Chapter 2

Interpretability in Machine Learning

> Magic's just science that we don't understand yet.
>
> Arthur C. Clarke

Abstract

Demand for intrinsically interpretable, transparent ML models are rising as increasing number of anthropocentric applications are using ML/AI for decision making. In this thesis we developed a set of intrinsically interpretable ML classifiers. This chapter reviews the terms associated with explainable, interpretable, transparent and fair AI or ML techniques. We hope this gives the reader has a fair idea whether our claim of having developed explainable and intrinsically interpretable classifiers in the later chapters is valid. Additionally with this chapter the author hopes to convey to readers why interpretability, explainability, and trustworthiness of ML/AI techniques has become so crucial.

2.1 Introduction

With the emergence of better and affordable sensors and other data collection tools in various domains there has been an explosion of data. Application of machine learning (ML)techniques in these domains accelerate in-depth analysis of the collected data. Machine Learning (ML) techniques are being increasingly used in healthcare, judiciary, insurance, logistics, and finance, among other anthropocentric sectors (Ghosh et al. 2020). There is a rise in demand from various stakeholders, for transparency in the AI/ML methods making a decision affecting them. For some fields, such as precision medicine it is more than just ethics which requires model transparency(Carvalho et al. 2019). Medical experts require more than just a crisp label as output for supporting their diagnosis and for strategic treatment planning. By focusing on performance as the sole criteria the models become increasingly black-box or opaque (Bibal and Frénay 2016), (Arrieta et al. 2020). In this era the results and performance are of higher interest to most researchers. However, science and society require much more than just performance metrics of an ML model to be

able to adopt it for large scale implementations in the real world, with accountability, fairness, and transparency at the core (Arrieta et al. 2020). Contrarily (Wang et al. 2020),(Holzinger et al. 2017) raises an interesting observation that we tend to hold AI to a harsher explanatory criteria than we do for drugs and clinicians. The motivation behind this is clinicians sometimes cannot explain the reason for arriving at a particular diagnosis: a decision which is intuitive to them but might not actually be explainable. Similarly certain effective drugs had been used widely even before their working mechanism was understood (Wang et al. 2020).

An interpretable model can help improve the model's implementability, ensure relatively more impartial decision-making, and promote robustness by exposing the potential adversarial attacks which are likely to affect the model's predictions. This has consequently ushered in the era of Explainable ML or Explainable Artificial Intelligence (XAI) (Carvalho et al. 2019). This chapter discusses interpretable ML models and XAI, where these terminologies are interchangeable and where not, the various related taxonomies and definitions, desired properties of an XAI or an interpretable model, issues of quantifying interpretability and explainability, different levels of transparency, different types of explainability and different categories of interpretable models. Thereafter some classifiers are discussed in terms of how interpretable or explainable they are. Finally the chapter discusses the current challenges of XAI and the trade-off between performance and interpretability.

2.2 Taxonomy

(Carvalho et al. 2019) states that interpretability lacks a mathematical definition. The mentioned publication however provides a number of non-mathematical definitions of interpretability. They have broadly divided interpretability into two categories: (a) the intrinsically interpretable models, and (b) explanation methods which can extract local explanations from opaque or black-box methods. Next we review the concept of XAI (Carvalho et al. 2019). Explainability is related to the idea of a link between humans and the decision-making model, and simultaneously an accurate approximation of the decision-making model while being lucid to humans (Arrieta et al. 2020). According to (Arrieta et al. 2020) XAI is defined as "Given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand." Thus any technique of model simplification or reducing of model complexity should be regarded as an XAI approach (Arrieta et al. 2020). At this point explaninability and interpretability both seem interchangeable terminologies. Some publications have also discussed related terminologies which are used as replacement of interpretability and/ or explainability. Therefore in this section we shall review some of these terms, as explained in (Arrieta et al. 2020), (Carvalho et al. 2019) and (Bibal and Frénay 2016).

- **Understandability / intelligibility:** The characteristic of a model to enable a human to understand how it works, without having to go into the detailed logic of its algorithm. Understandability has two aspects: model understandability and human understandability (Arrieta et al. 2020).

