Model-specific Explainable Artificial Intelligence techniques: State-of-the-art, Advantages and Limitations.

Arghem Khan
Supervisor(s): Chhagan Lal, Mauro Conti
EEMCS, Delft University of Technology, The Netherlands
22-6-2022

A Dissertation Submitted to EEMCS faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

Abstract

Artificial Intelligence (AI) and Machine learning (ML) applications are being widely used to solve different problems in different sectors. These applications have enabled the human-effort and involvement to be very low. The AI/ML systems make their own predictions and do not require a great deal of human help. However, over the last few years several incidents of the developed systems have led to questions regarding the transparency of those AI/ML systems. Without expertise, it is not always as straightforward to understand certain predictions. This pressing issue has led to the emerging topic of Explainable Artificial Intelligence (XAI). In this research, we will present the current work on a specific type of XAI, namely model-specific XAI. Model-specific XAI techniques are particular to certain types of ML techniques. We will look into several recent model-specific XAI techniques and provide the advantages and disadvantages. Within similarities we find that there is a set of general requirements that the techniques should adhere to (expertise, bias, time, privacy and performance). We characterize the techniques in feature-based, concept-based and logic-based. With regard to future work, there is room for improvement on several areas. For example, this includes work from exploring hybrid techniques to investigating how current techniques can improve the privacy.

1 Introduction

Within the Digital Age companies are rapidly growing in technology and the world is growing along. Artificial Intelligence (AI) and Machine Learning (ML) have been in the nucleus of the growth. Applications and systems in a wide variety of fields have expanded their use of ML. For example in healthcare, finance, agriculture and criminal justice [1]. However, what goes on at an implementation level of these models, has been characterized as black-box. This has led to catastrophic issues and failures. For example, the system failure of a self-driving car leading to deadly accidents [2] or uncensored sounds being played while intending to play a children's song [3]. The trillion dollar crash, caused by stock market AI software for trading, is another example on how influential these issues have been [4]. The pervasive influence of these black-box systems has raised questions on the transparency, bias, trust and ethics of these systems. Who remains accountable when things go wrong? But conversely, who knows why things go right? Is there room for improvement? Can systems be fully trusted if there is no transparency? There's no form of accountability and there's ground being lost for models that could be improved even more. The solution has been proposed in the form of Explainable Artificial Intelligence (XAI).

XAI moves towards gaining explainability and interpretability. In this way models and their predictions can be better understood, well analyzed and improved. A system needs to be transparent in its working. For an AI model this means that the predictions need to be clear. Human-understandable explanations need to be provided that explain the working of a system. The inner-workings of the system need to show why a certain decision has been made. For example, in a Decision Tree problem, the path can be shown. In Neural Networks (NN) the importance of some features can be highlighted. We will discuss these workings in more detail, in Section 2.

In order to be able to define, categorize and analyze methodologies for XAI, it would be helpful to understand what explainability entails. Even though there is not one single definition for explainability [5], we can measure the level of explainability to some extent. Namely, the level to which one can grasp the reason of the decision that has been made [6]. Secondly, the level to which one can consistently predict the outcome of a model [7].

Within XAI an ontology of the domain has been formulated that distinguishes two interpretability techniques [8]. The two main categories here are model-agnostic and model-specific. The latter one will be the topic of this research. Model-specific XAI techniques are the techniques that apply to a specific type of AI model or technique. On the other hand, model-agnostic techniques hold in general and are thus not bound to a model. Since this research focuses on model-specific technique, we paper will cover techniques that benefit Deep Learning (i.e. Deep Neural Networks or Convolutional Neural Networks).

The main question of this research is to review the current status on model-specific XAI. This includes the comparison between different interpretability methods for deep learning methods, leading to a characterization. This research is organized as follows. Section 2 provides an insight into some of the existing model-specific XAI approaches. This is the set of approaches on which this research is based. Section 3 presents the requirements for a good model-specific XAI technique, alongside a characterization of the foundings in Section 2. Moreover, Section 4 describes some ethical aspects of this research. Lastly, section 5 provides some direction for future work, followed by the conclusion in Section 6.

2 Background & Preliminaries

XAI Overview

Even though the large chunk of publications on XAI are from the last couple of years, the topic is by no means a new one. The term XAI itself is used since 2004, but the concept of it dates back to 1958 [9]. Going further in time, the popularity decreased. The need of expert knowledge created a stigma in the early 90's. Moving along, Neural Networks (NN) were used more over time in the mid 90's. It was not until 1992, where several methods were introduced to visualize data regarding the decision making of NN's [10].

In the upcoming years after 1992, several contributions have been made, under which the research from the Defence

Advanced Research Projects Agency (DARPA), with its research on XAI in 2017. However, the most ground breaking introduction came in 2018 in the form of the European General Data Protection Regulation (GDPR) with the "right of explanation". This strong regulation mentions that it is a right to explain the output of an algorithm. After this, new frameworks were developed, focusing more on accountability, gaining trust and transparency. Lastly, the latest important addition was the research from a more social perspective point of view. The different aspects of psychology and interaction have been stated to be added to the frameworks for XAI [6].

