

A greedy constructive algorithm for the optimization of neural network architectures

Massimiliano Lupo Pasini ¹, Junqi Yin ², Ying Wai Li ³, Markus Eisenbach ⁴

Abstract

In this work we propose a new method to optimize the architecture of an artificial neural network. The algorithm proposed, called Greedy Search for Neural Network Architecture, aims to minimize the complexity of the architecture search and the complexity of the final model selected without compromising the predictive performance. The reduction of the computational cost makes this approach appealing for two reasons. Firstly, there is a need from domain scientists to easily interpret predictions returned by a deep learning model and this tends to be cumbersome when neural networks have complex structures. Secondly, the use of neural networks is challenging in situations with compute/memory limitations. Promising numerical results show that our method is competitive against other hyperparameter optimization algorithms for attainable performance and computational cost. We also generalize the definition of adjusted score from linear regression models to neural networks. Numerical experiments are presented to show that the adjusted score can boost the greedy search to favor smaller architectures over larger ones without compromising the predictive performance.

This manuscript has been authored in part by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (<http://energy.gov/downloads/doe-public-access-plan>).

Introduction

Deep neural networks are nonlinear models used to approximate unknown functions based on observational data [27, 29, 33, 34] in deep learning (DL). Their broad applicability derives from a complex structure, which allows these techniques to reconstruct complex relations between quantities selected as inputs and outputs of the model [16]. From a mathematical perspective, a neural network is a directed acyclic graph where the nodes (also called neurons) are organized in layers. The type of connectivity between different layers is essential for the neural network to model complex dynamics between inputs and outputs. The structure of a graph is called architecture and is mainly summarized by the number of layers in the graph, the number of nodes at each layer and the connectivity between nodes of adjacent layers.

The performance of a neural network is very sensitive to the choice of the architecture for multiple reasons. Firstly, the architecture strongly impacts the prediction computed by a neural network. Indeed, neural networks with different structures can produce different outputs for the same input. On the one

¹ Oak Ridge National Laboratory, National Center for Computational Sciences, Oak Ridge, TN, USA, 37831, email: lupopasini@ornl.gov

² Oak Ridge National Laboratory, National Center for Computational Sciences, Oak Ridge, TN, USA, 37831, email: yinj@ornl.gov

³ Los Alamos National Laboratory, Computer, Computational and Statistical Sciences Division, Los Alamos, NM, USA, 87545, email: yingwaili@lanl.gov

⁴ Oak Ridge National Laboratory, National Center for Computational Sciences, Oak Ridge, TN, USA, 37831, email: eisenbachm@ornl.gov

hand, too simple structures may not be articulate enough to reproduce complex relations. This may result in underfitting the data with high bias and low variance in the predictions. On the other hand, too complex architectures may cause numerical artifacts such as overfitting, leading to predictions with low bias and high variance. Secondly, the topology of a neural network affects the computational complexity of the model. Indeed, an increase of layers and nodes leads to an increase of floating point operations to train the model and to make predictions. Therefore, identifying an appropriate architecture is an important step that can heavily impact the predictive power and the computational complexity of the model. However, the space of neural network architectures is too large for an exhaustive search. In fact, the number of architectures grows exponentially with the number of layers, the number of neurons per layer and the connections between layers. This motivated the study and development of optimization algorithms to automatize the selection of an appropriate architecture design.

Several approaches have been proposed in the literature for hyperparameter optimization [2, 3, 5, 9, 11, 12, 14, 25, 26, 35, 43]. Grid Search (GS), or parameter sweep, searches exhaustively through a specified subset of hyperparameters. The subset of hyperparameters and the bounds in the search space are specified manually. Moreover, the search for continuous hyperparameters requires a manually prescribed discretization policy. Although this technique is straightforwardly parallelizable, it becomes more and more prohibitive for computational time and resources when the number of hyperparameters increases. Therefore, attempts to reduce the number of model evaluations are preferred. Random Search (RS) [4] differs from GS mainly in that it explores hyperparameters randomly instead of exhaustively, since close points in the search space likely lead to similar models. Therefore, only randomly selected models are evaluated across different regions of the hyperparameter space. The major benefit resulting from this approach is a decreased processing time without compromising the performance attained. RS is likely to outperform GS, especially when only a small number of hyperparameters affects the final predictive power of DL model. However, the drawback of RS is unnecessarily high variance, as the method is entirely random. Moreover, a blind-folded approach for the selection of models to evaluate may lead to very expensive models to train and test. Sequential Model-Based Optimization (SMBO) algorithms [3] have been used in many applications where evaluation of the fitness function is expensive. An example of SMBO algorithms is Bayesian Optimization (BO) [35, 36], which rely on all the information available from previously evaluated models to guide the choice of models to evaluate in following steps. This generally reduces the actual number of neural networks built and trained. In addition, BO provides an assessment of uncertainty incorporating the effect of data scarcity. However, results are highly sensitive to the choice of the prior distribution on the hyperparameter space as well as the acquisition function to select points in the hyperparameter space. Another class of hyperparameter optimization methods is represented by genetic algorithms [8, 15, 19, 20, 21, 41] and evolutionary algorithms (EA) [42, 48], which evolve the topology of a neural network by alternatively adding or dropping nodes and connections based on results attained by previous neural networks. Earlier evaluated neural networks are treated as parents that generate new architectures treated as a generational offspring. However, genetic and evolutionary algorithms can suffer from the restricted areas explored in the hyperparameter space. Indeed, the small changes induced between architectures of successive generation can cause the algorithm to locally stagnate and not explore significant regions of the hyperparameter space.

All the approaches described above adopt powerful expedients to overcome theoretical and computational barriers [7, 18] in the search for an optimal neural network architecture under some optimality criterion. However, none of these methods fully exploits traditional statistical tools to perform on-line model diagnostics. Moreover, some of the aforementioned algorithms select neural networks with complex architectures. This can cause expensive computations that cannot be afforded in absence of large scale computers, as well as the results of complex models are difficult to interpret from the perspective of domain scientists.

In this work we propose a novel neural network architecture optimization with the goal to identify a neural network that attains a desired performance with the minimal structural complexity. We will refer to this method as *Greedy Search for Neural Network Architecture*. This approach minimizes the number

of hidden layers needed in the neural network, which would generally result in a simpler architecture. The algorithm iteratively enriches the architecture of the neural network by expanding the number of hidden layers in an adaptive fashion. At each iteration the number of hidden layers is fixed at a specific number. For a fixed number of hidden layers, RS is performed to identify the number of nodes per layer as well as for other hyperparameters. The information about selected values of hyperparameters per layer is transferred across the iterations, so that the new neural networks are built by recycling the hyperparameter selection already performed for previous hidden layers during earlier iterations. Adaptive algorithms for neural network architectures have already been studied in the literature [6, 22, 24] and share common features with adaptive methods for other types of regression models [10]. However, our method performs RS restricted to one hidden layer at each iteration, whereas other adaptive algorithms select the hyperparameters via gradient methods to minimize an objective function. Greedy Search for Neural Network Architecture has appealing properties in terms of algorithmic and computational scalability, because RS is confined to a fixed number of hyperparameters at each iteration. On the one hand, the recycling of information from previously evaluated models guarantees a fine level of *exploitation*. On the other hand, the random nature of the algorithm at each iteration still guarantees a thorough (albeit not exhaustive) *exploration* of the hyperparameter space to prevent stagnations at local minima. In order to favor computationally cheap models that are also easy to interpret, we introduce a definition of the *adjusted score* for neural networks that generalizes the definition of adjusted score for linear models [38]. The goal is to correct the score that measures the performance of a neural networks with information about the structural complexity of the DL model.

The paper is organized in five sections. Section 1 introduces the statistical background. Section 2 characterizes the adjusted score for neural networks. Section 3 explains our novel optimization algorithm for the architecture of neural networks. Section 4 presents numerical experiments where we compare the performance of our hyperparameter optimization algorithm with other approaches, as well as we study how generalization of adjusted score for neural networks can benefit the hyperparameter optimization algorithm select smaller architectures. Section 5 summarizes the results presented describes future directions to possibly pursue.

1 Statistical background

The goal of a regression or classification model is to approximate some unknown function f of the form

$$\mathbf{y} = f(\mathbf{x}), \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^p$, $\mathbf{y} \in \mathbb{R}^b$ and $f : \mathbb{R}^p \rightarrow \mathbb{R}^b$. This very general formulation incorporates situations where some components of the feature vector \mathbf{x} and the output vector \mathbf{y} may attain only discrete values. If the set of possible values for the output is finite, then the problem falls into a classification paradigm, whereas cases with infinite many possible values for the output are treated as regression problems. Statistical models aim to empirically reconstruct an approximation of f using a set of data points that correspond to specific input values of \mathbf{x} and related values of \mathbf{y} . The quantity \mathbf{x} is generally referred to as *input*, *predictor* or *regressor*, whereas \mathbf{y} is generally referred to as *output*, *response* or *target*.

Let us assume that we have a collection of n data points $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$. The data set used may contain inaccurate evaluations of f at specific values \mathbf{x} and the inaccuracy in the measurements may be due to various factors (e.g. human error in collecting data, inaccuracy of a measurement device). One way to statistically model the presence of errors in the data points is by adding a term to formula (1) as follows:

$$\mathbf{y}_i = f(\mathbf{x}_i) + \boldsymbol{\varepsilon}_i, \quad i = 1, \dots, n. \tag{2}$$

The term $\boldsymbol{\varepsilon}_i$ represents the error associated with the measurement of the i th data point.

