

Interpretable Neural Networks using EAGGA

Simon Stürzebecher

SIMON.STUERZEBECHER@CAMPUS.LMU.DE

Abstract

- tabular data still difficult for NNs, where it's still outperformed by other ML model classes
 - research suggests NN performance benefits from heavy regularisation - using EAGGA, we regularise an NN and achieve both comparable performance as XGB on EAGGA as well as interpretability - for this, we propose a network architecture specifically suited for the EAGGA algorithm

Keywords: tabular data, multi-objective optimization, interpretability, deep learning

1 Introduction

Tabular Data - most common type of data - still difficult for neural networks - (Borisov et al., 2024, p. 7499) - recent research suggests that strong regularisation is beneficial to NN performance on tab data (Kadra et al., 2021, 8)

- we know regularisation from linear models can come with improvements in interpretability * e.g. LM-LASSO (L1) regularisation as feature selection - reduces # features used in model by setting some coeffs to 0 (Tibshirani, 2018, p. 267) - other forms of regularisation improve performance * e.g. LM-Ridge (L2) on multi-collinear data (Hoerl and Kennard, 1970, p. 55) * e.g. NN dropout, reduces co-adaptation (Hinton et al., 2012, p. 1) * e.g. NN early stopping, reduces overfitting on training data (Finnoff et al., 1993, p. 778f)

- we want to explore if we can use NN regularisation to tackle both interpretability and improved performance (i.e. “comparable” to XGBoost) on tab data - using EAGGA framework, which proved it can improve performance while keeping (already high) performance of XGBoost on tab data - see if interpretability improvements translate to NN + performance can also be on par

2 Background and Related Works

2.1 Interpretability

As there is no clear definition for interpretability, we will consider it as “the ability to provide *explanations* in *understandable terms* to a human”, where *explanations* are logical decision rules and *understandable terms* relate to commonly used terms in the domain of the problem, as suggested by Zhang et al. (2021, chap. 1). Further, we use the term “explainability” in an exchangeable manner with “interpretability”, as is commonly done.

Explainability of a model's reasoning is in many ways desirable. Zach (2019, pp. 3-4) gives a range of examples, amongst which are *gaining trust*, e.g. when doctors rely on medical diagnosis predictions, avoiding *subconscious biases* by making sure loan approvals are non-discriminatory, or *regulatory* reasons, most notably the EU's “Right to Explanation” warranted by the GDPR (Antamis et al., 2024, p. 1) or for approval of drugs discovered

using machine learning models (Zhang et al., 2021, 1B). It can further prove helpful for explaining unexpected drops in model performance, which could arise from optimising a model with respect to loss and then judging its performance on a different metric such as accuracy, a practice known as *model debugging* (Zhang et al., 2021, 1B) or for *scientific understanding* in domains where only models can make sense of increasingly complex data anymore and learnt knowledge encoded in a model needs to be made accessible to humans to be used reliably (Antamis et al., 2024, p. 1).

Commonly, methods for model interpretation are divided into intrinsic methods, where the search space only comprises models with a structure simple enough to be considered “explainable” (such as tree-based or simple linear models), and post-hoc methods, where interpretation methods are applied after model training. Amongst post-hoc techniques, we can further divide the space into model-specific (such as analysing GLM coefficients) and model-agnostic (e.g. partial dependence plots, ALE) methods (Molnar, 2022, chap. 3.2).

Zhang et al. (2021, chap. 2) extend this distinction to a three-dimensional taxonomy, allowing for better categorisation of neural networks (NNs), a model class that in its fully-connected feedforward form is inherently non-interpretable.

Passive vs Active Approaches, where *passive* are all post-hoc methods and *active* methods actively change either the architecture or training process to increase model interpretability.

Type of Explanations, distinguishing between *example* methods, providing concrete examples of what leads to a desired output, *attribution* methods that attribute the effect on the output for a specific feature, *hidden semantic* methods, which explain the types of inputs particular neurons or layers pick up on, and logical *rules*, such as if-then clauses or tree-induced rules.

Local vs Global Interpretability, ranging from *local* methods providing explanations based on individual samples, *semi-local* methods, explaining model behaviour for sets of samples grouped by some criterion, to *global* methods, which explain the network as a whole.

Given its unclear definition, evaluating interpretability can be challenging. Doshi-Velez and Kim (2017, 3) propose a taxonomy to categorise possible evaluation methods based on their rigorosity.

Application-grounded evaluation evaluates interpretations directly with respect to the task, by having human experts evaluate the outcome and is therefore the most expensive and time-consuming of the three approaches.

Human-grounded metrics is similar to application-grounded evaluation in that a human still evaluates the interpretations, but tries to simplify the task so that a layperson can do it. Human-grounded approaches are especially suitable if it’s sufficient to validate the general concepts of a task. A typical evaluation set-up in this category is binary forced choice, where the human evaluator chooses, which of two generated explanations he prefers.

Functionally-grounded evaluation is the least rigorous, but easiest to implement of the three. It assess explanatory quality according to some formally defined proxy for interpretability and is particularly useful in ranking different models if their model-class is already identified (e.g. via human-grounded evaluation) to be interpretable. The main challenge for functionally-grounded evaluation is finding a good proxy.

2.2 Hyperparameter Optimization (HPO)

In contrast to model parameters, which are optimized during training, hyperparameters (HPs) are those describing the machine learning algorithm and are fixed before training. Because they usually have a significant impact on the trained model’s performance, HPs are often optimized, too. Let $\mathcal{D} \subseteq \mathcal{X} \times \mathcal{Y}$ be a dataset consisting of n tuples drawn from the data-generating distribution, i.e. $(\mathbf{x}^{(i)}, y^{(i)}) \stackrel{i.i.d.}{\sim} \mathbb{P}_{\mathbf{x}y}, \forall i = 1, \dots, n$. Furthermore, let $\mathcal{I} : (\mathbb{D} \times \mathbf{\Lambda}) \rightarrow \mathcal{H}, (\mathcal{D}, \boldsymbol{\lambda}) \mapsto \hat{f}$ be an algorithm that maps a given dataset \mathcal{D} and HP configuration $\boldsymbol{\lambda} \in \mathbf{\Lambda}$ to a model \hat{f} . Denote with $\mathcal{I}_{\boldsymbol{\lambda}}$ an algorithm with fixed HP configuration a chosen loss function with L . The goal of model training is, for fixed $\boldsymbol{\lambda}$, to have $\mathcal{I}_{\boldsymbol{\lambda}}$ find model \hat{f} minimising the expected generalisation error $GE(\mathcal{I}_{\boldsymbol{\lambda}}, \mathcal{D}, L) = \mathbb{E}_{(\mathbf{x}, y) \sim \mathbb{P}_{\mathbf{x}y}}[L(y, \mathcal{I}_{\boldsymbol{\lambda}}(\mathcal{D})(\mathbf{x}))]$. The goal of HPO, on the other hand, is to find $\boldsymbol{\lambda}$ minimising the expected generalisation error, i.e. $\arg \min_{\boldsymbol{\lambda} \in \mathbf{\Lambda}} GE(\mathcal{I}_{\boldsymbol{\lambda}}, \mathcal{D}, L)$. For this, there is no analytical expression available, making HPO a blackbox optimization problem. (Karl et al., 2023, pp. 2f)

We will first outline model-free methods to approach blackbox optimization problems, before presenting a model-based variant.

