapter 2, a brief background about DL methods used for genetic disorder classification, visual explanation methods and comparison of eye-tracking data to visual explanation methods are discussed. The various available explanation methods are discussed in this chapter briefly. Chapter 3 gives a more detailed description of the selected DL method and XAI method. Chapter 4, discussed the method used for solving this problem of this research work. Discussed in detail about the methodology chosen for the research work. Following that, Chapter 5 gives the solution for the proposed problem statement. The chosen methods and their

implementation are discussed in detail in this chapter. The evaluation strategies used to evaluate the research work are discussed in Chapter 6. In this chapter, the comparison of eye-tracking heat maps to the XAI heat maps is discussed. Finally, chapter 7, concludes the contributions and possible future work of this work.

2

Background

In this research work we follow three different aspects: a) Finding the best suitable DL model for dysmorphic and non-dysmorphic faces classification, b) Finding the most suitable XAI method for eliminating the black box in the DL model predictions, c) comparison methods available for comparing the XAI saliency maps with the ey tracking saliency maps.

2.1 DL methods for genetic disorder classification

Different DL techniques have been widely applied to different areas of medical research applications [35]. In this research work, we try to compare the DL models regarding facial analysis in terms of genetic disorder identification [35]. Identification of genetic disorders from the faces is difficult in cases where the genetic conditions are extremely rare. An approach presented by Theyazn et al. [1], compared various deep learning-based approaches for classifying autism spectrum disorder. The approach uses facial images with and without autism disorder from the Kaggle dataset. The approach uses CNN for classifying the images as autism disorder images or not. The proposed architecture contains the CNN with a pooling layer followed by a fully connected layer to detect the features in a facial image. Finally, the activation function is implemented to the fully connected layer to detect and categorize as syndromic or non-syndromic images [49]. The proposed approach was able to get an accuracy of 0.94% on the test and was able to correctly identify and classify the images [1]. They have also built a web application, where we upload the images and the application tells us, whether it's a syndromic face or not. The approach uses CNN combined with transfer learning [1]. This approach compares various methods such as MobileNet [56], Xception [10], and InceptionV3 [60], and found MobileNet

with higher prediction accuracy and is used in the web application. The proposed approach can only be applied to specific applications, in this case, to detect autism disorder or not [44].

Another approach presented by Maciej et al. [17] uses a DL based approach for screening the facial images and telling if they are genetic syndrome or not. The paper compares four different face recognition models such as DeepFace [64], ArcFace [28], DeepID [41], FaceNet [48]. There is a reasonable difference in how these models generate the faces. The better results in classifying the multi-class problem were using the ArcFace model. For the two-class classification DeepFace model gave better accuracy results [17]. The approach used only 15 different rare genetic syndromes. The approach can be scaled to identify more disorders. The drawback of the approach is that some of the genetic syndromes remain unidentified [5]. Another approach presented in [15] classifies images into two different syndromes. The approach uses EfficientNet-B4 for classifying the images into syndromic or not. The approach uses the FairFace dataset and a custom dataset for training the network. StyleGAN face detector is applied to the images to detect the faces in the images. The obtained results from the classifier are compared to clinician analysis. The proposed approach outperforms the results of clinicians. An approach presented by Pan et al. [42], develops a deep convolutional network for recognising the facial features of the Turner Syndrome (TS). The proposed approach was used in the clinical setting and achieved higher accuracy in terms of detecting the syndrome correctly. This approach is application-specific and can only be applied to a particular syndrome. The authors in [30] compared various CNN architectures such as VGG-16, VGG-19, ResNet-10, and MobileNet-V2 for the identification of Williams Beuren syndrome. The authors say that the VGG-19 architecture obtained higher accuracy, whereas MobileNet-V2 achieved the worst classification accuracy. The results from these models were also compared with clinicians in these fields, and the VGG-19 model outperformed the clinicians in predictions. The authors state that the VGG-19 model architecture can perform well in the diagnosis of Williams-Beuren syndrome. In the deep learning approach proposed by [54], the authors were able to classify various syndromic and non-syndromic faces correctly. This model can be applied by clinicians for an initial diagnosis of a few syndromes. The model used a small dataset of 1126 images for seven different syndromes, the approach presented was able to

detect and classify most of the syndromes correctly. The GestaltMatcher-Arc model presented by [24] which is trained on the GMDB, is the most prominent one for various disorder classifications that is available up to date. The GestaltMatcher-Arc trained on syndromic faces will be used in this research work for identifying and classifying rare disorders.

2.2 Visual explanation methods

Visual explanation methods help the end-users to easily identify the predictions of a model. The main properties of visual explanation methods are described by [46] as follows:

- **Translucency** - Translucency describes the relevance of explanations of a model based on the internal structure parameters such as structure and weights. Highly translucent models are easily understandable models such as linear regression models, which are model-specific. The advantage of highly translucent methods is that they can rely on more information in generating the explanations. Whereas, low translucent methods are those in which the input is manipulated and we look at the predictions of the model to obtain predictions. These low translucent models are more robust to apply to various applications and are considered as the black box models.

- **Expressive power** - Expressive power is defined as the structure or the language of how these explanations are created by the methods [9]. The expressive power defines how the model is able to make decisions such as decision trees, weighted sum, and IF-THEN [9].

- **Portability** - Portability is defined as the range to which the descriptions are generated from the ML models [9]. Low translucent models have higher portability because these models are considered black boxes and can be employed in different real-time applications [9]. An example of a higher portability property is the surrogate models with the visual saliency methods [2]. Methods that only work for a particular application such as recurrent neural networks have very low portability [2].

- **Algorithmic complexity** - Algorithmic complexity refers to the algorithmic complexity of the various approaches that generate the explanations [9]. This is one of the important properties to consider when the time used to compute is bigger to generate the explanations [2].

One of the most popular approaches is to make use of saliency maps [55] [8] [19] for the visualizations. The saliency-based approaches can be mainly classified into three categories: a) perturbation-based methods, b) backpropagation-based methods, and c) approximation-based methods.

2.2.1 Perturbation-based methods

Perturbation-based methods aim at observing the changes in the output by changing the input of the DL model under consideration [47]. The changes in the model's output indicate the parts of the incoming signal are the most prominent features that influence the input [47]. In the survey on perturbation-based methods by Ivanos et al. [25], they have compared various methods for DL networks. All these perturbation-based approaches can be applied to image data and can be used for various applications for understanding the model predictions [36]. Perturbation-based methods thus find the regions in the image, when the input is perturbed [36]. A summary of perturbation-based methods is shown in Table 2.1.

Randomized Input Sampling for Explanation (RISE) [43]

Some of the perturbation-based methods use the changes occurring at the intermediate level [43]. An example of such an approach is RISE [43]. The RISE approach works on the black-box models for generating the saliency maps by combining the input with the randomly masked versions for obtaining the outputs [43].

LIME

As the name suggests, LIME [14] uses local explanations for making predictions. In lime, local explanation refers to that, the explanations are based on the behaviour of the model at the instance when the values are predicted. The explanations generated by these models are useless when humans cannot interpret these explanations. LIME

[45] approach creates various versions of the input image and only selects those which have some attributes of the original image [2]. LIME creates a new dataset from the new observations consisting of perturbed samples and then gives the corresponding predictions to the model [45]. A toy example of the process of LIME is shown in Figure 2.1. The blue/pink background represents the model's decision function, and as observed from the figure, the values are non-linear. The bright red cross is the time when the explanations are made. Instances around this Red Cross are generated and weighed according to their relevance towards the Red Cross. The explanations generated by these will not be globally relevant, but effective in local predictions.

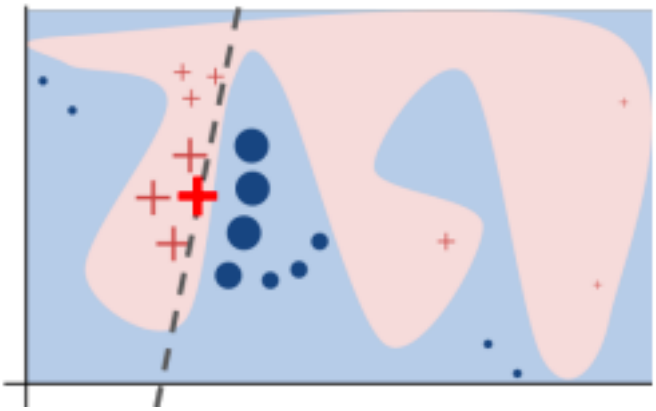


Figure 2.1: Toy example of LIME. Image source: [45]

Occlusion sensitivity maps

Occlusion sensitivity maps are used for identifying the snippets of the input image that are most prominent while making decisions for the deep neural network [62]. Occlusion maps [62] are also the frequent common approach used for visualizing the regions that are most important while making decisions [2]. These heat maps are created by repeatedly replacing different patches of the image that are considered with grey-coloured squares and then observing the predicted class labels [2]. These maps are then checked, if the model's prediction is somehow aligned with the object location or anywhere nearby the object [2]. These maps are created from the last convolutional layer based on the model's predicted probability for the correct class [2].

Real-time saliency

Another approach is real-time saliency [12], in which a masking model is created for manipulating the classifier scores and the approach masks some parts of the image for saliency detection. This approach is a better option for predicting unseen images with a single-forward passing of the images and can be applied to real-time systems.

Method	Data type	Applications	Evaluation method
Occlusion [62]	images	face recognition [58], object tracking [20]	qualitative analysis
LIME [45] [14]	images	text and image classification	human evaluations
Real-time saliency [12]	images	real-time applications	localization error, saliency metric
RISE[43]	images	image classification	pixel insertion and deletion scores

Table 2.1: Summary of perturbation-based approaches.

2.2.2 Backpropagation-based methods

Backpropagation-based methods were one of the earliest methods for highlighting the regions in an input that affect the prediction results. The backpropagation-based methods backpropagate the signals from the output through the layers to the input, thereby making an efficient prediction. A summary of backpropagation-based approaches is shown in Table 2.2. In our research work, we use one explanation method as Grad-CAM which is explained in detail in the next chapter [15].

Layer-Wise Relevance Propagation

An example of a backpropagation-based approach includes Layer-Wise Relevance Propagation [37] [4]. In backpropagation-based approaches, the feature activation maps generated at any layer are considered for the generation of saliency maps, for example, the saliency map generated from the last convolutional layer.

Deep Learning Important FeaTures (DeepLIFT)

Another example of a backpropagation-based approach is DeepLIFT [53], in which the generated descriptions of the subject under study are based on fixed features likely weights, activation and bias [9]. DeepLIFT method is a model-dependant, in which the visual saliency maps to the predictions are obtained using analysing the contributions to respective neurons in the DL network by backpropagating the input values [9].

CAM

Another example of this approach is CAM [27] [63], which generates visual explanations from the activation maps using the linear combination of these activation maps. The CAM based approaches are mostly applied for CNN models. The idea of CAM replaces the need for fully connected layers and uses convolutional layers and global average pooling instead.

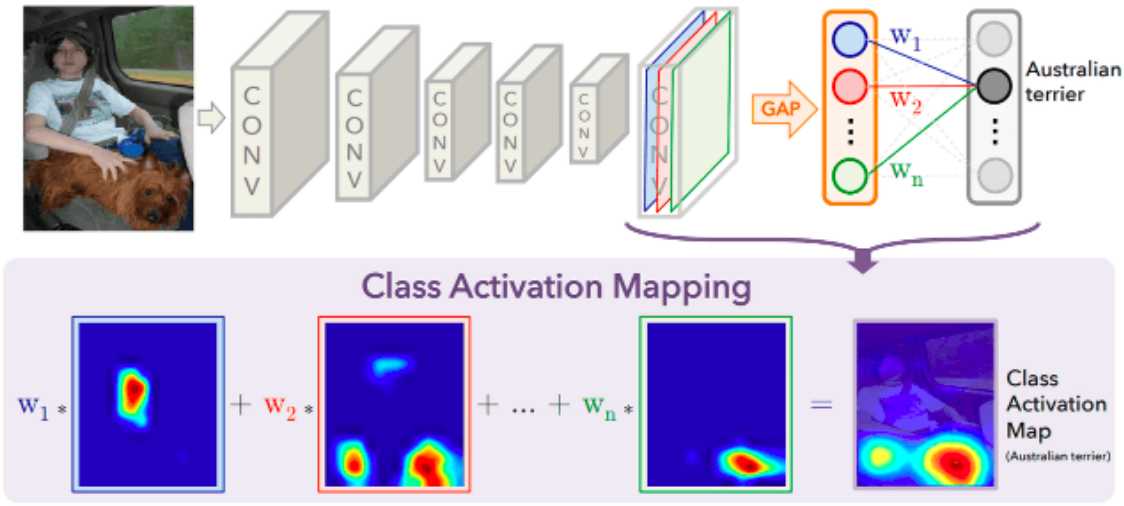


Figure 2.2: Generation of saliency maps using CAM. Image source: [63]

Grad-CAM [51]

When we use gradients applied to the output of the DL models along with the cam, Grad-CAM visual saliency maps are generated[50]. The Grad-CAM based

approaches are most common for the images as the input data and generate the visual explanations by applying gradients to the predictions of the DL model. In the current research work, we have focused on generating visualizations for the class-specific application, by applying a gradient to the output to generate these visual saliency maps [50].