In order to reconstruct an accurate approximation of f in Formula (1), several approaches are viable and they differ from each other on two aspects. Firstly, they differ for the assumptions made on the

complexity of the unknown f that they attempt to reconstruct. Secondly, they differ on the assumptions made about the measurement error. For the discussion in this paper, we focus on the former, as the latter plays a role only for statistical inference and uncertainty quantification which goes beyond the scope of this work.

1.1 Linear regression models

The simplest statistical approach that uses data samples to model an unknown function f is *linear regression*. Its use is mainly restricted to situations where the output variable \mathbf{y} is continuous. The goal is to identify a set of coefficients $\mathbf{w} \in \mathbb{R}^p$ to express a linear relation between \mathbf{x} and \mathbf{y} . If the data set made of n samples is collected experimentally, \mathbf{x} and \mathbf{y} are subject to measurement errors. Therefore, an experimentally guided linear relation between \mathbf{x} and \mathbf{y} is

$$\mathbf{y}_i^T = \mathbf{x}_i^T \mathbf{w} + \boldsymbol{\varepsilon}_i^T, \quad i = 1, \dots, n, \quad (3)$$

where $\mathbf{w} \in \mathbb{R}^{p \times b}$ is the tensor of regression weights and $\boldsymbol{\varepsilon}_i \in \mathbb{R}^b$ is a the tensor that models experimental errors in the measurement of \mathbf{y} . If we restructure the \mathbf{x}_i 's, \mathbf{y}_i 's and $\boldsymbol{\varepsilon}_i$'s over global quantities as follows:

$$Y = \begin{bmatrix} \mathbf{y}_1^T \\ \vdots \\ \mathbf{y}_n^T \end{bmatrix} \in \mathbb{R}^{n \times b}, \quad X = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} \in \mathbb{R}^{n \times p}, \quad E = \begin{bmatrix} \boldsymbol{\varepsilon}_1^T \\ \vdots \\ \boldsymbol{\varepsilon}_n^T \end{bmatrix} \in \mathbb{R}^{n \times b} \quad (4)$$

we obtain

$$Y = X\mathbf{w} + E. \quad (5)$$

The traditional way to select \mathbf{w} is by solving an ordinary least squares problem to minimize the discrepancy between predicted values and observations in (5). Therefore, denoting with $\|\cdot\|_2$ the ℓ_2 -norm in a Euclidean space, the coefficient vector \mathbf{w} is selected as follows:

$$\hat{\mathbf{w}} = \underset{\mathbf{w} \in \mathbb{R}^{p \times b}}{\operatorname{argmin}} \|Y - X\mathbf{w}\|_2. \quad (6)$$

A statistical analysis leading to interval estimates of $\hat{\mathbf{w}}$ requires assumptions on the nature of the errors $\boldsymbol{\varepsilon}_i$. However, such a discussion would go beyond the scope of this paper. Therefore we refer to [28] for further details in this respect.

We denote with $\bar{\mathbf{y}} \in \mathbb{R}^b$ the sample mean of the output variable over the data set

$$\bar{\mathbf{y}} = \frac{1}{n} \sum_{i=1}^n \mathbf{y}_i. \quad (7)$$

The prediction $\hat{\mathbf{y}}_i$ performed by the linear regression model for a data point \mathbf{x}_i is computed as

$$\hat{\mathbf{y}}_i^T = \mathbf{x}_i^T \hat{\mathbf{w}}. \quad (8)$$

If we denote with $\hat{Y} \in \mathbb{R}^{n \times b}$ the set of predictions for the set of inputs X , we have that

$$\hat{Y} = \begin{bmatrix} \hat{\mathbf{y}}_1^T \\ \vdots \\ \hat{\mathbf{y}}_n^T \end{bmatrix} \in \mathbb{R}^{n \times b}, \quad \hat{Y} = X\hat{\mathbf{w}}. \quad (9)$$

The performance of a regression model can be measured by monitoring the discrepancy between observed data Y and predicted values \hat{Y} . To this goal, we introduce new quantities to measure the discrepancy. The first quantity we introduce is the *sum of squared errors in the model*:

$$SSM = \sum_{i=1}^n \|\hat{\mathbf{y}}_i - \bar{\mathbf{y}}\|_2^2. \quad (10)$$

SSM expresses the variability of the data set that the statistical model is able to capture. The discrepancy between observations and predicted values reconstructed with the model is related to the portion of data variability that the model is not able to describe. Such a discrepancy is measured by the *sum of residual squared errors*:

$$SSR = \sum_{i=1}^n \|\mathbf{y}_i - \hat{\mathbf{y}}_i\|_2^2 \quad (11)$$

The *total sum of squared errors* captures the entire variability of the data set and is defined as

$$SST = \sum_{i=1}^n \|\mathbf{y}_i - \bar{\mathbf{y}}\|_2^2 \quad (12)$$

The following relation between SSM, SSR and SST holds for linear regression models:

$$SST = SSM + SSR, \quad (13)$$

which can be described as a statistical reinterpretation of the Pythagorean theorem. These quantities can be used to measure the efficiency of the statistical model in describing the relation between inputs and outputs. This leads to the definition of *coefficient of determination* or R^2 to measure the performance of a regression model:

$$R^2 = \frac{SSM}{SST} = 1 - \frac{SSR}{SST}, \quad 0 \leq R^2 \leq 1. \quad (14)$$

The R^2 attains values between 0 and 1 and the more the value attained by R^2 approaches 1, the higher is the predictive power of the model. However, an improvement in predictive power may require a significant increase in computational complexity. This would translate into increasing the value of p , that is the number of regressors used to explain the trend of the target \mathbf{y} . The complexity of the statistical model may result in relations between \mathbf{x} and \mathbf{y} that are difficult to interpret from the perspective of domain scientists. Therefore, it is recommendable to prefer simpler models over complex ones, especially if the gain in performance is negligible. To counterbalance the predictive power of a linear regression model with its complexity, a correction of the coefficient of determination was proposed in [38], named *adjusted coefficient of determination* or *adjusted- R^2*

$$R_{adj}^2 = 1 - (1 - R^2) \frac{n-1}{n-p} = 1 - \frac{SSR}{SST} \left(\frac{n-1}{n-p} \right). \quad (15)$$

The corrective term $\frac{n-1}{n-p}$ penalizes complex architectures over simpler ones and allows the following relation between standard and adjusted coefficient of determination:

$$R_{adj}^2 \leq R^2 \leq 1. \quad (16)$$

1.2 Nonlinear regression models

Although linear regression conveniently relates monitored quantities in an easy way to communicate and understand, sometimes linear models are too simplistic and cannot capture complex dynamics. This shortcoming justified the introduction of nonlinear regression models in the literature.

1.2.1 Logistic regression for classification

A renown nonlinear regression model is the *logistic regression*, which is used for classification problems. In this section we focus on a situation where there are only two categories, labeled with 0 and 1. In this

case, the output y_i is one-dimensional binary quantity and it is treated as a Bernoulli random variable with expected value π_i , that is

$$y_i \sim Be(\pi_i), \quad \pi_i = P(y_i = 1), \quad E(y_i) = \pi_i. \quad (17)$$

It also follows that $P(y_i = 0) = 1 - \pi_i$. For a binary classification problem, the linear regression model in Equation (3) is modified as follows:

$$\pi_i = g(\mathbf{x}_i^T \mathbf{w}) + \varepsilon_i, \quad i = 1, \dots, n, \quad (18)$$

where $\mathbf{x}_i \in \mathbb{R}^p$, $\mathbf{w} \in \mathbb{R}^p$, $\pi_i \in [0, 1]$ and $\varepsilon_i \in \mathbb{R}$ still represents a possible error in the measurement of the response. The nonlinear function g is chosen as follows:

$$g(s) = \frac{\exp(s)}{1 + \exp(s)}. \quad (19)$$

The nonlinear regression model is transformed into a linear regression model through an auxiliary variable called linear predictor η_i that is defined by the transformation

$$\eta_i = g^{-1}(\pi_i) = \ln \frac{\pi_i}{1 - \pi_i}. \quad (20)$$

The transformation is often called *logit transformation* of the probability π_i , and the ratio $\frac{\pi_i}{(1 - \pi_i)}$ in the transformation is called the *odds*. This mean that the relation between observations y_i and regression coefficients \mathbf{w} is nonlinear

$$\pi_i = g(\eta_i) = \frac{\exp(\mathbf{x}_i^T \mathbf{w})}{1 + \exp(\mathbf{x}_i^T \mathbf{w})} + \varepsilon_i. \quad (21)$$

In this scope, the goal is to compute a regression coefficient vector $\hat{\mathbf{w}}$ that would allow the logistic model to accurately predict the binary outcome y for any given set of features \mathbf{x} . Due to the nonlinear relation between inputs and outputs, computing the regression coefficient vector $\hat{\mathbf{w}}$ leads to an optimization problem with a solution that does not have closed analytic form. We refer to [28, Chapter 13 - Section 13.2] for more details about the optimization algorithm to compute $\hat{\mathbf{w}}$ in linear regression, other models for binary response data and generalizations of the classification problem from binary to multi-class data.