- **Comprehensibility:** It is the ability of a model to illustrate its acquired knowledge in a manner which humans can understand. This characteristic is generally associated with model complexity (Arrieta et al. 2020),(Bibal and Frénay 2016).

- **Transparency:** A model is defined as transparent if it is intrinsically interpretable or understandable. The different degrees of understandability in a transparent model varies (Arrieta et al. 2020), and are explained in subsection 2.4.3.

- **Explainability** is generally associated with post-hoc explainability of a model, since it includes techniques which can render even non-interpretable models into at least locally explainable ones. For e.g., using techniques such as LIME and DeepView on opaque ML models (Ribeiro et al. 2016), (Schulz et al. 2015), (Arrieta et al. 2020).

Understandability is the central concept pf XAI, and both transparency and interpretability are closely linked to it. Due to this, the target audience is the cornerstone of XAI.

2.3 Properties of an XAI

Some of the goals and properties of XAI, which we try to address in this thesis, are explained in the following paragraphs.

Trustworthiness: It might be considered as the confidence of whether a model will function as was intended to a given problem. It is not easily quantifiable. Also, every trustworthy model might not be explainable. The target audience include domain experts and stake-holders (Arrieta et al. 2020).

Causality: To verify a XAI model for having this property, one requires significant prior knowledge to establish that observed effects are indeed causal. Even though ML models can find correlations between data variables based on the data it learns from, that might not be enough to establish a causal relationship. However, when the ML model is explainable it can provide intuitive causal relationships within the data it has learned from. Target audience for this property are the domain experts, managers and executive board members, regulatory bodies (Arrieta et al. 2020).

Transferability: When a user is able to understand the inner relations and working within of a model, the user is able to apply this extracted knowledge on a different problem as well. Lack of proper understanding of the working logic of the model might cause the user to develop incorrect assumptions about it and that eventually is likely to lead to fatal consequences. However, not every transferable model is explainable. Domain experts and data scientists are the target audience for this property (Arrieta et al. 2020).

Informativeness: Information is required to substantiate the user's decision based on the ML model's suggestion. It is therefore expected of an explainable ML model to be able to provide information about the problem it was applied on (Arrieta et al. 2020).

Confidence: For an ML model to be reliable its confidence should be evaluated. This involves finding how robust and stable it is. Trustworthiness and other properties of a model become meaningless if the model is unstable. An explainable model should provide information about the confidence of its final decision. The interested audience are domain experts, developers, managers, and regulatory bodies (Arrieta et al. 2020). However as mentioned in (Wang et al. 2020), (Holzinger et al. 2017), when a diagnosis is made by a human doctor, even if they are unable or unwilling to share their decision-making process, their decisions enjoy far more confidence by the stakeholders (such as patients, family of patients and insurance companies) than when it is a diagnosis by an ML algorithm.

Fairness: An explainable ML should be able to illustrate how the decision was made by the model, thus allowing for analysis of the model's fairness and ethics. Additionally it should be able to identify whether bias was present in the data. The target audience for this trait are the stakeholder and regulatory bodies (Arrieta et al. 2020). Again, when it is a human expert, such as a doctor making the decision, even if their decision-making process cannot be perfectly elucidated, society is currently still more biased towards having greater faith in that decision with respect being fair, than when the same decision is reached by an inanimate object such as an ML classifier (Wang et al. 2020),(Holzinger et al. 2017).

For other properties and their stakeholders, such as accessibility, interactivity, and privacy awareness please refer to (Arrieta et al. 2020).

2.4 Types of explainability and interpretability and transparency

There are different criteria for categorising techniques making a model interpretable. According to (Carvalho et al. 2019) there are three such categories: (1) pre-model vs in-model vs post-model interpretability; (2) intrinsic vs post-hoc interpretabil-

ity; and (3) model-specific vs model agnostic interpretability. On the other hand, (Arrieta et al. 2020) discusses the three levels of transparency, and the types of post-hoc explainability for ML models. In this section we shall first discuss the categories of interpretability as explained in (Carvalho et al. 2019), follow up by post-hoc explainability and end with the different levels of transparency.

