Supplementary material for: A machine learning-based method for the prediction of pseudo-realistic pediatric abdominal phantoms for radiation dose reconstruction

Marco Virgolin[a,*], **Ziyuan Wang**[b], **Tanja Alderliesten**[b], **Peter A. N. Bosman**[a,c]
[a]Life Sciences and Health Group, Centrum Wiskunde & Informatica, Science Park 123, Amsterdam, the Netherlands
[b]Department of Radiation Oncology, Amsterdam UMC, University of Amsterdam, Meibergdreef 9, Amsterdam, the Netherlands
[c]Algorithmics Group, Delft University of Technology, Mekelweg 5, Delft, the Netherlands

A. Hyper-parameters and tuning

Table 1 shows the hyper-parameters used by the Machine Learning (ML) algorithms. For LARS and LASSO we optimize λ to penalize complex models. Note that both algorithms perform squared-error minimization (along with penalization handling) via cyclical coordinate descent[1]. However, the hyper-parameter tuning grid-search cross-validation procedure we use is aligned to MAE minimization, and this procedure is wrapped around the squared residual-based optimization.

For Random Forest (RF), we use the default, relatively large number of trees (given the datasets at hand) of 500 as advised by literature.[1] We optimize how many features are randomly chosen when building the nodes that compose each decision tree ("mtry"), and the minimum number of data samples a node should represent ("min. node size"), the same way we optimize λ for LARS and LASSO.

For GP-Trad and GP-GOMEA, the function set $\mathcal{F}$ defines which functions to use as model components (tree nodes). The division operator $\div_{\text{A}}$ is the analytic quotient,[2] which not only guarantees that the divider can never be null, but also ensures smoothness (in contrast with the protected division operator,[3] which can harm generalization[2]). The logarithm operator is protected to avoid infeasible computations.[4] The Ephemeral Random Constants (ERC) are constants for which the value is set by uniform sampling from a defined interval.[3] Mathematical expressions are encoded as parse trees in GP-Trad and GP-GOMEA. We set a small tree height to keep the resulting mathematical expressions short and readable (we found that larger three heights can result in hard to read expressions), and to prevent overfitting.

The number of candidate expressions to evolve, i.e., the population size, is a sensitive parameter for GP algorithms. We run GP-Trad and GP-GOMEA using the Interleaved Multistart Scheme (IMS), a method that interleaves multiple runs with increasing population size. We set the number of sub-iterations between runs, g_{IMS}, to 4 as has been reported to work well on benchmark problems.[5,6] Since the IMS can in principle run forever, we set a time limit of 60 seconds. We found this limit to be reasonable because the datasets are small and evaluations are fast, and because the other ML algorithms take only a few seconds to execute. We also preliminarily observed that increasing the time limit (e.g., to 5 or 10 minutes) does not alter the results in a significant way. For further details on GP-Trad, GP-GOMEA, the IMS, and other hyper-parameters, the reader is referred to the seminal paper on GP-GOMEA for regression.[6]

[1] https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html

Table 1 Hyper-parameters of the ML algorithms. The subscript "tune" means that the hyper-parameter setting is subject to optimization with 5-fold cross-validation grid-search among the listed values.

Algorithm	Hyper-parameter	Settings
LARS and LASSO	λ_{tune}	$10^{-10}, 10^{-9}, \ldots, 10^{10}$
RF	nr. trees	500
	min. node size$_{\text{tune}}$	$5, 10, \ldots, 20, 25$
	mtry$_{\text{tune}}$	$1, 2, \ldots, \frac{\#\text{ features}}{2}$
GP-Trad and GP-GOMEA	$\mathcal{F}$	$\{+, -, \times, \div_{\text{A}}, \exp, \log_{\text{P}}\}$
	ERC	$\mathbb{U}[-10, 10]$
	tree height	2
	g_{IMS}	4
	time limit	60s