Several metrics can be used to measure the performance of a classification model. We describe some of them in the following. The *accuracy* (*ACC*) of a classification model quantifies the percentage of data points that are correctly labeled and it is defined as:

$$ACC = \frac{\sum_{i=1}^n \mathcal{I}(\hat{y}_i = y_i)}{n} = \frac{\text{true positives} + \text{true negatives}}{\text{positives} + \text{negatives}}. \quad (22)$$

However, the accuracy is not recommended to measure the performance if the classes are represented through unbalanced data. To account for unbalanced classes, other metrics are usually preferred. The *precision* or *positive predicted value* (*PPV*) is

$$PPV = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \quad (23)$$

and the *sensitivity*, *recall*, *hit rate*, or *true positive rate* (*TPR*) is

$$TPR = \frac{\text{true positives}}{\text{positives}}. \quad (24)$$

PPV and *TPR* can be combined to better describe the performance of a classification model in case of unbalanced class representations. To this goal, the *F1 score* is defined as the harmonic mean between *PPV* and *TPR*:

$$F1 = 2 \frac{PPV \cdot TPR}{PPV + TPR}, \quad 0 \leq F1 \leq 1. \quad (25)$$

As for the coefficient of determination for regression problems, also in the case of classification problems one can introduce a definition of the F1 score that penalizes the model with respect to the number of predictors. We refer to this quantity as the *adjusted-F1 score*:

$$F1_{adj} = 1 - (1 - F1) \left(\frac{n-1}{n-p} \right), \quad F1_{adj} < F1. \quad (26)$$

Similarly to linear regression, also logistic regression has strong modeling limitations that prevent it from efficiently describing complex input-output relations. Therefore, the discussion of this paper mainly focuses on neural networks, a more versatile class of nonlinear regression models that can be used both for regression and classification. The linear regression model and the logistic regression described in Section 1.1 and 1.2.1 are particular cases of neural networks, as we show in the next section.

1.2.2 Dense feedforward neural networks (multilayer perceptron)

A *deep feedforward network*, also called *feedforward neural network*, or *multilayer perceptron* (MLP) [11, 17] is a predictive statistical model to approximate some unknown function f as in (1). In particular, feedforward neural networks compose together many different functions such as

$$\hat{f}(\mathbf{x}) = f_L(\cdots f_{\ell+1}(f_\ell(f_{\ell-1}(\cdots f_0(\mathbf{x}))))), \quad (27)$$

where $\hat{f} : \mathbb{R}^p \rightarrow \mathbb{R}^b$, $f_0 : \mathbb{R}^a \rightarrow \mathbb{R}^{p_1}$, $f_L : \mathbb{R}^{p_L} \rightarrow \mathbb{R}^b$ and $f_\ell : \mathbb{R}^{p_\ell} \rightarrow \mathbb{R}^{p_{\ell+1}}$ for $\ell = 1, \dots, L-1$. The goal is to identify the proper number ℓ so that the composition in Equation (27) resembles the unknown function f in (1). The composition in Equation (27) is modeled via a directed acyclic graph describing how the functions are composed together. The number L that quantifies the complexity of the composition is equal to the number of hidden layers in the neural network. We refer to the input layer as the layer with index $\ell = 0$. The indexing for hidden layers of the deep neural networks starts with $\ell = 1$. In this section we consider a neural network with a total of L hidden layers. The symbol p_ℓ is used to denote the number of neurons at the ℓ th hidden layer. Therefore, p_0 coincides with the dimensionality of the input, that is $p_0 = p$. The very last layer with index $L+1$ represents the output layer, meaning that $p_{L+1} = b$ coincides with the dimensionality of the output. We refer to $\mathbf{w} \in \mathbb{R}^{N_{tot}}$ as the total number of regression coefficients. Following this notation, the function f_0 corresponds to the first layer of the neural network, f_1 is the second layer (first hidden layer) up to f_L that represents the last layer (output layer). In other words, deep feedforward networks are nonlinear regression models and the nonlinearity is given by the composition in Equation (27) to describe the relation between predictors \mathbf{x} and targets \mathbf{y} .

This approach can be reinterpreted as searching for a mapping that minimizes the discrepancy between values $\hat{\mathbf{y}}$ predicted by the model and given observations \mathbf{y} . The statistical model is described by a set of parameters that we represent as $\mathbf{w} \in \mathbb{R}^{N_{tot}}$. Given a dataset with m data points, the process of predicting the outputs for given inputs via a feedforward neural network can thus be formulated as

$$\mathbf{y} = F(\mathbf{x}, \mathbf{w}) + \boldsymbol{\varepsilon}, \quad (28)$$

where the operator $F : \mathbb{R}^{p_0} \times \mathbb{R}^{N_{tot}} \rightarrow \mathbb{R}^b$ is

$$F(\mathbf{x}, \mathbf{w}) = \varphi_{L+1} \left(\sum_{k_{L+1}} w_{k_{L+1}k_L} \varphi_L \left(\sum_{k_L} w_{k_Lk_{L-1}} \varphi_{L-1} \left(\cdots \varphi_1 \left(\sum_{i=1} w_{k_1i} x_i \right) \right) \right) \right) \quad (29)$$

and the vector $\boldsymbol{\varepsilon} \in \mathbb{R}^b$ is used to model measurements errors. Using the matrix notation for the weights connecting adjacent layers as

$$W_{\ell,\ell-1} \in \mathbb{R}^{p_\ell \times p_{\ell-1}} \quad (30)$$

we can rewrite (29) as

$$F(\mathbf{x}, \mathbf{w}) = \varphi_{L+1} \left(W_{L+1,L} \left(\varphi_L \left(\dots \left(\varphi_1 \left(W_{1,0} \mathbf{x} \right) \right) \right) \right) \right). \quad (31)$$

The notation in (31) highlights that N_{tot} is the total number of regression weights used by the neural network. This value must account for all the entries in $W_{\ell,\ell-1}$'s matrices, that is

$$N_{tot} = \sum_{\ell=1}^{L+1} p_{\ell} p_{\ell-1}. \quad (32)$$

If the target values are continuous quantities, the very last layer is usually chosen to be linear. Therefore, φ_{L+1} is picked as the identity function. If the target values are categorical, then φ_{L+1} is usually set to be the logit function in Formula (19). If the number of hidden layers is set to $L = 0$ and φ_1 is set to be the identity function, then the statistical model becomes a classical linear regression model. If the number of hidden layers is set to $L = 0$ and φ_1 is set to be the logit function, then the statistical model becomes a logistic regression model.

Also for nonlinear models the weight matrices $W_{\ell,\ell-1}$'s are computed to minimize the discrepancy between predictions and observations. If we reorganize \mathbf{x}_i 's, \mathbf{y}_i 's and $\boldsymbol{\varepsilon}_i$'s over global quantities as follows:

$$Y = \begin{bmatrix} \mathbf{y}_1^T \\ \vdots \\ \mathbf{y}_n^T \end{bmatrix} \in \mathbb{R}^{n \times b}, \quad X = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} \in \mathbb{R}^{n \times p_0}, \quad E = \begin{bmatrix} \boldsymbol{\varepsilon}_1^T \\ \vdots \\ \boldsymbol{\varepsilon}_n^T \end{bmatrix} \in \mathbb{R}^{n \times b} \quad (33)$$

we obtain

$$Y = \hat{F}(X, \mathbf{w}) + E, \quad \hat{F} : \mathbb{R}^{n \times p_0} \times \mathbb{R}^{N_{tot}} \rightarrow \mathbb{R}^{n \times b}. \quad (34)$$

Given a specific configuration for the coefficients $\hat{\mathbf{w}} \in \mathbb{R}^{N_{tot}}$, we can refer to the prediction $\hat{\mathbf{y}}_i \in \mathbb{R}^b$ associated with the \mathbf{x}_i as follows

$$\hat{\mathbf{y}}_i = F(\mathbf{x}_i, \hat{\mathbf{w}}). \quad (35)$$

Therefore, the set of predictions $\hat{Y} \in \mathbb{R}^{n \times b}$ generated by the neural networks on the set of inputs X is

$$\hat{Y} = \begin{bmatrix} \hat{\mathbf{y}}_1^T \\ \vdots \\ \hat{\mathbf{y}}_n^T \end{bmatrix}, \quad \hat{Y} = \hat{F}(X, \hat{\mathbf{w}}). \quad (36)$$

The computation of $\hat{\mathbf{w}}$ requires solving a nonlinear optimization problem to minimize an objective function that measures the discrepancy between the set of observations Y and the set of predictions \hat{Y} . The choice of the objective function to minimize depends on the nature of the observations. In the case of regression problems, the most common objective functions are the *mean squared error* (MSE) or the *mean absolute error* (MAE). For classification problems, the *Kullback-Leibler divergence* is the most common choice. We do not provide more details on these herein as this discussion would go beyond the aim of this work. We refer to [16, Chapter 3 - Sections 3.3, 3.4, 3.5, 4.16, 10.5] for further details. As for linear regression models and logistic regression, it is also important for neural networks to define quantities that can measure the predictive power of the model. As we have discussed, the coefficient of determination and adjusted coefficient of determination are useful indicators that measure the performance of a linear regression model and logistic regression. The generalization and use of the former to neural networks is straightforward. We highlight the fact that the coefficient of determination for nonlinear regression models is not necessarily nonnegative as it was for linear regression models. Indeed, the coefficient of determination compares the fit of the chosen model with that of a horizontal straight line. If the chosen model fits worse than a horizontal line, then the coefficient of determination is negative. This situation cannot happen for linear regression

models by definition, but it can occur for nonlinear regression models. So the coefficient of determination can attain negative values without violating any mathematical rules. The coefficient of determination is negative only when the chosen model does not follow the trend of the data, so it fits worse than a horizontal line. The definitions of SST, SSM, SSR, ACC, PPV and TPR are still valid for neural networks. However, the relation (13) between SST, SSR and SSM does not hold for nonlinear regression models.

