Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

Master's Thesis

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

Aswinkumar Vijayananth

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

Supervised by

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

November 2022

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

_____ _____

Date Aswinkumar Vijayananth

Abstract

Your abstract

Acknowledgements

Thanks to ....

Contents

List of Figures

List of Tables

1

Introduction

"Artificial Neural Networks (ANNs) are increasingly applied for medical image diagnostics" [1]. Medical image data such as scans produced from imaging devices like x-rays [2], Magnetic Resonance Imaging (MRI) [3], Computed Tomography (CT) [4], and ultra sound [5], waveforms produced from procedures like ElectroCardioGraphy (ECG) [6] and ElectroEncephaloGraphy (EEG), histological images [7], images of body parts, and admission notes [8] are fed into ANNs to perform tasks such as segmentation [9], classification [10] and abnormality detection [7].

Predictions made by neural network models are typically intended to be used in a clinical setting, to aid medical practitioners in diagnosing their patients. As a result, it can help in an overall reduction of "misdiagnosis", which is one of the most severe problems in health care [8]. Besides, the adoption of ANN based Artificial Intelligence (AI) systems facilitates early screening and identification of life-threatening conditions such as cancer in population with limited or no access to sub-specialty trained clinicians.

ANN models such as the ones used in AI systems for intracranial haemorrhage classification [4] and breast cancer prediction [11] are claimed to outperform clinicians in their respective tasks. In spite of their success in terms of predictive performance, the black box nature of ANNs restrains them from getting deployed in a clinical setting. Inherently, ANNs lack transparency and therefore are considered to be less dependable for high stake applications like medical diagnosis and autonomous driving. Besides, regulatory bodies such as "US FDA (United States Food and Drug Administration) require any clinical decision support software to explain the rationale or support for its decisions to enable the users to independently review the basis of their recommendations" [4]. Such contradictions further restrict the deployment of ANN based AI systems for performing medical diagnoses in a clinical setting.

"Explainable Artificial Intelligence (XAI) is a set of processes and methods that allows human users to comprehend and trust the results and output created by machine learning algorithms" [12]. Application of XAI techniques to Machine Learning (ML) models enables its human users to better understand their behavior and rationale behind their predictions. As a result, they improve transparency of models like ANNs, in turn making them more trustworthy for applications like medical diagnosis.

In the recent times, there has been a steep rise in the adoption of XAI methods to explain ANN models for medical diagnoses. A considerable number existing works focus on explaining neural network models that use radiological image data to perform tasks such as COVID classification [13], breast cancer risk assessment [5] and haemorrhage detection [4] respectively. A few other leverage XAI techniques to explain

neural network models used for detecting cancer from histopathological [7] and skin images [14].

The objective of this thesis is to use XAI to explain one such neural network model called GestaltMatcher [15], which surpasses the performance of clinical practitioners in the identification of certain rare genetic syndromes, from frontal facial images of patients. The findings of this work shall make the GestaltMatcher model more transparent and dependable, thereby taking it a step closer to be deployed in a clinical setting.

1.1 Motivation

"Rare genetic disorders affect more than 6.2% of global population" [15]. A significant fraction of the population with certain such disorders are characterized by facial abnormalities, which make up their respective facial phenotypes. The facial phenotypic information is used by clinical geneticists along with results from other laboratory tests such as molecular, to reach a diagnosis. However, the rarity in occurrence of such disorders combined with the lack of distinctive traits for a subset of them, makes their diagnosis a challenging task even for experienced medical practitioners.

ANN models such as the ones presented in DeepGestalt [10] and GestaltMatcher [15] comprise a promising step forward in using AI for the task of recognizing rare genetic disorders from facial phenotypes. Such works rely on databases like Face2Gene [16] and GestaltMatcher database (GMDB) [17] which offer a valuable collection of frontal facial images and other medical data of the patients with such rare disorders. The GestaltMatcher [15] model surpasses human expert's performance in the task of recognizing certain syndromes and their sub types. Therefore, deployment of such models in a clinical setting has the potential to significantly improve the speed and accuracy of diagnoses.

