TRUSTAI

TRANSPARENT, RELIABLE
& UNBIASED SMART TOOL

Use Case 1 – Health Care

D5.2: Evaluation with healthcare experts of learned models

Centrum Wiskunde & Informatica (CWI)
and
Leids Universitair Medisch Centrum (LUMC)

Leids Universitair Medisch Centrum

CWI
Centrum Wiskunde & Informatica

September 30, 2022

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant agreement No 952060.

DOCUMENT CONTROL PAGE

DOCUMENT	**D5.2 : Evaluation with healthcare experts of learned models**
TYPE	**Report**
DISTRIBUTION LEVEL	Public
DUE DELIVERY DATE	30/09/2022
DATE OF DELIVERY	30/09/2022
VERSION	1.0
DELIVERABLE RESPONSIBLE	CWI
AUTHOR (S)	Evi Sijben (CWI), Tanja Alderliesten (LUMC), Peter Bosman (CWI)
CLINICAL EXPERT	Jeroen Jansen (LUMC)
OFFICIAL REVIEWER/s	LTPLABS

DOCUMENT HISTORY

VERSION	AUTHORS	DATE	CONTENT AND CHANGES
0.1	Evi Sijben (CWI)	30/05/2022	First draft (without expert validation)
0.2	Evi Sijben (CWI)	26/07/2022	First complete draft
0.3	Tanja Alderliesten (LUMC)	18/08/2022	Feedback on version 0.2
0.4	Evi Sijben (CWI)	18/08/2022	Revision of version 0.3 based on feedback
0.5	Peter Bosman (CWI)	24/08/2022	Feedback version 0.4
0.6	Evi Sijben (CWI)	24/08/2022	Revision of version 0.5 based on feedback
0.7	Evi Sijben (CWI), Tanja Alderliesten (LUMC), Peter Bosman (CWI)	25/08/2022	Final check before submission
0.8	LTPlabs	27/09/2022	Review
1.0	Evi Sijben (CWI), Tanja Alderliesten (LUMC), Peter Bosman (CWI)	30/09/2022	Revision based on review, final version

DISCLAIMER:

The sole responsibility for the content lies with the authors. It does not necessarily reflect the opinion of the CNECT or the European Commission (EC). CNECT or the EC are not responsible for any use that may be made of the information contained therein.

Executive Summary

This deliverable concerns a formal validation of the AI models developed for the first (simplified) version of the healthcare problem. These models will be designed by CWI and validated by medical experts from LUMC. The report will present the first insights on the results obtained with the model results and suggestions for modifications.

Table of Contents

Abbreviations and Acronyms

AI	Artificial Intelligence
EC	European Commission
EU	European Union
HCXAI	Human-centered Explainable AI
KPI	Key Performance Indicators
MS	Milestones
PM	Person Month
PR	Press Release
SMEs	Small and Medium-sized Enterprises
WP	Work Package
XAI	Explainable Artificial Intelligence

1. Introduction

In this use case we focus on paraganglioma (i.e., a type of tumors) in the head and neck area. Although they are usually benign and slow growing, they can cause severe complaints such as cranial nerve dysfunction and hearing loss. On the one hand, if the tumors stay small over the lifetime of the patient and the patient does not experience any complaints from the tumor, the patient does not benefit from treatment such as radiotherapy and surgery. In this case, the treatment could do more harm than good, because of the risks associated with the treatment. On the other hand, if the tumor grows big over the lifetime of the patient, the patient is more likely to develop severe complaints and one would have wanted to intervene as early as possible (especially since complaints can be irreversible).

As of right now, it is hard to predict how the tumor will develop. This imposes a difficult dilemma; should we treat this patient? If so, when? If we were able to predict the tumor development for a patient, and maybe even how this development contributes to future complaints due to the tumor, this could support on the one hand giving inevitable treatment early on (with likely less risk and potential complications) and on the other hand avoiding unnecessary, stressful, and costly follow-ups.

In [1], different known functions for describing tumor growth were fitted on 77 paraganglioma tumors. These functions concern the linear function, the exponential function, the Mendelsohn function. the Gompertz function, the logistic function, the Spratt function and the Bertalanffy function. The general shape of these functions is described in Figure 1.

Per tumor a total of 3 volume measurements were available. It was concluded that the s-shaped functions are the best fit. However, for all the fitted functions, there are instances of tumor growths where the predicted volume at conception and/or the volume at the end of the life of the patient is unrealistic. Therefore, we want to use eXplainable Artificial Intelligence (XAI) to automatically find a well-fitting tumor growth function that adheres to constraints, such as having a realistic volume at the end of the life of the patient, without making any assumptions about the specific function structure. Moreover, using XAI it should be possible to arrive at a function that is still interpretable by humans. By making very few assumptions about the function structure, it should be possible to find a tumor growth function that fits the data well. The resulting function could possibly be one of the 6 pre-existing functions, or a new function altogether.