As for the adjusted coefficient of determination, its generalization to DL requires some insight in the architecture and computational complexity of the neural network.

2 Adjusted coefficient of determination and adjusted $F1$ score for neural networks

Replicating the discussion for linear models and logistic regression, we propose a generalization of adjusted- R^2 and adjusted- $F1$ that combine metrics for the performance with metrics for the complexity of the neural network. More specifically, we consider the number of nodes per layers and the total number of hidden layers in a neural network as parameters that describe the complexity of the architecture. Consider a dataset with n sample points and an MLP with L hidden layers and p_ℓ neurons at the ℓ th hidden layer for $\ell = 1, \dots, L$. The generalization of R_{adj}^2 to neural networks that we propose is

$$R_{adj}^2 = 1 - (1 - R^2) \left[\frac{n - 1}{n - \max_{\ell=0, \dots, L} p_\ell} \right] \left[\frac{n - 1}{n - (L + 1)} \right]. \quad (37)$$

Analogously, the generalization of $F1_{adj}$ to neural networks that we propose is

$$F1_{adj} = 1 - (1 - F1) \left[\frac{n - 1}{n - \max_{\ell=0, \dots, L} p_\ell} \right] \left[\frac{n - 1}{n - (L + 1)} \right]. \quad (38)$$

As the absence of hidden layers in a neural networks reduces it to either a linear regression or a logistic regression model, the generalization of R_{adj}^2 and $F1_{adj}$ proposed above should account for this behavior. In fact, their definitions reduce to the standard definitions of R_{adj}^2 and $F1_{adj}$ for linear regression and logistic regression when $L = 0$. The relations

$$|R_{adj}^2| < |R^2| \quad (39)$$

and

$$F1_{adj} < F1 \quad (40)$$

still hold. The adjusted- R^2 and adjusted- $F1$ penalize neural networks with complex structure. Indeed, if two neural networks attain the same predictive performance and one has less neurons per layers and/or less hidden layers than the other, the simpler neural networks is awarded with a higher R_{adj}^2 or $F1_{adj}$. The definitions of R_{adj}^2 and $F1_{adj}$ that we just introduced can thus guide hyperparameter optimization algorithms to avoid complex structures in neural networks that can cause overfitting.

3 Adaptive selection of the number of hidden layers

The search of a proper architecture for a neural network is a challenging task and hyperparameter optimization algorithms are still object of study. In this section we describe a novel approach to perform architecture optimization of neural networks, called *Greedy Search for Neural Network Architecture*. A pseudocode that describes the procedure is presented in Algorithm 1. The method relies on discriminating the hyperparameters of a neural network architecture based on their statistical interpretation. In particular, hidden layers model complex features to reconstruct articulate relations between input and output that a simpler model such as linear regression would not be able to reproduce. It is thus reasonable to

think that more complex relations between inputs and output would require more hidden layers. A neuron in a hidden layer can be interpreted as an artificial regressor that relates different levels of nonlinearity to each other. In fact, neurons of a hidden layer are topologically connected with neurons residing on adjacent layers. The number of neurons needed to accurately reconstruct the nonlinearity may vary from layer to layer. It is thus possible that the neural network may have to alternatively expand and contract across the hidden layers to properly model the nonlinear relations between inputs and outputs.

A hyperparameter optimization algorithm that treats number of hidden layers and number of nodes per layer as two generic hyperparameters usually builds complex neural networks at intermediate steps to explore the hyperparameter space. However, building complex models is not recommendable unless strictly necessary, especially if such models are only an intermediate step and are later discarded to favor other models with a better predictive performance. Therefore, it may be beneficial to minimize the construction of complex architectures to only those situations where it is worthwhile, that is when an increase in complexity leads to a significant improvement in the predictions. To this goal, we exploit the interpretation of hidden layers explained above to build a greedy algorithm that optimizes the architecture of neural networks. The user needs to provide the maximum number of iterations (maximum number of hidden layers) and the range that must be spanned for any other hyperparameter to optimize. The approach we propose is greedy with respect to the number of hidden layers required for a prescribed performance, whereas it performs a stratified RS on the remaining hyperparameters to determine. Although RS can simultaneously optimize multiple hyperparameters, discussion in this section is limited to the hyperparameters that impact the computational complexity of the neural network. Therefore, we consider the case when RS is performed only over the number of neurons per layer. The method starts performing RS over neural networks with one hidden layer and it selects the neural network that attains the highest validation score. If the validation score of the selected neural network satisfies the performance requirements, such a neural networks is returned to the user and the algorithm stops. Otherwise, the number of hidden layers is increased by one, new neural networks with two hidden layers are built and a new RS takes place. The number of neurons at the first hidden layer is not object of optimization, as the structure of first hidden layers is recycled from the first iteration. Therefore, the neural networks built at the second iteration have the same number of neurons in the first hidden layers, whereas the number of neurons in the second hidden layer randomly changes across the neural networks due to the RS performed. The same rationale is applied to successive iterations until either the desired performance is obtained or the maximum number of hidden layers is reached. This means that the neural networks evaluated at each iteration are built by prolonging the best neural network of the previous iteration with an extra hidden layer at the end. Therefore, RS performed at each iteration only involves the number of neurons at the last hidden layers, since the structure of the previous hidden layers is inherited from the previous iterations. More specifically, the structure of the previous hidden layers is shared between all the neural networks evaluated at a given iteration. An illustration that explains how Greedy Search for Neural Network Architecture proceeds is shown in Figure 1. The idea of recycling the structure of neural networks across successive iterations was already explored in previous works [30, 44]. However, we need to highlight that our approach only transfers information about the architecture, not the value of the regression weights. Opposed to what is done in transfer learning [31, 32], our approach trains neural networks from scratch at each iteration by recomputing the regression weights across the entire architecture.

Incremental approaches like the one proposed in this paper are opposed to other regularizing techniques, where a complex structure is simplified via pruning the graph with a deletion of edges. Some of these techniques are pruning [39, 45] or random dropout [37]. Constructive or incremental approaches [40] have some advantages over pruning approaches. First, for constructive algorithms, it is straightforward to specify an initial network, whereas for pruning algorithms, one does not know in practice how big the initial network should be. Second, constructive algorithms always search for small network solutions first. They are thus more computationally convenient than pruning algorithms, in which the majority of the training time is spent on networks larger than necessary. Third, as many networks with different sizes may be capable of implementing acceptable solutions, constructive algorithms are likely to find smaller

network solutions than pruning algorithms. Smaller networks are more computationally efficient and can be described by simpler rules. Moreover, by searching for small networks, the amount of training data required for good generalization may be reduced. Some constructive algorithms also have hurdles that they struggle to overcome. For instance, constructive greedy algorithms may be suboptimal in some cases. This is due to the gradient descent techniques employed to identify the proper number of units per per hidden layer. However, the random search performed in our new approach should guarantee enough exhaustive exploration of the hyperparameter space to overcome this problem. Moreover, the localized random search performed at each iteration guarantees a certain level of parallelization, since the neural networks evaluated at each step can be trained concurrently.

Algorithm 1 Greedy Search for Neural Network Architecture

1: **Require:**

- L = maximum number of hidden layers
- N_{max_nodes} = maximum number of nodes (neurons) per layer
- $score_{threshold}$ = threshold on the final performance prescribed
- $model_eval_iter$ = number of model evaluations per iteration

2: Set number of hidden layers $\ell = 1$

3: Set `best_model` as linear regression (for regression problems) or logistic regression (for classification problems)

4: Compute `score`

5: **while** `score` < $score_{threshold}$ & $\ell \leq L$ **do**

6: Build `model_eval_iter` neural networks with ℓ hidden layers each:

- Set number of nodes and activation functions for first $(\ell - 1)$ hidden layers as in `best_model`
- Perform random search for number of nodes in the last hidden layer and for the remaining hyperparameters

7: Select `best_model` as the neural network with best performance

8: Retrieve `best_model` and store info about number of nodes and activation functions per layer

