Model-Agnostic XAI Models: Benefits, Limitations and Research Directions

Mikolaj Knap
Supervisor(s): Chhagan Lal, Mauro Conti
EEMCS, Delft University of Technology, The Netherlands

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

June 19, 2022

Abstract

The ever increasing presence of Machine Learning (ML) algorithms and Artificial Intelligence (AI) agents in safety-critical and sensitive fields over the past few years has spurred massive amounts of research in Explainable Artificial Intelligence (XAI) techniques (models). This new frontier of AI research aims to resolve some of the fundamental issues that accompany the usage of ML algorithms in sensitive fields such as medicine or criminology. For ML algorithm's to be implemented and used within fields such as medicine, it is not simply enough that they are proficient and effective tool at solving their assigned task (such as classifying whether a patient has Covid-19 or not). These ML techniques lack the ability to allow their human counterparts the possibility of understanding why they have made such a prediction, therefore not allowing the human supervisor a peak into the black box. This black box problem is one of the underlying difficulties that currently prevent the widespread adoption of ML/AI algorithms into these safety-critical fields. XAI techniques aim to solve this very issue and in particular Model-Agnostic XAI techniques aim to generate explanations on the predictions of any ML or AI algorithm. In this paper we will be exploring and investigating the different Model-Agnostic XAI techniques and looking into their individual advantages and disadvantages. After we've analyzed each individual technique, we will take a more global view into the different characteristics and how the XAI implementations compare to each other using different metrics for comparison. Finally we will propose future improvements and extensions that can be made to the various investigated XAI techniques.

1 Introduction

In recent years, the explosion of popularity that the fields of Machine Learning (ML) and Artificial Intelligence (AI) have experienced has not gone unnoticed by regulatory forces. The ever-increasing presence of ML and AI into practical applications in fields such as health services, counter-terrorism, criminal justice, and credit/insurance risk assessment has drawn the eye of public and government organizations to establish forms of transparency and accountability behind the application of ML/AI systems into these often sensitive and critical fields [1].

Unfortunately while ML/AI systems have come a long way in their usefulness and application, most of algorithms behind these systems exhibit black-box behavior, meaning that there is a no transparency within the predictions made by ML/AI system system's [1]. This lack of explainability a behind the system's decision making process has discouraged organizations from the widespread practical adoption of ML/AI systems [1]. Without any explainability or transparency built into these algorithms, there is no way for the users to debug and take a peak into these black-box systems (hence the name black-box). This lack of human oversight, can lead to AI failures where critical functionalities within these systems have unnoticed issues or biases that can lead to faulty decision making. One particular high-profile system failure was Amazon's recruitment AI having an unintended bias against female candidates [2]. In order to account for the possibility of AI systems having unintended bias's or issues, it is imperative to introduce an element of explainability to these systems, therefore allowing users to identify potential flaws and errors in the AI's judgment [2]. Beyond helping to account for and fix critical system flaws, introducing explainability will also result in an increased level of trust in these systems as users are able to understand how a system arrived at a final choice [2].

This is where eXplainable Artificial Intelligence (XAI) techniques (or models) come in, which can be defined as types of AIs that aim to introduce the attribute of explainability into these black-box AI/ML algorithms. These XAI techniques aim to eliminate the limitations of ML/AI systems mentioned previously, adding an element of explainability into the processes of the AI/ML algorithms [2].

Organizations such as the Defence Advanced Research Projects Agency (DARPA) in the US have even set up standalone research programs, to analyze and track XAI's current development and maturity level [3]. DARPA's interest in XAI technologies stem from the challenges that the Department of Defense (DoD) faced when implementing ML techniques into their autonomous systems [3]. The DoD found that the best preforming ML systems, would be the ones most opaque and most likely to exhibit black box behavior in their operation [3]. This meant that extracting any explanation from these ML systems would be impossible for any operator. While this might be acceptable for applications that aren't safety critical nor ethically complex, many of the promising state-of-the-art ML applications within fields such as defense or medicine require a form of accountability and transparency behind their processes and reasoning [4].

One practical usage of ML algorithms that comes with the requirement of a form of explainability before the full practical adoption of this technology is the usage of artificial neural networks in brain-machine interfaces for performing medical tasks [5]. Interpretability and explainability are seen as necessary requirements for this application as since this is a medical application there are stringent restrictions as to who is held accountable for each decision [5]. XAI techniques aim to solve this issue by providing a form of transparency through the use of explainable models, there allowing for a degree of accountability even with the usage of an ML algorithm since medical professionals will be able to oversee the decision making process used by the ML algorithm [5].

This paper will be investigating and exploring the different limitations, benefits, and advantages of various state-of-the-art XAI techniques. For the initial two sections of the paper we will introduce XAI techniques, what they are and their primary objectives, as well as what category of XAI techniques we will target for investigation. After this initial introduction into the relevant background information, the methodology section will lay out the specifics behind how our paper will investigate and compare the selected XAI techniques.

2 Background Information

2.1 General Overview on XAI Techniques

Research on explainable models isn't a new field, it has been active since the 1980s but the newfound popularity of ML/AI practical applications has lead to a large variety of XAI techniques being developed in recent years [2]. This recent uptick in the amount of XAI techniques has resulted in the need for the categorization of the various models, as well as clarification as to what the goals of a general XAI technique are. Our previous definition of what a XAI technique ("A type of AI that aims to introduce explainability into an AI system") is a bit vague and could benefit from further elaboration into the terms used within the definition as well as what the main objectives of any XAI technique are.

