Results

# Original

The results are shown in Table 2, on the two measures of concern: reconstruction ability and complexity. As we are interested in achieving the best reconstruction ability, while also achieving the simplest representation, we compute the number of times a method was dominated across datasets. Here, the definition of non-dominated is no other method achieves either a simpler representation with the same (or improved) recreation ability, i.e., fewer dominations is a good sign.

The number of times dominated was chosen as the comparison methods present only a single solution, therefore we cannot compare hypervolumes of resulting frontiers (despite the proposed method returning a frontier), yet we still wish to compare the two objectives without making assumptions about the importance of either (i.e. without computing a scalar value). As the proposed method returns a frontier of solutions, only the resulting solution with the highest reconstruction ability (as measured by the fitness function on the training data, not on the unseen data), was used to represent the performance. This result is shown in Fig. 5.

GP (the proposed method) was not dominated on any of the datasets. One caveat is that analysing the dominated counts alone is not a comprehensive indicator of performance, since if the simplest possible model was used (i.e. just use the majority class, which would give a complexity of 0 as no rules were learnt), then this would never be dominated. Likewise, using the black-box model itself would achieve maximal reconstruction ability, and even though the complexity would be far greater this would still never be dominated. Another argument could be that since GP was the only method which simultaneously balanced these objectives, this measure can be seen as biased towards the proposed. This is true, but also shows the ability/usefulness of population-based techniques such as GP as they can effectively optimise multiple objectives simultaneously. This shows that multi-objective optimisation is a good choice for IML, as the objectives were optimised better than the existing approaches (in terms of dominance).

In addition to the dominated counts, the relative performance of the methods must also be considered. Friedman testing paired with Nemenyi post-hoc analysis is performed to compare whether the difference in the resulting accuracies was statistically significant across datasets. The resulting p-values are visualised in Fig. 6, where the only statistically significant differences in recreation ability are between the proposed method and the standard decision tree, and Bayesian rule lists and the decision tree. For complexity, the proposed method was significantly simpler than all comparisons methods.

To analyse the results on specific results, a per dataset breakdown of the reconstruction ability vs complexity is given in Fig. 7, where the ideal position is the top left of the chart (i.e. minimal complexity, maximal reconstruction ability). From this, we are able to conclude the proposed GP-based method consistently produces compact rules, while achieving statistically equivalent accuracy to the more complex approaches. The one exception to this was the decision tree, which was, however, on average 15× more complex than the proposed approach.

# Condensed

We compare methods based on the number of times each method was dominated. Where domination is defined as another method achieves a simpler representation with the same (or improved) recreation ability. We cannot compare hypervolumes of frontiers as the comparison methods only return a single solution therefore we select the solution from our resultant frontier with the highest reconstruction ability to represent our methods performance.

GP (the proposed method) was not dominated on any of the datasets. One caveat is that analysing the dominated counts alone is not a comprehensive indicator of performance, since both the simplest possible model (majority class) and the black-box model itself would never be dominated. Another argument is that since GP was the only method which simultaneously balanced these objectives, the measure can be seen as biased towards the proposed method. This is true, but also shows the ability/usefulness of population-based techniques such as GP as they can effectively optimise multiple objectives simultaneously. This shows that multi-objective optimisation is a good choice for IML, as the objectives were optimised better than the existing approaches (in terms of dominance).

Friedman testing paired with Nemenyi post-hoc analysis is performed to compare whether the difference in accuracies was statistically significant across datasets. The resulting p-values are visualised in Fig. 6.