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

In general, ANN models are capable of learning novel discriminative features or regions from the training data, that are relevant to the task at hand. Associating such learned features or regions with the real world knowledge (with respect to the dataset) can provide new insights about the data and task at hand. In the case of genetic disorder classification, associating attention regions of SOTA models like DeepGestalt [10] and GestaltMatcher [15] in their frontal facial input images can enhance human kind's understanding about facial phenotypes of genetic disorders. Such a study to associate the attention regions of genetic disorder classifiers with the facial phenotypic information known to the medical community, is yet to be conducted.

1.2 Importance

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

The knowledge about facial phenotypes of rare genetic disorders are contained in resources like Human Phenotype Ontology (HPO) [18] and relevant medical literature. However, due to rare occurrences of such disorders, not all variations in their phenotypes are known to the medical community. Analyzing the attention regions of genetic disorder classifier models offers a way to discover facial regions that contain novel phenotypic traits, there by enhancing human kind's understanding of such rare medical conditions.

Datasets like GMDB [17] only contain a fractional number of instances per class, when compared with sizes of general purpose image classification datasets like ImageNet [19]. Most of the disorder classes contains samples in the order of tens and a few in the order of hundreds. This in turn demands use of highly effective data pre-processing and model training techniques to learn the most from a dataset. Analyzing attention regions of a classifier model trained on such datasets, enables an ML practitioner to identify any possible biases that the model could have learned, and consecutively enables him to adopt suitable techniques to remove them. This eventually makes the model more generalized and possibly increases its predictive performance.

1.3 Challenges and Difficulties

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

- **Dataset size and imbalance:** As briefly mentioned in the previous section, datasets like GMDB [15] are small-sized, and also are characterized by problems such as dataset imbalance and low image resolution. Figure **??** depicts the class-wise distribution of samples in GMDB. Such problems are caused by various factors like inherent rarity in occurrence of genetic syndromes, data-privacy constraints and lack of openly available datasets. The above listed issues of genetic syndrome datasets often have consequential effects on the performance and explainability of machine learning models they are trained with.

- **Low predictive performance of the classifier:** Although, SOTA genetic syndrome classifier models such as DeepGestalt [10] and GestaltMatcher [15] surpass human-level performance in diagnosing rare genetic disorders, their predictive perfomances are exceptionally low when compared with that of top classification models trained on large general purpose datasets. Besides, over-fitting is a common problem experienced by these models. In most of the existing works on XAI methods for neural networks, the research community has benchmarked them on high performance models. This raises doubts about the quality of explanations generated by XAI methods, when applied to low performance models.

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

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

1.4 Problem Statement

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

The selected set of techniques are applied to GestaltMatcher [15], in order to generate explanations associated with the model's predictions for every patient image in the GMDB [15] dataset. The generated explanations are then analyzed with intents to understand the behaviours of both the model and the chosen set of XAI techniques with different input categories. Besides analyzing the model's regions of interest in individual patient images, this work also studies its characteristic attention regions on a class level for specific syndromes. Such characteristic representations are produced by combining patient-wise attention maps of individual classes.

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

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

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

Research Questions

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

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

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

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

1.5 Structure

This report consists of seven chapters (including this chapter) with each discussing different aspects of the conducted research work. The first chapter introduced the reader to the research topic, and discusses the scope and significance of this work. Chapter 2 gives the necessary background knowledge of rare genetic syndromes and their diagnoses using GestaltMatcher [15], which is necessary to appreciate this work. The chapter also briefly introduces the reader to the field of XAI and discusses the taxonomy of explanation methods. In Chapter 3, a literature review of SOTA explanation methods considered for this research work is provided. Chapter 4 describes the systematic approach taken to address the problem at hand. The chapter gives the rationale behind the design of experiments and other choices made. The fifth chapter discusses implementation details and provides specifications of datasets and models used for experimentation. The results obtained from the conducted experiments are listed and analyzed in Chapter 6. Besides, it also presents and describes the formulated evaluation questionnaire. Finally, Chapter 7 concludes the report by summarizing the contributions of this project, and also discusses the possible future research directions.