Before we establish what the main goals of an XAI technique consists of, we should clear up what the definition of explainability in the scope of AI research is. While explainability as a concept is relatively straightforward we will in this paper use the nomenclature found in [6], where explainability is defined as the ability of a system to provide the context and reason for an AI system's decision in such a manner that it is understandable to a human. With explainability clearly defined, we can move on to what XAI techniques main objectives are, which DARPA's XAI initiative outlines as:

1. The technique should not to sacrifice performance when introducing explainability into the ML models [6].
2. The technique should aim to not only produce accurate explanations to the user, but it should convey these explanations in a concise and clear manner that enables end-users to understand the AI/ML models. [6]

These two goals outline the main objectives of XAI techniques, and to combine these two objectives into one statement, we can say that the primary function of any XAI implementation is to increase the explainability of an AI/ML system to the end user while not sacrificing performance to obtain this objective [6]. Now that we have a clear definition of what an XAI technique is, and what it's main goals are we can start to delve a bit deeper into the taxonomy of this field and the different subcategories of XAI techniques.

2.2 Model-Agnostic/Specific XAI Models

As mentioned previously the sheer amount of ML/AI applications in the modern world has resulted in an equally massive amount of proposed XAI techniques and implementations. With this volume of research coming through in the last few years, researchers have begun to classify these algorithms into their different subcategories. Researchers have come to a consensus that XAI techniques can currently be split into two main sub-categories, model-agnostic and model-specific techniques [1].

Model-specific techniques are a subsection of XAIs that only function on a specific ML/AI model[1]. These implementations are limited to their individual targeted model, and lack any form of model-flexability. However with this limitation there is the significant advantage that model-specific XAI techniques can focus singularly on their chosen model, and optimize their explanation generation strategy off their knowledge of the model's internal structure [1].

The other subcategory, model-agnostic techniques are not reliant upon the model that they are attempting to explain. They operate on a black-box methodology regarding their input model, only having access to the inputs and outputs of the provided model [1]. This approach has the advantage of being much more flexible into what models it can explain. However, this does mean that model-agnostic techniques lack the ability to gain insights from the model's internal structure as they treat each ML model as a black box function [1].

For our research paper, we will be focusing solely model-agnostic XAI techniques, as they are more directly comparable and by focusing on one category of XAI techniques we ensure our paper is concise and has a clear focus.

3 Method

The primary research goal in this paper is to deliver a comprehensive study and analysis on various model-agnostic XAI models and their characteristics. In the previous sections we've outlined background information on XAI research and the main goals of XAI implementations. In the next section (4) we will take a deeper dive into the inner workings of the individual techniques investigated and their advantages/disadvantages. Once we've investigated each technique, we will move onto a comparison between the various XAI techniques where we will use a variety of metrics and classifications to underscore the differences between the studied techniques. After this comparison, the paper will finish with proposed future directions for improvements in the individual XAI techniques studied with the goal to improve and alleviate some of the limitations and disadvantages we've discovered in our analysis. Concluding our paper we will also propose possible future directions for research within this field, in order to further evaluate the possible limitations present in model-agnostic XAI techniques.

3.1 Related Work and Literature Search

For searching literature and relevant papers, we will use tools such as Scopus (a scientific literature database) as well as arXiv (an open-access database of pre and post print papers).

In our literature research phase we collected both general model-agnostic XAI survey papers and the specific technique proposal papers. The XAI proposal papers outline the initial implementation details that was proposed by the original authors of the XAI technique as well the author's evaluation of the technique. Papers such as [7], [8], or [9] are examples of the original proposal papers for the XAI techniques of LIME,SHAP and Anchors respectively. These papers will be helpful in fully understanding the inner workings of each XAI technique as well as useful in identifying potential advantages/disadvantages, it is important to note however that these are not the only implementation papers gathered.

General XAI survey papers collected will be used for the extraction of the comparison metrics that we will use at the end of the paper to directly compare and contrast the investigated techniques. Papers such as [1]. [4], [6] are examples of papers that will be used to find what relevant metrics we can use to compare the different XAI techniques.

4 Model-Agnostic XAI Techniques

4.1 LIME

LIME or Local Interpretable Model-Agnostic Explanations was initially proposed by Marco Tulio Ribeiro et al in 2016 [7]. It proposes an explanation system LIME whose goal is to provide explanations for the decision of any ML model in a localized setting[7]. This means that the explanations provided by LIME would be on a local level, and therefore apply only to individual predictions made by the ML model being targeted. For example within a practical case where an ML model predicts a medical diagnosis on the likelihood of specific patient having the flu based off of their history of symptoms, LIME presents key symptoms displayed by the patient which LIME predicts were the leading factors behind the ML models prediction. This is visualized in figure 1, where LIME takes the inputs/outputs for the ML model (symptoms/prediction) and presents a visual explanation for the prediction of the ML model. Within this example LIME assigns positive (green) or negative (red) feature weights to the input symptoms to provide an explanation to the end user.

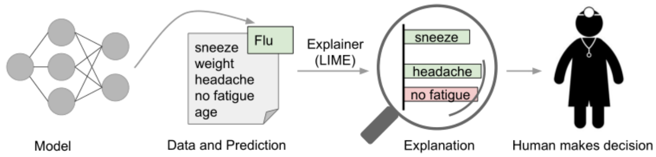
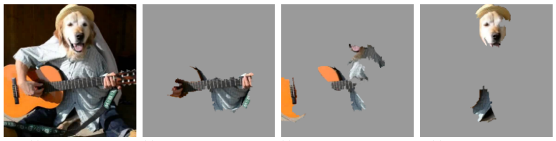


Figure 1: LIME's explanation for model's diagnosis [7]

The inputs that LIME can accept and provide explanations for are not limited to a textual form, and as can be seen in figure 2, LIME is capable of conforming to the input and output data provided by a model to generate its explanations. Within this example LIME is able to segregate and identify unique portions of an image classification, and provide explanations for the ML model's prediction. LIME does not rely at all on the target ML model treating it as a black box, which allows LIME to be model-agnostic [7].