2.4.1 Types of interpretability

Pre-model interpretability techniques are model-independent and depend only on the data. They comprise of descriptive statistical techniques to exploratory analysis techniques to data visualization techniques, such as filtering, t-Distributed Stochastic Neighbour Embedding (tSNE), and principal component analysis (PCA), k-means clustering, among others (Carvalho et al. 2019). **In-model** interpretability is the trait of **intrinsically interpretable** ML models. Thus any model which is not in-model interpretable, automatically needs post-hoc interpretable techniques. **Post-model interpretability** is synonymous to **post-hoc explainability**, which is explained in subsection 2.4.2. **Model-specific** interpretability techniques are usually associated with the intrinsically interpretable models (in-model interpretability). They can also extend to some models which are not intrinsically interpretable, however they would be based on the specifics of the model's internal working. **Model-agnostic** techniques on the other hand can be applied to any ML model, including the opaque black box ones. These cannot access the working logic of the models they are trying to interpret (Carvalho et al. 2019).

There have also been debates about how to compare interpretability of different classifiers, especially when comparing models of distinct types (Ghosh et al. 2020). To tackle this problem Backhaus and Seiffert proposed 3 criteria (Backhaus and Seiffert 2014) , (Bibal and Frénay 2016):
(1) the model's ability to perform feature selection from the input pattern,
(2) the model's ability to provide typical data points representing a class, and
(3) model parameters having information about the decision boundary directly encoded.
Later in the thesis we will discuss the interpretability of some models based on these criteria.

2.4.2 Post-hoc explainability

Post-hoc models can applied to intrinsically interpretable models as well since they are model-agnostic by definition. Thus they can extract additional information from the intrinsically interpretable models (Carvalho et al. 2019). However the main targets are obviously the opaque models which are not explainable by themselves. Post-hoc explainability include the following six methods: text explanation, visual

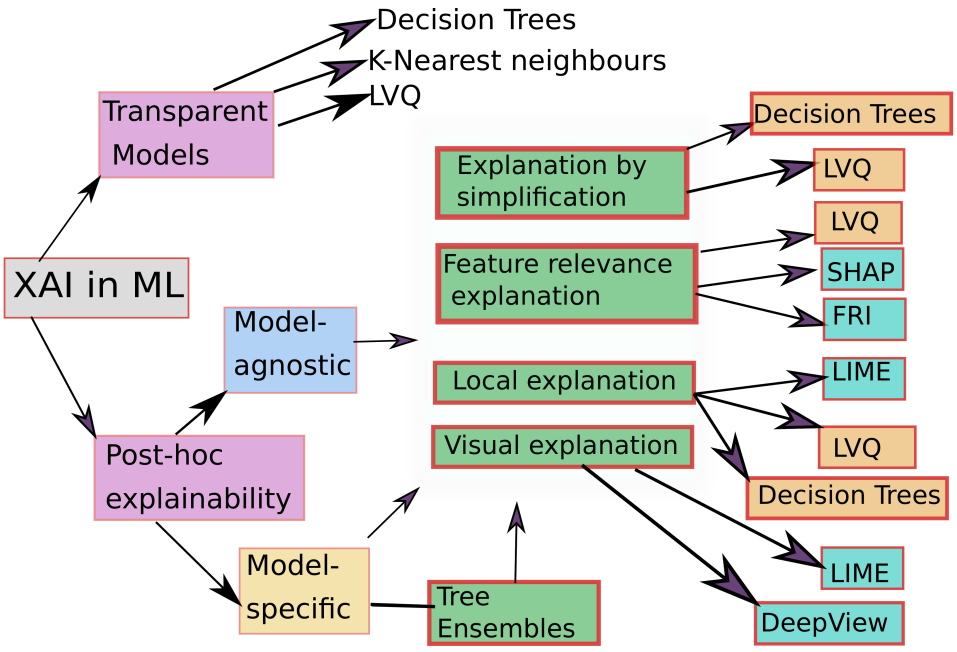

Figure 2.1: A colour-coded summary of the types of explainability and under which category each XAI technique fall. It is inspired from figure 6 of XAI:Concepts,Taxonomies, Opportunities and Challenges towards Responsible AI.

explanation, local explanation, explanation by example, explanation by simplification, and feature relevance explanation (Arrieta et al. 2020). Local Interpretable Model-Agnostic Explanations (LIME) and all its variant (Ribeiro et al. 2016), and DeepView (Schulz et al. 2015) are among techniques which provide explanation by simplification and also provide visualizations (Schulz et al. 2015). Feature Relevance Information (FRI) (Pfannschmidt et al. 2019) and SHapley Additive exPLanations (SHAP) (Lundberg and Lee 2017) are among techniques which provide post-hoc explainability in terms of feature relevance explanation.