Methods	Data type	Applications	Evaluation method
CAM [27]	image	object detection, image classification	visual inspection, ground-truth comparison
Grad-CAM	image	object detection	visual inspection, comparison to other baselines
Layer-Wise Relevance Propagation [37]	image, text	object detection, text identification	relevance score, visual inspection
DeepLIFT	image	genomics, text interpretation	visual inspection, domain expert feedback
Deep SHapley Additive exPlanations (Deep SHAP) [31]	images	computer vision, natural language processing	importance scores, area under the attribution curve
Guided backpropagation [57]	images	computer vision, DL model interpretation	human inspection

Table 2.2: Summary of backpropagation-based approaches.

2.2.3 Approximation-based methods

Approximation-based approaches for visual explanations use decision trees or linear regression-based approaches to solve complex black-box models. Table 2.3 shows the summary of approximation-based approaches.

SHapley Additive exPlanations (SHAP)

An example of such an approach is SHAP [11], which gives the visual saliency maps for the predictions using Shapely values. The approach of SHAP can be scaled up or down to any black-box model and can compute the explanations or the influences on any classes more efficiently [9]. This approach is established on the additive feature

attribution method, where the feature attribution depicts the change in the class probability concerning the average prediction probability [9].

SIDU

Another approach presented by [39], calculates the "Similarity difference" and "uniqueness" scores for giving importance to the features that are more valuable to give predictions. This approach is gradient-free and can be applied to localize the entire region in an image to give predictions. The SIDU is another state-of-the-art method used in this research work as the another XAI method.

Methods	Data type	Applications	Evaluation method
SIDU [38]	images	image classification, object detection	visual inspection, insertion and deletion scores
SHAP [11]	images, text	healthcare, medical diagnosis	domain expert evaluations

Table 2.3: Summary of approximation-based approaches.

2.3 Eye-tracking analysis and comparison to visual explanations

Eye-tracking data has been widely used in the medical domain for various studies. The eye-tracking study can be used to yield additional information in an image that human observers may not identify. One example of comparison of eye-tracking study to the XAI method has been conducted by [38]. The authors have compared the human evaluations from the eye-tracking data to the XAI method for knowing how humans and AI evaluate different features in an image [2]. The approach can compare how accurate the results from the ML methods [38]. Another approach presented by [33] has also compared eye-tracking data to XAI methods, and reached a conclusion that XAI method and humans look at various features in an image. The authors of [40] have suggested an eye-tracking study for comparison of AI decisions to human evaluations. In this paper, the authors were able to decide if the recommendations given by the AI to decide on safe or unsafe conditions are appreciable by the usage of an eye-tracking study. The paper is good for comparing the eye-tracking study to

the AI method, but reproducing the results from this paper is difficult.

3

State of the art

This chapter explains various attribution methods considered in this research work. The unsupervised techniques tend to detect all the features in more complex images [59]. To tackle the issues with unsupervised learning approaches, we will use supervised learning-based saliency detection methods in our case. Ullah et al. [59] have briefly classified image saliency methods based on classic learning-based and deep learning-based methods.

3.1 Classic learning-based saliency detection methods

The classical learning-based approaches are mainly supervised or semi-supervised methods, that use high-level features in images to get higher accurate saliency map predictions [59]. A Support Vector Machine (SVM) classifier has been used by Judd et al. [26] for eye-fixation prediction. The approach uses the eye-fixation location of the different participants as the training dataset. This approach compares the human saliency map with the saliency map obtained from the learning method and the paper states that the human and model look at different aspects in an image. The approach uses high-level, mid-level and low-level features from the eye-tracking data to train the SVM model. By combining all these features, the model was able to give a prediction that is almost similar to the human eye-tracking data. The approach shows the difference in viewing the images when they are cropped, and scaled. By performing such augmentation techniques, An approach proposed by Borji et al. [7] combines low-level features such as orientation, intensity, and colour with high-level features such as human faces, cars, and animals to train an AdaBoost classifier. In the approach, the authors proposed a direct mapping from the features to the eye-fixations. The approach uses SVM and AdaBoost classifier to have the

predictions. The AdaBoost classifier got the highest accuracy when compared with the human eye fixations [7]. The proposed approach also detected salient objects in an image with higher accuracy. By combining high-level features with low-level features the AdaBoost classifier was able to close the gap between the human eye-fixation obtained from an eye-tracker with the prediction model [7]. The classic learning-based algorithms use prior knowledge available such as labelled datasets, or some features available. By using this prior knowledge the classic learning approaches can boost the saliency detection approaches [7]. These approaches tend to reduce the performance of the models if they are not correctly used. Recent developments of DL-based approaches tend to outperform the existing classic learning approaches. These methods give more promising results in comparison to the classic learning approaches.

3.2 DL-based saliency methods

In this section, we will be introducing the deep learning-based approaches with a main focus on CNN, GestaltMatcher-Arc and ResNet-50.

3.2.1 GestaltMatcher-Arc

The GestaltMatcher-Arc model uses DeepGestalt [21] as the base model. The architecture of the DeepGestalt model is shown in Figure 3.1. The DeepGestalt is a deep convolutional neural network with ten convolutional layers. The network's last layer is followed by Batch Normalization and Rectified linear unit (ReLU). A max-pooling layer is applied after each pair of the convolutional layers and an average pooling is after the third layer. The fifth layer is a fully connected layer with the softmax function and a dropout of 0.5 applied. The proposed approach compares various low-level features with high-level features. The input to the model cropped facial features, which is basically used for training the framework [21]. The approach is used for the classification of the various genetic syndromes, but the approach is unable to identify and classify ultra-rare genetic disorders [21]. To identify ultra-rare genetic disorders, the GestaltMatcher-Arc model is used. The GestaltMatcher model is an extension of DeepGestalt that uses the last 320-dimensional fully connected layer before the classification layer [24]. This classification layer is employed as the

feature layer in the case of this application [24]. This feature layer serves as an encoder that has a feature representation vector for a 320-dimensional vector [24].

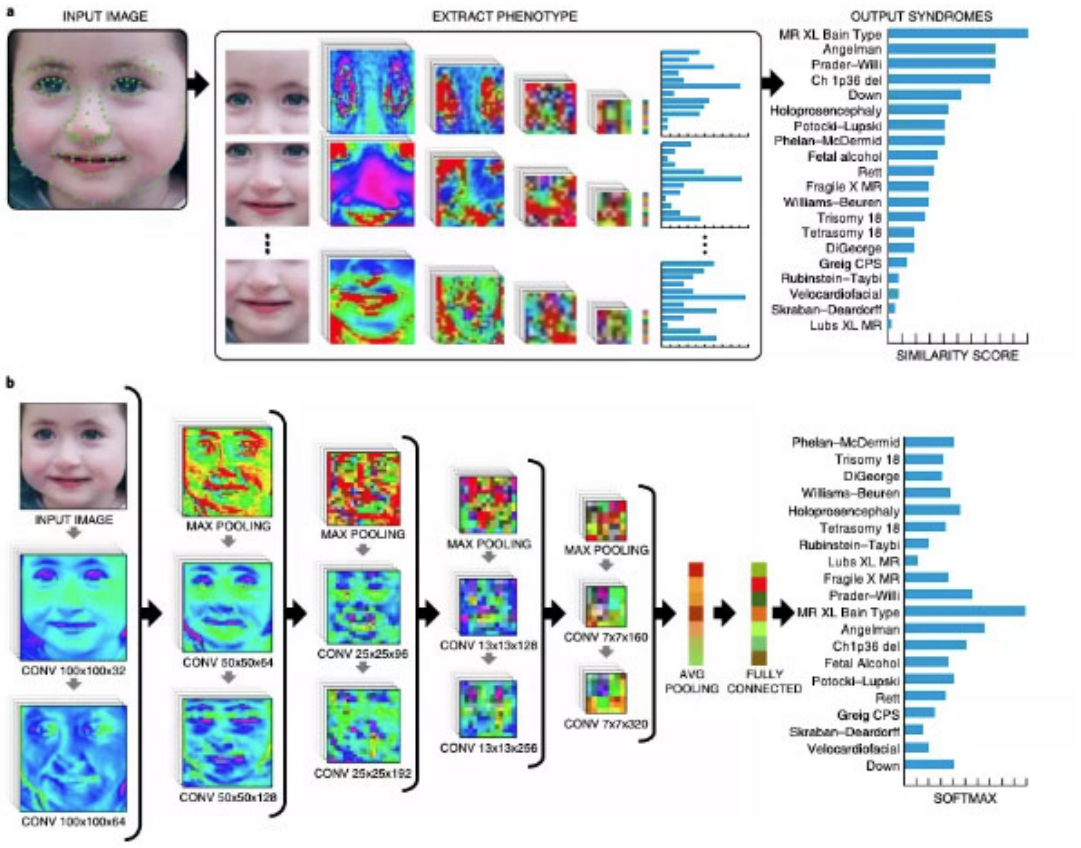


Figure 3.1: Architecture of DeepGestalt. Image source [21]

The representation vectors are distributed across Clinical Face Phenotype Space (CFPS). By using the CFPS, we can match the ultra-rare disorder patients with the patients with similar facial features. The GestaltMatcher model uses different approaches to compare the patient's similarity. The GestaltMatcher model is the deep convolutional neural network that learns from the different features of the input image from the GMDB to detect ultra-rare disorders [24]. The model can identify the ultra-rare disorders that are unseen during the training. By using this approach the researchers can compare patient-patient or syndrome-syndrome similarity [24]. In our case, we will be using the GestaltMatcher model because this is the most

promising model close to the recognition of the ultra-rare disorder and classification of the ultra-rare disorder [24]. The basic design of the GestaltMatcher model is shown in Figure 3.2.

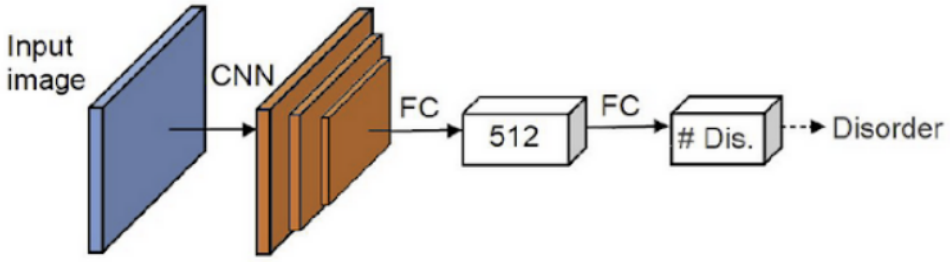


Figure 3.2: Model architecture of GestaltMatcher-Arc.

3.2.2 ResNet-50 Architecture

ResNet-50 is a variant of ResNet [22] CNN architecture with 50 layers [32]. The architecture of ResNet-50 is shown in Figure 3.4. The architecture of ResNet-50 introduced by [23], consists of 48 convolutional layers, with a MaxPool layer at the start and an Average Pool layer at the conclusion point of the network. The basic ResNet architecture solves the issue of vanishing gradients, whereas the ResNet-50 architecture solves the problem of 23 million parameters [32]. The performance of this ResNet-50 architecture is much higher compared to the various existing architectures. The ResNet architecture mainly consists of two blocks a convolutional block and an identity block. The ResNet-50 architecture mainly uses bottleneck design as the building block. The design is called a ”bottleneck”, because the residual block uses 1×1 convolutions, which cut back on the parameters and matrix multiplications [32]. This helps in reducing the training time used by each layer of the network [32].

An in-depth explanation of the layers of the ResNet-50 architecture is as follows [32]. First is a 7×7 convolutional kernel with 2 sized strides, followed by a max pooling layer with 2 strides. Followed by the max pool layer is 9 layers with the size of 3×3 with 64 kernel convolution for the first and 1×1 with 64 kernels, and the next with 1×1 with 256 kernels and these layers repeat for three times [32]. Followed by this are 12 more layers of size 1×1 and 3×3 with 128 kernels and 1×1 with 512 kernels, which is repeated for a batch of four. Followed by this layer are 18

more layers which are iterated over 6 times with layers of size $1 \times 1 \times 256$ and two cores of $3 \times 3 \times 256$ and a final layer with size $1 \times 1 \times 1024$ [32]. Followed by this layer are 9 more layers that are iterated over three times with the size of $1 \times 1 \times 512$, $3 \times 3 \times 512$ and the final layer with size $1 \times 1 \times 2048$. The above-mentioned layers all together form the 50 layers of the ResNet-50 architecture [32]. Followed by these 50 layers is an Average pooling layer, a fully connected layer and a softmax activation function [32]. The performance takes its cue from the residual blocks present in the network [32]. Let's consider the neural network block with input as x and we are required to find the distribution $H(x)$ [32]. First, consider the residual between the blocks as, $R(x) = Output - Input$, $R(x) = H(x) - x$. From this, we can acquire the true distribution $H(x)$ as, $H(x) = R(x) + x$ [32].