Addressing the goals for XAI in general can be difficult. A different field might have different requirements [11]. The writers of the article mention that different research areas can have a very different set of priorities because of the different specializations that are required. Therefore, different researchers from different domains are developing their own research on XAI.

Moreover, the research community overall is adapting an algorithm-centric approach [12]. This means that assumptions are made on how interpretable an algorithm is. Researchers and developers assume that an algorithm is understandable without the necessary human tests being taken for this. Additionally, not all the different researchers from different areas have the same mathematical knowledge and expertise. Hence, there is a disconnection between the mathematical formulation and the explainability.

Study of Model-Specific XAI Techniques

As mentioned before, XAI is developing to be an area of greater interest. This means that different approaches are being developed and used. Therefore, this research does not aim to cover all the possible model-specific XAI techniques. Rather, this research will cover some of the model-specific XAI techniques that are used in Deep Learning, listed in Table 1. Deep Learning is a specific type of ML technique. The reason why Deep Learning is much used, comes from the ability of these type of models to learn which features to focus on. Because of this, these systems can solve rather complex problems without much guidance being required from the programmer. Deep Learning builds so called Neural Networks (NN). NN's are designed in such a way that they mimic the structure of a brain. The output of one layer is the input of the second layer. Therefore, the computational load can be high. Internally, this is how the human brain works as well.

Perturbation

Perturbation methods essentially modify (perturbate) the input values and observe what changes there are to the output. The starting point for perturbating any input is by looking at which variations of the input can be applied, but more importantly, input variations that make sense. Not all variations of input variables are realistic [23]. This article shows that the goal is to find regions that are maximally informative. Another reason for finding meaningful perturbation inputs is because of the influence unmeaningful perturbations have on the artifact of AI systems. Because you are tweaking the inputs with inputs that do not necessarily make sense, the model might learn things that are not relevant and are therefore not needed.

Table 1: The Model-Specific XAI techniques that benefit Deep Learning methods. These are the methods for which the comparison will be performed in this research.

Model-Specific XAI Technique	Reference	Year
DeepLIFT	[13]	2020
xNN	[14]	2017
SIDU	[15]	2018
ACE	[16]	2021
DeepRED	[17]	2019
Net2Vec	[18]	2016
Grad-CAM	[19]	2018
Concept Embedding & ILP	[20]	2017
Integrated Gradients	[21]	2020
TCAV	[22]	2017
Perturbation	[23]	2018
NBDT	[24]	2017

Looking at the time efficiency, perturbation methods are not efficient. This is because for each input change, a whole forward propagation through the network is needed to compute the output. This means that for each input, the whole network can be traversed.

Even though this research does not aim to go into the internal workings of the different techniques, some background knowledge is necessary to grasp, in order to compare different techniques. One of these concepts is the *saturation problem*. As mentioned before, the perturbation methods changes the inputs to see if there is a change in the output prediction. However, this reasoning can lead to faulty conclusions. Saturated nodes in a NN are nodes that are not affected by change. Small changes in the hidden layers weights does not really reflect in the final weight [25]. However, perturbation techniques do not account for such nodes and overlook this. This can cause faulty scores for the inputs.

Another problem is the *thresholding problem*. When gradients are discontinuous, sudden jumps of gradients lead to misleading scores for the input relevance. Again, the actual mathematical working of the technique is not explained here. The important take-away is the fact that there is a thresholding problem that leads to misleading importance scores for input values and perturbation methods do not account for this.

DeepLIFT

The idea of DeepLIFT is to start with the output of the NN. Consecutively, each output prediction is decomposed to the specific input, in a back propagation manner [13]. This means that scores get assigned to an input feature, depending on the contribution of that particular feature. This is called an attribution method. The article mentions that at the end, each input gets such an attribution. This input importance score is dependant on the output that we calculate the importance for.

The fact that DeepLIFT goes over the model in a backpropagation manner makes it already more time efficient. The reason for this is that with backpropagation, one gets all the importance scores in one single pass. As mentioned before,

perturbation requires to go over the whole network, for each input. But in general, this technique is still considered time inefficient [8]. The reasoning behind this does not come from the inefficiency in the working of DeepLIFT. But rather, most of the mentioned techniques are already rather time inefficient since they need to extract information over the whole network. More on this will be explained in Section 3. Moreover, the internal working of DeepLIFT is mathematically sound. Therefore, one does not need to know anything about the internal workings of DeepLIFT to use it. Some other techniques require the need of some internal information (e.g. Integrated Gradients [21]).

Looking at the positives of DeepLIFT, it does solve the saturation and thresholding problem. However, the reason why DeepLIFT solves these problems lies deeper into the workings of the method. But in short, both problems are solved by the difference from reference activation. This is a way to keep giving the saturated nodes importance in defining the input scores and not have sudden jumps in gradients, that lead to irrelevant scores. On the negative side, DeepLIFT is not deemed consistent. Two equivalent networks do not always assign the same importance to the input features. More on this will follow in Section 3.