In the recent years another set of post-hoc explainable techniques which became well-known as model-agnostic methods for extracting explanations from opaque models are counterfactual explanations or counterfactuals in short. Counterfactuals provide explanation of which are the minimum number of features which need minimum amount of modification in order for an instance to be predicted with the target label (Artelt and Hammer 2019), (Verma et al. 2020). The counterfactual in-

volve a transparent surrogate model which trains on the same set of input features as the opaque model was trained on and cost-functions which ensure that the counterfactual is as close to the original instance as possible, and yet the output label is the desired label. Certain features cannot be modified in a certain direction. Let us suppose that for a health insurance of a 70 year old suppose the counterfactual indicates that prior hospital admission needed to be less than a certain number or age lesser than 65 years. These changes are infeasible. So counterfactuals need to be computed by keeping certain features constant, so as to get a realistic guidance to reach the target outcome. Further details about this interesting model-agnostic post-hoc explainability tool can be found in literature dedicated to counterfactual explanations, such as (Artelt and Hammer 2019), (Verma et al. 2020).

2.4.3 Levels of transparency

When a model is transparent it is self-interpretable to some extent. Models deemed as transparent are further divided in terms of the domain in which they are interpretable. These domains are, algorithmic transparency, decomposability, and simulatibility (Arrieta et al. 2020). If a model can be simulated strictly in the human mind then it is **simulatable**. If the model can be decomposed into various parts and each of those parts could be easily explained, then the model checks the criteria of **decomposability**. However, for decomposability to be achieved by a model, the features themselves should be interpretable, i.e., the features after transformation to a different feature space such as in PCA, or probabilistic PCA, will hinder the interpretability of the features. When a user can follow the working logic of the model and why it produced a certain output from its input data the model is said to have **algorithmic transparency**. However for algorithmic transparency in a model it is required that it is fully explorable by mathematical analysis methods(Arrieta et al. 2020).

2.5 Some transparent models

In this section we discuss some of the intrinsically interpretable ML models, i.e., ML models which have in-model interpretability, or otherwise models which are transparent to different degrees. We will discuss only those models which we have explored later in the thesis as well.

2.5.1 Decision trees

These check all the criteria for transparency, however depending on the depth of the tree. If the tree is short enough and can be thought of in the mind of a human, without any mathematical background, then it is simulatable. However more complex trees with more nodes and greater depth can render them decomposable (if the

input features are self explanatory), or algorithmically transparent otherwise. Post-hoc analysis is not required for decision trees (DTs) (Arrieta et al. 2020). However DTs are prone to overfitting and are weak learners by themselves (Parsons 2005). Tree ensembles, the most popular among which is the random forest (RF) is therefore used to mitigate the issue of poor generalization of DTs. Ensembling however compromises the transparency of the model(Arrieta et al. 2020)(Parsons 2005).

2.5.2 K Nearest Neighbours

K-Nearest Neighbours (kNN) predict the class label of a test sample based on the class to which majority of its k nearest neighbours belong. The nearest neighbours are picked, based on the selected dissimilarity or distance measure, the most common of which are Euclidean distance, and Manhattan distance (Parsons 2005). This type of decision-making is also very intuitive for humans, which is why this classifier falls under transparent models. When the data is high dimensional, and/or the dissimilarity measure is too complicated to be easily simulated in human minds then the classifier's transparency changes from simulatable to decomposable. When the hyperparameters such as type of dissimilarity measure, and number of neighbours, are required to be set using statistical analysis tools, and/or when the dissimilarity measures themselves are not decomposable, then the classifier is rendered algorithmically transparent (Arrieta et al. 2020).