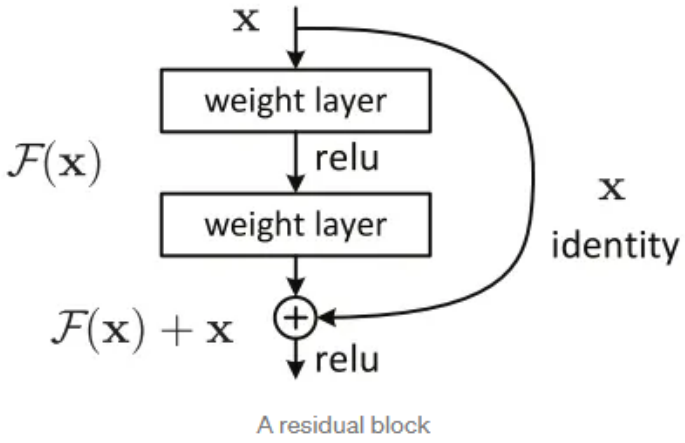


Figure 3.3: Residual Learning block of the ResNet-50 architecture. Image source [23]

Figure 3.3 shows an example of such a residual block of ResNet-50 architecture. The layers of the residual network are learning the difference in the input features and the output features, whereas the traditional ResNet network tries to learn the true distribution $H(x)$ [32]. Here by using residual block, we can learn input features and output features, whereas the traditional ones only consider the features of the input [32]. The residual block allows us to reuse the results of preceding layers activations [32]. In the approach proposed in this work, we will be using

Figure 3.4: ResNet-50 architecture. Image source [23]

the ResNet-50 model pretrained on ImageNet to classify the images into different syndromes and non-syndromic faces.

3.3 Explanations for the predictions

Most of the DL models are "black box" methods. This means we are not sure about where the network is looking at while making predictions which parts of a network are responsible for making predictions or how the model arrived at the final prediction results.

3.3.1 Grad-CAM [51]

Grad-CAM is an approach proposed by Selvaraju et al. [51] for understanding where a model is looking in an image to make predictions. Grad-CAM solves the problem of black box in DL model, by giving an idea about where the model is looking at to give the predictions. Grad-CAM uses gradients to create visualizations highlighting the regions that are important in an image [51]. Grad-CAM uses the last stage of the convolutional network for predictions [51]. The Figure 3.5 shows the overview of Grad-CAM. As shown in the Grad-CAM overview, initially the input image is fed into the CNN architecture, which identifies the image and classifies the image into some class based on the prediction [51]. A raw score is calculated for the correctly predicted class. A gradient of 1 is set to the right classification and 0 for the other classes [51]. This gradient value is then backpropagated to the ReLU and generates the heat map based in accordance with the model's forecasts [50]. This heat map is multiplied point-wise with the guided backpropagation to produce the guided Grad-CAM [52].

The authors [6] have shown that the CNN can capture a higher level of visual interest from the images. The convolutional layers of the CNN network mostly hold the spatial information, which is lost in most of the fully-connected layers [51]. Therefore, a piece of detailed spatial information and higher-level semantics are expected to be within the final convolutional layers [51]. The different aspects related to the predicted class are found by the neurons in these convolutional layers [51]. Grad-CAM approach uses the gradient information of the predicted class to give the last convolutional layer some input so that assigning higher values for the

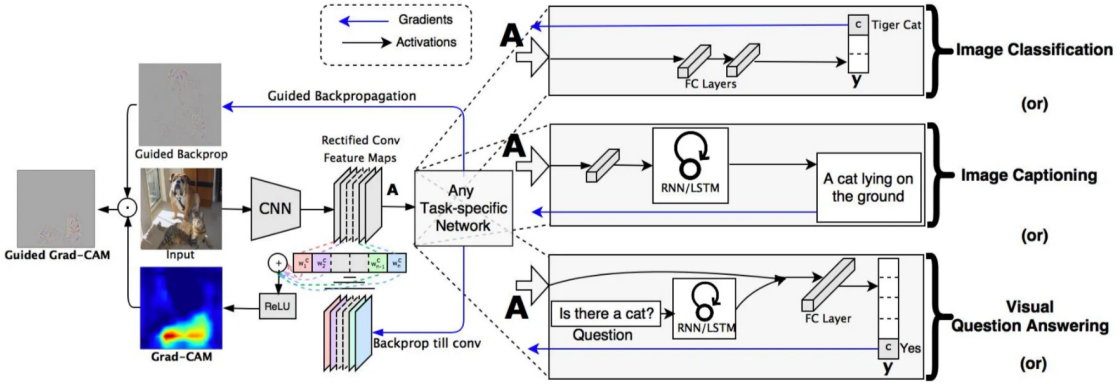


Figure 3.5: Grad-CAM predictions. Image source [52]

predominant class category under consideration [51]. Grad-CAM approach uses only the information obtained from the ultimate convolutional section of the neural network to make predictions [50]. As shown in figure 3.5 Grad-CAM, for acquiring the class-specific heat map, we first compute the gradient values for each class and give 1 for the right classification, which is calculated before the softmax layer [50]. The gradients are calculated concerning the feature map activation's A^k. The score for each class is given as y^c, in cases where c represents the class [52]. From the feature activation representations and the score, the gradient is calculated as $\frac{\partial y^c}{\partial A^k}$ [51]. The neuron importance weights are calculated by applying global average pooling across both width and height dimensions for the gradients obtained [51]. The neuron importance weights a_k^c is calculated as,

$$a_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{i,j}^k}. \tag{3.1}$$

During this backpropagation of gradients, the weight matrices for each class and the gradient values till the last convolutional layer are backpropagated. The neuron importance weights a_k^c, calculate the importance parameters in the feature map k for predicted class c. The final Grad-CAM is calculated as the weighted combination

of forward activation maps followed by the ReLU and is given as,

$$L^c_{Grad-CAM} = ReLU(\sum_k \alpha^c_k A^k) \tag{3.2}$$

This calculation results in a heat map as the dimensions of the last convolutional layer feature maps [52]. A Linear combination is applied to the values because we need only the values that contribute to the prediction of a particular class [51]. A positive intensity is applied to the class of interest, and negative pixels belong to the other classes [50]. In some cases, the ReLU sometimes gives the worst localization for the pixels in the image, by highlighting all the regions in an image, which may not be important in prediction [50]. To conclude, the Grad-CAM uses the gradients to highlight the regions within an image, that are influencing the prediction of a particular class [51].

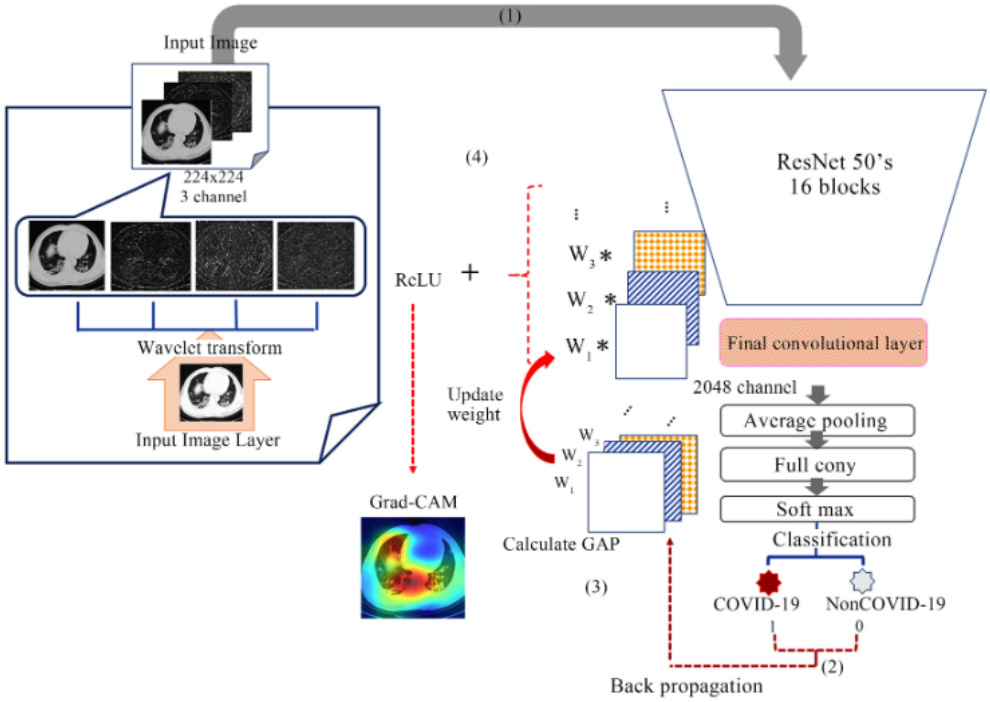


Figure 3.6: ResNet architecture with Grad-CAM. Image source [34]

Figure 3.6 shows an example of a Grad-CAM used along with the utmost layer of the ResNet-50 model. The architecture is employed for the categorization of COVID-

19-infected or non-infected Chest CT images [34]. The input to the model is chest CT images. In this application, the procedure of generating the Grad-CAM remains the same as defined before. The classification module contains the fully connected layer and the features extracted from these layers are converted into probability scores at the softmax layer. The final prediction of the correct class is based on these scores and the topmost probability value is chosen to give the prediction [34]. The gradient is calculated from these feature patterns of the concluding layer of the convolutional network to highlight the most important region while giving the predictions [34]. The most highlighted region, which is marked by red colour is the most significant AOI while making the predictions [34]. The method has correctly identified the regions in chest CT images for classifying the COVID-19 scans correctly.

3.3.2 SIDU [39]

SIDU is an XAI framework outlined by Satya et al. [39]. The mask generated in the approach is generated from the last activation mask of CNN [39]. Similarity Difference and uniqueness scores are computed from these generated masks to generate an explanation for the image provided [39]. The proposed approach can provide a better localization for the object classes present. The figure 3.7 shows the overall architecture of SIDU approach. The approach is mainly divided into three steps: generating feature activation masks, computing similarity differences and uniqueness and finally the interpretations for the predictions [39]. Each of these steps will be delineated comprehensively in further sections.

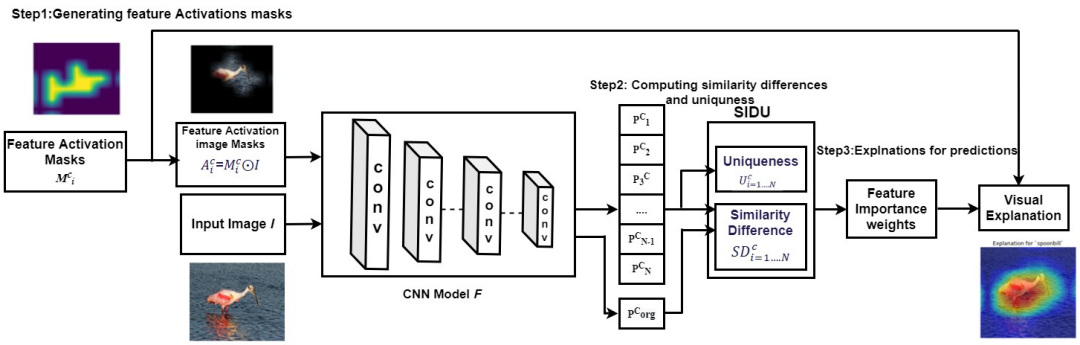


Figure 3.7: SIDU block diagram. Image source [38]

Generating feature activation masks

The first step is to obtain the feature representations that are being created from the final convolutional layer of the DL model under consideration for creating the visual explanations for the outcomes expected from the model [39]. Consider the DL model as F, the dimensions of the concluding convolutional layer as $n \times n \times N$ [39]. Where n depicts the dimensions of the final convolutional layer and N encapsulates the entire feature activations count for each class [39]. The feature activations for the class are of the form $f^c = [f_1^c, f_2^c, ...f_N^c]$. For example, if we have the dimensions of the concluding convolutional layer as $5 \times 5 \times 2048$, then we produce an aggregate of 2048 feature activations of the size 5×5 [38]. From this, we can see that the image class predictions are generated on the basis of activation masks generated [38]. These activation maps are then converted into binary masks using thresholding applied to each value [39]. The binary mask is given as $B_{i=1.N}^C = f_{i=1.N}^c > \tau$ [38]. The threshold values are chosen randomly and do not disrupt the final generation of the heat map [38]. Binary interpolation is applied to binary masks to have values between $[0, 1]$ [39]. These binary masks after applying binary interpolation are called feature activation masks. The feature activation mask is calculated by applying point-wise multiplication to the Image and binary mask. The feature activation mask A_c^i is calculated as,

$$A_c^i = F(I \odot M_i^c) \tag{3.3}$$

The F in the equation denotes the DL model. Figure 3.8 shows the feature-generating procedure for different feature representation maps. These feature representation maps generated in this step are utilized for acquiring a prediction score, which is subsequently employed for the computation of the similarity and difference and is covered in the next section [38].