The most basic model-free strategies for blackbox optimization are grid and random search. In the latter, the user defines a range of interest for each HP and then evaluating points along a grid within the cartesian product of those. This has the drawback of scaling extremely poorly in both number of HP dimensions and number of query points per range of interest. Random search, on the other hand, randomly samples a value for each HP until it runs out of budget. Its explorative nature, ease of use, and no assumptions about the model makes it a good baseline. It is also superior to grid search in cases where one or more HPs have little impact on model performance, as for a given budget B , each HP will likely be queried with B different values, whereas for grid search each HP will only be queried with $B^{1/N}$ different values for N HPs. (Feurer and Hutter, 2019, chap. 1.3)

Another popular class of model-free optimizers are evolutionary algorithms (EAs), which are conceptually simple and can handle even complex parameter spaces, given appropriate implementation of operators. EAs are iteratively evolving a population of individuals (an individual is simply a hyperparameter configuration) of size μ , where in each iteration (generation), λ^1 offspring are generated via three operators (Goldberg, 1989, pp. 10-14).

Reproduction samples an individual from the population to reproduce with probability proportional to its fitness. For HPO, fitness usually refers to the performance metric the model resulting from the individual’s HP configuration is evaluated on.

Crossover selects two “parents” from the pool of reproducing individuals. An index k is sampled at random and all HP values from index k on are swapped, yielding two new “children” individuals.

Mutation randomly changes single HP values of an individual chosen for reproduction, e.g. by adding Gaussian noise to real values or flipping bits on binary values.

At the end of each generation, the EA keeps the best μ individuals, either only from the offspring (“ (μ, λ) -selection”) or more commonly from population and offspring (“ $(\mu + \lambda)$ -selection”), which is also called an “elitist” strategy, as it’s guaranteed to keep the best individual. (Feurer and Hutter, 2019, chap. 1.3)

1. We distinguish between $\lambda \in \mathbb{N}$ when referring to the offspring size of EAs and $\boldsymbol{\lambda} \in \mathbf{\Lambda}$ when referring to a hyperparameter configuration.

One of the most popular EA implementations is the “Covariance Matrix Adaptation Evolution Strategy” (CMA-ES). It employs a multivariate normal distribution to generate offspring, for which the mean is a weighted average of the previous generation’s individuals and the covariance matrix similarly is the weighted covariance of the previous generation. The weighing scheme for both is done in a way as to reproduce previously successful (i.e. selected) steps (Hansen, 2023, p. 8-11).

Another popular EA approach is Differential Evolution. Its population is randomly initialised so the entire search space is covered. In each generation g , for each $\lambda_{i,g}$ with $i = 1, \dots, n$ it creates “mutant vectors” $\nu_{i,g+1} = \lambda_{r_1,g} + F \cdot (\lambda_{r_2,g} - \lambda_{r_3,g})$, where $r_1, r_2, r_3 \in \{1, \dots, n\}$ are mutually different random indices that also different from i and $F \in [0, 2]$ is constant. For crossover, it then generates a “trial vector” $\mathbf{v}_{i,g+1} = (v_{i,g+1}^{(1)}, v_{i,g+1}^{(2)}, \dots, v_{i,g+1}^{(N)})^T$ with

$$v_{i,g+1}^{(j)} = \begin{cases} \nu_{i,g+1}^{(j)} & \text{if } u \leq \text{CR or } j = R \\ \lambda_{i,g}^{(j)} & \text{else} \end{cases} \quad (1)$$

for a random index $R \in \{1, 2, \dots, D\}$, sampled $u \sim U[0, 1]$, and crossover constant $\text{CR} \in [0, 1]$. Selection is done by picking the better of $\mathbf{v}_{i,g+1}$ and $\lambda_{i,g}$ with respect to fitness. (Storn and Price, 1997, p. 343)

Contrasting to model-free approaches, model-based methods fit a surrogate model on the target function and optimize the surrogate. Currently, one of the most popular model-based approaches is Bayesian Optimization (BO). BO is an iterative algorithm comprising a probabilistic surrogate model \tilde{f} for the blackbox problem f and an acquisition function. It maintains and continually extends a set of points it already evaluated on the target function. In each iteration, the posterior predictive distribution is determined by fitting the surrogate model on this set of points. The acquisition function then retrieves the highest utility point from the surrogate model and evaluates it on the target function, after which it adds the point to the set of queried points (Feurer and Hutter, 2019, chap. 1.3.2) and (Frazier, 2018, pp. 2f). Evidently, neither the target function nor the surrogate model are optimized directly. Rather, the acquisition function trades-off exploration and exploitation of the surrogate and is maximised to yield the next query point most likely to optimize the two objectives as defined by the acquisition function. Common choices for the acquisition function are Expected Improvement (Eq. 2) and Thompson Sampling (Eq. 3).

$$\text{EI}_t(\lambda) := \mathbb{E}_t[(f(\lambda) - f_t^*)^+] \quad (2)$$

Expected Improvement computes the expected minimisation over the incumbent f_t^* for evaluating the target function f at query point λ . \mathbb{E}_t denotes the expectation under the posterior predictive distribution with t query points, similarly, the current incumbent is the minimum value of the target function from previous evaluations $(\lambda_{1:t}, f(\lambda_{1:t}))$. (Frazier, 2018, p. 7)

$$f^{(t)} \sim \tilde{f}_t \quad (3)$$

Thompson Sampling, on the other hand, can conceptually be described as drawing a candidate function $f^{(t)}$ from the posterior predictive distribution and minimising it to obtain the next query point for the target function. (Shahriari et al., 2016, p. 161)

Composing the other part of BO, surrogate models need to be able to model mean and variance of its target function estimate. Commonly chosen models are Gaussian Processes (Eq. 4) and Random Forests.

$$\text{GP}(m(\boldsymbol{\lambda}), k(\boldsymbol{\lambda}, \boldsymbol{\lambda}')) \quad (4)$$

Gaussian Processes (GP) are fully specified by their mean and covariance functions and have closed-form solutions. The mean $m(\boldsymbol{\lambda})$ fits all points already evaluated on the target function, while for the covariance or kernel function $k(\boldsymbol{\lambda}, \boldsymbol{\lambda}')$ it is desirable that points close to each other in HP space $\boldsymbol{\Lambda}$ exhibit stronger correlation than those far apart. The kernel function can be freely chosen, a common choice is the Gaussian (Eq. 5) kernel with hyperparameter α_0 . (Frazier, 2018, p. 5)

$$k(\boldsymbol{\lambda}, \boldsymbol{\lambda}') = \alpha_0 \exp(-\|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|_2^2) \quad (5)$$

A drawback of Gaussian Processes is its poor scalability in number of data points and number of hyperparameters, although there exist workarounds approximating the full GP with only a subset of the samples. (Feurer and Hutter, 2019, chap. 1.3.2)

Random Forests, unlike GPs, can handle complex hyperparameter spaces even including categorical and hierarchical HPs. Their computational complexity is also far less, making them a popular alternative to Gaussian Processes. (Feurer and Hutter, 2019, chap. 1.3.2)

2.3 Neural Architecture Search

A special case of HPO is Neural Architecture Search (NAS). Because of the flexibility of neural network architectures (number of hidden layers, strength of dropout, each layer can have a different number of neurons, each layer can have a different activation function, etc.), traditional HPO methods cannot efficiently explore the entire space. In our extension of the EAGGA approach we don't employ NAS as research focusses on the NLP and image domains, where features (i.e. tokens or pixels, respectively) exhibit strong correlation amongst themselves, which is usually not the case for tabular data (Borisov et al., 2024, p. 7499). Still, we want to outline two notable approaches for neural architecture search in this paper, one of which are **cell search spaces**. These modularise an NN into a chain-structure of cells, where each cell is a basic building block of the respective architecture. For regular feedforward neural networks, this could be a linear layer of fixed size and with a specific activation function, for CNNs this could be a specific convolutional operation or pooling function, for instance. Instead of tuning every hyperparameter, the sequential placement of the cells is then optimized during NAS, e.g. via random search or Bayesian Optimization. (Elsken et al., 2019, chap. 3.2)

Treating a neural network as a sequence of cells instead of one fixed entity further allows exploiting this structure with special sequential optimization techniques. Zoph and Le (2017, p. 3) and Zoph et al. (2018, pp. 2-4) propose a cell search method using RNNs and reinforcement learning. The cells in their approach are not fixed, but have hyperparameters themselves, such as the filter size and the stride size of a convolutional layer. These hyperparameters are then predicted sequentially, given all the already predicted hyperparameters of previous layers. The RNN is trained using reinforcement learning, where the actions are the different values the network can predict for a given hyperparameter and the reward function is the RNN's performance on held-out validation data.