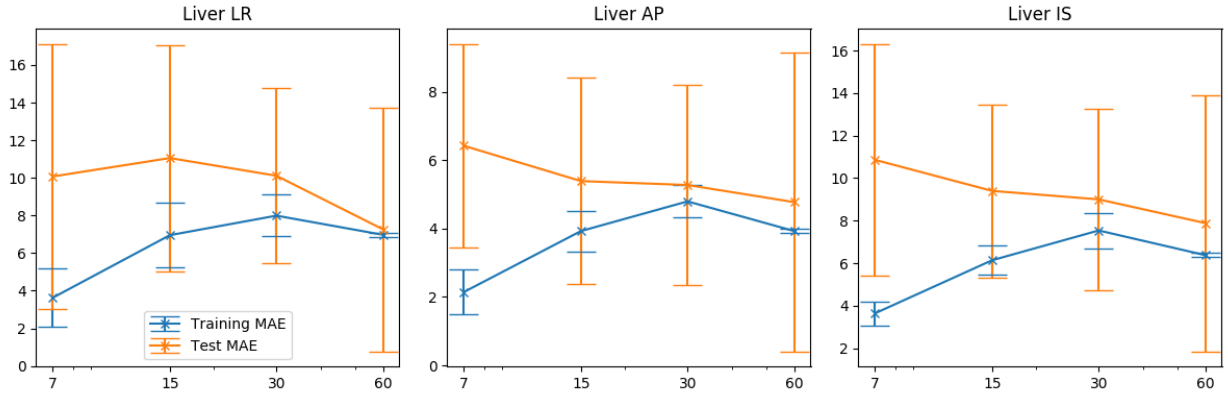

Fig 1 Effect of increasing the database size on learning the liver position using GP-GOMEA. Center points represent the mean, and error bars represent the standard deviation, of the MAEs obtained from the average (across 10 repetitions) leave-one-out cross-validation. The error bars represent the standard deviation of the mean MAEs.

For LARS, LASSO, and RF, we perform grid-search hyper-parameter tuning with 5-fold cross-validation upon the training data, to determine the best hyper-parameter values. We use the R package caret for this purpose,[7] focused on minimizing the mean absolute error. Once the best hyper-parameter settings are found, we train the ML algorithm on the training set using those settings, and test it on the test set. For GP-Trad and GP-GOMEA, we take the best expression found by the interleaved runs started by the IMS.

B. ML performance for increasing database size

Perhaps the biggest limitation of our study is the limited size of the database, consisting of only 60 patients. To put this number in perspective, and understand whether increasing the database size will likely improve ML performance, we simulated the effect of having smaller databases.

Figure 1 shows the effect of using GP-GOMEA (the overall best performing algorithm) on databases of different sizes when learning ML models of the liver position. The patients considered in each database (except for the one that contains all 60 patients) are sampled at random. This is repeated 10 times for each size to account for stochasticity.

Mean (of the 10 repetitions) test MAEs behave as can be expected: the larger the database size, the lower the MAE. The standard deviation behaves less intuitively, as it is not only large when

only 7 patients are included, it is also large for when all 60 patients are considered. This is likely the effect of patients that could be considered outliers being now included consistently.

Mean training MAEs exhibit a parabolic shape: they increase first, decrease then. Low training MAEs for smaller databases are simply a symptom of overfitting: it is relatively easy to find a model that fits a small number of points well. However, clearly these models tend to overfit. It is only at 60 patients that sufficient information exists for the ML models to decrease both training and test MAE again. The reason why the training MAE variance shrinks abruptly for the 60 patients-case is that the database is always the same (containing all 60 patients).

As the curves exhibit no convergent behavior, we can confidently conclude that the inclusion of more patients will result in further improving the performance of ML.

C. Examples of models found by GP-GOMEA

Examples of the mathematical expression models found by GP-GOMEA for the automatic construction of pediatric abdominal phantoms are reported below. Each model pertains to a particular metric, and has been manually re-arranged to aid readability. The features and the target metrics are normalized by z-scoring. To apply the model in practice, each feature needs to be de-normalized, i.e., scaled by the standard deviation and translated by the mean. Furthermore, the output of the model needs to be de-normalized as well, i.e., scaled and translated by the standard deviation and the mean of the target variable, respectively (this information is available in Table 1 of the article). We do not include these de-normalization coefficients in the models for the sake of readability. Also, since normalized features are approximately in the same scale, the magnitude of the coefficients included in the model can be associated with feature relevance.