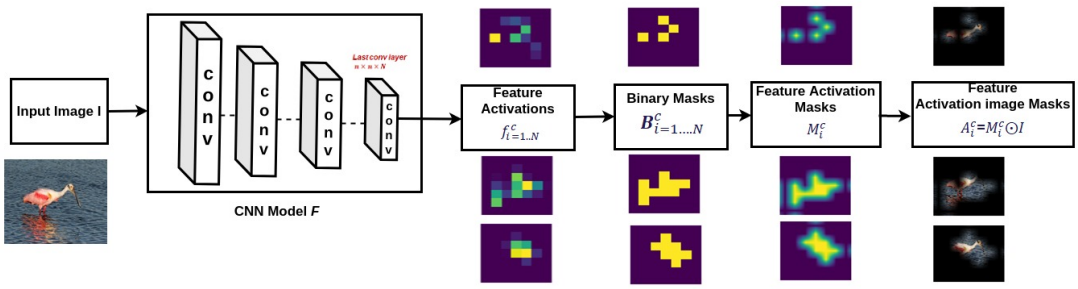


Figure 3.8: Generating visual feature mappings from the final convolutional layer of CNN. Image source [39]

Computing feature important weights

The step involves the computation of similarity and contrast between the feature masks generated to calculate the prediction score [39]. The amount of feature representation masks is equivalent to the total count of these representations present in the final convolutional layer [39]. Consider the dimensions of the final convolutional layer of the DL model as $n \times n \times N$ [**?**]. Here N represents the total count of feature representation masks [39]. A probability score is calculated for all of these feature representation segmentation masks for corresponding to individual classes present in the DL model classifier. The feature representation mask for each class is given as $A^C = [A_1^c, ...A_N^c]$ and the prediction probability score as P_i^c [38]. Here the P_i^c represents the probability score for image classification for each image and for each image, I, the probability score is denoted as P_{org}^c [38]. The vector dimensions of the prediction probability score are influenced by the class count present while training the DL model [39]. In our case, we have trained the DL model on the custom dataset with 11 different classes. So to define the prediction score in our case, a vector P_i^c for each feature generation mask is 1×11, where $i = 1....N$. Figure 3.9 depicts the overall procedure of calculation of the prediction score.

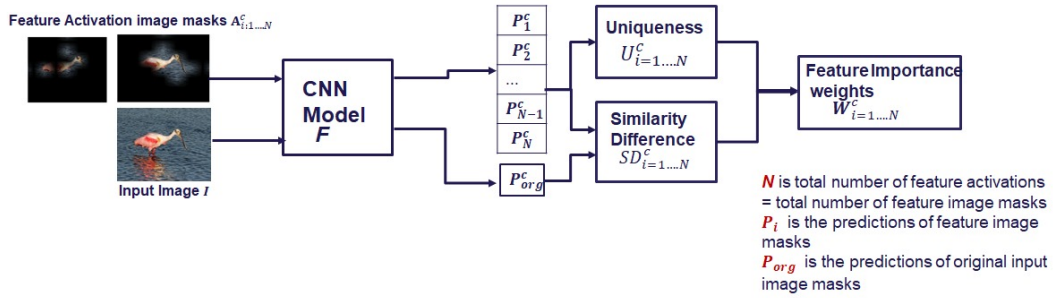


Figure 3.9: Calculating feature significant weights for an image. Image source [39]

This procedure of calculating the prediction score repeats for all of the feature activation masks and the input image. After we have all the prediction scores, the similarity and difference score is calculated for each of the prediction scores. Similarity difference, $SD^c_{i=1...N}$ is calculated between the original image and probability score [38]. This calculation of similarity difference gives the relevant feature action image masks to compare how the probability score changes when we are unaware of the features [38]. *"The similarity difference measures the similarity between the images and prediction score, to get an idea about the prediction score when the features are not known to the DL model"*[38]. The enhanced relevance of the feature representation mask indicates the forecasted class is correct and the lower value indicates the predictions are not the same [39]. The similarity measure is calculated as,

$$SD^c_i = exp(-\frac{|P^c_{org} - P^c_i|}{2\sigma^2}). \tag{3.4}$$

In the above equation, σ is the controlling parameter for detecting the correct predicted class [39]. In the above equation, P^c_i is the prediction score generated from the concluding layer of the neural network layer [39]. Similar to the similarity measure, the uniqueness score is also calculated. *"Uniqueness score U^c is calculated from the feature activation masks of the prediction score"* [39]. The uniqueness measure is calculated as,

$$U^c_i = \sum_{j=1}^{N} |P^c_i - P^c_j|. \tag{3.5}$$

The N in the above equation represents the complete number of the feature representation masks generated from the model [39]. A feature relevance weights score metric W_i^c is calculated as the dot multiplication of the similarity and uniqueness measure and is formulated as [38],

$$W_i^c = SD_i^c \dot{U}_i^c [38]. \tag{3.6}$$

The feature important weights generated have a size of a total count of masks N [38]. The higher values of the feature relevance weight score signify the higher influence of a feature in predicting the outputs and the lower value conveys the lower significance in predicting such a class [38]. Following this is a visual saliency map generation for the outcomes expected from the model which is further elaborated in the next subsection.

Visual explanation

The final heat map of the outcomes is computed by summing the feature representation masks and the feature relevance weights score [38]. The heat map fabricated is indicated as S_c and is computed for the given class as,

$$S_c = \frac{1}{N} \sum_{i=1}^{N} W_i^c \dot{M}_i^c. \tag{3.7}$$

The visual saliency map generation from the feature important weights and feature activation masks are shown in Figure 3.10 [38].

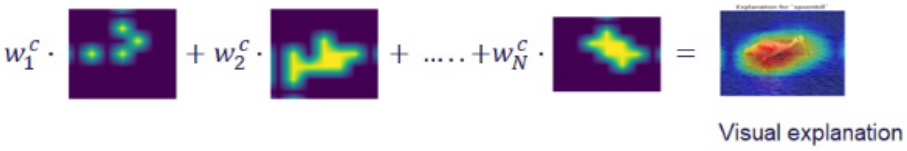


Figure 3.10: Visual representation for the class predicted using SIDU approach. Image source [39]

To summarise the overall procedure of the saliency map using SIDU approach as shown in Figure 3.11, firstly feature representations are constructed from the

final convolutional layer of the DL model [39]. Binary masks of these feature representation masks are constructed using binary interpolation, and from these feature representation masks are constructed, followed by a feature representation image mask calculated as the dot multiplication between the original image and feature representation mask. From these feature representation masks similarity and difference values of the probability scores of outcomes are calculated [38]. From these score features important weights are calculated, which gives the higher value to the most important feature in an image [38]. The final saliency map is created from these weights and the feature representation masks, which highlight the AOI of the image that is more relevant to the classification prediction [38].

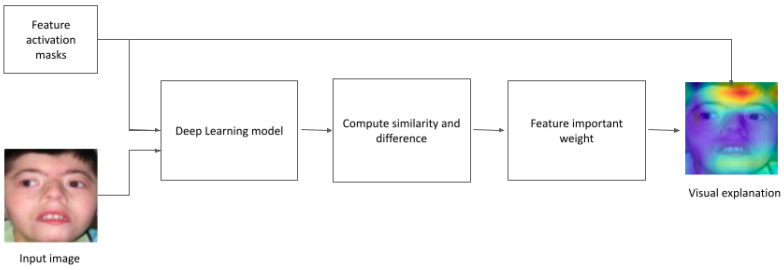


Figure 3.11: Overall architecture of SIDU in our case.

4

Methodology

The previous chapter discussed the literature review of the selected methods for rare-disorder classification and explanation generation methods. To make the models more interpretable, it is better to use XAI methods with the prediction of models. This chapter discusses the methodology adapted to solve the problem of this thesis. In this chapter, we will be discussing the selection of these methods and the combinations used for generating different combinations in our thesis. This chapter discusses the methodological approach used to address the research problem.

4.1 Selection of methods

In this section, we will be explaining how the DL methods and XAI methods are chosen in this particular research work.

4.1.1 Rare disorder identification and classification methods

Amongst the different methods discussed in Chapters 2 and 3, we took a total of 2 different rare-disorder classification models. The method and the reason for selecting this method is as follows:

- **GestaltMatcher-Arc** - This method is the most promising approach available till data for rare-disorder identification and classification. As discussed in the previous chapter, the model is trained to identify faces in an image and is trained on rare-disorder dataset GMDB. The GestaltMatcher-Arc model is chosen, because this model is more accurate in classifying and predicting rare-genetic disorders based on frontal facial images and outperformed the clinicians in diagnosing the genetic conditions.

- **ResNet-50** model is the most common model that is employed for image classification [52]. The model is learned from the features on Imagenet with 1000 different classes. In our case, we have chosen, ResNet-50 as another model for classifying rare disorders because of its higher capability of classifying the objects on which it is trained [2].

4.1.2 Explanation methods

In our thesis, we have chosen two XAI methods such as GradCam and SIDU [39]. These two methods are expected to give the most promising prediction for the problem at hand. The reasons for choosing these methods are as follows:

- **Grad-CAM** is a gradient-based approach that gives visual saliency maps with respect to the final layer of the neural network considered [51]. This method is the most promising and simple approach for identifying specific regions in the model's anticipated outcomes.
- **SIDU** method calculates the similarity and difference between the subject under consideration and the prediction masks for creating the visual explanations [38]. The proposed approach of SIDU as mentioned in [39] outperforms the existing XAI methods.

4.2 Proposed approach

4.2.1 Pipeline of the project

The methods are selected based on the evaluations discussed in section 4.1.1. The methods used for the further discussion are as follows:

- Rare disorder identification and categorization methods - GestaltMatcher-Arc, ResNet-50
- Explanation methods - Grad-CAM, SIDU

The overall findings of this research work are visualized in Figure 4.1. Each of the different stages of the pipeline will be elucidated extensively in detail as follows:

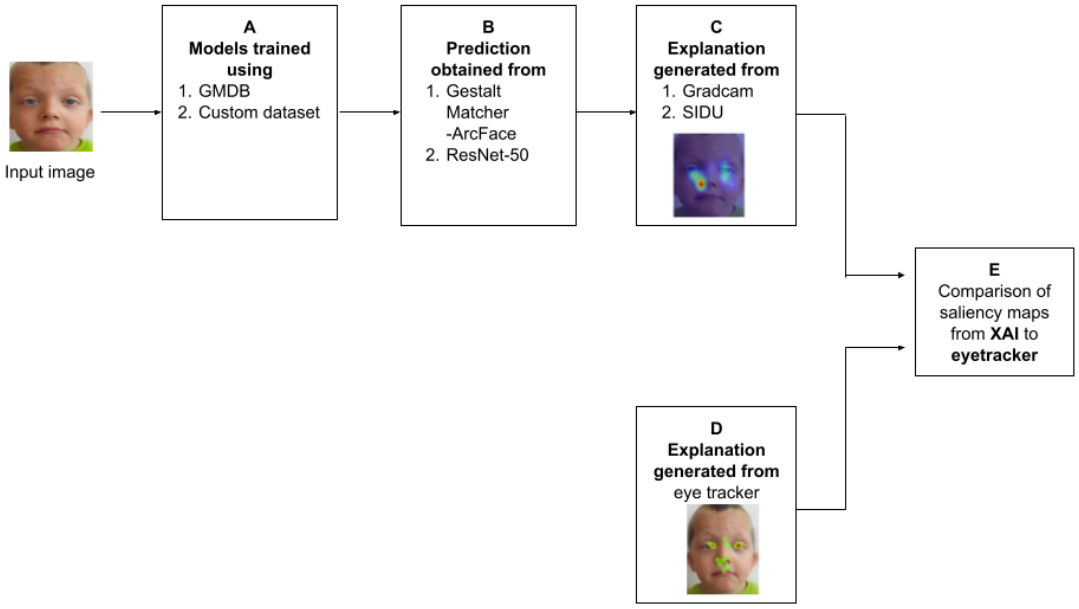


Figure 4.1: Overview of proposed approach.

Stage A - Training the neural network using datasets

To start with, a facial image prediction model is chosen. The model is trained using two different datasets GMDB and a custom dataset. The models considered here are GestaltMatcher-Arc and ResNet-50. The data fed into to the model is crop-aligned facial images and the output is features that the DL model considers [24].