The second approach to NAS is the **one-shot model**, which trains a “fabric”, which can be understood as a DAG comprising all architectures of a given search space. Each path through the DAG is one specific network architecture, where the nodes represent a layer of neurons and the edges in-between are operations on the layers’ values. The entire DAG has one designated input and one output node. It is then trained just like a regular NN, after which the optimal path, i.e. network architecture, is selected from the fabric. This has the advantage of, albeit being slower than training a single model, being far more efficient and less expensive than training all architectures encoded in the fabric. (Saxena and Verbeek, 2017, pp. 1-2, p.8)

2.4 Multi-Objective Optimization

In most applications, practitioners don’t just want to optimize for performance exclusively but, for instance, also desire an interpretable model, which they may measure via some proxy metric. Let $c_1, c_2, \dots, c_m = \mathbf{c} : \Lambda \rightarrow \mathbb{R}^m$ be the vector of m objectives, the goal of multi-objective optimization (MOO) is to minimise this vector (Karl et al., 2023, p. 11).

2.4.1 PARETO-OPTIMALITY

Optimizing \mathbf{c} usually comes with the challenge of conflicting objectives: improving one objective means deteriorating another. We thus seek to find a set of trade-off solutions, so called *non-dominated* points. A point λ dominates another point λ' ($\lambda \prec \lambda'$) if there is no other point that is strictly better in at least one objective and better or equal in the remaining ones. Formally,

$$\forall i \in \{1, \dots, m\} : c_i(\lambda) \leq c_i(\lambda') \wedge \exists j \in \{1, \dots, m\} : c_j(\lambda) < c_j(\lambda') \quad (6)$$

as defined in (Karl et al., 2023, pp. 7f) and (Goldberg, 1989, pp. 198f).

The set of non-dominated or *Pareto-optimal* points $\mathcal{P} := \{\lambda \in \Lambda \mid \nexists \lambda' \in \Lambda \text{ s.t. } \lambda' \prec \lambda\}$ is commonly referred to as the *Pareto set* and its image as the *Pareto front*. The goal of MOO is to find a set of non-dominated points $\hat{\mathcal{P}}$ approximating the true Pareto set \mathcal{P} well.

There are different techniques to evaluate an estimated Pareto-front. If we have knowledge over the true Pareto-front, we can compute the distance between the approximation and the true front. In most cases, this knowledge cannot be assumed. Then, usually volume-based approaches are used for evaluation, most notably the hypervolume. The hypervolume, or S-metric, of a Pareto-front, computes the volume between the estimated front and some chosen reference point, which is usually the worst point in objective space. The larger the hypervolume, the closer the estimate is to the true Pareto-front. (Karl et al., 2023, pp. 8-10)

2.4.2 A-PRIORI

For optimizing a multi-objective problem, there are largely two general approaches. One of those are *a-priori* methods, of which we will briefly outline two popular techniques.

Scalarization approaches impose an implicit order of priority on the objectives. The simplest one is optimizing a weighted sum of objective functions

$$\arg \min_{\lambda \in \Lambda} \sum_{i=1}^m w_i c_i(\lambda) \text{ s.t. } \sum_{i=1}^m w_i = 1 \wedge w_i > 0, \forall i = 1, \dots, m \quad (7)$$

Weights are chosen a-priori by the user and the solution is very sensitive to them, making this method very reliant on different users’ preferences. (Karl et al., 2023, p. 11) and (Srinivas and Deb, 1994, chap. 3.1)

Another form of scalarization is the ϵ -constraint, which translates all but one objective into constraints and then optimizes the remaining objective subject to the constraints. Without loss of generality, the first constraint can be optimized

$$\arg \min_{\lambda \in \Lambda} c_1(\lambda) \text{ s.t. } c_2(\lambda) \leq \epsilon_2, \dots, c_m(\lambda) \leq \epsilon_m \quad (8)$$

This method is, similarly to the weighted sum, conceptually simple yet also very sensitive to the chosen constraints. (Karl et al., 2023, p. 12)

Alternative to scalarization is the **lexicographic method**. For it, the user assigns a priority to each objective and then greedily optimizes each objective in order of priority, constrained to the solutions of the already optimized, higher-priority objectives. (Riera et al., 2023, p. 13749) Again, solutions are very dependent on the user-defined prioritisation.

2.4.3 A-POSTERIORI

- problem a priori * either restrict search space to “enforce our will” (e.g. only use 50% of features), optimise only prediction performance + use this * or leave search space unrestricted but adjust loss function to incorporate multiple objectives + take optimum from there * in either case: no knowledge of interplay between HP config + performance on all objectives * in practical applications, it is oftentimes useful to make the decision of which point of the Pareto set to use a-posteriori, e.g. if only a slight decrease in one objective translates to a significant improvement in another that would have been missed if the problem was optimised using a-priori methods * a-posteriori evaluates configs just as a-priori, but keeps track of multiple “best” (non-dominated) solutions - makes relationship between HP config + objectives visible - aside from usual baselines grid and random search there are multi-objective BO adaptations, mainly using either of two approaches (1) fit single surrogate model on scalarised objectives, e.g. ParEGO, which employs the augmented Tchebycheff function as scalarisation to ensure the Pareto front is explored sufficiently (Karl et al., 2023, pp. 15f) + (Knowles, 2006, pp. 54-56) (2) fit one surrogate per objective, then - either one acquisition function per objective - return set of promising next candidates - or one overall acquisition function aggregating surrogates, e.g. *EHI*, maximises expected improvement of hypervolume (Karl et al., 2023, p. 16) + (Emmerich et al., 2006, pp. 8f) - lastly, there is also multi-objective EA (MOEA) algorithms to explore Pareto front, one of which is NSGA-II (Nondominated Sorting Genetic Algorithm) * improves upon predecessor NSGA (Srinivas and Deb, 1994) by making it parameterless, ensuring elitism, and reducing computational complexity of ranking of individuals (Deb et al., 2002, p. 182) * uses all the regular operators, i.e. reproduction, mutation, crossover can be used from single-objective

EA (SOEA) * difference to SOEA: ranking of individuals (SOEA uses scalar fitness, MOEA multiple objectives) - multi-objective ranking mechanism based on 2 parts * non-dominated sorting, ensures elitism - iterative procedure to rank individuals in population by their fronts (1) determine pareto front, assign rank 0 (2) remove pareto front from population (3) if any individuals left, repeat from 1, increment rank # - (Goldberg, 1989, p. 201) + (Deb et al., 2002, pp. 183f) * crowding distance, ensures sufficient diversity, i.e. exploration of pareto front - assigns score depending on how crowded area around individual in objective space is - crowding distance of an individual = mean side length of cuboid spanned by its nearest neighbours as vertices in objective space - individuals without two neighbours in a dimension are assigned infinitely high distance value - (Deb et al., 2002, p. 185) - the less the crowding distance, the more “crowded” an area is by other individuals - rank individuals by nds front rank (ascending) + crowding distance (descending) as tie breaker * ranking used for selecting μ best individuals to keep for next generation + for reproduction: via binary tournament selection: select two random individuals, pick best w.r.t. nds + cd for offspring creation (i.e. mutation / crossover) * using nds for ranking / fitness makes intuitive sense, but why cd? - goal of MOO: not simply optimise HV, but approx true Pareto front well - only having individuals representing one area in objective space makes Pareto front estimate very “unstable”: removing this area (i.e. solutions from that area) would “collapse” entire front - instead having multiple areas be represented makes the estimate “stable”: removing one area wouldn’t impact Pareto front estimate a lot - (Goldberg, 1989, p. 185, pp. 189-192)