Integrated Gradients

In a similar way to DeepLIFT, the Integrated Gradients method also attributes importance to the input according to the output predictions. This is why it also solves the thresholding and saturation problem. [21]. The difference internally between the two is that Integrated Gradients take a different approach by using the integral of each gradient. The writers of this article show that a gradient represents the model scores of each input, with respect to the output. Involving calculations (i.e. integrals) takes more time. Hence, this approach is less time efficient than DeepLIFT.

Also, this technique still gives misleading results. This comes from the fact that such feature-based techniques are hard to evaluate [13]. Therefore, the authors of this article have not focused on actually evaluating this technique. Instead, they identify two axioms that each attribution method should have [21]. The authors argue that they solve their error prone approach by adhering to the axioms. Practically, for pixel detection, this means that wrong pixels get highlighted.

Explainable Neural Networks (xNN)

The Explainable Neural Network (xNN) is a particular design of a NN to learn features that can be explained [14]. In this article, a specific structure of the NN is provided in order to extract features that can be explained. Also, the article mentions that feed forward NN's usually have fully connected layers. This means that the output of a layer i is connected to all inputs of layer $i+1$.

The way xNN approaches their structure is by removing and limiting some connections between nodes, and hence tracking away from the fully connected network. Their structure is based on a so called Additive Index Model. The architecture of the xNN is shown in Figure 1.

The first layer is the same as the normal feed forward NN. This is called the projection layer. Then the difference lies within the first hidden layer and the last hidden layer. The

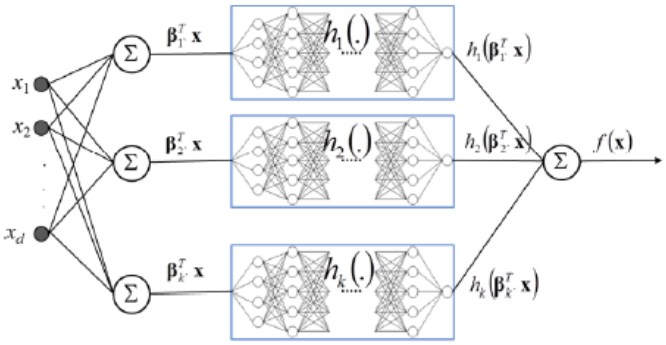


Figure 1: The architecture of xNN

layers in between are the subnetwork. The subnetwork itself is a fully connected NN. The first hidden layer has a linear transformation which is the input to each subnetwork. Finally, the output of the subnetwork provides a final output of the network with a linear combination. The weights that it has learned here are corresponding to the input variables since that structure was also linear. In this way, we gather the values of the inputs to this output.

The advantage of the xNN, is that one could use it as a surrogate model from the start and not only use it *a-posteriori*. This means that one could use the xNN afterwards to analyze, but could also adhere to the hyper parameters and construction of the xNN, to train the model from the get go. However, the article mentions that in practice this technique has very good performance on simpler models. Handling more complex examples is still subject to research. This means that this approach is not very practical, yet.

Another limitation is the computational cost of using the xNN. This comes from the fact that each subnetwork itself is a completely connected NN. This means that depending on how many input variables k there are, there are also k subnetworks that need to be processed.

SIDU

SIDU is a perturbation-based method for visual explanation. Here, the input and inner workings are perturbated to see changes in the output [15]. SIDU aims for visual explanation to classify objects in an image. It does this by defining a heatmap of the original image to show what objects of the images are important. Therefore, it functions on Convolutional Neural Networks (CNN). The article mentions that the method generates a visual map that generalizes the most important object as a whole in the image. Other methods, based on such heat maps classify objects too general or too specific.

The way it generates this heat map is different from other visual based methods (e.g. GRAD-CAM [19]). They both use the properties of the CNN, by extracting the features from the last layer. In general, the weights of this layer provide which pixels in the image are more important than others. The difference in accuracy of the different approaches lies in how they process this last layer of the CNN. SIDU does this by computing the so called Similarity Difference and Uniqueness scores to find out what the important regions are in a picture, to define an object.

ACE

Previously, we have discussed techniques that in one way or the other extract importance to the input values. These are feature-based explanations. However, these methods in general have drawbacks. By assigning importance scores to the input variables, you can still not summarize an explanation automatically, since these are not actual humanly readable explanations. Rather, it is just a score that shows which inputs are more important than the other.
Therefore, another technique is concept-based [16]. One example of a concept-based technique is ACE. The goal is to create explanations that are human-understandable. This is done by highlighting important concepts. A simple example of this is given for finding a police van in an image. A police van can be detected by wheels and a police logo. These are the two concepts that could detect a police van. The concepts are not detected, but rather labeled beforehand by human work.