2

Background

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit

ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

2.1 Rare Genetic Syndromes

2.2 Phenotypes

2.3 Human Phenotype Ontology

3

State of the Art

<What did the previous chapter explain?>. This chapter emphasizes on Explainable Artificial Intelligence (XAI) methods developed for convolutional neural network (CNN) models such as GestaltMatcher. The set of methods considered for this research work are categorized and explained in a concise manner.

3.1 Explainability Methods for Neural Networks

3.2 Input Attribution Methods

3.2.1 Occlusion Sensitivity Maps

Occlusion sensitivity mapping [20] is a model-agnostic perturbation based method, which generates explanations by manipulating parts of the input image. The approach is computationally expensive, O(#simultaneous occlusions * #features * #ablations_per_eval * 1/#strides), and is included in this work to verify if Gestalt Matcher model is focusing on key facial features, or simply using the surrounding context to produce predictions. This is achieved by systematically occluding different portions of the input image with a black square or rectangular mask, and computing the difference in outputs (logit scores of the target class). In this work, we use a black square mask of dimensions 10x10. Important portions of the input when occluded, result in relatively larger logit score differences, than the trivial ones. The differences are plotted on the image, yielding the occlusion sensitivity maps.

3.2.2 Deconvolution

Zeiler and Fergus proposed the "Deconvolution" [20] approach to visualize and provide insights into the functions learned by intermediate layers of a CNN. It is one of the earliest attribution techniques, which produces visualizations based on computing gradient of loss function with respect to a given input. The work acts a baseline till date for development and evaluation of new pixel attribution techniques.
The method uses a deconvolution counterpart for every building block of a CNN, to obtain reverse mapping from features to input pixels. The idea of deconvolution was first introduced by Zeiler et al. [21], as a way to perform unsupervised learning. In order to obtain attribution maps using the Deconvolution approach, the first step is to attach each block of convnet with its deconvolution counterpart as shown in the figure

1. Every activation except the ones belonging to the class of interest is set to zero. The activation value is then backpropagated through the deconvolution blocks such as unpooling, rectification and transposed convolution, all the way to the input layer. Deconvolution blocks act according to a pre-defined set of rules. The transposed convolution block performs the inverse of convolution operation by using transposed versions of the same filters. This is equivalent to flipping a given filter both in vertical and horizontal directions. In order to backtrack activations through max-pooling layer (.i.e. using the unpooling layer), indices corresponding to maximum activations in every layer, are first stored during the forward pass and later retrieved during the back propagation phase. However, the use of indices or switches from the forward pass, constrains the visualization on the input image [22].
Authors test their method on an AlexNet [23] trained on the ImageNet [23], Caltech-101 [24] and Caltech-256 [25] and PASCAL2012 [26] datasets. As a first step, they visualize the top 9 feature maps of the each of the first five layers, to show the proportional increase of complexity in features with respect to their receptive fields. The visualizations are obtained by backtracking the strongest activation of a feature map for most of the data samples, all the way until a given input, using the deconvolution rules.The paper also discusses about the proportionality between the time taken for a given layer to learn features and its corresponding depth. Further, it shows that features learned by top layers are more invariant to transformations like translations, rotations and scale changes.
The work evaluates itself by qualitatively comparing its resulting attribution maps with occlusion sensitivity maps. Occlusion sensitivity maps are obtained by systematically occluding portion of an image and analyzing the given classifier's output, to determine the most discriminative regions as shown in the first image of Figure 3.1.