3 EAGGA

- NSGA-II inspired evolutionary + genetic algorithm - uses AUC-ROC as performance metric and 3 interpretation metrics * NF: rel. # of features used in model * NI: rel. # of pairwise feature interaction effects * NNM: rel. # of non-monotone feature effects - aims to find high performing models (AUC) with low NF, NI, and NNM - ensures better interpretability than simply optimising for performance but creates very large objective space $\tilde{\Lambda}$ to be optimised over with tuples $(\lambda, s, I_s, m_{I_s}) = \tilde{\lambda} \in \tilde{\Lambda}$ with * $\lambda \in \Lambda$ “regular” model HP config * $s \in \{0, 1\}^p$ vector denoting feature usage * $I_s \in \{0, 1\}^{p \times p}$ matrix denoting pairwise interactions * $m_{I_s} \in \{-1, 0, 1\}^p$ vector denoting monotonicity constraint of each feature (-1 decreasing, 0 none, 1 increasing) - thus would require to solve $\arg \min_{\tilde{\lambda} \in \tilde{\Lambda}} (GE(\mathcal{I}_{\tilde{\lambda}}, \mathcal{D}), NF(\hat{f}_{\mathcal{D}, \tilde{\lambda}}), NI(\hat{f}_{\mathcal{D}, \tilde{\lambda}}), NNM(\hat{f}_{\mathcal{D}, \tilde{\lambda}}))$ - (Schneider et al., 2023, pp. 540f) - introduces equivalence relation R “allowed to interact” on features that significantly reduces search space for more efficient optimization - thus: augmented search space $\tilde{\Lambda} = \Lambda \times \mathcal{G}$ comprising model HP space and group structure space * each group structure $G \in \mathcal{G}$ consists of g groups: - G_1 set of excluded features - $\forall k = 2, \dots, g : G_k = (E_k, M_{E_k})$ tuple * E_k set of features allowed to interact with each other * M_{E_k} monotonicity constraint of entire group k - EA on model hyperparameters Λ - GGA on group structure space \mathcal{G} with adapted mutation and crossover operators - (Schneider et al., 2023, pp. 541f) - further introduces special initialisation of group structures to increase sample-efficiency of EAGGA algorithm compared to random init - (Schneider et al., 2023, pp. 542f)

4 Extension to neural networks

- original EAGGA applied to XGBoost model proved to vastly outperform union of competitor models w.r.t. dominated hypervolume and comparable or better than ParEGO on extended search space $\tilde{\Lambda}$ (Schneider et al., 2023, pp. 543-545) - why extend it to NNs? * NNs notoriously uninterpretable due to complex transformation of the feature space * due to (lin alg) non-linear activation functions, interactions can be modelled * no monotonicity guarantees - our approach (general algorithm) * EAGGA algorithm implemented largely as described in the paper * HP init - NN total layers $\in \{3, \dots, 10\}$, i.e. hidden layers $\in \{1, \dots, 8\}$, init from trunc geom with prob=0.5 - NN nodes per hidden layer (for each group) $\in \{3, \dots, 20\}$, init from trunc geom with prob=0.5 - NN dropout % $\in [0, 1]$, init from trunc gamma with shape=2, scale=0.15 * group structure init - feature detector * as described in paper * but instead of fitting 10 trees + taking rel. # of features used as prob for trunc geom, we simply use 0.5 (sklearn dectree examination not straightforward) * also in preliminary experiments we found the sampled # of features to be used occassionally to be i # of non-0 values in normalised information gain filter (e.g. if former is = total # of features and one feature is indep. from target) in these edge cases we simply use all features with non-0 filter values - interaction detector * as described in paper * same as for feat detect: use prob=0.5 to sample # of interactions used instead of 10 dec trees * also for FAST algorithm don't use RSS (lin reg metric) but mean accuracy, as we fit log reg - monotonicity detector * as described in paper * use default HPs (max depth 30, minsplitted = 20) of mlr3 classification tree implementation, as this is the library the paper uses - i both interaction + monotonicity detectors fit their models on 80% of train split from holdout + eval on remaining 20% (implementation details) * group structures + datasets - group structures implemented as described in the paper with additional list encoding sign of monotonicity of the individual features as detected by the monotonicity detector - included features are passed to dataset, dataset outputs included features, multiplied by the features' individual monotonicity (-1 / 1), so that group monotonicity can be encoded with only $\{0, 1\}$, as proposed in paper - entire group structure is passed to NN, where architecture is built accordingly * neural network - "instead of XGB as in the original paper, we apply the method to NNs to examine whether this type of regularisation can achieve interpretability on NNs while outperforming EAGGA on XGB (original paper), ..., this requires special architecture, etc. pp." - NN * comprises of "sub-NNs", one for each equivalence relation - i basically non-fully connected NN * hidden layers use ReLU activation and dropout afterwards, then a shared output layer with concatenated sub-NNs' outputs as input and sigmoid activation * optimizer is AdamW with default params, loss is binary cross entropy loss (sigmoid thus implicit in the loss function for better numerical stability, NN itself outputs "logits", which refers to pre-activation output in pytorch) - feature sparsity thus achieved by only training on included feature groups - i goes somewhat against notion of deep learning where model is supposed to decide itself, which feature to "use" / put importance on - i ELABORATE - feature interaction achieved by grouping, max-operation in ReLU induces interaction effect (different kind of interaction than e.g. in LM with multiplication) - i equation why the interacting features need to be grouped together when using ReLU, cf. photos * given layer input \mathbf{x} and w.l.o.g. omitting bias * $\text{ReLU}(\mathbf{x}) = \max(0, \sum_{j=1}^p w_j x_j)$ * interaction not of the form $x_j \cdot x_{j'}$ as used to from LMs but