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 2: LIME's explanation for Google Image Classifier [7]

The main algorithm behind LIME to generate it's explanations is a minimization problem as can be seen below [7].

$$\xi(x) = \underset{g\in G}{\operatorname{argmin}} \mathcal{L}\left(f, g, \pi_x\right) + \Omega(g)$$

Based on an original input x, the algorithm aims to minimize the loss function $\mathcal{L}\left(f, g, \pi_x\right)$ which represents "how unfaithful *g* is in approximating *f* in the locality defined by by π_x [7].

The local explanation model with this algorithm is represented by *g* and it determines the explanation behind the final result of LIME[7]. Therefore by solving the minimization problem, the LIME algorithm provides as accurate a local explanation model *g* as possible which should result in a more accurate local explanation model for the prediction of the target ML model *f* on the input x [7].

For training this local explanation model, LIME uses a perturbation strategy for optimization of the loss function. The algorithm generates new instances x' around the original input point x within a range of π_x [7]. This new data-set of perturbed instances is then fed back into the local explanation model *g* so that LIME can further improve the accuracy of *g* within this local domain [7]. This algorithm results in a locally faithful explanation model *g* that is limited to a local area within the target ML model.

The $\Omega(g)$ term measures the general complexity behind the explanation function and is included in the minimization problem to ensure that the explainable models *g* are as simple as possible [7]. Keeping these explanation simple increases usability and interpretability for end-users[7].

Advantages of LIME

- Can take a multitude of data types, ranging from textual, tabular to even image data as input, and generate explanations for any ML model's prediction behind it [7]
- LIME's explanations for an ML model's prediction are optimized to be as simple as possible [7]

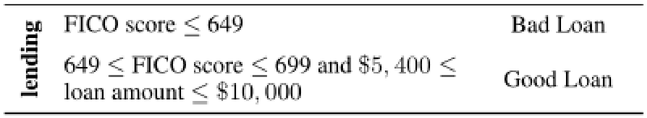
Disadvantages of LIME

- LIME only remains faithful in it's predictions for an ML model on a localized level [7]
- LIME doesn't convey it limitations in regarding to locality to end-users [10]
- Despite the best efforts of LIME's developers to simplify the explanations generated by this algorithm, the interpretability of the final results is still mediocre for non-expert users [10]
- LIME has consistency issues, often changing it's explanations completely despite minimal changes being made to the input and output of the ML model [11]
- LIME is vulnerable to adversarial attacks that attempt to modify the explanations provided by LIME [12]

4.2 Anchors

Anchors is another model-agnostic XAI technique that was proposed by Ribeiro et al in 2018 [9]. The technique was created with the aim of resolving some of the issues that are present in the technique proposed by Ribeiro et al in 2016 LIME paper [7]. The first main problem that Anchors aimed to iron out was the lack of a positive user response regarding the interpretability of LIME's results. Analysis into LIME's usability and clarity done in [10] has shown that users without ML knowledge experienced difficulty in using LIME's explanations.The other issue present in LIME that Anchors aims to fix was the lack of clarity regarding the locality where the explanations generated by LIME would remain accurate [9].

Anchors attempt to provide explanations for any target ML model by relying on a process of generating various anchors to provide an explanation to the user. These anchors are defined in the paper as explanations which "sufficiently anchors the prediction locally such that changes to the rest of the feature values of the instance do not matter" [9]. In more concrete terms, these anchors can be viewed as a set of if-then rules, which upon being full-filled, effectively guarantee the classification of an prediction [9]. One practical application of anchors can be seen in the figure 3 below, here we can see that two if-then rules that check the applicants FICO credit scores are generated by this technique to determine if the loan will be classified as either good or bad.



lending	FICO score ≤ 649	Bad Loan
	$649 \leq$ FICO score ≤ 699 and $\$5,400 \leq$ loan amount $\leq \$10,000$	Good Loan

Figure 3: Anchors generated from Tabular Data sets, If Rule (Left) and Prediction(Right) [9]

Each Anchor A is only considered a viable anchor if it fulfills the requirements below [9]

$$E_{\mathcal{D}(z|A)}\left[1_{f(x)=f(z)}\right] \geq \tau, A(x) = 1$$

Within this formula to be considered an Anchor, A must fulfil the condition A(x) = 1, which ensures that the anchor correctly provides an explanation for the output of the ML model with the initial input x [9].

The other part of the equation enforces a precision value τ onto the anchor A within a neighbourhood of the initial input x, which is indicated by all the values under D. Therefore this part of the equation ensures that the Anchor A remains locally faithful to the ML model *f(x)* within this domain distribution D, meaning that it can accurately provide explanations for an ML model as long as they fit within this neighbourhood of inputs [9].

Similar to LIME, Anchors themselves are created using a perturbation strategy from an the original input x [13]. However, since for each anchor A it is necessary to evaluate it's performance for all the neighbours or perturbations in D around x, it is computationally unfeasible to check whether anchor A meets the precision requirements for all perturbations within Domain D [13]. To get around this issue the authors formulated an complex candidate generation system where a modified beam search algorithm and a Multi-Armed Candidate algorithm are combined to generate the best set of anchors A for the domain within a reasonable time-frame [13].

The anchor generation algorithm also allows for the calculation of the coverage of each anchor. Coverage can be defined as the area where an anchor is able to provide an accurate explanation [13]. This notion of measuring an anchor's coverage is critically important to this technique as it solves of the main problems that were present in the previously explored technique of LIME. Since LIME only provides locally faithful explanations, this presents an issue to the end-user as they are unable to determine where LIME's explanations will remain accurate [7]. Anchors tackle this issue through the use of coverage, which convey the area within which an anchor will remain accurate.