We will tackle this problem of developing a new, human-interpretable, growth function using evolutionary algorithms, in particular genetic programming.

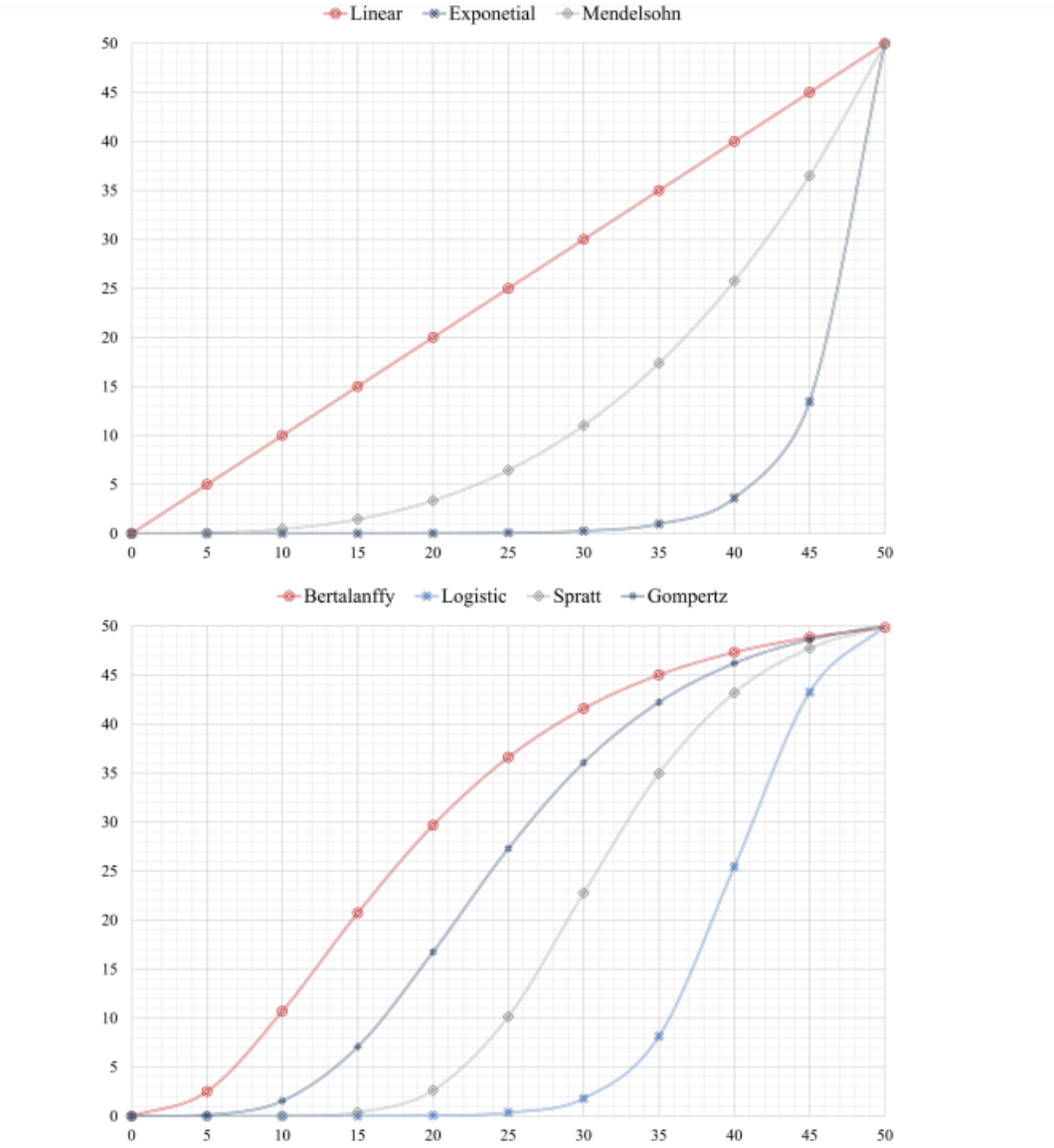

Figure 1: General shape of known tumor growth functions. Source:[1]. The x-axis represents the time (years) and the y-axis represents the volume (cc).

2. Problem formalization

Firstly, we will develop a model, or equivalently said, a *function*, that predicts tumor growth. The tumor growth function f_T of a tumor T predicts the volume of the tumor $V_{T,t}$ at a specific point l_t in the lifetime of the patient carrying the tumor. Therefore, a function predicting the tumor volume over time can be formalized as: $V_{T,t} = f_T(l_t)$. With this we can, for example, predict for a specific patient the expected tumor volume at different time points during their lifetime. This gives a lot of information on the scale of the growth of the tumor. However, knowing the function class f of the growth functions of all tumors of all patients is relevant information as well. If, for example, we know that f is a logistic function, we would know that all tumors at some point will get to a plateau, whereas with a linear function this would not be the case. The parameters of the function class can then be fit to each specific patient, using patient-specific information, to get a prediction function.