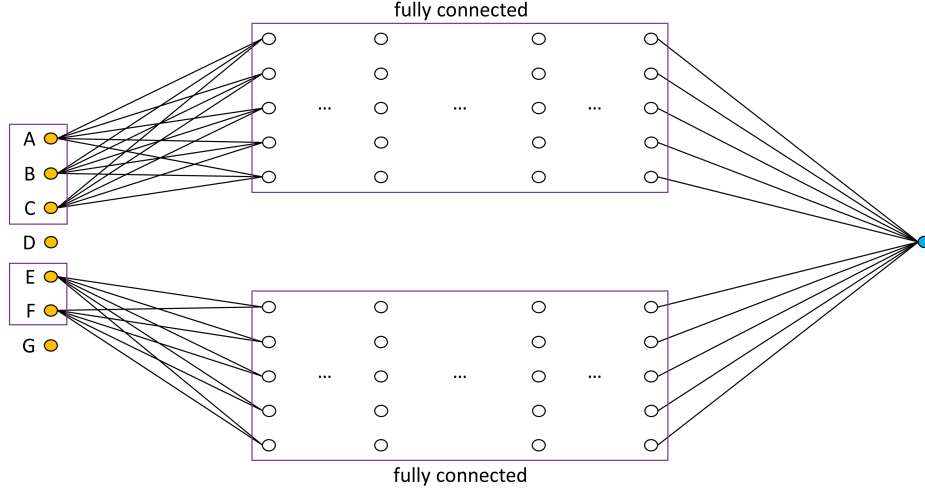


Figure 1: TODO: caption

of a form where the sum $w_j x_j + w_{j'} x_{j'}$ decides whether $x_j, x_{j'}$ pass through the activation or not * in other words: $\max(0, w_j x_j + w_{j'} x_{j'}) \neq \max(0, w_j x_j) + \max(0, w_{j'} x_{j'})$ and thus there is an interaction - monotonicity constraint achieved by clipping weights to $[0, \infty)$ after each epoch for restricted equivalence relations (monotonic decrease achieved via dataset object multiplying features with their individual signs), bias clipping not necessary (constant additive term) - equation how monotonicity is enforced with this * consider vector form computation of hidden neuron $\mathbf{o}_{\text{in}}^{(l)} \in \mathbb{R}^{p^{(l)}}$ of layer l - $\mathbf{o}_{\text{in}}^{(l)} = \mathbf{W} \mathbf{o}_{\text{out}}^{(l-1)} + \mathbf{b}$ - where $\mathbf{o}_{\text{in}}^{(l)}$ refers to pre-activation values, $\mathbf{o}_{\text{out}}^{(l)}$ to value after applying ReLU - weight matrix $\mathbf{W} \in \mathbb{R}^{p^{(l)} \times p^{(l-1)}}$ - bias $\mathbf{b} \in \mathbb{R}^{p^{(l)}}$ * definition of monotonically increasing function f : for $x_1 \leq x_2$ it must hold that $f(x_1) \leq f(x_2)$ * $\mathbf{W} \mathbf{o}_{\text{out}}^{(l-1)} + \mathbf{b}$ clearly only monotonic if and only if $\mathbf{W} \in \mathbb{R}_{0+}^{p^{(l)} \times p^{(l-1)}}$ - where $\mathbb{R}_{0+} = [0, \infty)$ - hence weight clipping necessary, bias clipping not because only constant additive term * $\mathbf{o}_{\text{out}}^{(l)} = \text{ReLU}(\mathbf{o}_{\text{in}}^{(l)}) = \max(0, \mathbf{o}_{\text{in}}^{(l)})$ monotonic * evaluation, holdout, cv, early stopping - evaluation via dominated hypervolume along AUC-ROC, NF, NI, NNM as defined in original paper * NF simply rel. # of included features * NI = sum of all possible pairwise interactions in each group over all possible pairwise interactions among all features = $\frac{\sum_g \binom{p_g}{2}}{\binom{p}{2}}$ * NNM = rel. # unconstrained features - outer holdout split: 2/3 train, 1/3 test (as in paper) * run EAGGA on holdout train split, i.e. train + select best $\mu = 100$ individuals based on non-dom-sorting (ascending) + crowding distance (descending) as tie breaker - inner CV split on outer train portion: 5-fold (as in paper) * in each fold fit model on 80% of CV-train portion, use remaining 20% of CV-train for early stopping - early stopping criterion * for each fold, always train for min 200 epochs, keep track of model with lowest loss * after that, use patience of 100: if current model's loss is $\bar{\ell}$ mean of last 100

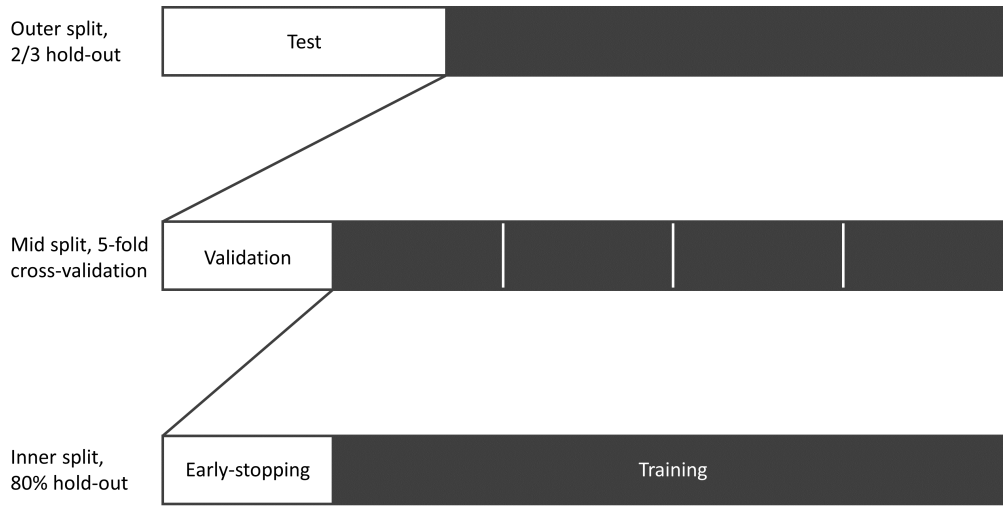


Figure 2: TODO: caption

epochs' losses -i stop early * if no early stopping, train each fold for max 10 minutes * then stop training, return model with lowest loss -i refer Figure 2 - then evaluate on last remaining from CV + keep best μ individuals + generate $\lambda = 10$ offspring for new generation + repeat - final evaluation of individuals of pareto set * for each individual, train model on training set of hold-out for max # of epochs taken during CV-training * decided on max # epochs instead of mean after looking at loss graphs in preliminary experiments, also refer Figure 4 * those show that there is no uptick in losses on the early stopping portion (disjunct from training portion) of the set, hence max is reasonable * hardware: training on Sagemaker Notebook instance - initial experiments on ml.g4dn.xlarge instance (2 vCPUs, 16GiB RAM, 1 NVIDIA T4, cf. <https://aws.amazon.com/de/ec2/instance-types/g4/>) using cuda not much faster than on ml.t3.medium (2 Intel Xeon 8000 vCPUs, 4GiB RAM, cf. <https://aws.amazon.com/de/ec2/instance-types/t3/>) - thus decided for more economic + ressourcen-schonend t3.medium * did not train on philippine and gina datasets (308, 970 features, respectively), as they ran out of memory when computing interaction detectors * known bugs - in rare cases (anecdotally once every 5-10 datasets), gga_mutate seems to be generating -1 as monotonicity attribute, despite np.random.randint(low=0, high=2, size=1) * loop crashes at group.structure creation, for these cases subtract previous runtime from 8hrs + load from last generation in output via load_population * so far only happened for madeline in after gen-8.json after 5hrs 45mins (at start of gen-9, which wasn't exported yet) -i loaded this and ran for another 2hrs 15mins