2.5.3 Learning Vector Quantization

Learning Vector Quantization (LVQ) is a nearest prototype based classifier (NPC). It is similar to KNN in terms of being a dissimilarity based technique. Similar to DTs and KNNs, LVQs can check all the criteria for transparency. In LVQ each class is represented by one or more representative point called a prototype. LVQ techniques discussed in (Schneider et al. 2009), (Bunte et al. n.d.), (Hammer et al. 2005), (Hammer and Villmann 2002), (Ghosh et al. 2020) among others, use adaptive dissimilarity instead of absolute dissimilarity between a newly presented sample and the prototypes to evaluate the class to which the sample should be assigned, based on the winner takes all scheme. Each of the features gets assigned a certain relevance during training, which makes adaptive dissimilarity computation possible. This method is explained in detail in a forthcoming chapter. The LVQ method can be decomposable or algorithmically transparent based on the type of dataset, variant of LVQ (discussed in later chapters) one is using. The LVQ models also check all the criteria of interpretability given by Backhaus and Scheiffert (Ghosh et al. 2020).

2.6 Models needing post-hoc explainability

Among models which require post-hoc explainability or model-agnostic techniques we will take the example of only the shallow ML method which we have also investigated later in this thesis: Random Forest.

2.6.1 Random forest

Random forest is a well-known tree ensembles introduced by Leo Breiman in (Breiman 2001). Even though RF has a good generalization ability when there are enough number of trees in the ensemble, it loses out on the intrinsic interpretability which the DTs provided. Breiman investigated the variable feature importance within the forest using mean decrease accuracy (MDA) or mean increase error (MIE) of the forest performance when a certain feature was dropped or added. For tree ensembles simplification and feature relevance techniques are found to be the most frequently used ones (Arrieta et al. 2020).

2.7 Trade-off: interpretability vs performance

There are several challenges in XAI, such as (1) having no metric to quantify explainability or interpretability, (2) of trying to achieve explainability in deep learning techniques, (3) XAI and adversarial ML pertaining to security issues with AI, (4) the confidence in the output of a XAI technique, and so on. However this thesis mainly deals with the issue of the trade-off between performance and interpretability. Therefore this issue will be discussed in this section. For the other issues please see (Arrieta et al. 2020) and (Carvalho et al. 2019).

Model complexity and accuracy do not necessarily go hand in hand. It might hold true when the provided dataset is clean and features are of high quality. The more complex a model is, the more flexibility it has at its disposal, such as more hyperparameter tuning. However when performance and model complexity are positively correlated to each other then interpretability had to take a step back, until some time ago. Recently though, research in XAI has led to development of some post-hoc interpretability techniques which could be applied on increasingly complex models as well (Arrieta et al. 2020). However these model-agnostic post-hoc explainability techniques still remain only locally interpretable and as mentioned earlier, cannot access the working of logic of the model (Doshi-Velez and Kim 2017) (Carvalho et al. 2019) (Ghosh et al. 2020). Thus some of the desired properties of an XAI, such as transferability, and accessibility are not met.

2.8 Conclusion

After considering the performance interpretability trade-off one should prefer interpretable techniques whenever possible. When a model is interpretable and explainable it is possible to check whether its decision making was fair and unbiased, thus leading us towards responsible AI (Arrieta et al. 2020). When a XAI technique reveals the factors based on which a model made its decision one can remove the sensitive information containing features which could marginalize people from certain post codes or could lead to bias against certain minorities. A fair XAI model should ensure that no individual or group is discriminated against and everyone has a fair chance. The model is however not the only factor. For fair and unbiased results from even the most transparent and high performing models we need unskewed, untainted data in which the class imbalance (if present) is handled carefully, and proxy features of sensitive information carrying features are removed (Arrieta et al. 2020)(Carvalho et al. 2019).

This thesis introduces a set of intrinsically interpretable classifiers of the LVQ family. These newly developed classifiers are competitive in terms of their performance and interpretability. Use of such classifiers especially in interdisciplinary research helps in useful knowledge extraction from the dataset which can enrich the domain from which the dataset was obtained.