9: $\ell = \ell + 1$

return `best_model`

3.1 Reduction of dimensionality in the hyperparameter search

The information transferred from smaller to bigger neural networks across successive iterations reduces the dimension of the hyperparameter space to explore. In this section we compare the cardinality (number of elements in a set) of the hyperparameter space explored by a standard hyperparameter optimization algorithm (e.g. GS, RS, SMBO, EA) with the cardinality of the hyperparameter space explored by Greedy Search for Neural Network Architecture. For the sake of simplicity, we consider only the number of hidden layers and the number of neurons per layer as hyperparameters to tune. Denote with $[1, N_{max_nodes}]$ the range of neurons per hidden layer and denote with L the maximum number of hidden layers. Denote with $\mathcal{S}_{Standard}$ the hyperparameter space explored by a standard hyperparameter optimization algorithm and denote with $\#(\cdot)$ the cardinality of a discrete set. For a standard hyperparameter optimization algorithm, the total number of architectures contained in $\mathcal{S}_{Standard}$ increases exponentially with the number of hidden layers, that is

$$\#(\mathcal{S}_{Standard}) = N_{max_nodes}^L. \quad (41)$$

The exponential increase of the cardinality of $\mathcal{S}_{Standard}$ RS with respect to the number of hidden layers is due to the *curse of dimensionality*. However, Greedy Search for Neural Network Architecture performs a

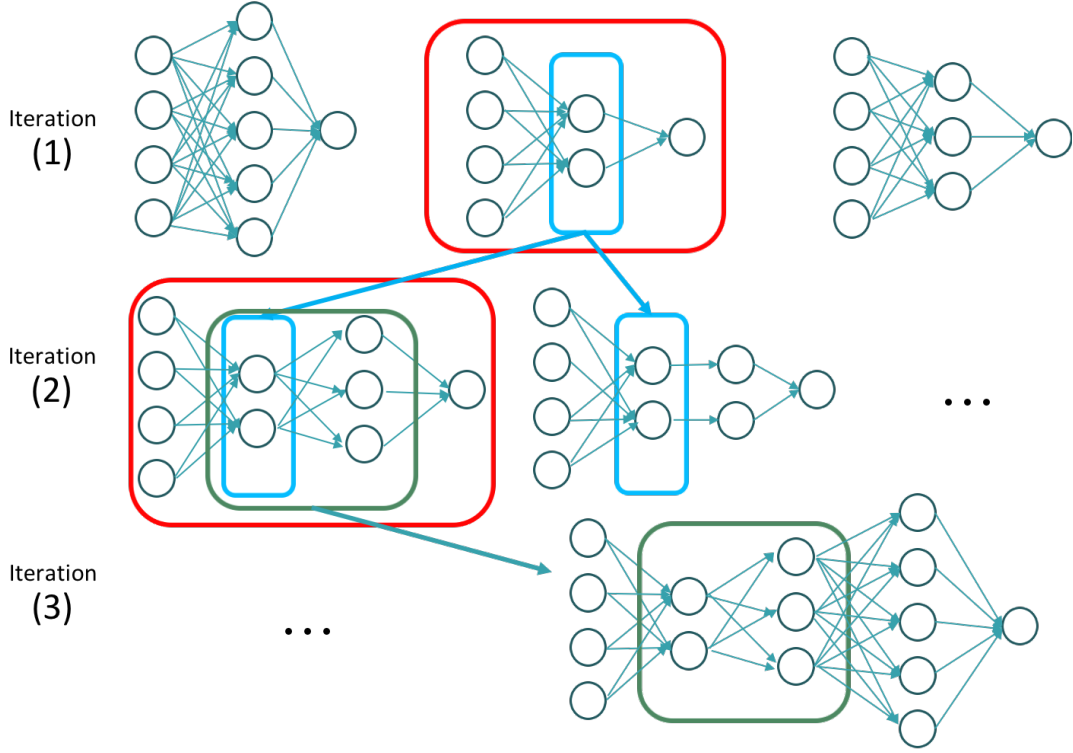


Figure 1: Illustration of Greedy Search for Neural Network Architecture. The illustration explains how the architecture of the neural network is enriched at each iteration. The neural networks built at iteration (1) have only one hidden layer and the number of neurons inside the hidden layers is chosen via RS. Every neural network is trained and the validation score is measured. The neural network with the highest validation score is selected (circled in red). If the validation score meets the desired performance, the algorithm stops and returns the selected neural network. Otherwise, the number of neurons contained in the first hidden layer of the selected neural network is transferred to iteration (2). The neural networks built at iteration (2) have the same number of neurons at the first hidden layers as the best neural network from iteration (1), whereas the number of neurons at the second hidden layer is chosen randomly with another stratified RS. The neural networks are trained and the validation scores from each neural network are collected. The neural network with the highest validation score is selected (circled in red). If the validation score meets the desired performance, the algorithm stops and returns the selected neural network. Otherwise, the information about the numbers of neurons at the first and second hidden layer are transferred to iteration (3), so that another stratified RS takes place on the number of neurons inside the third hidden layer.

RS on a single layer (the last hidden layer) at each iteration. This leads to a significant reduction of the cardinality of the hyperparameter space. In fact, we have

$$\#(\mathcal{S}_{\text{Greedy}}) = N_{\text{max_nodes}}. \quad (42)$$

The reduction of the cardinality allows the greedy search to efficiently explore $\mathcal{S}_{\text{Greedy}}$.

4 Numerical results

In this section we restrict our experiments to fully connected neural networks. The extension to other types of architectures is going to be considered in future works. We first describe some numerical experiments that compare the Greedy Search for Neural Network Architecture with state of the art hyperparameter optimization algorithms. We also present some numerical experiments that validate the effectiveness of the generalization of adjusted- R^2 and adjusted- $F1$ to neural networks and how this can benefit the selection of smaller architecture in the hyperparameter search.

4.1 Dataset description

We describe the data sets used for numerical experiments as follows. For each data set employed, we specify the source or how it has been generated, the number of features, the nature of the target (regression or classification) and the size of the dataset.

Artificially generated data

- **Eggbox**

Type of problem: regression.

Description: each point of the dataset is made of two-dimensional coordinates and the scalar value that the function has at those coordinates.

Input features: (x, y) coordinates

Target: eggbox function - $f(x, y) = [2 + \cos(x/2) * \cos(y/2)]^5$

Number of data points: 4,000

Datasets from UCI - Machine Learning repository [52]

- **Computer hardware** [1, 52]

Type of problem: regression.

Description: relative CPU performance data, described in terms of its cycle time, memory size, etc.

Input features:

1. vendor name: 30 (adviser, amdahl,apollo, basf, bti, burroughs, c.r.d, cambex, cdc, dec, dg, formation, four-phase, gould, honeywell, hp, ibm, ipl, magnuson, microdata, nas, ncr, nixdorf, perkin-elmer, prime, siemens, sperry, sratus, wang)
2. model name: many unique symbols
3. MYCT: machine cycle time in nanoseconds (integer)
4. MMIN: minimum main memory in kilobytes (integer)
5. MMAX: maximum main memory in kilobytes (integer)
6. CACH: cache memory in kilobytes (integer)
7. CHMIN: minimum channels in units (integer)
8. CHMAX: maximum channels in units (integer)
9. PRP: published relative performance (integer)

Target: estimated relative performance (integer) from the original article [1]. The relative performance metrics are taken from the Published Relative Performance benchmarks (PRP), from the influential BYTE magazine, for 209 FDA Approved CPUs active on the market today. See [1] for more details on how the relative performance values were set.

Number of data points: 209

- **Phishing websites** [52]

Type of problem: classification.

Description: this dataset collected mainly from: PhishTank archive, MillerSmiles archive, Google's searching operators.

Input features:

1. using the IP Address
2. long URL to hide the suspicious part
3. using URL shortening services "TinyURL"
4. URL's having "@" symbol
5. redirecting using "//"
6. adding prefix or suffix separated by (-) to the domain
7. sub domain and multi sub domains
8. HTTPS (hyper text transfer protocol with secure sockets layer)
9. domain registration length
10. favicon
11. using non-standard port
12. the existence of "HTTPS" token in the domain part of the URL
13. request URL
14. URL of anchor
15. links in <Meta>, <Script> and <Link> tags
16. server form handler (SFH)
17. submitting information to email
18. abnormal URL
19. website forwarding
20. status bar customization
21. disabling right click
22. using pop-up window
23. IFrame redirection
24. age of domain
25. website traffic
26. PageRank
27. Google index
28. number of links pointing to page
29. statistical-reports based feature

Target: Binary variable, phishing (Yes = 1, No = -1)

Number of data points: 11,055

Dataset from Kaggle [46]

- **Graduate admission**

Type of problem: regression.

<https://www.kaggle.com/mohansacharya/graduate-admissions>

Input features: GRE Score, TOEFL Score, University Rating, SOP, LOR, CGPA, Research (treated as a binary variable)

Target: chance of admit (real value between 0 and 1)

Number of data points: 400

4.2 Definition of the hyperparameter space

The hyperparameter search is performed over the number of hidden layers, the number of neurons per layer, the type of nonlinear activation function at each hidden layer and the batch size used to train the model with a first-order optimization algorithm. A discrete set of values is chosen to bound the hyperparameter search. The number of hidden layers spans from 1 to 5 and the discrete range of the number of neurons per layer spans from 1 to \sqrt{n} , where n is the number of sample points inside the dataset. The set of activation functions is made of the sigmoid function, the hyperbolic tangent and the rectified linear unit function. The discrete range for the batch size spans from 10 to the closest integer to $\frac{n}{10}$. These ranges for the hyperparameters is fixed for every hyperparameter optimization algorithm used for the study.