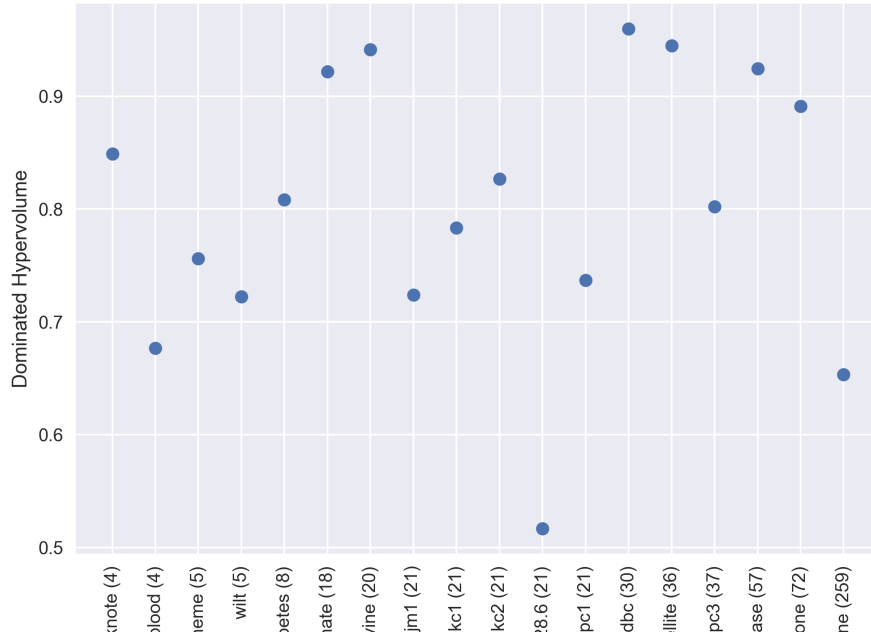


Figure 3: TODO: caption

5 Experimental Results and Discussion

- preliminary * our implementation of feature + interaction detectors likely tend to include more features + interactions for high p datasets - less for low p datasets - remember: original paper samples # of features included from truncated geometric distribution, what is not mentioned is probability of this distribution is determined from fitting 10 trees + looking at relative # of features used (cf. their github repo /R/TunerEAGGA.R function `get_n_selected_rpart()`) - \downarrow mlr3 default decision tree max depth 30 * i.e. datasets with $p \leq 30$ might use all features, which translates to trunc geom prob = 1 * vice-versa for $p \gg 30$ datasets relative # features ≤ 0.5 (our trunc geom prob) * similar reasoning for pairwise interactions * individuals not in Pareto sets occasionally exhibit AUC ≤ 0.5 - this is not by mistake, usually AUC on binary target can simply be inverted (simply predict the opposite class) - for us this cannot easily be inverted due to monotonicity constraints (weights cannot simply be multiplied by -1 if constraint) - could happen e.g. due to AdamW and weight clipping (employed to enforce monotonicity) * AdamW uses momentum * possible scenario: large momentum would yield negative weights, but after epoch they're clipped to $[0, \infty)$ + training stops due to early stopping * weight clipping very imperfect way of enforcing monotonicity, but currently in pytorch unfortunately only way to implement this

- Figure 3 * suggests comparable performance to XGBoost EAGGA * did not compare to unrestricted NN, would not be sensible, as NF, NI, NNM would be 1, hence hypervolume = 0, as values along 3 dimensions would be at reference point * as XGBoost EAGGA, NN EAGGA consistently outperforms union of competitors as evident from (Schneider et al.,

2023, Figure 2) * unfortunately due to time constraints no own union comparison with NN instead of XGBoost - overview over pareto sets can be accessed on our github repo at /code/export/*.csv * NN HPs - total layers mostly 3-4, most commonly 3, goes as high as 7 (Satellite (36), diabetes (8)) - nodes per hidden layer mostly 3-6, goes as high as 12 (diabetes (8)) - p dropout goes as high as 0.7 (climate (18), spambase (57)), but mostly in the 0.1 to 0.3 range * group structures - great diversity in NF across datasets * phoneme (5) up to 1 * blood (4) up to 1 * banknote (4) up to 0.75 * diabetes (8) up to 0.5 * climate (18) up to 0.67 - \downarrow low p datasets in the benchmark tend to have higher NF - possible consequence of shorter evaluation time - low p datasets have much higher # generations - group structure space much more likely to be exhausted, i.e. more exploration of NN hps - NI, NNM consequently (bounded by NF, can never be more than max # for respective # included features) rather low * dhv contributions - measures contribution of an individual to the hypervolume, i.e. difference between hypervolume of entire pareto front vs hypervolume of pareto front without individual λ : $\text{CON}_{\mathcal{P}}(\lambda) = \text{HYP}(\mathcal{P}) - \text{HYP}(\mathcal{P} \setminus \{\lambda\})$, where $\text{HYP}(\mathcal{P})$ denotes the hypervolume induced by the pareto set \mathcal{P} (Bringmann and Friedrich, 2010, p. 384) - predominately low for fitted models - mostly highest for featureless learner predicting majority class - \downarrow sign of good exploration of pareto front + stable estimate, refer Section 2.4.3 - loss graphs Figure 4 * suggests models could have benefitted from longer training on some datasets, as some loss curves haven't converged when stopping criterion hit * crit was likely triggered by short-term spike, thus could potentially be resolved by comparing average of last k losses against average of \downarrow patience \downarrow losses prior to that to not be as exposed to short-term spikes in loss * on other datasets, graphs suggest earlier stop would have been totally fine as the networks have long converged, but longer training likely not an issue as loss graphs come from early stopping dataset portion, which is disjunct from training set - dhv over generations Figure 5 * compute on val set, i.e. had 5 folds per individual - NF, NI, NNM always the same for each fold - but 5 different AUCs - \downarrow hypervolume of (mean(AUC 1, AUC 2, ..., AUC 5), NF, NI, NNM) * artifacts / drops along y likely due to inconsistent computation of Pareto front by third-party library - preliminary experiments on dummy pareto fronts with 10x4 metrics: noticed nds function returning different rankings for same front - returned ranking switched back and forth between two only slightly different options (only 1 or 2 indices were swapped) - not sure what caused this, as made same observation using another library that was planned as alternative - thus unfortunately not fixable for me * all but 3 datasets (2 of which only trained for 1 generation, anyway) show improvement of dominated hypervolume over generations * BUT: absolute as well as relative improvement almost negligible - \downarrow no large final dhv decrease would we just have evaluated the models gotten from the detectors - this was also the reason we didn't run EAGGA on philippine and gina simply without the detectors (as that's where they crashed) - \downarrow evidence suggested that detectors are vital to initial performance - original paper supports this assumption (Schneider et al., 2023, Fig. 4, p. 545)

6 Conclusion and Future Outlook

(Conclusion) - tabular data still difficult discipline for neural networks - research suggests heavy regularisation to improve performance on tabular data - EAGGA proved to be successful in making XGBoost more interpretable while keeping performance on par with un-

regularised XGBoost - we extend EAGGA to neural networks to see if we can utilise the regularisation induced by it to make NNs both interpretable and performant on tabular data - as consequence propose new architecture allowing to model equivalence relations of EAGGA - ... conforming to Zhang et al. (2021, chap. 2) taxonomy it fits as ..., ..., ... - found overall performance comparable to that of XGBoost fitted using EAGGA, which is a plus, but no outperformance

(Future Outlook) - MO BO on group structure space possible, perhaps via restricted BO?

Appendix A. Software used

for implementation we used openml Vanschoren et al. (2014), Feurer et al. (2021), numpy Harris et al. (2020), pandas pandas development team (2023), Wes McKinney (2010), pytorch Ansel et al. (2024), scikit-learn Pedregosa et al. (2011), scipy Virtanen et al. (2020), pymoo Blank and Deb (2020), and tqdm da Costa-Luis et al. (2024)

Appendix B. Plots

References

- Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, C. K. Luk, Bert Maher, Yunjie Pan, Christian Puhersch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Shunting Zhang, Michael Suo, Phil Tillet, Xu Zhao, Eikan Wang, Keren Zhou, Richard Zou, Xiaodong Wang, Ajit Mathews, William Wen, Gregory Chanan, Peng Wu, and Soumith Chintala. Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, ASPLOS '24, page 929–947, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703850. doi: 10.1145/3620665.3640366. URL <https://doi.org/10.1145/3620665.3640366>.
- Thanasis Antamis, Anastasis Drosou, Thanasis Vafeiadis, Alexandros Nizamis, Dimosthenis Ioannidis, and Dimitrios Tzovaras. Interpretability of deep neural networks: A review of methods, classification and hardware. *Neurocomputing*, 601:128204, 2024. ISSN 0925-2312. doi: <https://doi.org/10.1016/j.neucom.2024.128204>. URL <https://www.sciencedirect.com/science/article/pii/S0925231224009755>.
- J. Blank and K. Deb. pymoo: Multi-objective optimization in python. *IEEE Access*, 8: 89497–89509, 2020.
- Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6):7499–7519, June 2024. ISSN 2162-2388. doi: 10.1109/tnnls.2022.3229161. URL <http://dx.doi.org/10.1109/TNNLS.2022.3229161>.
- Karl Bringmann and Tobias Friedrich. An efficient algorithm for computing hypervolume contributions**. *Evol. Comput.*, 18(3):383–402, September 2010. ISSN 1063-6560.
- Casper da Costa-Luis, Stephen Karl Larroque, Kyle Altendorf, Hadrien Mary, richardsheridan, Mikhail Korobov, Noam Yorav-Raphael, Ivan Ivanov, Marcel Bargull, Nishant Rodrigues, Shawn, Mikhail Dektyarev, Michał Górny, mjstevens777, Matthew D.