Body S:

$$0.420 \times (\mathbf{ADAP} + \mathbf{ADLR} + \mathbf{SCIS}) .$$

Liver LR:

$$(\mathbf{GEND} + \mathbf{ADAP}) \frac{\sqrt{0.100 + \mathbf{RDLR}^2}}{8.264 + 0.100 \times \mathbf{RDLR}^2}$$

Liver AP:

$$\frac{0.169 \times \mathbf{ADAP}}{\sqrt{0.100 + (\mathbf{AGE} - \mathbf{HEIG})^2}}$$

Liver IS:

$$0.179 \times \mathbf{ADLR} + \mathbf{GEND} \times \mathbf{RDIS}$$

Liver S:

$$\frac{\mathbf{ADLR} + \mathbf{SCIS}}{(e\sqrt{0.100 + (4.389 + \mathbf{LDLR})^2}}$$

Spleen LR:

$$\frac{-0.821 \times \mathbf{LDLR}}{\sqrt{\mathbf{LDLR}^2 + 0.100 \times \mathbf{ICSC}^2 + 0.010}}$$

Spleen AP:

$$\frac{-0.285 \times \mathbf{ICSC}}{\sqrt{0.100 + (\mathbf{AGE} + \mathbf{ADAP})^2}}$$

Spleen IS:

$$\frac{\mathbf{AGE}}{\sqrt{(0.100 + \mathbf{HEIG}^2) \times (9.673 - 6.188 \times \mathbf{WEIG} + \mathbf{WEIG}^2)}}$$

Spleen S:

$$2.718^{\mathbf{AGE}} \times 0.057 \times \mathbf{SCIS}$$

Examples of interpretation: For the prediction of which body segmentation to retrieve (i.e., Body S), the model is particularly simple: it is the sum of the abdominal diameters (ADAP and ADLR) and of the spinal cord length (SCIS), scaled by a constant. Interestingly, these features were already found to be among the most relevant in a study about the feasibility of modeling notions of anatomical similarity using RF (see Virgolin et al. 2018). Arguably, the model is very reasonable: it predicts which body to retrieve based on its dimensions along the three axes: left-right (LR), anterior-posterior (AP), and inferior-superior (IS).

The other models found by GP-GOMEA are non-linear. The recurrence of square roots and squaring terms is due to the use of the analytic quotient, which prevents dividing by 0. For the sake of intuition, $\frac{x}{\sqrt{0.1+y^2}}$ can be considered as approximately $\frac{x}{|y|}$ ($|\cdot|$ takes the absolute value).

As further examples, consider the models that predict the LR position of the liver and of the spleen. The model for liver LR will always return a positive number, whereas the one for the spleen will always return a negative number. This is reasonable because the center of mass of the liver is normally on the right of the reference point we used to determine the organ position, i.e., the 2nd lumbar vertebra, and the opposite holds for the spleen. Moreover, for the prediction of the liver position along LR, the model combines the AP direction, by means of the abdominal diameter in AP (ADAP), and the LR direction, by means of the right diaphragm length along LR (RDLR). Note that, since GEND can either have value 0 (female) or 1 (male), the left multiplicand term is bigger for males than females. Thus, for males, the model predicts bigger shifts in LR. Differently from the liver, the LR position of the spleen relies on the LR size of the left diaphragm rather than on the right diaphragm. This aspect reflects the fact that the liver and spleen are respectively placed below the right and the left diaphragm.

References

1 P. Probst and A.-L. Boulesteix, "To tune or not to tune the number of trees in random forest," *Journal of Machine Learning Research* **18**(1), 6673–6690 (2017).

2 J. Ni, R. H. Drieberg, and P. I. Rockett, "The use of an analytic quotient operator in genetic programming," *IEEE Transactions on Evolutionary Computation* **17**(1), 146–152 (2013).

3 R. Poli, W. B. Langdon, and N. F. McPhee, *A Field Guide to Genetic Programming*, Lulu Enterprises, UK Ltd (2008).

4 J. R. Koza, "Genetic programming as a means for programming computers by natural selection," *Statistics and Computing* **4**(2), 87–112 (1994).

5 M. Virgolin, T. Alderliesten, C. Witteveen, *et al.*, "Scalable genetic programming by gene-pool optimal mixing and input-space entropy-based building-block learning," in *Proceedings of the Genetic and Evolutionary Computation Conference*, 1041–1048, ACM (2017).

6 M. Virgolin, T. Alderliesten, C. Witteveen, *et al.*, "Improving model-based genetic programming for symbolic regression of small expressions," *Accepted for publication in Evolutionary Computation. Preprint arXiv:1904.02050* (2019).

7 M. Kuhn, "The caret package," *Journal of Statistical Software* **28**(5), 1–26 (2008).