In general, ACE extracts concepts from the input. The concepts are labeled beforehand. ACE works generally with images. The article mentions that ACE first categorizes concepts more generally in each step. It starts with very detailed parts, like colors and textures and ends at actual objects in the image.

Secondly, several objects will be grouped that could be in the same concept. Concepts have a certain importance and hence certain objects get highlighted. Those objects are the explanation of the image. For highlighting the important concepts, TCAV is used. TCAV is a method itself, which we will discuss later in this section.

As mentioned before, the importance of the concepts is humanly dependent. This requires human work and introduces human bias. This means that this method is efficient once the concepts are known and ranked, but are quite cumbersome before. Moreover, ACE is performed on image data. However, there are different class types that require explanation (i.e. text). Therefore, a future line of work could be to introduce ACE into different types of inputs.

DeepRED

DeepRED (Deep neural network Rule Extraction via Decision tree induction) uses as the name says, extracts rules in the form of a decision tree [17]. This article mentions that previous research on such rule-based approaches only focused on NN's that have one hidden layer. However, in practice a NN has many more hidden layers. Otherwise, what we research would not be as black-box and no real problem would exist. Therefore, their aim was to develop a technique that also works for a Deep Neural Network (DNN), with many hidden layers.

The main idea of the technique is to extract rules for each layer, which is done in two steps. The first step is the rule extraction step. Within this step, they start in a back propagation manner, from the output. For the first hidden layer, starting backwards, simple rule extractions are performed. From this, simple decision trees are generated. This same step is done for each hidden layer and all layers, resulting into many rules for all the hidden layers. The second step is a simple merging step. Each rule is merged in such a way that it makes sense for a human. The end result is one rule that explains the outcome of the NN.

An advantage of DeepRED is that it uses a backpropagation in the first step, to gain all the individual explanations. As mentioned before, using backpropagation is computationally more efficient than perturbation. Also, DeepRED starts with a filtering at the start which looks at the inputs that have no influence to the outcome. This brings the advantage that some rules can be ignored in the final outcome.

Even though the explanation of this technique seems straight forward, there are some downsides. The fact that the technique is very straightforward means there is a trade-off between simplicity and understandability [26]. In general, rule-based techniques are easier to understand. However, this impacts the accuracy of the explanation. Lastly, this approach only works for a classification problem, so far. Future work of this approach looks at applying the algorithm to different types of AI models.

Net2Vec

The Net2Vec technique looks at a combination of NN filters. The question that this framework answers is whether or not one filter is representative enough for an object [18]. They argue that using multiple filter activations within a layer gives overall better performance, instead of just using one filter to represent a single object. This is because a concept can not be linked to a filter. The article mentions that a concept could be scattered over multiple filters of the CNN.

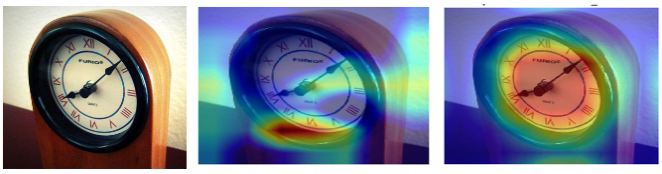
The reasoning behind this is quite intuitive. The amount of layers and nodes in the CNN do not correspond to the amount of concepts in the image. Hence you can not match semantic concept to filters. The amount of filters tends to be a smaller number.

Finally, the authors also argue that other techniques that only look at maximal filters, are not accurate enough. This means that just looking at maximal filters does not cover the whole semantic concept. Hence, one should look at a combination of filters instead.

Grad-CAM

Grad-CAM (Gradient-weighted Class Activation Mapping), is a gradient-based method for visual explanations [19]. We have discussed SIDU before. SIDU also is a visual explanation. However, SIDU uses backpropagation and processes the final layer scores of the CNN differently.
As explained for SIDU, the last layer of the CNN is used to extract information on the important regions of the image. This layer is a fully connected layer, that is a bout classification.

The previous layer does some feature extraction. Simply explained, these final layers define more concrete objects, layer by layer. This means that the first layer might see lines and shapes, whereas the second last layer extracts the head of a dog, the face of a person etc. Since the last layer classifies what the image is, this is the layer from which we can extract an explanation. This is done by using the pixel importance of that picture. Grad-CAM computes a weight value of every pixel of the image and applies it to the original image. The way these weights are calculated is by the use of the gradients



(a) Original Image (b) GRAD-CAM (c) SIDU

Figure 2: Grad-Cam inconsistent heatmap compared to SIDU

and the feature map. This leads to a heatmap that highlights the pixels that classify the image.

This gradient-based methods performs worse than SIDU. This comparison is based on time and accuracy of the explanation [15]. As shown in Figure 2, SIDU has a more accurate heatmap than GRAD-CAM. In terms of time complexity, Grad-CAM is more computationally expensive than SIDU.