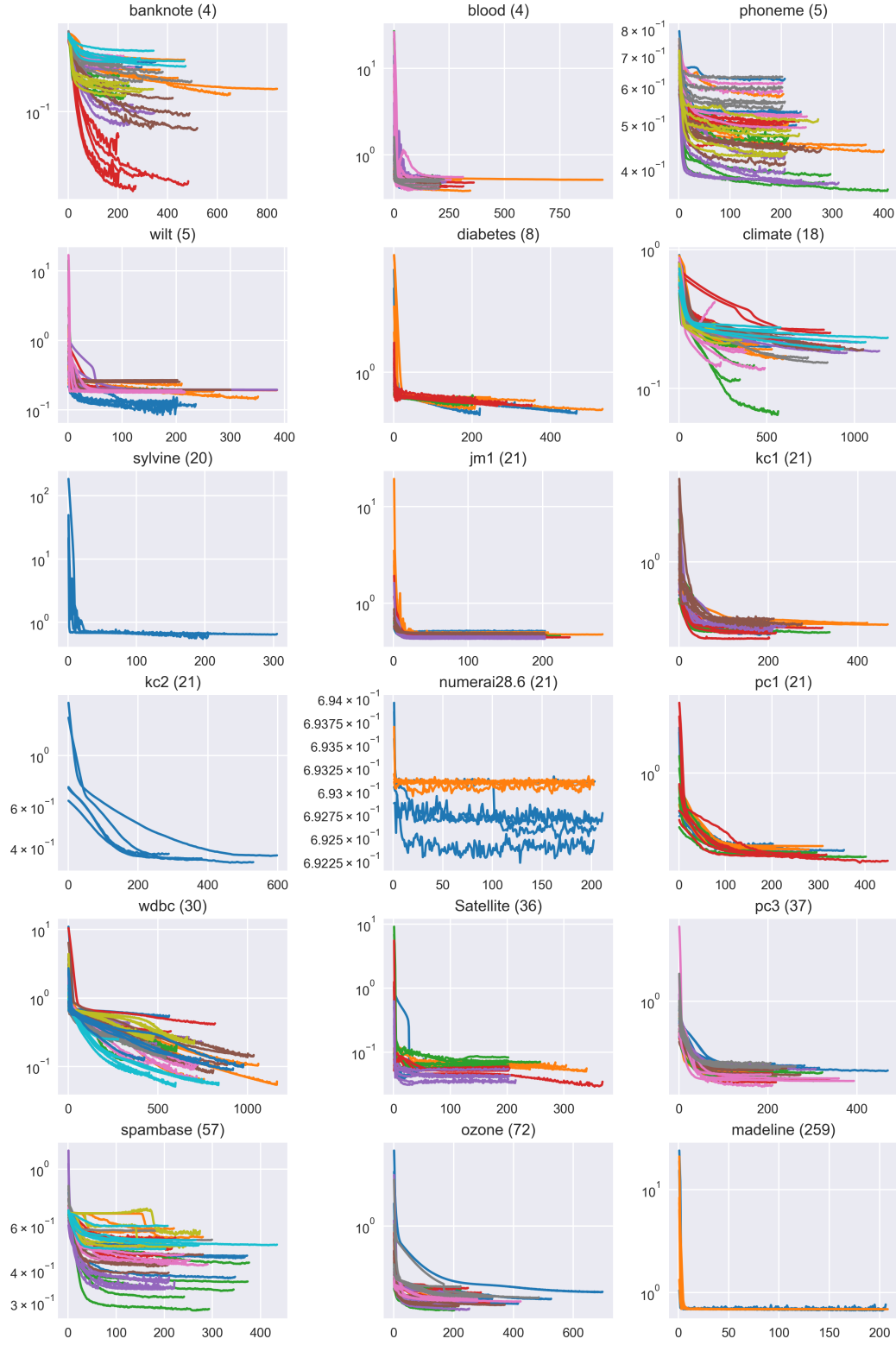


Figure 4: Pareto set loss of all datasets, evaluated on early stopping set. Same colours denote losses coming from folds of the same individual. x-axis portrait epochs, y-axis binary cross entropy loss.