4.3 Details about model evaluations

The datasets are split in three components: the training set, the validation set and the test set. The training set is used to train every model, the validation set is used to select the best performing model at each iteration and the test set is used at the end to measure the predictive power of the neural network selected by each hyperparameter optimization algorithm. The test set is 10% of the original dataset, the remaining portion is partitioned into training and validation in the percentage of 90% and 10% respectively. For classification problems, a stratified splitting is performed to ensure that the proportion between classes is preserved across training, validation and test sets. The optimizer used to train the model is the Adam method [13] with an initial learning rate of 0.001.

4.4 Definition of the hyperparameter search algorithms

The code is implemented in `python` and the neural networks are built using `Keras.io` [47]. We compare the Greedy Search for Neural Network Architecture described in this paper with the Tree-structured Parzen estimator (TPE) and Bayesian optimization (BO). The version of Greedy Search for Neural Network Architecture that we implemented performs concurrent model evaluations for the RS at each step with a distributed memory parallelization paradigm that uses `mpi4py` [49]. The version of TPE and BO used are provided by the `Ray Tune` library [50] through the routines named `HyperOptSearch` and `BayesOptSearch` respectively. The version of Ray Tune used is 0.3.1. As to `BayesOptSearch`, the utility function is defined by setting `utility_kwargs="kind": 'ucb', "kappa": 2.5, "xi": 0.0)`. For both `HyperOptSearch` and `BayesOptSearch`, the model selection and evaluations are scheduled using the asynchronous version of HyperBand [23] called `AsyncHyperBandScheduler`. The time attribute for the scheduler is the training iteration and the reward attribute is the validation score of the neural network. The validation score is also used as reward attribute for the stopping criterion of the hyperparameter optimization algorithm.

4.5 Hardware description

The numerical experiments are performed using Summit [51], a supercomputer provided by the Oak Ridge Leadership Computing Facility (OLCF) at Oak Ridge National Laboratory. Summit has a hybrid architecture, and each node contains two IBM POWER9 CPUs and six NVIDIA Volta GPUs all connected together with NVIDIA’s high-speed NVLink. Each node has over half a terabyte of coherent memory (high bandwidth memory + DDR4) addressable by all CPUs and GPUs plus 800GB of non-volatile RAM that can be used as a burst buffer or as extended memory. To provide a high rate of I/O throughput, the nodes are connected in a non-blocking fat-tree using a dual-rail Mellanox EDR InfiniBand interconnect.

4.6 Comparison for predictive performance and computational time

The first set of numerical experiments compares the predictive power of the Greedy Search for Neural Networks Architecture with TPE and BO. The attributes used for the model selection are the R^2 validation score for regression problems and the $F1$ validation score for the classification problem. The number of concurrent model evaluations per iteration is set to 10, 50 and 100 for all the three hyperparameter optimization algorithms. The maximum number of iterations is set to 5 and the threshold imposed on the validation score as stopping criterion is 0.99. The test score of the neural network identified as best and the computational time in wall clock seconds are reported for each hyperparameter search algorithm in Figure 2. The results presented are averaged over 10 runs with 95% confidence intervals for the mean value.

The experiments with the **Eggbox** dataset exhibit better results for the greedy search with respect to TPE and BO in terms of predictive power achieved. Moreover, we notice that the confidence band for the greedy search narrows as the number of concurrent evaluations increases. This happens because the inference on the attainable predictive performance becomes more accurate with a higher number of random samples for the stratified RS. A different trend is shown for the confidence band of TPE and BO. In this case, the confidence band does not narrow down by increasing the number of concurrent model evaluations. This highlights the benefit of using a stratified RS as performed by the Greedy Search for Neural Network Architecture, which limits the uncertainty of the random optimization by reducing the dimensionality of the search space. With regards to the **Graduate admission** dataset, the performance obtained with the three hyperparameter optimization algorithms is very similar. However, the performance in terms of computational time still shows that Greedy Search for Neural Network Architecture outperforms TPE and BO. As to the **Computer hardware** dataset, results in terms of performance still display an improvement using the greedy approach with respect to TPE and BO. The confidence band for the greedy search exhibits a similar trend to what experienced for the **Eggbox** dataset, as the confidence band narrows by increasing the number of concurrent model evaluations. Moreover, the computational time spent to perform the greedy search is significantly less than the one required for the other hyperparameter optimization algorithms. The three hyperparameter optimization algorithms reach a similar performance in terms of attainable performance for the **Phishing websites** dataset. However, the greedy search displays a better scaling for an increasing number of concurrent model evaluations. The scaling of Greedy Search for Neural Network Architecture proves that the stratified RS limits the computational cost of each iterations. Moreover, the parallelizability of the stratified RS enables to keep the computational time almost constant with respect to the number of model evaluations.

4.7 Sensitivity of greedy search to the number of concurrent model evaluations

In Figure 3 we show the performance obtained with the greedy search on the **Eggbox** dataset and the **Computer hardware** dataset as a function of the number of hidden layers for different numbers of concurrent model evaluations (10, 50 and 100). For both the experiments it is clear that the use of a small number

of concurrent model evaluations leads to significant fluctuations in the score, as the stratified RS does not explore enough architectures for a fixed number of hidden layers. However, a progressive increase of the concurrent model evaluations leads to a better inference. This happens because an exhaustive exploration of the stratified hyperparameter space reduces the uncertainty in the attainable best performance of the model. Moreover, a sufficient exploration of the stratified hyperparameter space enables to highlight the dependence between the maximum attainable performance of the neural network and the total number of hidden layers. Indeed, the examples displayed in Figure 3 confirm that nonlinear input-output relations can benefit from a higher number of hidden layers.

4.8 Efficacy of adjusted score in penalizing complex architectures

In this section we consider the `Graduate admission` dataset and we compare the value attained by the R^2 and the adjusted- R^2 across the iterations of the Greedy Search for Neural Network Architecture. The score threshold is not used for early stopping in this case and all experiments are run for a total of five iterations, meaning that the neural networks built at the end of the iterative process have 5 hidden layers. Figure 4 shows that the increase of hidden layers through successive iterations does not benefit the R^2 which remains almost constant. Because the performance of the model does not improve with more hidden layers, the adjusted- R^2 decreases when the number of hidden layers is increased. Indeed, the adjusted- R^2 penalizes larger neural networks over smaller neural networks when the performance is the same. Moreover, the selection of smaller neural networks with the adjusted- R^2 does not compromise the performance of the selected model, as shown by the regular test score that does not significantly change. Therefore, the adjusted score caps the dimensionality of the parameter space in the hyperparameter selection so that the complexity of the predictive model does not exceed what is justified by the amount of data available.

4.9 Use of adjusted score with Greedy Search for Neural Network Architecture

In this section we describe a set of experiments that compare the total number of parameters for the neural networks selected by the Greedy Search for Neural Network Architectures in two different cases: one uses the validation score as a criterion to select the model, whereas the other uses the validation adjusted score. We aim to prove that the adjusted score can help the hyperparameter search to favor simpler models over more complex ones by reducing the number of neurons per hidden layers (and therefore the model parameters). The result of the experiments are shown in Table 1 as an average over 10 runs. The number of concurrent model evaluations per iteration is set to 100 and the code is forced to perform the maximum number of iterations, so that the performance isolates the impact of the adjusted score over the hyperparameter selection. The use of the adjusted- R^2 does not lead to significantly smaller architectures for the `Graduate admission` and the `Computer hardware` datasets because the number of data samples is small. Therefore, the architecture selected is relatively small regardless of whether the greedy search is guided by the regular score or by the adjusted score. However, different performances are noticed for the `Eggbox` dataset and for the `Phishing websites` dataset. In these cases, a larger number of data samples allows the hyperparameter search to explore a wider set of architectures and the combination of the greedy search with the adjusted score benefits the reduction in complexity of the model selected. Indeed, the total number of regression weights selected with the adjusted score is significantly lower. This means that the neural networks selected with the adjusted score have less neurons per hidden layers than the ones selected by the traditional definition of score.

	Greedy Search + Score		Greedy Search + Adjusted Score		
	Score	Nb. parameters	Score	Adjusted Score	Nb. parameters
Eggbox	0.993	6,321	0.995	0.996	4,070
Graduate admission	0.831	1,250	0.845	0.803	1,115
Computer hardware	0.923	890	0.917	0.705	802
Phishing websites	0.920	15,749	0.916	0.905	5,803

Table 1: Comparison between performance and complexity of model selected by Greedy Search for Neural Network Architecture when the attribute used for the model selection is the standard definition of the score and when the adjusted score. Results are averaged over 10 runs.

5 Remarks and future developments

In this works we have presented a new greedy constructive algorithm for the selection of neural network architectures called Greedy Search for Neural Network Architecture. The method aims to identify the simplest architecture that obtains the desired performance. The algorithm adopts a greedy technique on the number of hidden layers, which can benefit the reduction of computational time and cost to perform the hyperparameter search. This makes the algorithm appealing to perform DL when computational and memory resources are limited. Moreover, small architectures, which obtain the desired performance, facilitate the interpretation of results when the outcome of the model is used by domain scientists to conduct analyses in their field of expertise. The recycling of hidden layer configurations disregards an exponential number of architectures in the hyperparameter space. However, having a smaller search space makes the optimization a tractable problem. Moreover, experimental results show that this does not result in significance loss in the attainable accuracy of the model search. We also generalized the definition of adjusted score to neural networks to penalize architectures according to their complexity. Numerical experiments on fully connected neural networks are presented, where Greedy Search for Neural Network Architecture outperforms Tree-structured Parzen Estimator and Bayesian Optimization to identify the best obtainable performance on datasets of different nature. The use of the adjusted score exhibits promising results in selecting small architectures that attain the desired predictive power.