Concept Embedding & ILP

Many approaches that we have discussed so far use visual explanation of the output, in form of heatmaps. Different techniques have different inner workings. However, another line of research moves towards using logic for your explanation (e.g. DeepRED). Another example of this is Concept Embedding & ILP. The goal is to derive more expressive explanations of DNN's [20].

Authors in [20] argue that there are certain aspects missing from the current way of work that limit expressiveness. This approach aims to take these into account. Heat maps do not identify the actual meaning of the object. For example, highlighting an eye is not helpful enough when you are identifying emotions. Eyes open or closed could be from a different type of emotion. Moreover, a person holding a flower in a picture does not deem them not being a terrorist. This is the problem of negation. This means that some conclusions can not be made by just highlighting one part of the image. but also, several concepts have same objects. You can not identify differences easily by highlighting an object. This is the problem of relations.

In conclusion, relations between features can not be made if one just highlights a part of the object. this approach aims to identify expressive and verbal explanations. The goal is to introduce symbolic values from the layers of a CNN.

The main process is to use Inductive Logic Programming (ILP). This is a ML technique that uses the theory of logic to create simple, but yet understandable explanation. An example of this is:

face(Example) :- contains(Example, Part), isa(Part, nose) [20]

These explanations are initially formed by Concept Embedding Analysis. The goal of this is to answer several questions about the visual concepts in either hidden layer results, or the predicted output (*whether*, *how well*, *how*, and *with what contribution to the reasoning*).

In terms of advantages, this technique has a way to evaluate how well the explanations are. This form of evaluation was missing from previously mentioned techniques, like DeepLIFT and Integrated Gradients. Moreover, this technique uses both the concept based approach and logic to form a cohesive yet understandable explanation. This could also be seen as a disadvantage since most of these rule-based approaches is that they are sometimes too simple in terms of the explanation. Even though this technique covers the best of both worlds, the disadvantages of both are also taken into account. For the concept part, the inaccurate heatmap id a disadvantage. The writers of [20] mention that their research has not looked at how well the concepts have been chosen.

TCAV

TCAV (Testing with Concept Activation Vectors) provides a quantitative explanation as to how much semantic concept matters in terms of the prediction, that a human has found. Again, this measure is after the model has been trained [22]. The concepts are the objects that are learned in the CNN. Just like other approaches, the model learns objects more specifically, layer by layer. This approach uses something called a Concept Activation Vector (CAV), to make sense of what the neural network has predicted, in terms of human understandability.

The next step is to use different examples to show how important the concept is that the human came up with. The output of this is a score that simply tells you how a change from the concept in the picture would affect the probability of that concept being the concept you are looking for. Of the same concept, different examples are taken and the scores are averaged out. This averaged computation is the TCAV score on the provided concept.

Within the goals of TCAV, this approach aims for high accessibility. Therefore, TCAV is relatively easy to use, without the need of expertise. Another advantage is that the user has control itself to identify how important an object is. Therefore, the human being involved refers to the goal of customization.

Simpler Neural Networks, with less layers are labeled as shallow. TCAV can perform badly in such cases since the concept can not be clearly identified with less layers [27]. Therefore, the concept can not be separated as well, leading to TCAV scores that are not representative.

Lastly, TCAV is only usable for images. However, future work could be towards using this approach for other types of data (e.g. text, tabular, audio)

NBDT

NBDT (Neural-Backed Decision Tree) aims to break the standards of most XAI techniques that lose accuracy while increasing explainability. The goal is to have a highly accurate model that is also highly explainable [24]. Therefore, this approach aims to capitalize on high accuracy from NN's and high interpretability from decision trees [17].

Most of the previously mentioned techniques use the last layer of the CNN to create heat maps, in different ways. this approach removes that last layer and connects a decision tree to it. The leaves of this decision tree are the input weights for the supposed final layer. The leaves are organized in such a way that similar weights are grouped together. Their parent node is the average of the two. As seen in Figure 3, both

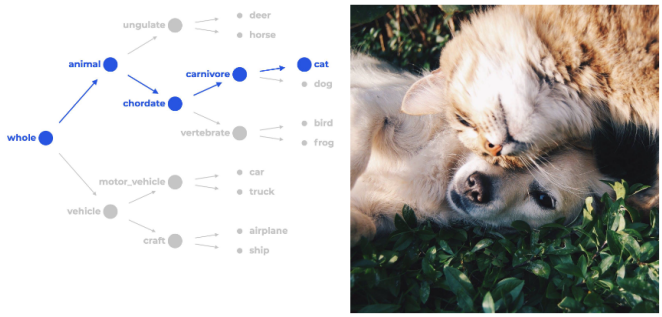


Figure 3: The added decision tree of NBDT, instead of the last layer. This shows that the cat is highlighted, but the dog is very close

the dog and cat are categorized together. This means that the probability that both are in the picture is high.

Even though this approach aims for both accuracy and explainability, it is very dependent on the intermediate nodes being explainable. If the intermediate nodes are not straightforward for the human, it is not explainable.