This is visualized in figure 4, where we can see that the anchor model provides an area of coverage within the complex binary classifier where it remains faithful, whilst LIME only approximates the decision boundary of the classifier locally and doesn't provide any information on it's locality.

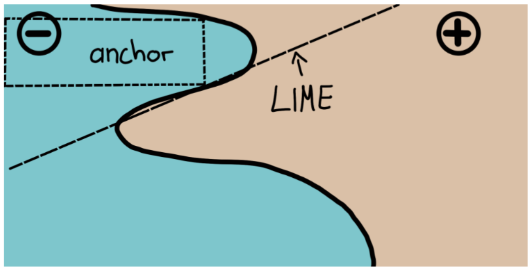


Figure 4: Lime's and Anchors Explanations [13]

Advantages of Anchors

- Anchor's interpretability for users has been shown to be better than other XAI techniques such as LIME [9]
- Anchor's coverage variable allows for the clarification of the space within which an anchor will provide accurate explanations for an ML model's prediction [9]
- Highly efficient due to the parallelizable nature of parts of anchor's calculation algorithm [13]

Disadvantages of Anchors

- Anchor's suffers to similar issues faced by LIME, where they are inconsistent in their explanations given small perturbation to their input [14]
- The algorithm can run into the issue where potentially conflicting anchors are created for the inputs [9]
- Rare predictions of an ML models classification may result in extremely complex anchors that provide low coverage and may not generalize well [9]
- Initial setup must be configured properly to attain respectable results [13]

4.3 SHAP

SHAP, or SHapley Additive explanations is a model agnostic explainable XAI technique that was proposed by Scott M. Lundberg in 2017[8]. The explanations generated from SHAP rely on the calculation of Shapley values for an individual input x. These Shapley values are a concept that emerged from game theory, and they allow for the calculation of specific feature weight and it's importance within the context of a ML model's prediction [8]. Assigning a Shapley value to each feature within an input allows the technique to explain what features had the highest influence on the ML model's prediction, therefore providing an explanation to the user as to what feature resulted in the ML model's prediction [8]. The formula that is used for calculating Shapley values is provided below with the final result being the Shapley value for feature i for the input x within model f [8].

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!\,(M - |z'| - 1)!}{M!}\left[f_x\left(z'\right) - f_x\left(z' \backslash i\right)\right]$$

The Shapley value $\phi_i(f,x)$ is calculated by iterating through all the possible subsets (*z'*) of the original feature space *x'* [8]. For each iteration, the formula calculates what the difference does the inclusion of a feature *i* have on the final output of the ML model's prediction, denoted by $f_x\left(z'\right) - f_x\left(z' \backslash i\right.$ [8]. The complex multiplier before this calculation within the iteration, accounts for the amount of features within a subset, naturally for small sets of features, the algorithm should account for individual features having a greater impact on the prediction generated by an ML model, and this factor is what allows the algorithm to account for extremely small and large feature sets [8].

Shapley value calculation has been a well known mathematical concept within the field of game theory, and as a byproduct Shapley Values have a few key beneficial properties that come alongside their inclusion in this algorithm [8]. The first and second property are less interesting and ensure that Shapley values are accurate when matching the original model f(x), and ensure that Shapley values do not change for a feature, if another feature is missing in the input [8]. The final property is the most interesting as it allows for consistency behind the calculated Shapley Values, (therefore possibly solving some of the issues present in other XAI techniques) [8].

Now that the calculation behind Shapley Values has been explored, we can finally take a peak under how SHAP uses these values. Each prediction of input *x* for a ML model *f* is explained in SHAP through the calculation of all the Shapley values of the input feature space. Once all the different feature values are calculated, the differences they make in the final outcome of the ML model can be easily portrayed to the user using weights. One visualization of the effect of different Shapley values can be seen in figure 5 where each feature's Shapley value adjusts the expected outcome of an input *x*, with the algorithm arriving at the final prediction point *f(x)* when all the Shapley Values are added to it [8]. Each Shapley value calculated aims to provide an explanation to the user how a specific feature effected the final outcome of an ML model.

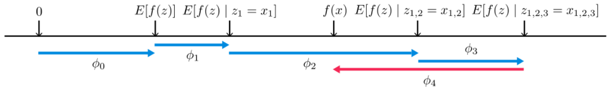


Figure 5: Shapley Values effects on the final outcome [8]

However this method of calculating the individual SHAP values for each feature within an input x is computationally unfeasible for complex models [8]. To bypass this issue the authors of SHAP proposed an approximation implementation called Kernel SHAP that combines the LIME algorithm with the classical Shapley value calculation [8]. This approach allows the approximation of Shapley values which decreases the computational time of the algorithm therefore making it possible to feasible calculate Shapley values for complex feature sets [8]. There are other approaches proposed within the paper of implementing SHAP but those are model-specific and for the rest of this paper each time we refer to SHAP we are referring to the Kernel SHAP version.