Stage B - Obtaining predictions from the trained models

Using the models trained in the previous stage, a total of n different predictions are made from the input image(I). The prediction probabilities for each class are obtained from the models. The higher probability towards a class indicates the image belongs to a particular class.

Stage C - Generating explanations based on the predicted probability

In this stage, explanations are created in accordance with the predictions. Explanations are generated using two different XAI methods such as SIDU and Grad-CAM. The explanations are created from the final layer of the convolutional network. For

SIDU approach similarity difference and uniqueness values are calculated for generating explanations. An explanation is generated for each of the predictions. For prediction P, one explanation per prediction is generated e_i.

Stage D - Generating explanations from the eye-tracker

The heat maps are generated from the Tobii- eye-tracker as explained in the next chapter. The eye-tracker heat maps of clinicians and non-clinicians who correctly identified the syndromes are taken.

Stage E - Comparison of XAI heat maps and eye-tracker heat maps

The heat maps generated from the XAI method are compared with the eye-tracker heat map. This comparison is performed to get an idea of whether human and AI are comparable. The heat maps are compared by calculating IoU between two heat maps. The values of IoU score are expected between 0 and 1. 0 represents no similarity between images and a value closer to 1 represents a higher similarity between images.

4.2.2 Combination of methods chosen

The various possible combination of DL method and XAI method used for this research is as follows:

$$\underbrace{\left\{\begin{array}{c} GestaltMatcher - Arc \\ ResNet - 50 \end{array}\right\}}_{\text{DL methods}} \times \underbrace{\left\{\begin{array}{c} Grad - CAM \\ SIDU \end{array}\right\}}_{\text{Visualization methods}} \tag{4.1}$$

5

Experimental Setup

The previous chapter provided an idea about the proposed pipeline of this thesis. In this chapter, we describe the datasets used for training various DL classifiers, eye-tracking experiments, architectures and the experiments conducted.

5.1 Datasets

In this thesis, we have focused on ultra-rare disorder identification and classification. This section discusses the datasets used to train the selected classifiers.

5.1.1 GMDB

GMDB [29] primarily consists of portraits of people affected with ultra-rare disorders. However, it also includes other data modalities such as portrait, X-ray, and fundoscopy images for medical applications. The dataset is available for clinicians and individuals performing research in such fields. The database consists of 7,533 images of 792 different disorders [29]. The images were obtained from 2,058 publications. The database also contains 1,018 frontal images collected from 498 unpublished works. Only 6 different disorders from this dataset are used for training the neural networks. These classes are chosen based on the disorders used for the eye-tracking survey. The disorders are Cornelia de Lange, Kabuki Syndrome, Williams-Beuren, Noonan and Beckwith-Wiedemann syndrome. Figure 5.1 shows a few images in the GMDB. As seen from the images, most of the images either do not have a higher resolution or have some other components in an image other than faces. Preprocessing is required to be conducted on these images to accurately

identify facial features in an image. A concise overview of the total image count per category (syndromes in our case) considered is shown in Table 5.1.

Figure 5.1: Few example of images in GMDB.

Syndrome Name	OMIM id	Number of images	Number of subjects
Cornelia de Lange	122470	377	316
Williams-Beuren	194050	250	227
Kabuki	147920	153	126
Noonan	605275	107	91
Rubenstein-Taybi	180849	103	90
Beckwith-Wiedemann	130650	26	22
Wolf-Hirschhorn	194190	11	6

Table 5.1: Total number of images in GMDB for the particular syndromes chosen. The images are chosen based on the categories used for the eye-tracking experiment. From the above dataset, Wolf-Hirschhorn images are eliminated, because of the few samples of the same in the dataset.

5.1.2 Custom dataset

A custom dataset is created with frontal facial images of 10 different disorders and 1 class of normal faces by National Institute of Health (NIH). These 10 disorders include 22q11.2 deletion, Beckwith-Wiedemann, Cornelia de Lange, Down, Kabuki, Noonan, Prader-Willi, Rubenstein-Taybi, Wolf-Hirschhorn, and Willaims-Beuren. This dataset is based on the disorders used for eye-tracking studies. A class with normal frontal facial images are also collected from different publications. Figure 5.2 shows a few examples of the images in the custom dataset. As observed from the images, the images are mostly not aligned and have a different orientation. Therefore, some preprocessing techniques such as cropping and aligning have to be performed on these images.

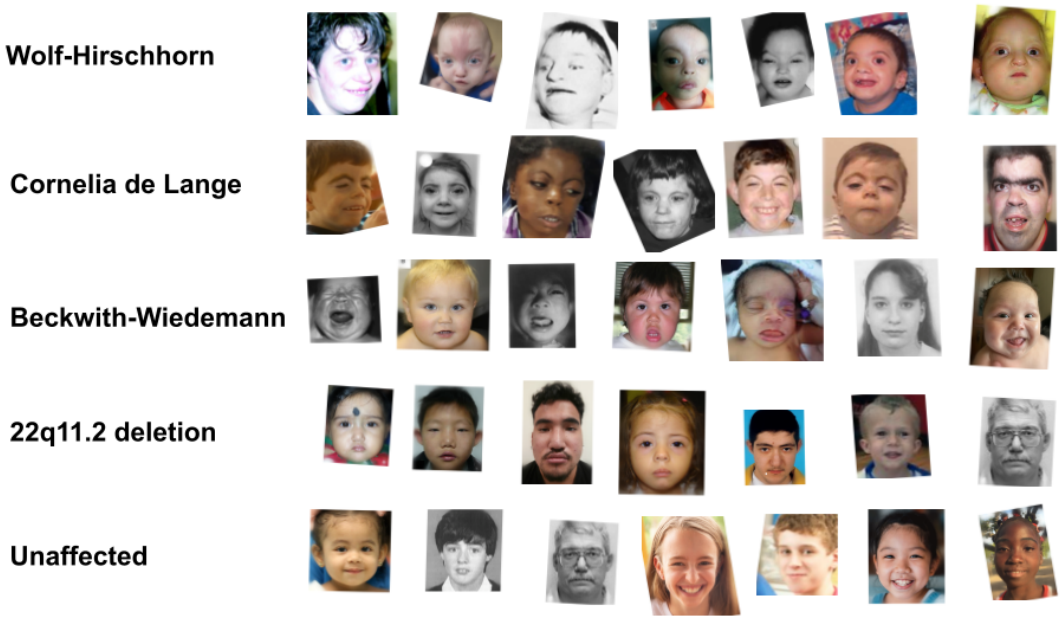


Figure 5.2: A couple of examples of the samples in the custom dataset created according to the eye-tracking study syndrome categories.

5.2 Eye-tracking experiment

The eye-tracking experiment has been completed on the Tobii Pro Fusion eye-tracker [16]. The experiment setup is shown in Figure 5.3. The experiment is designed

by NIH. The experiment is also performed in the Institute of Genomics Statistics and Statistics department as well. The eye-tracking experiment is as follows:

1. Initially a project is created for the Tobii Pro Fusion eye-tracking device. 17 different images are loaded into the device, which are either facial images of people with some genetic condition or normal faces. Followed by each image is a set of questionnaires asking,

   (a) *"Is this image of a person with a genetic condition?" (yes or no)*

   (b) *"How confident are you with your answer?" (1 - highly confident, 4 - not at all confident)*

   (c) *"Do you know the genetic condition?"*

   After each of these questions, a cross mark is shown on the screen to focus the eyes on the centre point of the screen. This whole step continues for 17 sets. Each of the images is shown for 7 seconds, with unlimited time for questionnaire slides, and 4 seconds for the cross mark.

2. To start the experiment, each of the participants will be seeing a calibration frame to calibrate the eyes. The participants are asked to stay in the same position throughout the experiment to have precise results. The distance between the participant and the eye-tracker device varies from person to person. The distance is adjusted according to the calibration step.

3. After the experiment is completed, we can save the recording and we can observe from the device, how the eyes are moved while analysing images [16].

4. We can download various types of visualizations from the device such as gaze plots (order or the pattern in which your eyes moved), and heat maps (saliency maps - where your eyes spent the most time). The default heat map setting from the eye-tracker produces a heat map with red (high), yellow, and green (low). These heat maps created are used in this research work to compare with the saliency maps created from the XAI methods [16].

The experiment is performed among clinicians and non-clinicians to know the difference in the pattern of analysing images by clinicians and non-clinicians [3].

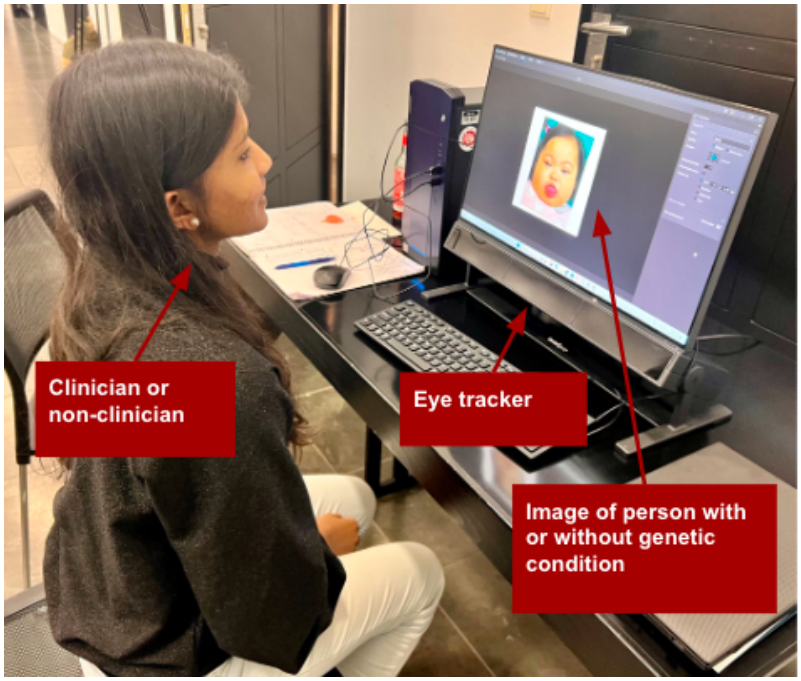


Figure 5.3: Eye-tracking experiment.

After analysing the patterns obtained from the eye-tracker, a conclusion is made based on the heat maps and gaze plots from the device [16]. The conclusion is that the pattern for staring at the faces changes from person to person, whether it is a clinician or non-clinician [16]. However, clinicians tend to look at the most prominent features of the face such as the eyes, nose, and mouth [16]. Whereas non-clinicians look at these features at random.

5.3 Network architectures

In this thesis, we have chosen the GestaltMatcher-Arc model and ResNet-50 for identifying and categorizing various rare-genetic syndromes. Two different datasets as mentioned in section 5.1 are used for the training of the chosen networks. The first dataset is GMDB and we have chosen 7 different syndromes from this dataset, which is based on the syndromes used for eye-tracking experiments [24]. The second dataset is a custom dataset with 11 different classes, also based on the syndromes used for the eye-tracking experiment [16]. All of these images were cropped to the size of 100×100 using GestaltEngine-FaceCropper [24]. The GestaltEngine-FaceCropper

Hyperparameters	Values	
	GestaltMatcher-Arc	ResNet-50
Input image size	100×100	100×100
Learning rate	0.001	0.01
Loss Function	Cross-entropy	Categorical cross-entropy
Epochs	50	50
Batch size	32	32

Table 5.2: Details of training of GestaltMatcher-Arc and ResNet-50.

uses RetinaFace [13] as the fundamental model, which is a single-stage face detector. The RetinaFace approach performs pixel-wise localization of faces to identify the faces in the image. The face detection using RetinaFace has few features as, the authors have manually annotated five facial landmarks from the WIDER Face dataset, which improved the detection more easily [24]. The authors have also used a self-supervised mesh-decoder, that is used for predicting the pixel-wise 3D shape of the face [13]. This approach was able to detect only faces in an image, thereby giving more accurate predictions on the input image. The GestaltEngine-FaceCropper crops align each image to give a focus to each face in an image. The range of different configurations used to train the chosen models is depicted in Table 5.2.