Lastly, another disadvantage is that this approach can not explain wrong or missed predictions. This is because it does not consider the hidden layers, but rather after a large chunk of the prediction is already made. Therefore, it can only explain what has been predicted.

3 Comparison

3.1 General Requirements

Even though some techniques differ in their workings, there are some requirements that hold in general for all the techniques. The division of the techniques, alongside the general requirements are explained in this section. Table 2 shows how each technique, mentioned in this research, adheres to the requirements.

Expertise refers to how easy it is to understand and apply the technique. Expertise can be divided in three stages [28]. The article refers to this as AI/Data novice, Data Expert and AI expert. This means that almost all techniques mentioned in this research are easily understandable for an AI/Data novice. An outlier here is the Net2Vec, which is better to understand if someone has a deeper understanding on how filters in NN's actually work. Even though this could be understood as well by an AI/Data novice, overall this technique is more difficult to grasp, compared to the other techniques. In conclusion, the techniques are relatively easy to grasp and apply after the predictions have been made.

Bias has proven to be a difficult topic in the field of XAI [28]. It is not easy to get rid of all bias in the data beforehand. However, a good XAI technique must be able to recognize bias and alleviate it, where possible. A couple of examples of goals mentioned in this article are fairness and bias avoidance. They argue that the bias avoidance goes hand in hand with fairness. If the technique is able to handle bias, or at least recognize it, the fairness increases as well. One could argue that one must look at the data beforehand and mitigate bias there. However, this is not always possible, hence good techniques should handle bias in one way or the other. Even though a better human like explanation already works towards recognizing bias [29], only a couple of techniques mention how they internally handle bias. Table 2 shows how well each technique explains their handling of bias.

Time refers to how computationally expensive each technique is. In general, not one technique is the fastest since this depends on the model. Since the techniques aim to explain DNN, time will not be the biggest trait. However, there are some techniques that are computationally less expensive, in comparison. One clear mentioned difference is the one of back propagating techniques and perturbation methods. Here, the back propagating techniques require less traversals over the model. In general, the techniques that use logic and are not based on saliency maps (i.e ILP and NBDT), are much faster. However, they compromise on accuracy of the explanation, as mentioned before. In conclusion, there are some traits that make the one technique faster than the other, but overall the techniques can be labeled as computationally expensive.

Privacy refers to the privacy awareness of the techniques. It is shown that not many techniques allude to design their techniques for privacy of data and information. Only SIDU has tests done on adversarial attacks, to see how vulnerable the technique is. XAI techniques must be designed in a way that they hide data from users that could potentially be sensitive [28]. The article mentions that there is a trade-off to some extent between explainability and privacy. On the one hand, the goal is to explain the reasoning towards the output. On the other hand, you do not want to give away private information. The table shows that most of the techniques are not implemented in a privacy aware manner, since they all obviously focus on the explainability. Perturbation methods and back propagation suffer the most from data leakage for the explanation [30].

Performance covers how well the technique contributes to understandability for the human and how the technique does comparatively to techniques of the same category. Even though all techniques have the aim to move towards interpretability, not every technique does that in the most effective manner, such that it is understandable for the human. For example, xNN has only proven so far to be effective on simpler models. This cuts of the explainability for more complex Neural Networks. Moreover, DeepRED focuses mainly on being more simple and humanly understandable but sacrifices on the accuracy of the explanation. Lastly, as shown in Figure 2, SIDU shows more correct heat maps compared to Grad-CAM and DeepLIFT outperforms Integrated Gradients. So even though Integrated Gradients is a well established technique on its own, comparatively a likewise technique like DeepLIFT performs better.

Visualization focuses on to what extent the methods ex-

Table 2: An overview of to which extent the different techniques hold to general requirements. The colors represent how well the technique does for that particular requirement. Green = Good, Orange = Average, Red = Bad.

Technique	Type	Expertise	Bias	Time	Privacy	Performance	Visualization
DeepLIFT	Feature-based						Global
Integrated Gradients	Feature-based						Global
Grad-CAM	Feature-based						Global
SIDU	Feature-based						Global
Perturbation	Feature-based						Both
xNN	Feature-based						Both
ACE	Concept-based						Global
Net2Vec	Concept-based						Global
TCAV	Concept-based						Both
Concept & ILP	Concept/Logic-based						Global
NBDT	Logic-based						Both
DeepRED	Logic-based						Global

plain the workings. Visualization can be categorized in global or local [8]. Some techniques can either have a more in depth local explanation or a general global explanation. Local refers to understanding more data points in depth towards the final output. On the other hand, a global view focuses on how the input is related to the output. For example, a perturbation method can explain for a specific input how that changes the output and have a local explanation. Conversely, the whole model can be explained by showing which inputs have more relevance towards the output.