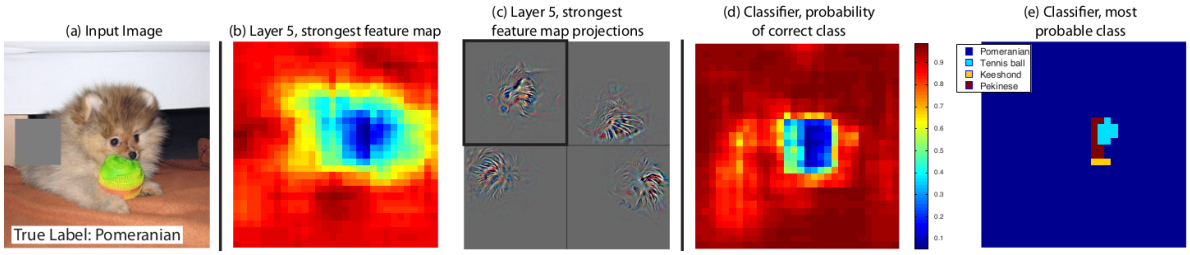

Figure 3.1: An illustration of occlusion sensitivity mapping

3.2.3 Saliency Maps

Simonyan et al. [27] propose two visualization techniques with intents to generate an image which maximizes the class score, and to compute a class-specific saliency map for a given input. The first technique numerically optimizes the input image while the other computes the spatial support of a given class in an input. This work is one of the earliest to leverage saliency maps for the task of weakly-supervised object segmentation. Authors demonstrate the proposed techniques by applying to a deep convnet trained on the ILSVRC-2013 dataset [28].

Class Model Visualization

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

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

Image-Specific Class Saliency

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

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

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

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

In order to obtain the saliency map for a multi-channel image, the maximum magnitude of gradient for a given position across channels is used. Class saliency maps thus produced are used as initial points to compute object segmentation mask using the GraphCut algorithm. (**??**raphcut). Foreground and background portions are considered as Gaussian Mixture Models and the former is estimated from pixels with saliency value higher than the 95% quantile of the image's saliency distribution.On the other hand, the latter is estimated from pixels with saliency smaller than 30% quantile.
The work evaluates its outputs on test split of the data set for the localization task in the ILSVRC-2013 [28] challenge, where it achieves 46.4% top-5 error in spite of its simplicity. In hindsight, apart from the strategy used to reverse the ReLU layer this approach is equivalent to Deconvnet [20].

3.2.4 Guided Backpropagation

Springenberg et al. proposed a new variant of the deconvolution approach in their work, as a means to analyze their "All Convolutional Net" architecture, which replaces max-pooling layers by convolutional layers with increased stride. The first objective of this work was to empirically prove the equivalence (in terms of predictive performance) between a max-pooling layer and a convolutional layer with an increased stride. This was achieved by evaluating a custom cnn model with max-pooling layers against its convolutional counterpart on three datasets: CIFAR-10, CIFAR-100 and ILSVRC 2012. In all cases,

performances of the all convolutional models were on par with their max pooling counterparts. The second objective was to determine the quality of representations learned by the intermediate layers of the all convolutional neural network models. In order to achieve this, authors proposed a visualization approach, which can be considered as a slight modification of the Deconvnet [20] technique.

‘

Back propagation through ReLU One of the most significant difference between Saliency Maps [27], Deconvnet [20] and Guided backprop [22] approaches is the strategy used by these methods to backpropagate gradients through the ReLU layer.

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

Figure b illustrates differences between back propagation strategies with help of an example feature map.

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

3.2.5 Deep LIFT

Shrikumar et al. propose an attribution technique, which assigns scores to input pixels based on their contribution to change in activation of each neuronal unit with respect to a reference value, which is chosen based on the problem in hand. Authors claim the technique to be computationally inexpensive, and yield meaningful representations in comparison with other methods. Besides, the technique is devised in a way that it is suitable for neural net variants apart from CNN like Recurrent Neural Networks (RNN). Intuitively, Deep LIFT seeks to explain the deviation in output from some reference output in terms of deviation in input from its respective reference. The motivation to use a reference value, arises from the need to handle the “saturation problem”.