5.3.1 Architecture of GestaltMatcher-Arc

In this segment, we will talk about the training of the GestaltMatcher-Arc model. The model is trained on two datasets as mentioned in section 5.1. Figure 5.4 shows the training done on the model. The example here shows the model being trained on the custom dataset with seven different classes with a total of 786 images in the training set. As observed from the figure, we got a top-5 accuracy for prediction as 90.23%.

```
Loaded dataset: gmdb (version v1.0.3) with image size 100x100 in RGB, while not retaining the
Training dataset size: 786, with 7 classes and distribution: [228, 196, 182, 60, 57, 48, 15]
Validation dataset size: 105, with distribution: [28, 24, 15, 13, 9, 8, 8]
Weighted cross entropy weights: tensor([0.5329, 0.5383, 0.5412, 0.6250, 0.6316, 0.6562, 1.0000
Loading pretrained weights from saved_models/glint360k_r100.onnx
Freezing model weights
Freezing model weights
Created frozen glint360k_r100 model with 3 in channels, 512d feature dimensionality and 7 clas
2023-10-19 11:03:18.352522: I tensorflow/core/platform/cpu_feature_guard.cc:193] This TensorFl
in performance-critical operations:  SSE4.1 SSE4.2 AVX AVX2 FMA
To enable them in other operations, rebuild TensorFlow with the appropriate compiler flags.
Overall average training loss: 1.488382
Average BCE Loss (1.2234680244502323) during validation
        Top-1 accuracy: 0.4571428596973419, Top-5 accuracy: 0.9238095283508301
        Mean Top-1 accuracy: 0.37426739926739927, Mean top-5 accuracy: 0.9023809523809524
Elapsed time during validation: 304.8s
Saving model in: s1_glint360k_r100_512d_gmdb__v1.0.3_bs128_size100_channels3_e1.pt
Overall average training loss: 0.555004
Average BCE Loss (0.9994617601560744) during validation
        Top-1 accuracy: 0.6285714507102966, Top-5 accuracy: 0.9523809552192688
        Mean Top-1 accuracy: 0.5479068550497123, Mean top-5 accuracy: 0.9357142857142857
```

Figure 5.4: Training output of GestaltMatcher-Arc.

5.4 Implementation details

From the DL models chosen, we have obtained predictions for different syndromes from the classifiers. The AOI are highlighted using two XAI methods: SIDU and Grad-CAM. Figure 5.5 shows the architecture of Grad-CAM applied to the last layers of the ResNet-50 model. The input image is cropped using GestaltEngine-FaceCropper to a size 100×100 and is given as the input to the various convolutional layers. The GestaltEngine-FaceCropper uses RetinaFace [13] for detecting faces in an image. RetinaFace is a single-stage face detector used for detecting the faces in an image. The approach uses a combination of extra-supervised and self-supervised multi-task learning to apply pixel-wise face localization on faces [13]. The approach annotates five facial landmarks in facial images and uses a self-supervised mesh decoder to predict the 3D shape of the image. By using this approach, the faces in the image can be detected correctly and the background noise in the images can be eliminated. After applying RetinaFace to each image, all the images are cropped to the size of 100×100 to have the images correctly aligned and cropped. These cropped images are further used to train the chosen neural network models.

Feature activation maps are created for this input image and the classifier makes the predictions on the basis of the inputs provided to the model. Figure 3.5 shows how the visual saliency maps are created from the ResNet-50 architecture for the XAI methods Grad-CAM and SIDU. As illustrated in Figure 3.5, a gradient is applied to the last convolutional layer of the ResNet-50 to get the Grad-CAM,

highlighting the regions the model looked at while giving predictions. In Grad-CAM gradients are calculated based on the network's last layer. For the SIDU approach, the last convolutional layer is used for calculating similarity and difference values for generating the visual explanations. After calculating the similarity difference and uniqueness score, feature important weights are assigned to get the final visual explanations. A similar approach is applied to the other state-of-the-art image classification model - GestaltMatcher-Arc model. The only difference is that the GestaltMatcher-Arc model uses 512 representation vectors. The visual representation maps are created from the Grad-CAM and SIDU based on the outcomes of the predictions and are obtained from the final convolutional layer of the network under consideration.

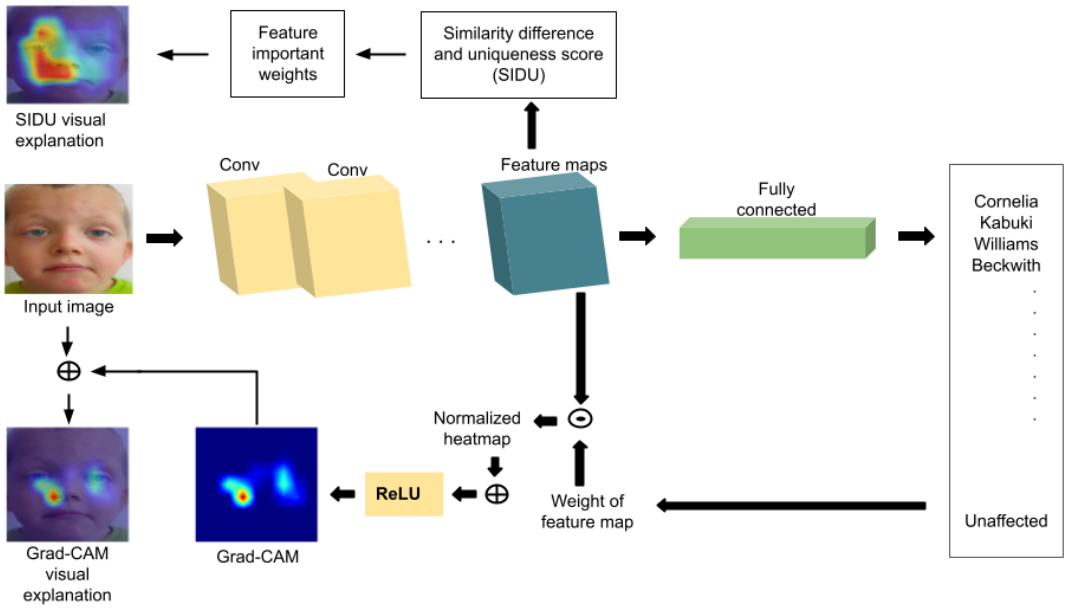


Figure 5.5: Grad-CAM and SIDU for GestaltMatcher-Arc and ResNet-50 in our research.

6

Evaluation and Results

Within this passage, we evaluate the performance of two XAI methods: namely SIDU and Grad-CAM. Also comparison of XAI heat maps to the eye-tracking heat maps.

6.1 Experiments

This section provides a detailed explanation of the experiments conducted in this thesis.

6.1.1 Experiment 1: Explanation generated by visual explanation methods

The explanations for the predictions are generated by two different approaches namely Grad-CAM and SIDU. Tables 6.2, 6.1 show the original image, eye-tracking data and prediction results along with generated visual explanations for different approaches. Table 6.2 shows the heat maps for correctly predicted classes and table 6.1 shows the heat maps for wrongly predicted classes.

Generated heat maps using Grad-CAM

The Grad-CAM heat maps are generated using PyTorch-Grad-CAM implemented in [18] and are created from the final layer of the neural network. The created heat maps are shown in Table 6.2. As we can observe from Table 6.2, Grad-CAM heat maps tend to highlight a few regions in the face and are more concise. As observed from the heat maps, the heat maps do not provide disorder-specific AOI in these images.

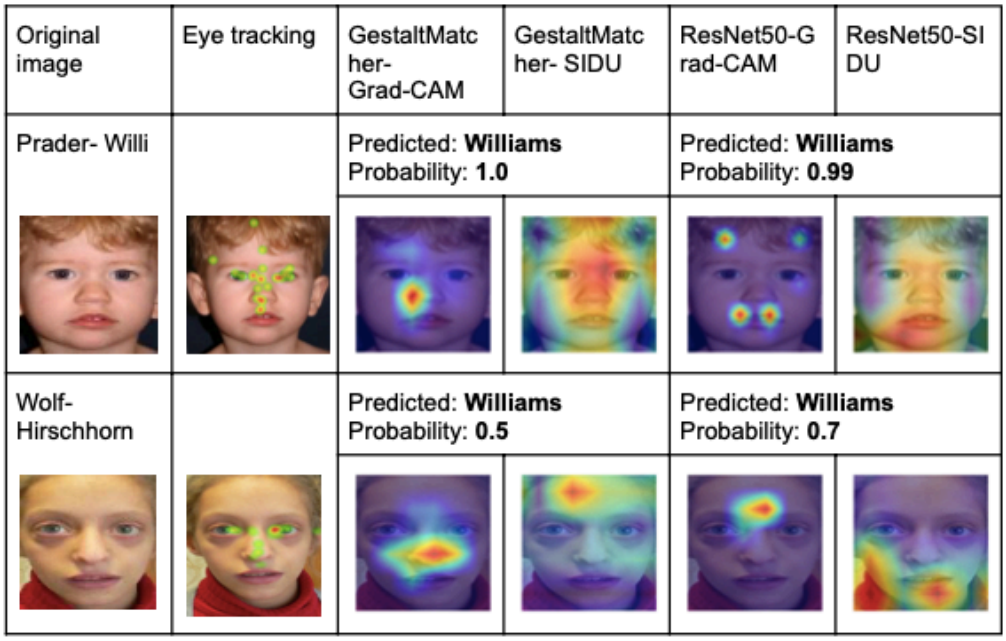


Figure 6.1: Heat maps of wrongly predicted classes.

Generated heat maps using SIDU

The SIDU heat maps were generated using the code from [38]. The SIDU heat maps tend to spread over the region in the face. The heat maps generated using SIDU tend to give the most indicative output for the 22q11.2 deletion syndrome. The approach fails to give a disorder-specific explanation. The approach highlights the regions in a face, that the model might think are important in giving a prediction, but these features are mostly irrelevant according to a clinical perspective. The proposed approach of SIDU has been used for classifying various objects such as cats, dogs, tables and so on. However, this approach has been used in the present thesis work for identifying different features of a face. That might be the reason why the approach is highlighting most of the AOI in a face, which might be even irrelevant to the present approach.

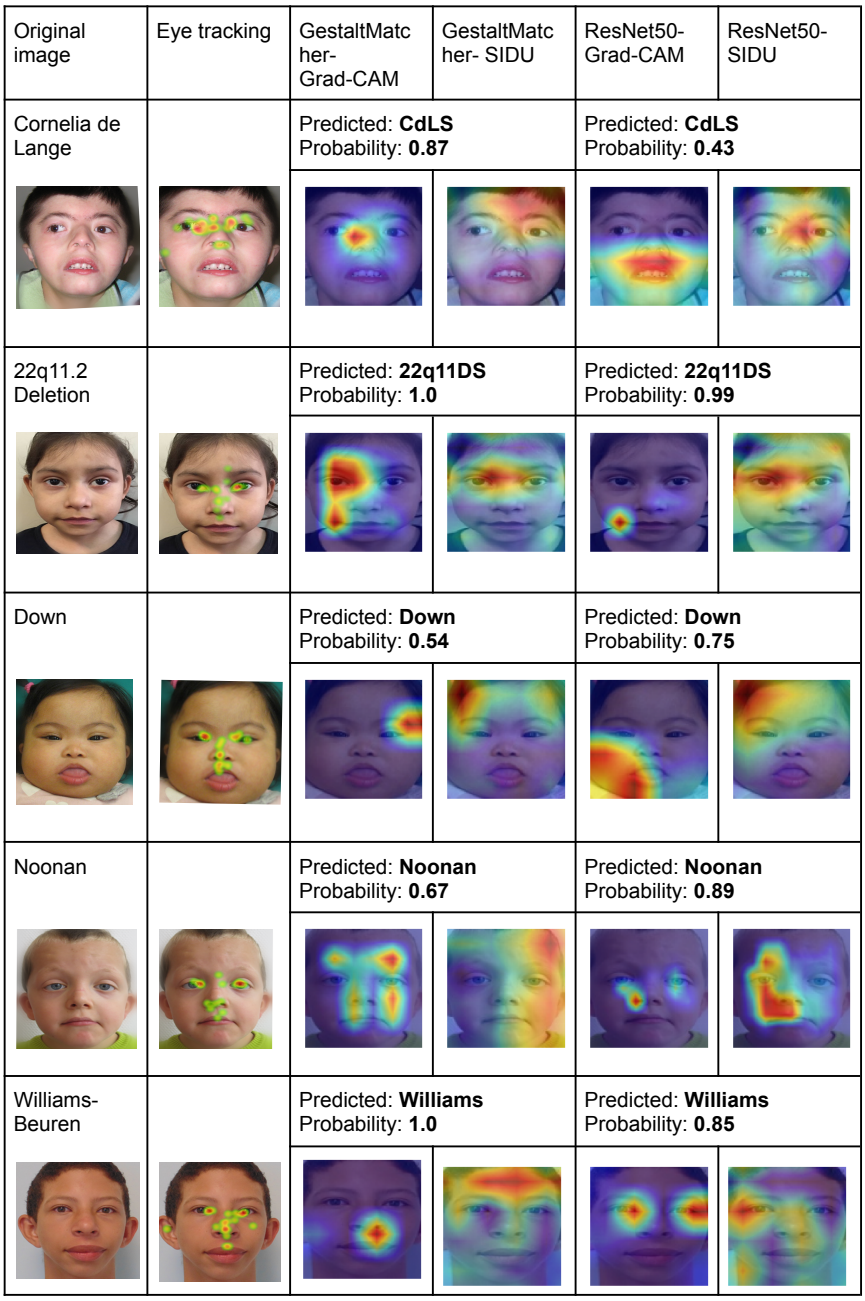


Figure 6.2: Heat maps of correctly predicted classes.