Advantages of SHAP

- As a mathematically enforced concept Shapley values have additional beneficial proprieties such as consistency and accuracy [8]
- Within the proposal paper [8], evaluation done on SHAP portrays that it evaluates feature weights in a way that fits human intuition better than LIME

Disadvantages of SHAP

- Despite the application of approximation of Shapley Values in Kernel SHAP, this XAI technique still is relatively slow in it's computational time [13]
- KernelSHAP doesn't consider the interaction between features when calculating weights [13]
- SHAP has been shown to be inconsistent and vulnerable to adversarial attacks despite the mathematical properties of Shapley Values [12] [14]

4.4 Counterfactual Explanations

Unlike the previously investigated XAI techniques where there is an attempt to generate a rule-set or numeric model to explain an ML model, the next investigated techniques take a more contrastive approach when generating their explanations. One form of contrastive explanations is the idea of providing counterfactuals for a decision. Counterfactuals have been shown to be an effective way of conveying explanations for humans, and studies have indicated that in general humans prefer counterfactual arguments over other explanation approaches such as case-based reasoning [15]. Within the context of a ML model prediction, a counterfactual explanation details the minimal possible change to the feature values of an input that would change the classification made by the ML algorithm [13].

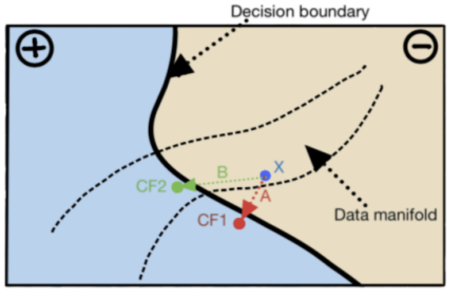


Figure 6: Input x with two counterfactuals [15]

A visualisation of how a counterfactual is generated can be seen in figure 6, where the ML model's decision boundary is denoted by the black line with an original input x (blue dot). Two separate counterfactual argument are generated in this example (red/green dots) and from the diagram it is clearly visible that despite the small changes to the counterfactuals generated, there final classification is different than that of the original input. This example outlines the main goal of any counterfactual generation algorithm: which is to change an original input x as minimally as possible, so that it results a change of classification by the ML model.

Wachter et al. proposes that a new counterfactual x' can be generated from an original input datapoint x by solving the minimization algorithm below [16].

$$\arg\min_{x'}\max_{\lambda} \lambda \left(f_w\left(x'\right) - y'\right)^2 + d\left(x_i, x'\right)$$

Here the first part of the equation calculates the distance between the machine learning model's classification of x' and the prediction y' (the intended classification). The second part of the equation calculates the distance between the new counterfactual x' and the original input [13]. Therefore by solving the minimization problem in this equation, we can simultaneously ensure our new counterfactual x' is close to the predicted goal y' and is as similar to the original input x as possible. It is important to note that for generating counterfactuals x' we randomly initialize different values for it [16].

This XAI technique, is different than the others explored previously, instead of relying upon a numeric or rule-set model, counterfactual explanations leverage the nature of humans to use counterfactuals as explanations [16]. Due to this nature, this does mean that counterfactuals are easier to comprehend and understand for end users in certain applications. Since these algorithms aim to produce counterfactuals that are as close to the input as possible, this creates easier to comprehend and understandable explanations for end users [17].

Advantages of Counterfactual Explanations

- Counterfactual Explanations can be efficiently generated for a variety of ML models [16]
- Counterfactual Explanations are particularly useful for explaining positive and negative decisions to users [16], making this approach particularly interpretable for non-expert users [17]

Disadvantages of Counterfactual Explanations

- Counterfactuals generation proposed by [16] is vulnerable to manipulation and small perturbations result in large changes to the counterfactuals generated [18]
- Counterfactual are not private, with [19] showing that it is possible to execute ML model-extraction attacks against current counterfactual generation methods [16]

4.5 Contrastive Explanations

Contrastive Explanations is a model-agnostic XAI technique proposed by Dhurandhar et al in [20]. Contrastive Explanations taps into another part of human nature regarding how we naturally look at explanations. As opposed to generating counterfactuals to aid in explaining a ML model, contrastive explanations create sets of contrastive facts that can explain a ML model's prediction [20]. The explanations that this method intends to produce can be stated in the form "An input x is classified in class y because features fi, ..., fk are present and because features fm, ..., fp are absent [20]".

Explanations that rely on certain aspects being present and certain aspects being missing are extremely common in critical fields such as medicine already [20]. Within those fields these contrastive characteristics are defined as Pertinent positives (PPs) and Pertinent Negatives (PNs) [20]. A Pertinent positive can be defined as a "factor whose presence is minimally sufficient in justifying the final classification" [20] meanwhile a Pertinent negative can be defined as a "factor whose absence is necessary in asserting the final classification" [20]. An example of this could be the diagnosis of a patient with Covid-19, a patient with a cold and a cough has the PPs that point towards either the flu or Covid-19. However if the patient does not exhibit a loss of taste, this means they show the correct PN to diagnose them as to not be sick with Covid.

The original paper proposed an XAI technique called contrastive explanations method (CEM) [20], which aims to identify both the PPs and PNs within a ML classifier. However this method only works within neural networks so for our paper we had to investigate another model-agnostic contrastive explanations method but an a visualisation from the original paper is still useful for portraying PPs and PNs [20]. Within Figure 7 we can see how CEM highlights the necessary PPs for identifying the figure as 3/7 indicated by the cyan color, and we can see what PNs are missing for a classification of 5 or 9 [20].

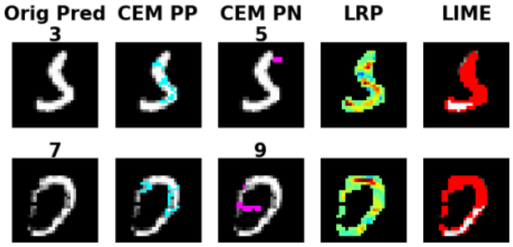


Figure 7: PPs and PNs identification for image classification [20]

After the original paper proposed these contrastive explanations, a model-agnostic method was created by Dhurandhar et al [21]. This paper proposed a new method MACEM or model-agnostic contrastive explanations method, with this new technique, we can feed any black box machine learning classifier into this technique and obtain an assortment of PPs and PNs [21].

