2 Background

# Original

There are several main approaches to IML/XAI, which are briefly introduced along with the limitations here. For a more in-depth discussion, please see [15].Firstly is exploratory data analysis [32] (EDA). While EDA can help analyse features and attributes of the data, it does not tell us anything about the model being used. For this reason, we do not consider EDA as part of IML/XAI. Rather, a pre-processing step for data explanation (and not model explanation). Likewise, features selection can also be considered as IML [34]. However, again here we consider this a potential pre-processing step only.

Next is to use explainable models directly, with methods such as decision trees, linear models, or simple classification rules. While it is true that these models can offer high interpretability, the performance is drastically lower than the current state-of-the-art methods in ML (e.g. neural networks, random forests and boosting). This is true for most classification algorithms, where there is a trade-off between the accuracy and interpretability of the models [8]. Often, this drop in predictive ability practical is too large to consider for the increased interpretability. For this reason, simple models on their own are not an ideal approach to IML.

Sparse models are another approach. However, while sparsity does simplify models, they are still not necessarily interpretable. For example consider applying an L1 penalty to a deep neural network, despite having zeroed out some weights. This resulting model is still far from interpretable.

For deep learning methods, there are various explanation methods which can be used, such as sensitivity analysis with partial derivatives, heat maps of activation’s, layer-wise methods [28], or deconvolutional networks [36]. The limitation with such methods is that they are only relevant to deep learning models (not arbitrary black-box models), and also are often local (applicable to a single prediction only) in their application.

There are other local model agnostic approaches to IML, such as LIME [27], which give information about particular predictions, but not on the global behaviour of a system. Local explanations can be useful but should be paired with a global approach to provide a fuller understanding.

A global model agnostic approach to IML is model extraction [1] (also called mimic models [18] or global surrogate models [23]). One of the trade-offs of using interpretable models was a drop in predictive performance, so one solution is to utilise two models – one black-box (complex) model for predictions, and a secondary simple model for describing the black-box. This is referred to as model extraction [1]. Rather than the secondary model being trained on the original outputs, the secondary model is trained on the predictions from the black-box model. This process is shown in Fig. 1.This is the approach we take in this work, due to the fact that the method is applicable to any black-box method and makes no further assumptions (such as gradient-based, or ability to apply sparsity).

It should also be mentioned that it is not always the case that interpretability is important, i.e. to prevent "gaming the system"[18], or in well-studied problems [14]. A model extraction approach means the standard models used and work-flows can remain the same, however, in cases where interpretability is required, a secondary model can be utilised to gather additional insights to the complex model.

# Condensed – REVISE THIS!

There are several approaches to IML/XAI some of which we introduce. We do not consider exploratory data analysis [32] (EDA) and features selection [34]. They are useful pre-processing for data explanation but not for model explanation.

Direct use of explainable models is another IML technique (decision trees, linear models or simple classification rules). While these models can be easily interpretable their performance is drastically lower than the current state-of-the-art ML methods (neural networks, random forests and boosting). For this reason simple models alone are not an ideal approach to IML.

Spare Models - TODO

LIME - TODO

Deep Learning explanation models -TODO

A global model agnostic approach to IML is model extraction [1] (aka mimic models [18] or global surrogate models [23]) where two models are utilised - one black-box (complex) model for predictions, and a secondary simple model for describing the black-box; that is trained on the black-box’s output/predictions (Shown in Fig. 1).

Previous approaches to model extraction include using: Decision Trees, Bayesian Rule Lists and Logistic Regression as simple interpretable models.

Model extraction is the approach we take in this work because the method is applicable to any black-box without affecting its work-flow, therefore, in cases where additional insights into a model are required model extraction can provide an unobtrusive interpretation.