6.1.2 Experiment 2: Human-grounded evaluation

Human-grounded evaluation is the most feasible evaluation approach for testing the explanations generated by DL models. For applications such as rare-genetic disorder identification, expert decisions from clinicians cannot give more expert diagnoses. For such applications, one excellent way is comparing the relation of the eye-tracker with the generated visual explanations by a model. Both approaches created heat maps that draw attention to essential regions in an image. The eye-tracking experiment and gathering data through that experiment is already described in Section 5.2. In our study, we have taken the average of heat maps of clinicians, who have correctly diagnosed the syndrome. The clinicians have looked at these features presented in Table 6.1 in faces for evaluating the generated heat maps.

Evaluations from clinicians based on the AOI in face

Condition	Facial features				
	Forehead	Eyes/Periorbital	Nose	Mouth	Ears
Cornelia de Lange		+	+	+	+
22q11.2 deletion			+		
Down		+	+	+	+
Noonan		+			+
Prader-Willi		+			
Wolf-Hirschhorn	+	+	+		
Williams-Beuren		+	+	+	

Table 6.1: Syndrome-specific facial features identified by a clinician. + sign shows the feature important in a face.

6.2 Comparison metrics

To evaluate the models with human fixations, we have used IoU score to compare the generated XAI and the eye-tracker heat maps.

6.2.1 Evaluation of eye-tracking to XAI methods

For the comparison of the eye-tracking heat map with XAI heat map,IoU score is used. The IoU score is calculated as,

$$IoU = \frac{\text{Area of Overlap}}{\text{Area of Union}} \tag{6.1}$$

For calculating IoU score between the eye-tracking heat maps and the heat maps obtained from the various XAI method, all the images are first converted to binary scale images (black and white) using the OpenCV library to have the same colour scheme. After converting all the images into binary scale, the IoU score is calculated for analysing how close the eye-tracking heat maps are to the XAI methods. The IoU scores for each method are given in that table 6.2. The value of IoU score that is close to 1 shows the higher comparability of the two images.

Syndrome	GestaltMatcher-Grad-CAM	GestaltMatcher-SIDU	ResNet50- Grad-CAM	ResNet50-SIDU
Cornelia de Lange	0.76	0.56	0.30	0.89
22q11.2 deletion	0.78	0.86	0.20	0.67
Down	0.34	0.20	0.34	0.43
Noonan	0.45	0.36	0.56	0.67
Prader-Willi	0.67	0.89	0.56	0.78
Wolf-Hirschhorn	0.56	0.76	0.65	0.10
Williams-Beuren	0.45	0.65	0.78	0.34
Average	**0.58**	**0.62**	**0.48**	**0.55**

Table 6.2: Comparison table for IoU score. The IoU values are calculated between the eye-tracker heat maps and different XAI heat maps considered.

As inferred from the table 6.2, it shows little correlation between the eye-tracking

heat maps and the XAI generated heat maps. This might be because the human and DL method might be looking at different features while giving predictions. On average, as observed from the table 6.2, the higher correlation between the XAI heat maps and eye tracker heat maps is for the GestaltMatcher-Arc model with SIDU approach. A higher correlation between the images is found for the GestaltMatcher-Arc model than for the ResNet-50 model. This might be because the GestaltMatcher-Arc model has a higher probability of identifying the disorders more accurately. A higher correlation between the features of the face is observed for 22q11.2 deletion syndrome for the GestaltMatcher-Arc model. The features highlighted by the model are mostly the eyes and periorbital region, which is mostly highlighted by the eye-tracking experiment as well, whereas this feature is less important for a clinician in identifying the syndrome. In the case of Cornelia de Lange syndrome, the eye-tracker highlighted the eyes and nose region, whereas important regions according to clinicians are the head, eyes, nose, mouth, and ears. The XAI models, especially SIDU have highlighted the regions most closely to the syndrome according to the clinician. The worst depicted heat maps are for Down syndrome. The heat maps generated by different XAI methods do not signify any syndrome-specific knowledge.

7

Discussion

In previous chapters, we have discussed the results of various experiments performed in this thesis. In this chapter, we conclude all the findings of this thesis work by summarising the findings, giving answers to the research questions, lessons learned and future work.

7.1 Revisiting the research questions

RQ1: What are the image saliency methods available to identify the AOI in a given image to make predictions?

In this thesis work, we have compared two different image saliency methods for explaining two different DL models. The selected approaches were suitable for identifying different features in an image concerning ultra-rare genetic disorders. The identified image saliency methods are Grad-CAM and SIDU. The Grad-CAM uses the final layer of the convolutional neural network for creating visual saliency maps, whereas SIDU calculates similarity differences and uniqueness for obtaining the visual saliency maps.SIDU approach was able to emphasize the crucial regions more accurately than the approach Grad-CAM. SIDU approach highlights the regions spread over some portion of the face, whereas, Grad-CAM approach highlights small regions in the face.

RQ2: What are the important AOI in a face to make predictions related to genetic disorders?

The important features that are specific to each genetic disorder considered in this approach are mentioned in Table 6.1. These facial features were identified by a clinician. We have asked the clinician to give a score from 1-5(1-facial

features are correctly focused on essential regions related to the syndrome, and 5-facial features related to that syndrome are not focused or irrelevant). According to the clinician, there is no genetic disorder specific AOI identified by the visual explanation method. According to the clinician, the SIDU based approach tends to highlight AOI in the images more specifically than the Grad-CAM based approach.

RQ3: How can we compare the attribution maps obtained from image saliency methods to heat maps from an eye-tracker? Do the saliency maps of clinicians and eye-tracker saliency maps align?

For the comparison of attribution maps from the image saliency method to heat maps from an eye-tracker IoU score is calculated between these heat maps as discussed in Section 6.2.1. The IoU score is expected between 0 and 1. Table 6.2 has depicted the obtained IoU score for various genetic disorders considered. As observed from the table, most of the values are between 0.2 to 0.8. This shows a high similarity between the features obtained from the eye-tracker and the image saliency approach for the Cornelia de Lange (CDL) from the ResNet-50 SIDU and eye-tracker heat map. Whereas a very low similarity in features between Down syndrome heat maps and eye-tracker heat maps. To conclude, there is a huge difference in the features observed by a model and a human while looking at features in an image.

RQ4: Do the clinicians and non-clinicians look at the same facial features?

For this comparison, the eye-tracking heat maps were compared. An average over the correctly identified heat maps for clinicians and non-clinicians was considered. It was observed from the heat maps, that there isn't much difference between the clinician's and non-clinicians heat maps. Even though for some syndromes, in which clinicians are more specific about a syndrome, the clinicians tend to look at other features excluding the obvious ones. Whereas non-clinicians always tend to look at the eye, nose, and mouth region.

7.2 Contributions

The main contribution of this work is as follows:

- We have conducted a literature review on the approaches that can be used for ultra-rare disorder identification and classification. A literature review of various explanation methods is also conducted.
- Implementation of the selected ultra-rare disorder classification approaches. Selected explanation methods are then applied to the predictions of the classifier and a satisfied output is obtained.
- A comparison of the saliency maps from the XAI methods and the eye-tracker heat maps was conducted. A conclusion is made based on the comparison of these heat maps.
- Clinicians with expertise in the field have quantified the results obtained and commented that the ultra-rare disorder-specific features are not identified by the model and the eye-tracker.
- An opportunity to present the idea of this work as a poster AGDev and is shown in Appendix A. Also, an opportunity to be part of a scientific paper that is currently in the review period. Details of this scientific paper are also discussed in Appendix A.

7.3 Future work

This research work can be extended or improved to the following in future.

- In this research work, we have chosen datasets similar to each other. Both datasets contained images of people with 10 rare genetic disorders. This can be extended to identifying and classifying many more rare genetic disorders. Also, the images of the normal class were similar to the genetic disorder. This might confuse the classifier. Need to consider a class with normal images as well.
- Eye-tracking survey can be extended by asking more questions to people taking the survey. Maybe ask, which feature you felt important while analysing the image.
- A more high-level approach is required to quantify the results obtained from the XAI method.

A

Contributions

An opportunity to be part of a scientific paper that is available in PubMed [16]. I got an opportunity to be part of the scientific paper entitled *"Human and computer attention in assessing genetic conditions"*. The paper compares the human attention maps generated from the eye tracker with some other machine-generated saliency maps. The idea is to compare humans and AI in analysing the syndromic faces. The ideas of this paper are also under peer review for the PLOS Genetics (`https://journals.plos.org/plosgenetics/` entitled as *"Comparison of clinical geneticist and computer visual attention in assessing genetic conditions"*.

The ideas of this thesis work have been presented as a poster in AGDev 2023 conference (`''https://www.agdev.de/en/''`). The poster is as follows:

Quantitative Comparison of Deep Learning Classifiers and Human Attention in Assessing Rare Disorders

Anna Rose Johny

Hochschule Bonn-Rhein-Sieg
University of Applied Sciences

ukb universitäts klinikumbonn

Institute for Genomic Statistics and Bioinformatics

Motivation and background

- Perform eye tracking analysis and to understand how clinicians diagnose patients.
- Comparing with explainable AI to understand difference between human and AI.

Datasets used

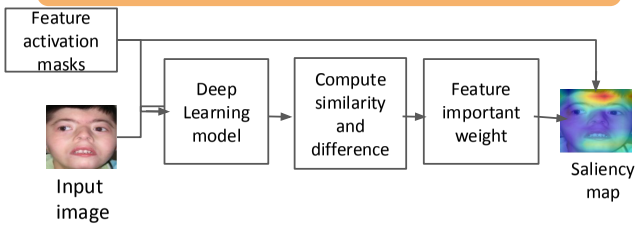
- **GestaltMatcher DataBase** (7 syndromes)
- **Custom dataset** (11 classes -10 syndromes, 1 unaffected

Deep Learning classifiers used

Deep Learning classifiers are used to classify the images into various syndromes. Deep Learning classifiers used in this work,

- **GestaltMatcher ArcFace**
- **ResNet50**

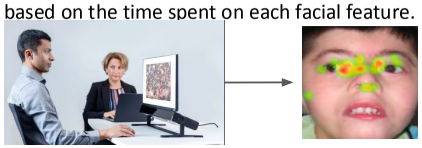
Explainable AI Methods



The regions responsible for prediction are highlighted using explainable AI method (**GradCam**, **SIDU**). The figure above represents the saliency map generation using SIDU.

Eye-tracking data using Tobii-Pro

A Tobii-eye tracker device is used for collecting data from clinicians and non-clinicians provided with syndromic and non-syndromic faces. Heatmaps are generated from the eyetracker based on the time spent on each facial feature.



Comparison to eye-tracking data

The eye tracking heatmaps are compared with the explainable AI heatmaps using KL-Divergence and got 0.31 for ResNet gradcam, 0.51 for ResNet SIDU approach.

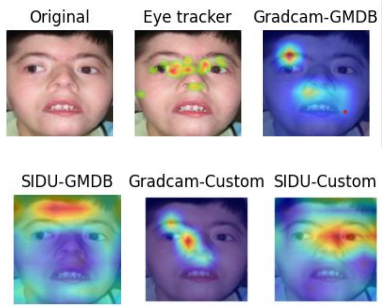


Figure 2: Comparison of eye tracker heatmaps with Gradcam and SIDU.

Conclusion and results

- AI and human look at different features in an image
- Difficult to compare AI and human for different syndromes

References

[1] Satya M. Muddamsetty, Visual explanation of black-box model: Similarity Difference and Uniqueness (SIDU) method, Pattern Recognition,2022

Figure A.1: Research poster presented at AGDev 2023.

B

Questionnaire

We have asked two clinicians to evaluate the results obtained from the eye-tracker on a scale of 1 to 5 (1-facial features are correctly focused on essential regions related to the syndrome, and 5-facial features related to that syndrome are not focused or irrelevant)). The red region gives higher intensity and the blue is the lowest. Based on the evaluations of the clinicians, a syndrome-specific AOI is not highlighted by any of these approaches. As we can observe from the figures, both clinicians agree on different facial features and their evaluations are different. Both clinicians have agreed to the same number for Noonan syndrome. Based on the clinician's review, the XAI models have been poorly performed for identifying rare-genetic disorders.

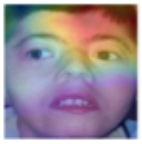
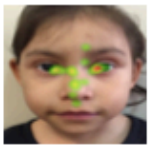
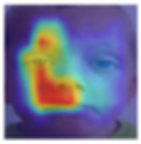
Original image	Eye tracking	GestaltMatcher-Grad-CAM	GestaltMatcher- SIDU	ResNet50-Grad-CAM	ResNet50-SIDU
1) Cornelia de Lange					
Clinician 1	3	3	3	2	3
Clinician 2	2	5	3	4	3
2) 22q 11.2 deletion					
Clinician 1	4	4	3	5	3
Clinician 2	3	4	1	5	3
3) Down					
Clinician 1	4	4	4	4	4
Clinician 2	2	5	5	5	3
4) Noonan					
Clinician 1	3	3	4	4	3
Clinician 2	3	3	4	4	3

Figure B.1: Evaluation from two clinicians.