For future developments we aim to extend the study to different types of architectures other than multilayer perceptrons, such as convolutional neural networks. Moreover, we are going to use Greedy Search for Neural Networks Architectures for specific problems by selecting customized attributes other than the score for the hyperparameter optimization. We are also going to conduct an uncertainty quantification analysis to estimate the sensitivity of the inference on the hyperparameters with respect to the dimension of the hyperparameter space and the number of concurrent model evaluations.

Acknowledgements

This work is supported in part by the Office of Science of the US Department of Energy (DOE) and by the LDRD Program of Oak Ridge National Laboratory. This work used resources of the Oak Ridge Leadership Computing Facility (OLCF), which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725.

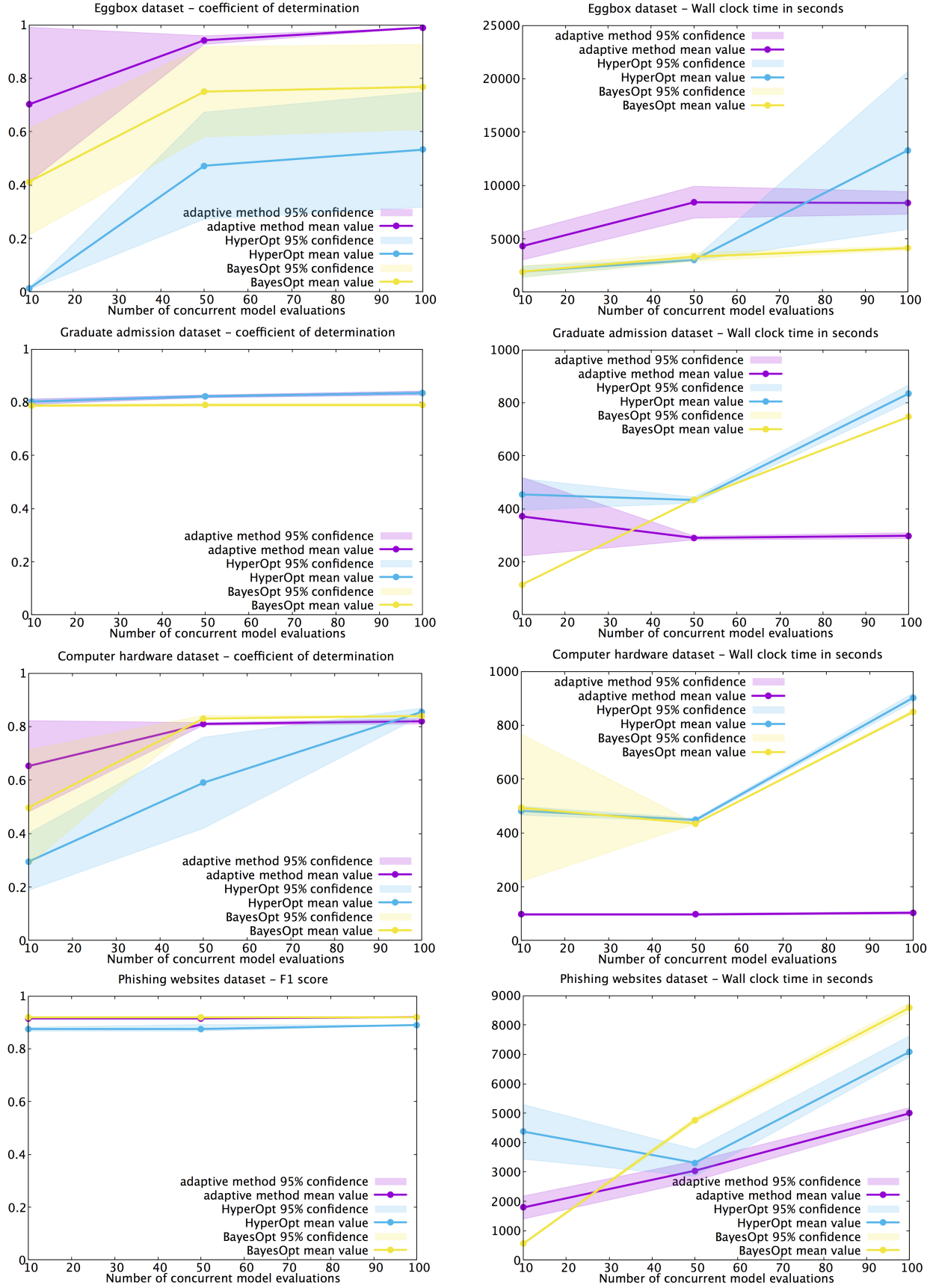


Figure 2: Comparison between Greedy search, HyperOptSearch and BayesOptSearch. The comparison is performed for four datasets: the **Eggbox** dataset (first from the top), the **Graduate admission** dataset (second from the top), the **Computer hardware** dataset (third from the top) and the phishing dataset (last). The graphs on the left show the performance obtained by the model selected by the hyperparameter search on the test set. The graphs on the right shows a comparison for computational time.

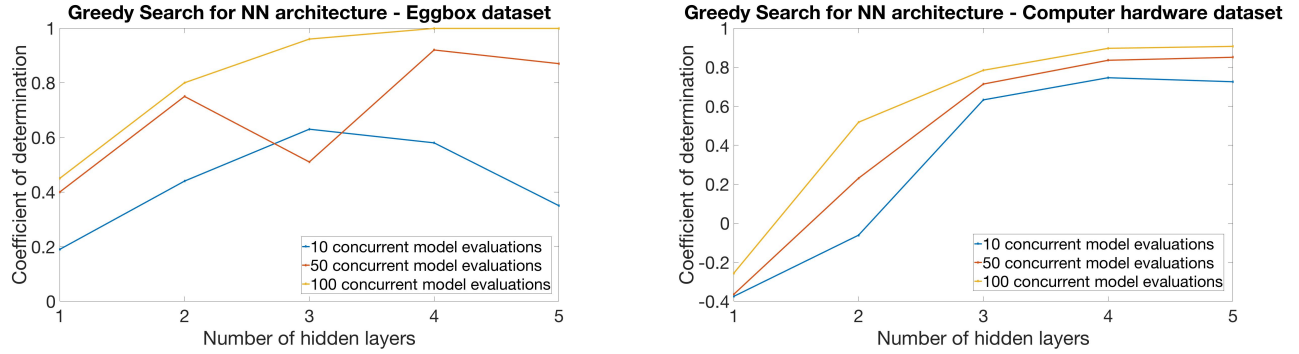


Figure 3: Greedy Search for Neural Network Architecture. Coefficient of determination expressed in terms of the number of hidden layers for **Eggbox** and **Computer hardware** datasets using 10, 50 and 100 concurrent model evaluations. Results are shown for a single run.

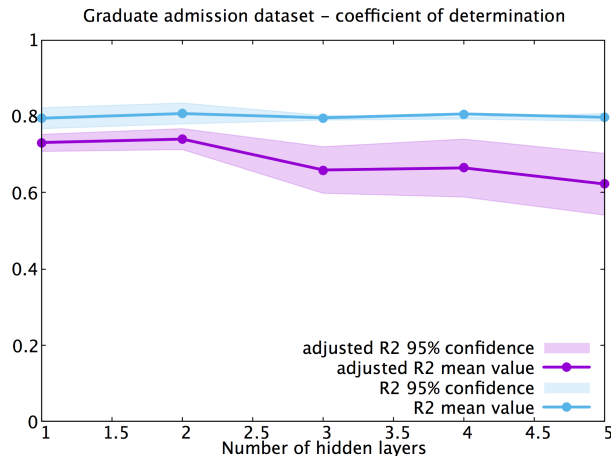


Figure 4: **Graduate admission** dataset. Comparison between R^2 and adjusted- R^2 for Greedy Search for Neural Network Architecture as a function of the number of hidden layers.

