3.1 Overall Algorithm

# Original

The overall training algorithm is shown in Fig. 2. The black-box classifier is trained once only on the original data (x and y values), then the evolutionary process is performed based on the resulting predictions (yˆ) from this black-box model (for a total 50 generations). The evolutionary algorithm never sees the original labels (y), as this is instead attempting to recreate the predicted labels yˆ. At the end of the evolutionary run, the result is a set of Pareto optimal models/trees which approximate the complex black-box model. Only the model with the highest reconstructive ability is used here (i.e. the largest f 1). The overall evolutionary process is similar to NSGAII. When selecting individuals, the non-dominated sorting in NSGA-II algorithm is used to rank the individuals. The two objectives are outlined in the next section.

# Condensed

The overall training algorithm is shown in Fig. 2. The black-box classifier is trained once only on the original data (x and y values), then the evolutionary process is performed based on the resulting predictions (yˆ) from this black-box model. The evolutionary algorithm never sees the original labels (y), as this is instead attempting to recreate the predicted labels yˆ. At the end of the evolutionary run, the result is a set of Pareto optimal models/trees which approximate the complex black-box model. Only the model with the highest reconstructive ability is used here (i.e. the largest f 1). The overall evolutionary process is similar to NSGAII. When selecting individuals, the non-dominated sorting in NSGA-II algorithm is used to rank the individuals, to ensure elitism NSGA-II sorts across both the parent and child populations combined.