3.2 Feature-based explanations

Feature-based explanations extract some kind of importance to the input values. These are techniques like DeepLIFT, Integrated Gradients, Perturbation, SIDU , Grad-CAM and xNN. From Table 2 it is visible that feature-based techniques are the ones that keep bias in mind, the most. This has to do with the fact that feature-based methods directly extract information from the inner layers. This is a known issue and is hence highlighted well by techniques like DeepLIFT, Integrated Gradients and Grad-CAM. In terms of time-efficiency, these techniques do not necessarily stand out, but they perform better in general. However, their performances are not always as consistent. Some approaches are not as humanly understandable and accurate as other feature-based methods.

Feature-based methods have two additional requirements: *Sensitivity* and *Implementation Invariance* [21]. Sensitivity refers to the saturation problem defined previously. It mentions the variables that the function does not necessarily depend on. These variables should not be nullified. We previously mentioned this as the *saturation problem*. Therefore, if a method is not sensitive, it does not correctly deal with all attributions that the method is trying to cover. Hence this is an important requirement for a feature-based technique. DeepLIFT and Integrated Gradients are the only ones that satisfy this requirement, since their inner workings account for sensitivity. Implementation Invariance is a requirement that two equivalent networks should always assign the same importance to the same input features. Neither the perturbation methods, including SIDU, DeepLIFT and Grad-CAM adhere to this and only Integrated Gradients does.

3.3 Concept-based explanations

Concept-based explanations extract concepts that are previously defined and categorize an image by using those concepts. Examples of these techniques are ACE, Net2Vec and TCAV. Even though these techniques are more understandable, since the concepts are predefined, this is less time-efficient. The human effort that comes in place, leads to a better human understanding but also contribute to more human-effort and human bias. Since ACE uses the TCAV score, it inherits the work that TCAV does to mitigate bias. However, the concept-based explanations showed that they performed better, compared to the other two types.

Moreover, the concept-based explanations have some additional requirements like *Negation* and *Feature Values*. Because the concepts are mostly predefined, some conclusions can be drawn that sometimes have no additional value to the explanation. Negation highlights that highlighting a concept does not mean that one could either make an assumption or negate the opposite [20]. A simple example is the one of a person holding a flower. Holding the flower does not conclude that the person is not a terrorist but is not sufficient to label the person as not a terrorist. Therefore, a good concept-based method should keep this reasoning in mind.

Feature values is the requirement that highlighting some feature is not always helpful. For example, highlighting the eye in a picture is not helpful when one wants an explanation on the facial expressions, in general. From the concept-based methods reviewed, only Concept & ILP accounts for this.

3.4 Logic-based explanations

Logic-based explanations are the methods that use logic to explain the model. Methods like Concept & ILP, NBDT and DeepRED are such methods. Where the previous methods were focused on highlighting parts of images, these methods account for human readability by using textual reasoning to explain the model. Even though the readability was already

part of the Performance, these methods have to be more human readable. Out of the three methods, DeepRED does this in the best way. However, it compromises on the accuracy and depth of the explanation. On the other hand, NBDT uses keywords but heavily depends on how well the inner layers are described. Lastly, Concept & ILP tries to combine the human readability of Logic-based explanations and the accuracy of concept-based explanations. Hence it offers the logical reasoning as support to the concept-based explanation.

4 Responsible Research

Accounting for the ethical responsibility is of utmost importance in every research. Since this research has been a literature study describing on what already exists, there are no real results to reproduce. However, the results regarding the advantages, disadvantages an room for improvement comes from the already existing research from all the mentioned techniques in Table 1. Therefore, this report is based on the online findings on the different techniques and different surveys of XAI methods. Thus, we must make this research available online so it is accessible for other people investigating the state-of-the-art. Therefore, the poster of this research and the paper will be available on the TU Delft repositories.

5 Future Research Directions

Future work on model-specific XAI is needed since there are some different ways to gain different types of understandability. Therefore, the limitations of each category has room be explored. Firstly, an important area of research is the compromise between accuracy and explainability. Even though a primary goal is for the human to understand the explanation, explanations must not be too easy. This means that the accuracy of the explanation should not be compromised on. We argue that there must be more research done on how we can find the right balance between accuracy and explainability.
Secondly, the use of hybrid techniques should be exploited. For example, NBDT combines the work of concept based techniques with logic by providing a decision tree that shows the reasoning behind it [24]. Even though this is not as humanly expressive as some of the other logic-based techniques, this is still a way in the right direction. NBDT has proven to gain more trust from humans compared to concept-based techniques that use saliency maps. Hybrid techniques are interesting to discover since they exploit the positives of both the techniques.

Feature-based techniques do not have a guideline in terms of evaluating the performance. Importance scores are extracted but there is no real way to find out how accurate these scores are. Concept-based techniques cover this in a way that they show their accuracy in terms of the saliency map. However, the evaluation on feature-based techniques is an area that can be exploited and is certainly worthwhile. Different feature-based techniques could be compared in more depth if there is a general way to evaluate the performance.