Saturation problem

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

Working

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

- Linear rule for linear layers such as the fully-connected and convolutions
- Rescale rule for non-linear transformations such as ReLU, tanh or sigmoid
- Reveal-cancel rule which enables the measurement of marginal effect of having an input over all possible orderings, similar to Shapely values [29].

The technique also takes in to account that a zero contribution could be due to cancellation of positive and negative contributions from a pair of entities. The application of DeepLIFT also depends on the choice of target neuron's layer, logit or softmax layer in a classifier network, for example. Authors demonstrate the approach by applying it to a CNN trained on MNIST dataset [30] and an RNN trained on simulated genomic data.
The benchmarking on MNIST classifier for various methods was performed on basis of the change in log-odds score, when a selected pixels of a given image belonging to class c0 were erased to convert it to an image of some target class ct. The upper section of figure illustrates the result of masking pixels ordered as the most significant for converting to target classes 3 and 6. As graphically represented in bottom part of figure, DeepLIFT outperformed other backpropagation based methods.

3.3 Layer Attribution Methods

3.3.1 GradCAM

Selvaraju et al. propose GradCAM [31], a pixel-attribution technique leveraging the gradients of a given target class with respect to a convolution layer. This approach can be considered as a generalization of CAM for CNNs with fully connected layers. Besides, the technique is applicable to neural network architectures for tasks such as image captioning and visual-question answering. The work also comes up with a means to convert class-agnostic fine grained visualization approaches like Guided-Backprop and Deconvolution to be class-discriminative in nature.
Grad-CAM is one of the most widely used approaches in obtaining attribution map based explanations, for neural network models used with medical data (provide references). Concisely put, Grad-CAM creates a visualization of which parts of an input image a convolutional layer "looks" for a certain class prediction. The working of Grad-CAM can be described in the following steps:

1. Perform a forward pass of the input image through CNN to obtain logit scores for all classes.
2. Except for the logit activation of the class of interest, set other activations to zero.
3. Backpropagate the gradient of the class of interest all the way to the chosen convolutional layer containing k feature maps Ak. $\partial y_c / \partial A_k$
4. Global pooling: For every feature map, weigh their pixels based on the gradient value, and obtain their weighted mean value scaled by constant Z where y_c refers to logit score for class c, A^k refers to the k^{th} activation map of dimensions ixj and α refers to the computed weightage.
5. Obtain the α weighted average of all pooled feature maps and apply a ReLU to filter out negative values, if any present. where Lc refers to the localization map produced by the Grad-CAM for the class of interest c and Ak refers to the kth feature map.
6. An element-wise multiplication operation is applied on the scaled version (input image dimensions) of Grad-CAM's maps, and outputs of fine-grained visualization approaches like Guided-Backpropagation and Deconvolution, to obtain fine-grained yet class discriminative visualizations. These combinations are termed as Guided-GradCAM and Guided-Deconvolution approaches respectively. Guided Grad-CAM visualizations are both high-resolution and class-discriminative.

Typically, last convolutional layers of a neural network model are chosen for Grad-CAM as they contain high-level semantics and detailed spatial information (**??**ame paper). This is because the attribution maps produced by the methods becomes progressively worse qualitatively, as we use earlier convolutional layers which have relatively lesser receptive fields. In a classification network, logit scores of the target class are used for gradient computations. However, any differentiable activation value can be backpropagated. The embedding based Grad-CAM discussed in the section, draws motivation from this fact. In the original implementation of the attribution method, a Rectified Linear Unit (ReLU) function is applied on heat

maps, to obtain only regions that positively affect the given prediction. However, to get better insights into prediction decisions, this work uses unfiltered heat maps, which consists of negatively correlated regions as well.
Authors evaluate the method on three different tasks: weakly-supervised localization, weakly-supervised segmentation and Pointing Game experiment [**?**]. The approach is also evaluated and benchmarked on its ability to generate discriminative, trust-worthy, faithful and interpretable attribution maps. A couple of neural network models with different architectures (AlexNet and VGG-16) are used for evaluation, to determine the method's performance consistency across architectures. Finally, the work also demonstrates the association of a given concept with a neuron, similar to the ones presented by (Visualizing and understanding convolutional networks, Object detectors emerge in deep scene CNNs.)
Authors also demonstrate it use in analyzing failure modes in neural network models and identifying bias in dataset. The approach is considered to be computationally in-expensive when compared to perturbation based methods such as LIME or SHAP, yet producing interpretable and discriminative attribution maps.