References

- [1] D. W. AHA, D. F. KIBLER, AND M. K. ALBERT, *Instance-based prediction of real-valued attributes*, Computational Intelligence, Volume 5, p. 51, 1989.
- [2] B. BAKER, O. GUPTA, N. NAHIK AND R. RASKAR, *Designing neural network architectures using performance prediction*, 2018 International Conference on Learning Representations, Workshop Track, 2018.
- [3] J. BERGSTRA, R. BARDENET, Y. BENGIO AND B. KÉGL, *Algorithms for hyper-parameter optimization*, Proceeding NIPS'11 Proceedings of the 24th International Conference on Neural Information Processing Systems, pp. 2546–2554, 2011.
- [4] J. BERGSTRA AND Y. BENGIO, *Random Search for hyper-parameter optimization*, Journal of Machine Learning Research, Vol. 13, pp. 281–305, 2012.
- [5] H. CAI, T. CHEN, W. ZHANG, Y. YU AND J. WANG, *Reinforcement learning for architecture search by network transformation*, arXiv:1707.04873, 2017.
- [6] C. CORTES, X. GONZALVO, V. KUZNETSOV, M. MOHRI AND S. YANG, *AdaNet: adaptive structural learning of Artificial Neural Networks* Proceedings of the 34th International Conference on Machine Learning, PMLR 70:874–883, 2017.
- [7] T. DOMHAN, J. T. SPRINGENBERG AND F. HUTTER, *Speeding up automatic hyperparameter optimization of deep neural networks by extrapolation of learning curves*, IJCAI'15 Proceedings of the 24th International Conference on Artificial Intelligence, pp. 3460–3468, 2016.
- [8] M. ETTOUIL, M. LAZAAR, AND Y. GHANOU, *Architecture optimization model for the multilayer perceptron and clustering*, Journal of Theoretical and Applied Information Technology, Volume 10, Volume 47, Number 1, 2013.
- [9] S. FAHLMAN AND C. LEBIERE, *The cascade-correlation learning architecture*, Advances in neural information processing system, pp. 524–532, 1990.
- [10] J. .H. FRIEDMAN, *Multivariate Adaptive Regression Splines*, Annals of Statistics, Vol. 19, No. 1, pp. 1–67, 1991.
- [11] I. GOODFELLOW, Y. BENGIO AND A. COURVILLE, *Deep Learning*, The MIT Press, Cambridge, Massachusetts and London, England, 2016.
- [12] W. GRATHWOHL, E. CREAGER, S. K. S. GHASEMPOUR AND R. ZEMEL, *Gradient-based optimization of neural network architecture*, ICLR 2018 Workshop Track, 2018.
- [13] D. P. KINGMA, AND J. L. BA, *Adam: a method for stochastic optimization*, Conference Paper at International Conference on Learning Representations 2015.
- [14] T. KUMAR GUPTA AND K. RAZA, *Optimizing deep neural network architecture: a tabu search based approach*, arXiv:1808.05979, 2018
- [15] T.K. GUPTA AND K. RAZA, *Optimizaiton of ANN architecture: A review on nature-inspired techniques*, Machine learning in Bio-signal and Diagnostic Imaging, Elsevier, 2018.
- [16] S. HAYKIN, *Neural Networks And Learning Machines, Third Edition*, Pearson Education Ltd, 2009.

- [17] S. HAYKIN, *Neural Networks and Learning Machines, Third Edition*, Pearson Education Inc., Upper Saddle River, New Jersey, 2009.
- [18] T. HINZ, N. NAVARRO-GUERRERO, S. MAGG AND S. WERMTER, *Speeding up the hyperparameter optimization of deep convolutional neural networks*, arXiv:1807.07362, 2018.
- [19] J. HOLLAND, *Genetic Algorithms, for the Science*, Scientific American Edition, Number 179, pp. 44–50, 1992.
- [20] H. KITANO, *Designing neural networks using genetic algorithms with graph generation system*, Complex Systems Journal, Vol. 4, pp. 461–476, 1990.
- [21] P. KOEHN, *Combining Genetic Algorithms and Neural Networks: The Encoding Problem*, Master of Science Thesis, University of Knoxville, Tennessee, USA, 2001.
- [22] T. Y. KWOK, AND D. Y. YEUNG, *Constructive algorithms for structure learning in feedforward neural networks for regression problems*, IEEE Transactions on Neural Networks, Volume 8, Issue 3, pp. 630–645, 1997.
- [23] L. LI, K. JAMIESON, G. DESALVO, A. ROSTAMIZADEH, AND A. TALWALKAR, *Hyperband: bandit-based configuration evaluation for hyperparameter optimization*, Published as a conference paper at ICLR 2017.
- [24] D. LIU, T. S. CHANG, AND Y. ZHANG, *A constructive algorithm for feedforward neural networks with incremental training*, IEEE Transactions on Circuits and Systems—I: Fundamental Theory and Applications, Volume 49, Number 12, 2002.
- [25] C. LIU, B. ZOPH, J. SHLEN, W. HUA, L. LI, L. FEI-FEI, A. YUILLE, J. HUANG AND K. MURPHY, *Progressive neural architecture search*, arXiv:1712.00559, 2017.
- [26] R. LUO, F. TIAN, T. QIN, E. CHEN AND T. LIU, *Neural architecture optimization*, arXiv:1808.07233, 2018.
- [27] M. L. MINSKY, *Some universal elements for finite automata*, In C. E Shannon & J. McCarthy (Eds.), Automata studies, Princeton: Princeton University Press, pp. 117–128, 1956.
- [28] D. C. MONTGOMERY, E. A. PECKS AND G. G. VINING, *Introduction to Linear Regression Analysis*, Wiley Series in Probability and Statistics, 2012.
- [29] J. VON NEUMANN, *The general and logical theory of automata*, In L. A. Jeffress (Ed.), Cerebral mechanisms in behavior, New York: Wiley, pp. 1–41, 1951.
- [30] H. PHAM, M. Y. GUAN, B. ZOPH, Q. V. LE AND J. DEAN, *Efficient neural architecture search via parameter sharing*, arXiv:1802.03268, 2018.
- [31] L. Y. PRATT, *Discriminability-based transfer between neural networks*, NIPS Conference: Advances in Neural Information Processing Systems 5, Morgan Kaufmann Publishers, pp. 204–211, 1993.
- [32] L. Y. PRATT, *Special Issue: Reuse of Neural Networks through Transfer*, Connection Science, Volume 8, Issue 2, 1996.
- [33] F. ROSENBLATT, *The perceptron: a probabilistic model for information storage and organization in the brain*, Psychological Review, Volume 65, Number 6, 1958.
- [34] F. ROSENBLATT, *The perceptron: a theory of statistical separability in cognitive systems*, Buffalo: Cornell Aeronautical Laboratory, Inc. Report Number VG-1196-G-1, 1958.

- [35] J. SNOEK, H. LAROCHELLE AND R. P. ADAMS, *Practical Bayesian optimization of Machine Learning algorithms*, Proceeding NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2, pp. 2951–2959, 2012.
- [36] J. SNOEK, O. RIPPEL, K. SWERSKY, R. KIROS, N. SATISH, N. SUNDARAM, M. M. A. PATWARY, PRABHAT AND R. P. ADAMS, *Scalable Bayesian optimization using deep neural networks*, <https://arxiv.org/pdf/1502.05700.pdf>
- [37] N. SRIVASTAVA, G. HINTON, A. KRIZHEVSKY, I. SUTSKEVER, AND R. SALAKHUTDINOV, *Dropout: a simple way to prevent neural networks from overfitting*, Journal of Machine Learning Research, Volume 15, pp. 1929–1958, 2014.
- [38] A. K. SRIVASTAVA, V. K. SRIVASTAVA AND AMAN ULLAH, *The coefficient of determination and its adjusted version in linear regression models*, Econometric Reviews, Volume 14, No. 2, 1995.
- [39] S. W. STEPNIIEWSKI AND A. J. KEANE, *Pruning backpropagation networks using modern stochastic optimization techniques*, Neural Computing and Applications, Vol. 5, No. 2, pp. 76–98, 1997.
- [40] N. K. TREADGOLD, AND T. D. GEDEON, *Exploring Constructive Cascade Networks*, IEEE Transactions on Neural Networks, Volume 10, Number 6, 1999.
- [41] J. T. TSAI, J. H. CHOU AND T. K. LIU, *Tuning the structure and parameters of a neural network by using hybrid Taguchi-genetic algorithm*, IEEE Trans. Neural Networks, Vol. 17, No. 1, pp. 69–80, 2006.
- [42] S. R. YOUNG, D. C. ROSE, T. P. KARNOWSKI, S. LIM AND R. M. PATTON, *Optimizing deep learning hyper-parameters through an evolutionary algorithm*, MLHPC '15 Proceedings of the Workshop on Machine Learning in High-Performance Computing Environments, Article No. 4, 2015.
- [43] B. ZOPH AND Q. V. LE, *Neural architecture with reinforcement learning*, arXiv:1611.01578, 2016.
- [44] B. ZOPH, V. VASUDEVAN, J. SHLENS AND A. V. LE, *Learning transferable architectures for scalable image recognition*, Proceedings of the IEEE conference on computer vision and pattern recognition, 2018.
- [45] *Compressing and regularizing deep neural networks*, <https://www.oreilly.com/ideas/compressing-and-regularizing-deep-neural-networks>
- [46] *Kaggle: Your Home for Data Science*, <https://www.kaggle.com>
- [47] *Keras: The Python Deep Learning library*, <https://keras.io>
- [48] *Multi-node Evolutionary Neural Networks for Deep Learning (MENNDL)*, <https://www.ornl.gov/division/csmd/projects/multi-node-evolutionary-neural-networks-deep-learning-menndl>, Oak Ridge National Laboratory.
- [49] *MPI for Python*, <https://mpi4py.readthedocs.io/en/stable/>
- [50] *Ray Tune: Hyperparameter Optimization Framework*, <https://ray.readthedocs.io/en/ray-0.3.1/tune.html>
- [51] *Summit - Oak Ridge National Laboratory's 200 petaflop supercomputer*, <https://www.olcf.ornl.gov/olcf-resources/compute-systems/summit/>
- [52] *UCI Machine Learning Repository*, <https://archive.ics.uci.edu/ml/index.php>