Advantages of Contrastive Explanations

- PPs and PNs align nicely with natural human explanation techniques within sensitive fields such as medicine [20], therefore improving the techniques interpretability

Disadvantages of Contrastive Explanations

- A local scoped XAI method, only able to provide PPs and PNs for individuals predictions as opposed to a global level [21]
- Computation of PPs and PNs may not be computationally efficient for model-agnostic methods such as MACEM [22]

5 Comparison

Within the comparison section of this report we will be comparing the investigated XAI techniques using a list of relevant metrics that will be outlined and explained in section 5.1. The metrics themselves aren't simply a different measure of the performance of the XAI technique within a specific task, some of the metrics such as the Scope, or method approach are a categorization of the individual XAI technique within a taxonomy rather than a direct comparison metric. Since the topic of the report calls for a detailed comparison between the XAI models, we thought it would be pertinent to include metrics that classify the investigated XAI techniques into their own separate categories. Further details surrounding each metric used are outlined in the section below.

5.1 Comparison Metrics Used

Local/Global Scope: Global Scope XAI techniques generate an explanation for a ML model on a global scale, creating an explainable model that encompasses the entire ML model being targeted [1]. Alternatively, XAI techniques can choose to focus on explaining individual predictions made by ML models, therefore focusing their attention on explaining the narrow area where this prediction occurred within the ML model (Local scope) [1]. Both approaches have their own advantages and disadvantages and come with their own difficulties. It is often assumed that the local scoped methods are computationally simpler to compute and generate more precise explanation but are limited in the area they cover [1].

Method Approaches (Perturbation and Contrastive): XAI techniques generally conform to using one of two approaches for generating an explanation for a ML model, either they use a perturbation or contrastive approach [1]. For the perturbation approach, the main strategy is to generate feature-based explanations that are created through the perturbation of an initial input's various features [1]. These strategies observe what the effect these small changes to the feature cause in the final prediction made by the target ML model [1]. Once they know the contribution of each feature, they base their explanations off this behavior [1]. On the other hand, the contrastive approach generates their explanations by focusing instead on providing as similar an instance to the initial input, but with the important detail that these newly generated instance results in the ML model classifying it differently [1]. These explanations are counterfactual in nature and aim to provide a contrastive look into how a classification may shift based off the minimal amount of changes to the initial input [1].

Consistency: Defined as the degree of change in explanations generated by the XAI techniques for similar ML model predictions [24]. Similar predictions should ideally result in the XAI technique providing consistent explanations between different runs, having a XAI technique give different explanations on similar outputs is an ongoing issue that many popular XAI techniques such as LIME/SHAP/Anchors struggle with [14]. Consistent XAI techniques provide similar explanations for alike predictions while inconsistent XAI techniques differ in their explanations drastically even when a similar prediction is provided [24]. Consistency is often referred to also as the stability or robustness of an XAI technique [14], [11]

Resistance to Adversarial Attacks (RAA): Relating heavily to the previously mentioned metric of consistency, the resistance to an adversarial attack is measure of how resistant an XAI technique is to an adversary attempting to modify the explanation given by a XAI technique through the slight perturbation of the input data [12]. Many of the investigated methods have been shown to be easily fooled into modifying their explanation despite minimal changes to the input [12] (Scale: None, Low, Medium, High). Techniques that haven't had their RAAs investigated will be marked with a ?.

Time: The computational run time that an XAI technique will require to generate an explanation for a ML model's prediction [24]. For normal algorithms, such a comparison metric is fairly straightforward but with the amount of differences between the investigated XAI models, comparing their computational time is nearly impossible and we will be relying on comparison survey's or direct comparison's between the techniques in the implementation papers to classify them into 3 categories: Low, Medium or High[24].

Interpretability: A degree of measurement of how well the explanation provided by a XAI technique can be understood by a human without any ML experience [6]. Can also be referred to as the Comprehensibility or Functional Understanding of a system [6]. (Scale: Low, Medium, High)

Model Privacy Despite the main focus of this paper being around AI that provide explanations behind an ML model's prediction, some XAI techniques have been engineered with the aim to provide a degree of privacy behind the ML model's internal workings [25]. There is often a trade-off between a techniques ability to provide explains and it's ability to conceal a ML model's inner workings [25]. This design focus isn't present in most of the techniques investigated in this paper, save for counterfactual generation [16] but even within this technique further evaluation shows this this supposed advantage isn't present [19] (Scale: None, Present)

Table 1: Comparison between the investigated XAI techniques

XAI Technique	Scope	Approach	Consistency	RAA	Time	Interpretability	Privacy
LIME (2016) [7]	Local	Perturbation	Inconsistent [23]	None [12]	Medium [7]	Medium [7]	None
Anchors (2018) [9]	Local	Perturbation	Inconsistent [14]	?	Medium [9]	High [9]	None
SHAP (2017) [8]	Local	Perturbation	Inconsistent [11]	None [12]	High [13]	Medium [8]	None
Counterfactual Explanations (2017) [16]	Local	Contrastive	Inconsistent [18]	None [18]	Low [16]	High [17]	None [19]
Contrastive Explanations (2019) [21]	Local	Contrastive	?	?	High [22]	High [20]	None

6 Conclusion and Future Work

6.1 Future Improvements for XAI Techniques

Regarding future improvements for the techniques, after the comparison was finished a few general improvement points across the spectrum of investigated techniques were identified. The first main issue identified is the lack of consistency/robustness present throughout many of the current XAI implementations. Out of 5 models investigated, 4 were found to be inconsistent [12] [14] [18]. This a major predicament for XAI techniques, as having explanations that vary massively between similar predictions is unacceptable for the usage of XAI in any critical field. For any future work into the explored XAI methods in this investigation, we would propose that authors include an evaluation section in their proposal paper detailing the consistency/robustness of the new XAI technique, as well as include possible countermeasures to deal with the issue of consistency.