3.3.2 HiResolution Class Activation Mapping (HiResCAM)

Draelos et. al propose HiResCAM [32] as a generalization of CAM, and the method aims to use feature maps of a layer directly for visualization without averaging unlike GradCAM. The authors demonstrate why the gradient averaging step in attribution methods like GradCAM limit them from being reliable to highlight a neural network model's regions of attention in an image. Besides, they mathematically show the conceptual relationship between HiResCAM and other CAM approaches. The faithfulness of explanations produced by HiResCAM is shown, by conducting certain experiments on natural images whose results were verified using crowd-sourced assessment.
HiResCAM can be seen as a modification of Grad-CAM, as it is primarily designed to address the later's limitation caused by its averaging step. As described in 3.3.1 and illustrated in the figure below, importance weight α^f for a feature map f, is computed by computing the mean of gradients over spatial dimensions. However, the process of averaging may cause the final attribution map not highlight locations within the image which the model used to make predictions.
The above figure illustrates the working of GradCAM and HiResCAM methods. Let s_m be the logit score of a CNN model for a given image input, with respect to class m. GradCAM computes the derivative $\partial s_m / \partial A$, which yields the gradient for every position of a feature map. The gradients are then averaged to yield ${\alpha_m}^f$ for every map. Following which, the gradient values are multiplied with feature activation values ($\alpha_m^f A^f$). While HiResCAM also computes $\partial s_m / \partial A$, they are multiplied directly with feature map activation values leaving out the averaging procedure. Feature map values across channels are then combined as sum to yield the chosen layer's attribution map. By skipping the averaging procedure, HiResCAM preserves the effect of the gradients across every feature map.
Similar to GradCAM, authors recommend using the last convolutional layers for visualization. Apart from proposing the method, the authors also argue that explaining a model doesn't correspond to weakly-supervised segmentation, as the objective of the former to reveal the working of the model while the latter is to localize an object of interest. The work reiterates the fact that every explanation method

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

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

3.3.3 Full-Grad

Full-grad [33] is a gradient-based attribution method, which produces saliency map explanations by providing attributions to both the input image and the neuronal units of intermediate layers of a given neural network. The method is layer-agnostic and produces a single attribution map for an input image passed through a network by model, considering different scales of features, starting from pixel-level to high-level features. This makes the method to produce sharper saliency maps than other techniques.
Throughout the work, the authors emphasize the importance of a saliency map based explainability method to satisfy two key properties:

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

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

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

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

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

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

2. RemOve And Retrain (ROAR) experiment [34]: In this experiment, the top-k pixels highlighted by an attribution method for an entire data set are removed, and the considered classifier is retrained on this modified data set. The attribution method corresponding to the least accurate classifier model is considered to be the best, as it indicates that the method correctly identified the salient pixels.

The work reports that the FullGrad method outperformed the other considered methods (Input-gradients [35], Integrated-gradients [36], Smooth Grad [37] and GradCAM [31]) in both the experiments. Apart from the quantitative evaluation, the authors also conducted a visual inspection in which they found the method to produce meaningful attribution maps which highlight both a given object's boundaries and interiors clearly. Figure

3.4 Summary

4

Methodology

Last chapter gave a literature review of different attribution methods available to explain neural network models like GestaltMatcher. This chapter explains the methodical approach taken to address the research problem at hand, by discussing the design of experiments and rationale behind different choices made to conduct them.