This leads to another issue that can be covered in terms of usability. Most of the feature-based techniques focus on DNN's and the concept-based techniques are mostly based on CNN's. However, this can be extended by using different data types. Some examples of different data types are audio, text, tabular or sequential data. This is an area that is not only general but also specific to each of the techniques.

Lastly, as mentioned in Section 3, not a lot of techniques have any work done regarding the privacy. Some existing mathematical properties of the techniques can be helpful in order to gain privacy. This has not been exploited since no real testing has been done for this. Privacy awareness is a pressing issue and can be explored for all the mentioned techniques. The amount of data that almost all the applications require does not only mean that the predictions should be well explained, but the data should also be protected. Therefore, there is room for work on the trade-off between privacy and explainability. We want to gain explainability, but not at the expense of losing privacy.

6 Conclusion

In this paper, we analyzed some of the recent model-specific XAI techniques. We did not only describe the advantages and disadvantages of the techniques but also looked at how some of the techniques differ or compare to each other. From here, we presented a categorization that shows the general characteristics of the techniques. The analyzed techniques can be divided into feature-based, concept-based and logic-based. Moreover, we presented some requirements that a good model-specific XAI technique should adhere to and compared the analyzed techniques on these requirements. We showed that some of the different categorizations have some additional requirements. From this, we visually presented how the techniques perform on those requirements. Concluding that in general, the concept-based explanations perform better, it is not hard to use these methods and there is not any emphasis on bias and privacy. Lastly, we showed that there are several aspects that pave the way for future work. This includes work on the trade-off between accuracy and explainability, exploring hybrid techniques, finding a generic way to evaluate feature-based techniques, extending the current work to more data types and finally, exploring privacy awareness.

References

[1] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[2] "Tesla driver killed while using autopilot was watching harry potter, witness says," Jul 2016.

[3] E. Staff, "Whoops, alexa plays porn instead of a kids song!," Jan 2017.

[4] "2010 flash crash," Dec 2019.

[5] C. Molnar, *Interpretable machine learning*. Lulu. com, 2020.

[6] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, vol. 267, pp. 1–38, 2019.

[7] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[8] S. Das, N. Agarwal, D. Venugopal, F. T. Sheldon, and S. Shiva, "Taxonomy and survey of interpretable machine learning method," in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 670–677, IEEE, 2020.

[9] V. Belle, "Logic meets probability: Towards explainable ai systems for uncertain worlds.," in *IJCAI*, pp. 5116–5120, 2017.

[10] M. W. Craven and J. W. Shavlik, "Visualizing learning and computation in artificial neural networks," *International journal on artificial intelligence tools*, vol. 1, no. 03, pp. 399–425, 1992.

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

[12] 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, 2020.

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

[14] J. Vaughan, A. Sudjianto, E. Brahimi, J. Chen, and V. N. Nair, "Explainable neural networks based on additive index models," *arXiv preprint arXiv:1806.01933*, 2018.

[15] S. M. Muddamsetty, M. N. Jahromi, A. E. Ciontos, L. M. Fenoy, and T. B. Moeslund, "Introducing and assessing the explainable ai (xai) method: Sidu," *arXiv preprint arXiv:2101.10710*, 2021.

[16] A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim, "Towards automatic concept-based explanations," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[17] J. R. Zilke, E. Loza Mencía, and F. Janssen, "Deepred–rule extraction from deep neural networks," in *International Conference on Discovery Science*, pp. 457–473, Springer, 2016.

[18] R. Fong and A. Vedaldi, "Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8730–8738, 2018.

[19] 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*, pp. 618–626, 2017.

[20] J. Rabold, G. Schwalbe, and U. Schmid, "Expressive explanations of dnns by combining concept analysis with ilp," in *German Conference on Artificial Intelligence (Künstliche Intelligenz)*, pp. 148–162, Springer, 2020.

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

[22] 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*, pp. 2668–2677, PMLR, 2018.

[23] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Proceedings of the IEEE international conference on computer vision*, pp. 3429–3437, 2017.

[24] A. Wan, L. Dunlap, D. Ho, J. Yin, S. Lee, H. Jin, S. Petryk, S. A. Bargal, and J. E. Gonzalez, "Nbdt: neural-backed decision trees," *arXiv preprint arXiv:2004.00221*, 2020.

[25] J. McCaffrey, "Neural network saturation," 2017.

[26] P. Jackson, "Introduction to expert systems," 1986.

[27] G. Alain and Y. Bengio, "Understanding intermediate layers using linear classifier probes," *arXiv preprint arXiv:1610.01644*, 2016.

[28] A. Rawal, J. McCoy, D. Rawat, B. Sadler, and R. Amant, "Recent advances in trustworthy explainable artificial intelligence: Status, challenges and perspectives," 2021.

[29] N. Kilbertus, M. Rojas Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf, "Avoiding discrimination through causal reasoning," *Advances in neural information processing systems*, vol. 30, 2017.

[30] R. Shokri, M. Strobel, and Y. Zick, "On the privacy risks of model explanations," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 231–241, 2021.