Another general problem present within 3 out of 5 techniques investigated, is the lack of any resistance to adversarial attacks, which relates to the previously explored issue of consistency in XAI techniques but with a more nefarious twist. Adversarial attacks against XAI methods aim to change the explanation provided through slight perturbations to the input data [10]. Currently, the amount of research done into performing adversarial attacks against XAI methods [12] [18] has not been matched by researchers attempting to guard against such attacks. In future research on the explored XAI models, there should be an evaluation section into how resistant a XAI technology is against attacks as well as what steps can be taken to increase the XAI model's resistance.

Focusing on more specific future improvements to individual XAI techniques, **LIME** suffers with an overestimation of it's interpretability by the original proposal authors[7]. Various studies done on the usability of LIME has shown that non-expert users had difficulty interpreting and understanding the explanations provided by LIME [10]. Therefore for any future additions to the LIME XAI model a more thorough and complete user evaluation study should be conducted. Another issue present is the lack of any indication for the locality of the model generated by LIME, and future extensions of LIME should conduct more research into an indication of the area covered by LIME's explainable model.

SHAP suffers to a similar problem as LIME since it also has been shown to have a low level of interpretability in practical applications for non-expert users [26]. A more accurate investigation into the usability of SHAP should considered for future research. Model-Agnostic SHAP implementations has one of the worst computational times out of all the investigated techniques [13], more research should be conducted into if it is possible to optimize this run time.

The final perturbation XAI technique explored **Anchors** shows no specific major issues besides the general issues described in the first two paragraphs. Minor problems such as conflicting anchors being generated or too specific anchors are easily fixable with minor modification to the existing implementation[13], the improvements for the future of this technique we mainly propose is to deal with the issue of consistency and a lack of resistance to adversarial attacks.

One unique case for future improvements was found in **Counterfactual Explanations**, where one of the listed major advantages of this XAI approach was that it was privacy focused and it wouldn't expose the inner workings of the ML model it was explaining [13]. This turned out to be untrue as shown in [19] where researchers discovered Counterfactuals can in fact be used to extract ML models. Further research into how to prevent such scenarios should be considered, since being privacy focused can be one of the main advantages of a Counterfactual XAI approach. Excluding this issue other problems to be tackled include the inconsistency of the method and its lack of resistance to attacks.

The final technique explored **Contrastive Explanations** is also the technique with the least amount of research done into it's advantages/disadvantages. There was lack of papers detailing anything regarding it's resistance, complexity or consistency, and with the only paper [24] mentioning the consistency of the CEM. Therefore further research directions for this technique would include more exploration into it's performance, time complexity and it's robustness.

6.2 Future Research Directions

Beyond the specific future research directions for improvements to XAI techniques, we will also propose future general directions for research within the field of XAI models. Within our research we noticed there is a definite lack in the amount of direct comparison and survey papers between the current state-of-the-art XAI models when compared to the amount of implementation papers. This leads to significant difficulty for the direct comparison of XAI techniques against each other due to this lack of surveys. The creation of this paper ran into these very issues as certain metrics we would have liked to include were simply unattainable due to the sparsity of general surveys or comparisons with the XAI field. While many of the proposal papers do include some evaluation between the new technique proposed and a few others, this is often a very narrow range of compared models and there are few reliable large-scale XAI surveys. For further research within this field, more analysis and focus should be put onto the creation of large scale surveys of XAI techniques that contain user evaluation studies, complexity/run-time analysis, practical application evaluation, and adversarial resistance testing.

6.3 Conclusion

For this paper, we have performed a detailed analysis and investigation into the selected 5 Model-Agnostic XAI techniques. After the short introduction to what XAI techniques are and their goals, the paper went into a detailed look into the individual XAI technique's inner workings and function as well as their inherent advantages/disadvantages. Once all 5 techniques were evaluated and investigated the paper moved onto the comparison section of the report. Here we listed the techniques against a table of metrics and compared the techniques amongst themselves. From this we gathered potential points for future improvements and for future research in the field. The main contributions of this paper is the detailed analysis into the advantages/disadvantages of each technique, as well as the future points of improvements and research directions for the XAI field outlined in the final section.

7 Responsible Research

The next two section will detail the two main requirements of responsible research which are: maintaining the Scientific Integrity of a paper (Proper source gathering and usage), and Reproducability (allowing for the replication of the results found within the paper)

7.1 Scientific Integrity

This paper is primarily a research-centered document with no experiment being conducted over the course of the research, therefore this means that all the information gathered and collected for analysis within this paper has been gathered from external sources and papers. This reliance upon external sources and paper for information required us to approach the process of our literature gathering and review with the utmost care to uphold the standard of scientific integrity within this paper.

Therefore even before we started gathering data and sources for our investigation into model-agnostic XAI techniques, the selection of the proper databases where we conduct our search was critical. In the end, and after the suggestion of our supervisor, we settled on using both Scopus and arXiv as our primary databases where we will gather pertinent research papers. Scopus is an academic database containing thousands of academic papers, and can be deemed as good database for reliable peer-reviewed scientific articles and papers. The other database used, arXiv, is an pre/post print research paper database that is owned by Cornell University. While not being fully peer-reviewed, the fast moving nature of XAI research meant that many of the useful papers we used in our research were located on this database. Both databases were identified as reliable sources for our research and were the primary locations where we gathered papers that we have used in the in the final investigation. The one notable exclusion from this data-set was an extremely useful book titled "Interpretable Machine Learning" that was determined during literature analysis as being a reliable source for information for this report.