4.1 Selection of Methods

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

- **Saturation problem:** As briefly described in 3.2.5, saturation problem occurs when the output of a neural network model gets saturated and its gradients with respect to inputs become insensitive, thereby affecting the quality of generated attribution maps. Shrikumar et al. [35] reported that occlusion sensitivity maps, layer-wise relevance propagation, saliency maps, guided-backpropation and guided-GradCAM methods suffer from this problem. Besides, they also report the failure of deconvolution approach, in the presence of discontinuos input gradients.

- **Sensitivity to changes in model and data:** Adebayo et al. [38] investigated the faithfulness of explanations produced by popular attribution methods. The chosen methods were evaluated based on their sensitivity to randomization induced in the considered neural network model's weights and labels of the training instances. Authors report that guided-backpropagation and guided-GradCAM remained insensitive to the changes in model and data, having functioned merely as edge detectors. "GradCAM" passed this sanity check.

- **Robustness of explanations:** An interpretability method is robust when it generate similar explanations for similar inputs. David Alvarez et al. [39] and Yi-han Sheu [40] report that DeepLIFT experiences robustness issues and produces inaccurate results in the presence of multiplicative interactions between features. Besides the lack of robustness, the success of DeepLIFT depends on factors such the choice of reference/baseline image, which is difficult to determine.

- **Relevance to the problem at hand:** Along with the above list dimensions, this work shortlists the methods based on their ability to produce meaningful explanations in the context of genetic

syndrome recognition. Figure **??** shows few sample attribution maps generated from applying both the considered sets of input and layer attribution maps to the GestaltMatcher model. It can be observed that contours of regions in attribution maps produced by layer attribution methods closely match the scales of phenotypic features of genetic syndromes and parts of human face, in general. This effect is possibly due to the inherent working principles of the two classes of attribution methods.

Layer attribution maps are obtained by scaling and superimposing feature map activations of the last convolutional layers of a CNN, which are responsible for extracting high-level features from a given input image. This nature makes them more suitable for the problem at hand than input attribution methods, which represent pixel-wise attribution methods.

4.1.1 More Reasons to Consider Layer Attribution Methods

In addition to the above discussed reasons, it is also observed that layer attribution approaches like GradCAM are widely used to explain neural network models for medical diagnosis [] [] []. However, the downside of GradCAM is it approaches attribution as weakly supervised segmenation problem and "sometimes higlight locations the model didnot actually use" [32]. HiResCAM [32], a recent successor of GradCAM overcomes this issue and produces more faithfulness explanations. Therefore, the method is included in the scope of this work.

GradCAM and HiResCAM are layer-specific and therefore their explanations are limited by the choice of the convolutional layers made. FullGrad overcomes this deficit by considering attributions across all neuronal units of a model and thus becomes a candidate for experimentation. Along with these CAM techniques, occlusion sensitivity mapping is included as a reference explanation method, to better understand the classifier model's regions of attention.

4.2 Datasets

4.3 Choice of Syndromes for Evaluation

4.4 Neural Network Models

4.5 Design of Experiments

4.5.1 Overview

4.5.2 Objectives

5

Solution

5.1 Implementation Details

5.1.1 Patient-wise Attribution Maps

5.1.2 Attribution Maps for Clinical Evaluation

5.1.3 Similarity Maps

5.1.4 Composite Faces

5.1.5 Generic Attribution Maps

5.2 Implementation details

6

Evaluation and Results

6.1 Metrics

Describe the experiments/evaluation you are performing to analyse your method.

6.2 Qualitative Analysis

6.2.1 Patient-wise Maps

6.2.2 Syndrome-wise Maps

6.2.3 Composite Faces

6.3 Effect of Class Imbalance on the Explainability of Models

6.4 Quantitative Analysis

6.4.1 Use of Occlusion Sensitivity Maps

6.4.2 Eye-tracking based Evaluation

6.4.3 Similarity Mapping

6.5 Evaluation Summary

Describe the results of your experiments in detail.

7

Conclusions

7.1 Contributions

7.2 Lessons learned

7.3 Future work

A

Design Details

Your first appendix

B

Parameters

Your second chapter appendix

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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