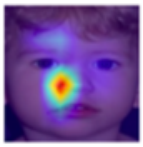
Original image	Eye tracking	GestaltMatcher-Grad-CAM	GestaltMatcher- SIDU	ResNet50-Grad-CAM	ResNet50-SIDU
5) Prader-Willi					
Clinician 1	4	4	3	4	3
Clinician 2	3	4	3	3	5
6) Wolf-Hirschhorn					
Clinician 1	3	4	4	4	4
Clinician 2	3	4	3	3	5
7) Williams-Beuren					
Clinician 1	4	4	4	3	3
Clinician 2	4	4	4	3	4

Figure B.2: Evaluation from two clinicians.

Bibliography

[1] Zeyad AT Ahmed, Theyazn HH Aldhyani, Mukti E Jadhav, Mohammed Y Alzahrani, Mohammad Eid Alzahrani, Maha M Althobaiti, Fawaz Alassery, Ahmed Alshaflut, Nouf Matar Alzahrani, Ali Mansour Al-Madani, et al. Facial features detection system to identify children with autism spectrum disorder: deep learning models. *Computational and Mathematical Methods in Medicine*, 2022, 2022.

[2] Laith Alzubaidi, Jinglan Zhang, Amjad J Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, José Santamaría, Mohammed A Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8:1–74, 2021.

[3] Akanksha Atrey, Kaleigh Clary, and David Jensen. Exploratory not explanatory: Counterfactual analysis of saliency maps for deep reinforcement learning. *arXiv preprint arXiv:1912.05743*, 2019.

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

[5] Florian Beck, Minh Nhat Vu, Christian Hartl-Nesic, and Andreas Kugi. Singularity avoidance with application to online trajectory optimization for serial manipulators. *IFAC-PapersOnLine*, 56(2):284–291, 2023.

[6] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[7] Ali Borji. Boosting bottom-up and top-down visual features for saliency estimation. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 438–445, 2012.

[8] Aidan Boyd, Kevin W Bowyer, and Adam Czajka. Human-aided saliency maps improve generalization of deep learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2735–2744, 2022.

[9] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.

[10] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[11] Michael Chromik. reshape: A framework for interactive explanations in xai based on shap. 2020.

[12] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers, 2017.

[13] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild, 2019.

[14] Jürgen Dieber and Sabrina Kirrane. Why model why? assessing the strengths and limitations of lime. *arXiv preprint arXiv:2012.00093*, 2020.

[15] Dat Duong, Ping Hu, Cedrik Tekendo-Ngongang, Suzanna E Ledgister Hanchard, Simon Liu, Benjamin D Solomon, and Rebekah L Waikel. Neural networks for classification and image generation of aging in genetic syndromes. *Frontiers in Genetics*, 13:864092, 2022.

[16] Dat Duong, Anna Rose Johny, Suzanna Ledgister Hanchard, Chris Fortney, Fabio Hellmann, Ping Hu, Behnam Javanmardi, Shahida Moosa, Tanviben Patel, Susan Persky, et al. Human and computer attention in assessing genetic conditions. *medRxiv*, 2023.

[17] Maciej Geremek and Krzysztof Szklanny. Deep learning-based analysis of face images as a screening tool for genetic syndromes. *Sensors*, 21(19):6595, 2021.

[18] Jacob Gildenblat and contributors. Pytorch library for cam methods. `https://github.com/jacobgil/pytorch-grad-cam`, 2021.

[19] Tristan Gomez, Thomas Fréour, and Harold Mouchère. Metrics for saliency map evaluation of deep learning explanation methods. In *International Conference on Pattern Recognition and Artificial Intelligence*, pages 84–95. Springer, 2022.

[20] Darrel Greenhill, J Renno, James Orwell, and Graeme A Jones. Occlusion analysis: Learning and utilising depth maps in object tracking. *Image and Vision Computing*, 26(3):430–441, 2008.

[21] Yaron Gurovich, Yair Hanani, Omri Bar, Nicole Fleischer, Dekel Gelbman, Lina Basel-Salmon, Peter M. Krawitz, Susanne B. Kamphausen, Martin Zenker, Lynne M. Bird, and Karen W. Gripp. Deepgestalt - identifying rare genetic syndromes using deep learning. *CoRR*, abs/1801.07637, 2018.

[22] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[24] Alexander Hustinx, Fabio Hellmann, Ömer Sümer, Behnam Javanmardi, Elisabeth André, Peter Krawitz, and Tzung-Chien Hsieh. Improving deep facial phenotyping for ultra-rare disorder verification using model ensembles. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5018–5028, 2023.

[25] Maksims Ivanovs, Roberts Kadikis, and Kaspars Ozols. Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognition Letters*, 150:228–234, 2021.

[26] Tilke Judd, Krista A. Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. *2009 IEEE 12th International Conference on Computer Vision*, pages 2106–2113, 2009.

[27] Hyungsik Jung and Youngrock Oh. Towards better explanations of class activation mapping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1336–1344, 2021.

[28] Arun Kumar, V Anoosh Solayappan, et al. Masked deep face recognition using arcface and ensemble learning. In *2021 IEEE 2nd International Conference on Technology, Engineering, Management for Societal impact using Marketing, Entrepreneurship and Talent (TEMSMET)*, pages 1–6. IEEE, 2021.

[29] Hellen Lesmann, Gholson J. Lyon, Pilar Caro, Ibrahim M. Abdelrazek, Shahida Moosa, Jean Tori Pantel, Merle ten Hagen, Stanislav Rosnev, Tom Kamphans, Wolfgang Meiswinkel, Jing-Mei Li, Hannah Klinkhammer, Alexander Hustinx, Behnam Javanmardi, Alexej Knaus, Annette Uwineza, Cordula Knopp, Elaine Marchi, Miriam Elbracht, Larissa Mattern, Rami Abou Jamra, Clara Velmans, Vincent Strehlow, Amira Nabil, Claudio Graziano, Borovikov Artem, Franziska Schnabel, Lara Heuft, Vera Herrmann, Matthias Höller, Khoshoua Alaaeldin, Aleksandra Jezela-Stanek, Amal Mohamed, Amaia Lasa-Aranzasti, Gehad Elmakkawy, Sylvia Safwat, Frédéric Ebstein, Sébastien Küry, Annabelle Arlt, Felix Marbach, Christian Netzer, Sophia Kaptain, Hannah Weiland, Koen Devriendt, Karen W. Gripp, Martin Mücke, Alain Verloes, Christian P. Schaaf, Christoffer Nellåker, Benjamin D. Solomon, Rebekah Waikel, Ebtesam Abdalla, Markus M. Nöthen, Peter M. Krawitz, and Tzung-Chien Hsieh. Gestaltmatcher database - a fair database for medical imaging data of rare disorders. *medRxiv*, 2023. doi: 10.1101/2023.06.06.23290887.

[30] Hui Liu, Zi-Hua Mo, Hang Yang, Zheng-Fu Zhang, Dian Hong, Long Wen, Min-Yin Lin, Ying-Yi Zheng, Zhi-Wei Zhang, Xiao-Wei Xu, et al. Automatic facial recognition of williams-beuren syndrome based on deep convolutional neural networks. *Frontiers in Pediatrics*, 9:648255, 2021.

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

[32] Bishwas Mandal, Adaeze Okeukwu, and Yihong Theis. Masked face recognition using resnet-50. *ArXiv*, abs/2104.08997, 2021.

[33] Miguel Angel Meza Martínez, Mario Nadj, Moritz Langner, Peyman Toreini, and Alexander Maedche. Does this explanation help? designing local model-agnostic

explanation representations and an experimental evaluation using eye-tracking technology. *ACM Transactions on Interactive Intelligent Systems*, 2023.

[34] Eri Matsuyama et al. A deep learning interpretable model for novel coronavirus disease (covid-19) screening with chest ct images. *Journal of Biomedical Science and Engineering*, 13(07):140, 2020.

[35] Maria Mikhailenko, Nadezhda Maksimenko, and Mikhail Kurushkin. Eye-tracking in immersive virtual reality for education: a review of the current progress and applications. In *Frontiers in Education*, volume 7, page 697032. Frontiers Media SA, 2022.

[36] Masahiro Mitsuhara, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. St-abn: Visual explanation taking into account spatio-temporal information for video recognition. *arXiv preprint arXiv:2110.15574*, 2021.

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

[38] Satya M Muddamsetty, Mohammad NS Jahromi, and Thomas B Moeslund. Expert level evaluations for explainable ai (xai) methods in the medical domain. In *International Conference on Pattern Recognition*, pages 35–46. Springer, 2021.

[39] Satya M Muddamsetty, Mohammad NS Jahromi, Andreea E Ciontos, Laura M Fenoy, and Thomas B Moeslund. Visual explanation of black-box model: Similarity difference and uniqueness (sidu) method. *Pattern recognition*, 127: 108604, 2022.

[40] Myura Nagendran, Paul Festor, Matthieu Komorowski, Anthony Gordon, and Aldo A. Faisal. Eye-tracking of clinician behaviour with explainable AI decision support: a high-fidelity simulation study. In *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*, 2023.

[41] Wanli Ouyang, Xiaogang Wang, Xingyu Zeng, Shi Qiu, Ping Luo, Yonglong Tian, Hongsheng Li, Shuo Yang, Zhe Wang, Chen-Change Loy, et al. Deepid-net: Deformable deep convolutional neural networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2403–2412, 2015.

[42] Zhouxian Pan, Zhen Shen, Huijuan Zhu, Yin Bao, Siyu Liang, Shirui Wang, Xiangying Li, Lulu Niu, Xisong Dong, Xiuqin Shang, et al. Clinical application of an automatic facial recognition system based on deep learning for diagnosis of turner syndrome. *Endocrine*, 72:865–873, 2021.

[43] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models, 2018.

[44] Lauro A Pradela-Filho, William B Veloso, Iana VS Arantes, Juliana LM Gongoni, Davi M de Farias, Diele AG Araujo, and Thiago RLC Paixão. based analytical devices for point-of-need applications. *Microchimica Acta*, 190(5):179, 2023.

[45] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016.

[46] M. Robnik-Sikonja and Marko Bohanec. Perturbation-based explanations of prediction models. In *Human and Machine Learning*, 2018.

[47] Sam Sattarzadeh, Mahesh Sudhakar, Konstantinos N Plataniotis, Jongseong Jang, Yeonjeong Jeong, and Hyunwoo Kim. Integrated grad-cam: Sensitivity-aware visual explanation of deep convolutional networks via integrated gradient-based scoring. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1775–1779. IEEE, 2021.

[48] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[49] Boran Sekeroglu and Ilker Ozsahin. ¡? covid19?¿ detection of covid-19 from chest x-ray images using convolutional neural networks. *SLAS TECHNOLOGY: Translating Life Sciences Innovation*, 25(6):553–565, 2020.

[50] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*, 2016.

[51] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that?, 2017.

[52] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, oct 2019. doi: 10.1007/s11263-019-01228-7.

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

[54] Pushkar Shukla, Tanu Gupta, Aradhya Saini, Priyanka Singh, and Raman Balasubramanian. A deep learning frame-work for recognizing developmental disorders.

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

[56] Debjyoti Sinha and Mohamed El-Sharkawy. Thin mobilenet: An enhanced mobilenet architecture. In *2019 IEEE 10th annual ubiquitous computing, electronics & mobile communication conference (UEMCON)*, pages 0280–0285. IEEE, 2019.

[57] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net, 2015.

[58] Daniel Sáez Trigueros, Li Meng, and Margaret Hartnett. Enhancing convolutional neural networks for face recognition with occlusion maps and batch triplet loss. *Image and Vision Computing*, 79:99–108, 2018.

[59] Inam Ullah, Muwei Jian, Sumaira Hussain, Jie Guo, Hui Yu, Xing Wang, and Yilong Yin. A brief survey of visual saliency detection. *Multimedia Tools and Applications*, 79:34605–34645, 2020.

[60] Xiaoling Xia, Cui Xu, and Bing Nan. Inception-v3 for flower classification. In *2017 2nd international conference on image, vision and computing (ICIVC)*, pages 783–787. IEEE, 2017.

[61] Samir S Yadav and Shivajirao M Jadhav. Deep convolutional neural network based medical image classification for disease diagnosis. *Journal of Big data*, 6 (1):1–18, 2019.

[62] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014.

[63] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization, 2015.

[64] Ning Zhu. Neural architecture search for deep face recognition. *arXiv preprint arXiv:1904.09523*, 2019.