5 FURTHER ANALYSIS

# Original

To give further insights on the resulting extracted models, a comparison is given on the hill-valley dataset across the methods. The hill-valley dataset was chosen as this is the most challenging (i.e. lowest average test accuracy on the black-box models), and the deep learning recreations are used. While we would like to present all results for all methods, this would compromise 900 diagrams per method. Logistic regression is not presented, as this only a list of up-to 100 coefficients for the features, and 1 value for the intercept. These comparisons are shown in Fig. 8.

From the examples (Fig. 8), we can see that the resulting tree from the proposed method and the Bayesian rule list are by far the simplest of the compared methods, condensing the approximate knowledge from a 200 layer neural network into small human readable form. A similar trend is seen across the datasets when comparing the complexities. In this case, the Bayesian rule-list actually just predicts 1 class, so is overly simplistic (shown by the differences in the f -measures). To get an idea of what GP has found, the evolved tree can be consulted (Fig. 8a). The evolved tree is attempting to split the data based on four features (or points) in the dataset.

This makes sense when we consider the hill-valley dataset, which "when plotted in order (from 1 through 100) as the Y coordinate, the points will create either a Hill (a ’bump’ in the terrain) or a Valley (a ’dip’ in the terrain)" [13]. We can see the tree is checking the first point, and comparing to the point at 30% (i.e. the 30th feature), or the point at 70%, where the tree is trying to distinguish between classes by finding the common points for the hills/valleys and checking if these are high or low relative to the training data (e.g. a high point at the start, a low point at 30%, then a high point at 57% indicates a valley based on this tree).

Across the board, the datasets which were most difficult to reconstruct the predictions on were: Autouniv-Au7-500, Eeg-Eye-State, Gesturephasesegmentationprocessed (GesturePhase), and MonksProblems-2.

Two of these datasets (Autouniv-Au7-500, GesturePhase) have 5 classes. One explanation here is that as the number of classes in a dataset increases, as does the complexity necessarily with treebased methods. For example, if we have 100 classes, we therefore require 100 leaf nodes to have a predictive branch for each class. This presents a potential area of future research, as the size of the trees could negate the explainability as the number of classes grows. Here the pressure for small trees was perhaps too strong, and this requirement would need to be relaxed in the case of a high number of classes.

Monks-Problems-2 is entirely categorical features. In the proposed method, a categorical node has a branch for each feature - this potentially overfits to the training data, and combining categorical features into a single branch should be considered in future work (for reference, this is done in the decision tree method, where we can see a significant improvement in reconstructive ability, and this is consistent across the datasets with all categorical features).

For eeg-eye-state, the data is sequential/time-series. The proposed method is not optimised/designed for such datasets, so this explains the lower performance.

To highlight another benefit the proposed method has over the existing IML approaches, a Pareto front is given in Fig. 9. This was the result for one of the runs on the kr-vs-kp dataset, but similar fronts are available for all datasets. In all cases, the model with the highest reconstruction ability was chosen, however, even simpler models could be used from the front if desired. Likewise, if models were overly simple, any restrictions on the height of evolved trees could be relaxed.

# Condensed

To give further insight into the resulting models from the compared IML methods, a comparison is given on the hill-valley dataset. The hill-valley dataset is considered the most challenging as the black-box models had lowest average test accuracy on this dataset. We can see that the proposed method’s resulting tree and the Bayesian rule list are by far the simplest interpretable models, both condense the 200 layer neural-network into small human readable form.

Looking further the Bayesian rule list just predicts 1 class so is considered overly simplistic. Looking into our evolved tree we can see its splitting points make sense when considering the hill-valley dataset, which "when plotted in order (from 1 through 100) as the Y coordinate, the points will create either a Hill (a ’bump’ in the terrain) or a Valley (a ’dip’ in the terrain)" [13]. We can see the tree is checking the first point, and comparing to the point at 30% (i.e. the 30th feature), or the point at 70%, where the tree is trying to distinguish between classes by finding the common points for the hills/valleys and checking if these are high or low relative to the training data (e.g. a high point at the start, a low point at 30%, then a high point at 57% indicates a valley based on this tree).

Below are the datasets that were most difficult for our method to reconstruct predictions for:

* The Autonuniv-Au7-500 and GesturePhase datasets have 5 classes. As the number of classes increases so does the complexity for tree-based methods. Here the push for smaller trees may need to be relaxed in cases where datasets contain a high number of classes.
* Monks-Problems-2 is entirely categorical features. In the proposed method, a categorical node has a branch for each feature – this potentially overfits to the training data. Combining categorical features into a single branch should be considered for future work.
* For eeg-eye-state, the data is sequential/time-series. The proposed method is not optimised/designed for such datasets, so this explains the lower performance.