After selecting two reliable databases where we gathered selected research papers, the literature review process formally began. After an initial set of papers were received by us from the supervisor, we then curated and extended this set of papers from sources we've gathered on Scopus/arXiv. This set of papers then underwent additional review by our supervisor, who determined whether they are pertinent to our investigation and reliable enough to include in our research. Finalizing our literature review, we got approval from our supervisor on the final used set of papers, as well as possible suggestion for additional sources.

This stringent process of literature review improved the scientific integrity of our paper since most of the sources used in this review came from reliable databases and were approved by our supervisor for inclusion in our final paper.

Beyond the literature review of portion of our paper, to ensure our paper had proper scientific integrity, we decided against the usage of singular survey papers when creating our comparison tables and sections of our model-agnostic XAI technique investigation. While it might have been easier to simply use a singular survey's final results to determine most of our classification and comparison of XAI techniques, doing so puts too much exposure for our scientific integrity on the reliability of a singular source. Therefore for many of the identified metrics within our paper, we used different papers and sources to back up claims such as a XAI technique being labeled inconsistent. This therefore increases our overall scientific integrity since we did not rely overly on one or two sources for the comparison of the XAI techniques.

Concluding, in order to ensure scientific integrity is held up within this paper, we decided to undergo a strict literature review process, to ensure that the sources gathered and used within this paper originate from accurate and reliable sources. We also focused on the variety of the papers we used as sources throughout the paper, ensuring that no one paper becomes the be-all end-all of our research.

7.2 Reproducability

To ensure that our paper has the factor of reproducability, we've included citations for each of the sources used at the end of research paper (reference style being used was IEEE). All the sources that were referenced within the paper, as well as the sources where we used figures from, can be found in the references section at the end of the report. This allows for the easy retrieval of the sources we've used within our paper, and therefore many of the details we've uncovered in our report are easily viewable within the cited sources. Since this paper focuses around further research and analysis based off of other research papers, not much else must be accounted for to ensure the reproducability of our research. The inclusion of the sources within the report, and the consistent citation of where our information originated from is enough to ensure that any future research is able to reproduce our findings in this report.

References

[1] A. Carrillo, L. F. Cantú, and A. Noriega, “Individual explanations in machine learning models: A survey for practitioners,” 2021.

[2] A. Rawal, J. McCoy, D. Rawat, B. Sadler, and R. Amant, “Recent Advances in Trustworthy Explainable Artificial Intelligence: Status, Challenges and Perspectives,” 11 2021.

[3] D. Gunning and D. Aha, “Darpa’s explainable artificial intelligence (xai) program,” *AI Magazine*, vol. 40, pp. 44–58, Jun. 2019.

[4] L. Chazette, W. Brunotte, and T. Speith, “Exploring explainability: A definition, a model, and a knowledge catalogue,” in *2021 IEEE 29th International Requirements Engineering Conference (RE)*, pp. 197–208, 2021.

[5] 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, pp. 4793–4813, nov 2021.

[6] G. Schwalbe and B. Finzel, “A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts,” 2021.

[7] M. T. Ribeiro, S. Singh, and C. Guestrin, “"why should i trust you?": Explaining the predictions of any classifier,” 2016.

[8] S. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” 2017.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, “Anchors: High-precision model-agnostic explanations,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, Apr. 2018.

[10] J. Dieber and S. Kirrane, “Why model why? assessing the strengths and limitations of lime,” 2020.

[11] D. Alvarez-Melis and T. S. Jaakkola, “On the robustness of interpretability methods,” 2018.

[12] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, “How can we fool lime and shap? adversarial attacks on post hoc explanation methods,” *ArXiv*, vol. abs/1911.02508, 2019.

[13] C. Molnar, *Interpretable Machine Learning*. 2 ed., 2022.

[14] D. Slack, S. Hilgard, S. Singh, and H. Lakkaraju, “Reliable post hoc explanations: Modeling uncertainty in explainability,” 2020.

[15] S. Verma, J. Dickerson, and K. Hines, “Counterfactual explanations for machine learning: A review,” 2020.

[16] S. Wachter, B. Mittelstadt, and C. Russell, “Counterfactual explanations without opening the black box: Automated decisions and the gdpr,” 2017.

[17] C. Fernández-Loría, F. Provost, and X. Han, “Explaining data-driven decisions made by ai systems: The counterfactual approach,” 2020.

[18] D. Slack, S. Hilgard, H. Lakkaraju, and S. Singh, “Counterfactual explanations can be manipulated,” 2021.

[19] U. Aïvodji, A. Bolot, and S. Gambs, “Model extraction from counterfactual explanations,” 2020.

[20] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das, “Explanations based on the missing: Towards contrastive explanations with pertinent negatives,” 2018.

[21] A. Dhurandhar, T. Pedapati, A. Balakrishnan, P.-Y. Chen, K. Shanmugam, and R. Puri, “Model agnostic contrastive explanations for structured data,” 2019.

[22] A. Artelt and B. Hammer, “Efficient computation of contrastive explanations,” 2020.

[23] G. Visani, E. Bagli, F. Chesani, A. Poluzzi, and D. Capuzzo, “Statistical stability indices for LIME: Obtaining reliable explanations for machine learning models,” *Journal of the Operational Research Society*, vol. 73, pp. 91–101, feb 2021.

[24] 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, 2020.

[25] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, “Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai,” 2019.

[26] H. J. P. Weerts, W. van Ipenburg, and M. Pechenizkiy, “A human-grounded evaluation of shap for alert processing,” 2019.