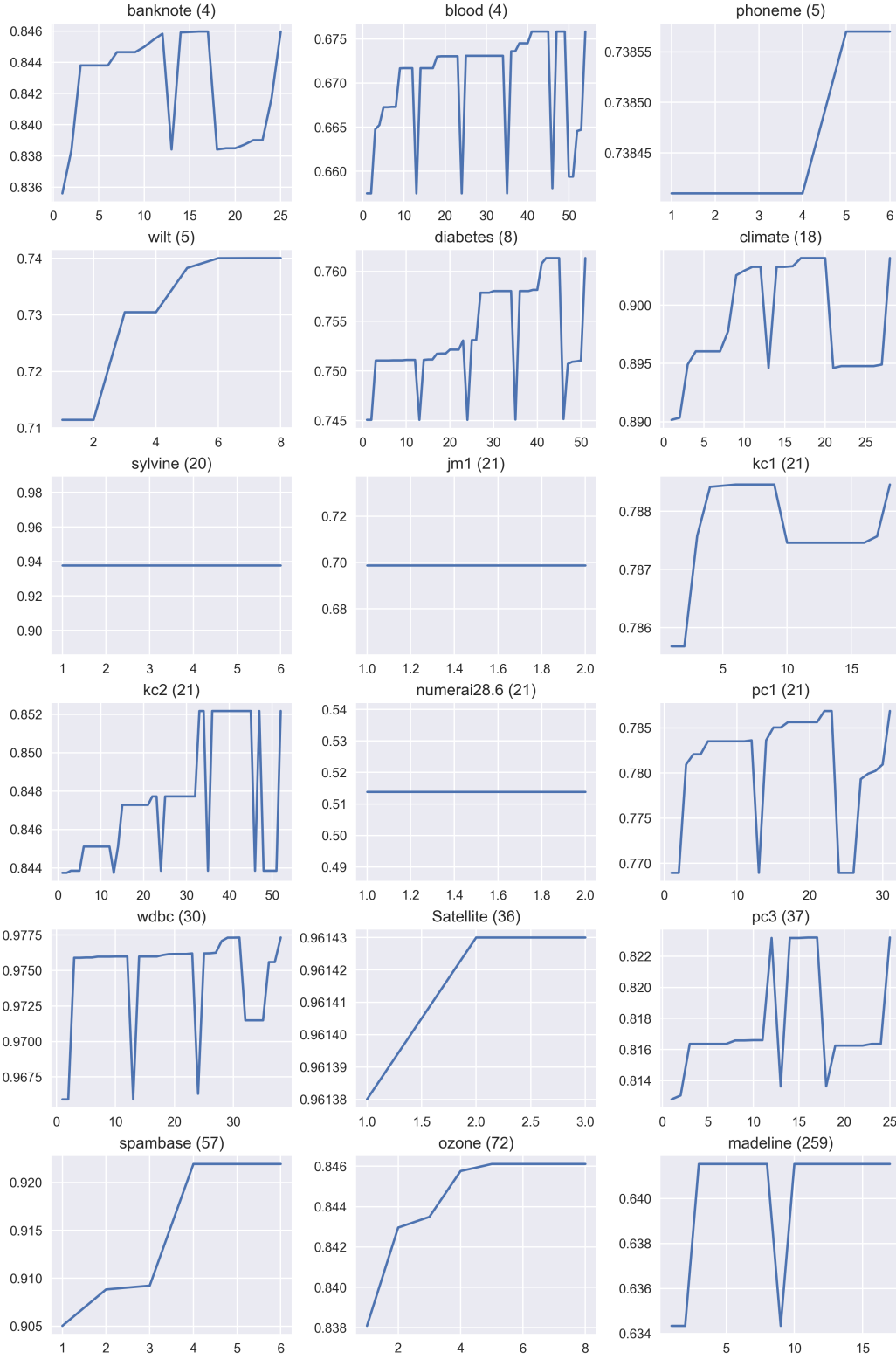


Figure 5: Dominated hypervolume over generations, evaluated on validation set. x-axis portrait generations, y-axis dominated hypervolume using mean AUC over folds.

- Pagel, Martin Zugnoni, JC, CrazyPython, Charles Newey, Antony Lee, pgajdos, Todd, Staffan Malmgren, redbug312, Orivej Desh, Nikolay Nechaev, Mike Boyle, Max Nordlund, MapleCCC, and Jack McCracken. tqdm: A fast, extensible progress bar for python and cli, November 2024. URL <https://doi.org/10.5281/zenodo.14231923>.
- K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, 2002. doi: 10.1109/4235.996017.
- Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv: Machine Learning*, 2017. URL <https://api.semanticscholar.org/CorpusID:11319376>.
- Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search. In Hutter et al. (2019), pages 69–86.
- M.T.M. Emmerich, K.C. Giannakoglou, and B. Naujoks. Single- and multiobjective evolutionary optimization assisted by gaussian random field metamodels. *IEEE Transactions on Evolutionary Computation*, 10(4):421–439, 2006. doi: 10.1109/TEVC.2005.859463.
- Matthias Feurer and Frank Hutter. Hyperparameter optimization. In Hutter et al. (2019), pages 3–38.
- Matthias Feurer, Jan N. van Rijn, Arlind Kadra, Pieter Gijsbers, Neeratyoy Mallik, Sahithya Ravi, Andreas Müller, Joaquin Vanschoren, and Frank Hutter. Openml-python: an extensible python api for openml. *Journal of Machine Learning Research*, 22(100):1–5, 2021. URL <http://jmlr.org/papers/v22/19-920.html>.
- William Finnoff, Ferdinand Hergert, and Hans Georg Zimmermann. Improving model selection by nonconvergent methods. *Neural Networks*, 6(6):771–783, 1993. ISSN 0893-6080. doi: [https://doi.org/10.1016/S0893-6080\(05\)80122-4](https://doi.org/10.1016/S0893-6080(05)80122-4). URL <https://www.sciencedirect.com/science/article/pii/S0893608005801224>.
- Peter I. Frazier. A tutorial on bayesian optimization, 2018. URL <https://arxiv.org/abs/1807.02811>.
- David E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., USA, 1st edition, 1989. ISBN 0201157675.
- Nikolaus Hansen. The cma evolution strategy: A tutorial, 2023. URL <https://arxiv.org/abs/1604.00772>.
- Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL <https://doi.org/10.1038/s41586-020-2649-2>.

- Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors, 2012. URL <https://arxiv.org/abs/1207.0580>.
- Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. ISSN 00401706. URL <http://www.jstor.org/stable/1267351>.
- Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren, editors. *Automatic Machine Learning: Methods, Systems, Challenges*. Springer, 2019.
- Arlind Kadra, Marius Lindauer, Frank Hutter, and Josif Grabocka. Well-tuned simple nets excel on tabular datasets. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 23928–23941. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/c902b497eb972281fb5b4e206db38ee6-Paper.pdf.
- Florian Karl, Tobias Pielok, Julia Moosbauer, Florian Pfisterer, Stefan Coors, Martin Binder, Lennart Schneider, Janek Thomas, Jakob Richter, Michel Lang, Eduardo C. Garrido-Merchán, Juergen Branke, and Bernd Bischl. Multi-objective hyperparameter optimization in machine learning—an overview. *ACM Trans. Evol. Learn. Optim.*, 3(4), December 2023. doi: 10.1145/3610536. URL <https://doi.org/10.1145/3610536>.
- J. Knowles. Parego: a hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems. *IEEE Transactions on Evolutionary Computation*, 10(1):50–66, 2006. doi: 10.1109/TEVC.2005.851274.
- Christoph Molnar. *Interpretable Machine Learning*. 2 edition, 2022. URL <https://christophm.github.io/interpretable-ml-book>.
- The pandas development team. pandas-dev/pandas: Pandas, January 2023. URL <https://doi.org/10.5281/zenodo.7549438>.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- Jefferson A. Riera, Ricardo M. Lima, and Omar M. Knio. A review of hydrogen production and supply chain modeling and optimization. *International Journal of Hydrogen Energy*, 48(37):13731–13755, 2023. ISSN 0360-3199. doi: <https://doi.org/10.1016/j.ijhydene.2022.12.242>. URL <https://www.sciencedirect.com/science/article/pii/S0360319922060505>.
- Shreyas Saxena and Jakob Verbeek. Convolutional neural fabrics, 2017. URL <https://arxiv.org/abs/1606.02492>.

- Lennart Schneider, Bernd Bischl, and Janek Thomas. Multi-objective optimization of performance and interpretability of tabular supervised machine learning models. In *Proceedings of the Genetic and Evolutionary Computation Conference, GECCO '23*, page 538–547, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701191. doi: 10.1145/3583131.3590380. URL <https://doi.org/10.1145/3583131.3590380>.
- Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando de Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016. doi: 10.1109/JPROC.2015.2494218.
- N. Srinivas and Kalyanmoy Deb. Multiobjective optimization using nondominated sorting in genetic algorithms. *Evol. Comput.*, 2(3):221–248, September 1994. ISSN 1063-6560. doi: 10.1162/evco.1994.2.3.221. URL <https://doi.org/10.1162/evco.1994.2.3.221>.
- Rainer Storn and Kenneth Price. Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11(4): 341–359, Dec 1997. ISSN 1573-2916. doi: 10.1023/A:1008202821328. URL <https://doi.org/10.1023/A:1008202821328>.
- Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 12 2018. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1996.tb02080.x. URL <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>.
- Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. Openml: networked science in machine learning. *SIGKDD Explor. Newsl.*, 15(2):49–60, June 2014. ISSN 1931-0145. doi: 10.1145/2641190.2641198. URL <https://doi.org/10.1145/2641190.2641198>.
- Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay May-
orov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.
- Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010. doi: 10.25080/Majora-92bf1922-00a.
- J. Zach. Interpretability of deep neural networks. 2019. URL <https://api.semanticscholar.org/CorpusID:198979458>.

Yu Zhang, Peter Tiño, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5): 726–742, 2021. doi: 10.1109/TETCI.2021.3100641.

Barret Zoph and Quoc V. Le. Neural architecture search with reinforcement learning, 2017. URL <https://arxiv.org/abs/1611.01578>.

Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.