It is possible that tumor growth can be divided over several function classes, for example, in the case where specific genetic mutations cause different growth functions. In this case, dividing the patients into groups and fitting a function class to each group would be sensible. This will be elaborated in the future work section.

For every patient we need tumor volume measurements at 3 moments in time. This is the minimum required number of measurements, since 1 or 2 measurements cannot capture accelerating or decelerating growth. Therefore, we have $l_t = [l_{t_1}, l_{t_2}, l_{t_3}]$ and $V_{T,t} = [V_{T,t_1}, V_{T,t_2}, V_{T,t_3}]$. If we would assume f to be linear, e.g., $V = a + b \cdot l$, we need to find constants a_t and b_t for the three measurements of every tumor such that the mean of the squared error between $f_t(l_t) = a_t + b_t \cdot l_t$ and the actual volume for all tumors is minimized.

In optimizing a function class f and its instances f_T, we consider the following constraints:

- In the literature, paraganglioma tumor volumes are assumed to be monotonically increasing [3].
- At conception there cannot be any tumor volume since the one cell existing cannot be a tumor cell.
- It is physically impossible that the tumor volume increases endlessly. The biggest paraganglioma tumor known in the LUMC is 1,190 cc. Therefore, we will assume that the volume cannot grow bigger than 1,500 cc.

3. Solution approach

We use the Real Valued Gene-Pool Optimal Mixing Evolutionary Algorithm (RV-GOMEA) [4], a version of the Gene-Pool Optimal Mixing Evolutionary Algorithm (GOMEA) specifically aimed at problems with real-valued variables, to fit the parameters of the function class per patient. We choose RV-GOMEA over a gradient-based approach because we have observed in preliminary experiments that function classes can be ill-conditioned, leading to exploding or vanishing gradients.

Since we do not yet know the function class of tumor growth of paraganglioma, we propose to first use the Genetic Programming Gene-Pool Optimal Mixing Evolutionary Algorithm (GP-GOMEA) [5] to find this function class, which is state-of-the-art for this kind of problems [6]. Specifically, we optimize the function class, where a solution is a function of mathematical operators, variable l (the age of the patient when measuring the tumor volume), and function class constants. To determine the quality of a function class, the function class is passed to RV-GOMEA, which optimizes the function class constants for each tumor, i.e., a tumor-specific fit is made for each tumor. Ultimately, the fitness of a function class then is calculated by taking the mean error taken over all tumors. The error *per* tumor is defined as the Mean Squared Error (MSE) of the RV-GOMEA tumor-specific-tuned function over the three measurements. In other words, the quality of a function class is given by its average adaptability (by fitting its parameters) to different patients.

In the function class, we use only one variable (lifetime of the patient l_t), and (potentially) multiple function class constants. A function class constant is treated as a terminal node by GP-GOMEA, while it can have a different value for different tumors (i.e., after it has been optimized with RV-GOMEA). Therefore, a small population size for GP-GOMEA will likely be sufficient. However, as we cannot be sure about the best possible population size, we use a starting population size of 30 individuals and the Interleaved Multi-start Scheme (IMS) scheme. We run the IMS for 10 meta-generations.

We use the mathematical operators $\times, \div_p, +, -, ^{-1}, \hat{}$, exp and a tree height of 4 (i.e., 31 nodes). We choose a maximum tree height of 4, because this was reported to be the maximum tree height at which functions likely can still be interpreted [5].

For RV-GOMEA we set the maximum number of evaluations to 50,000 (which generally takes less than a second in our case). In addition, we set the hard bounds for the parameter values at -1e+308 and 1e+308. For the initialization ranges, we first compute the biggest value b in the training data (i.e. the biggest age l_t) of all patients. We then set the initialization range to $[-5 \times b, 5 \times b]$.

The constraints will be implemented as follows:

- For every tumor, we check whether $V_{T,t_i} \leq V_{T,t_{i+1}}$ at 1,000 timepoints t_i equally spread between the moment of conception and the moment when the patient is 99 years old. If for any of the timepoints this does not hold, the constraint is violated.
- For every tumor, we check whether $V_{T,t_i} \leq 1{,}500\ cc$, with t_i the point in time at which the patient is 99 years old. If this is not the case, the constraint is violated
- For every tumor, we check whether $V_{T,t_i} \leq 0.01\ cc$, where t_i is the point in time at which the patient is conceived If this is not the case, the constraint is violated.

We use 0.01 as threshold instead of 0 since there are functions that can only be bigger than 0 such as exponential functions. Here we see 0.01 cc as negligibly small.

In the algorithm, a change in a solution is only accepted if there are less or equal violated constraints, and the fitness is equal or better.

An overview of the proposed algorithm is shown in figure 2.

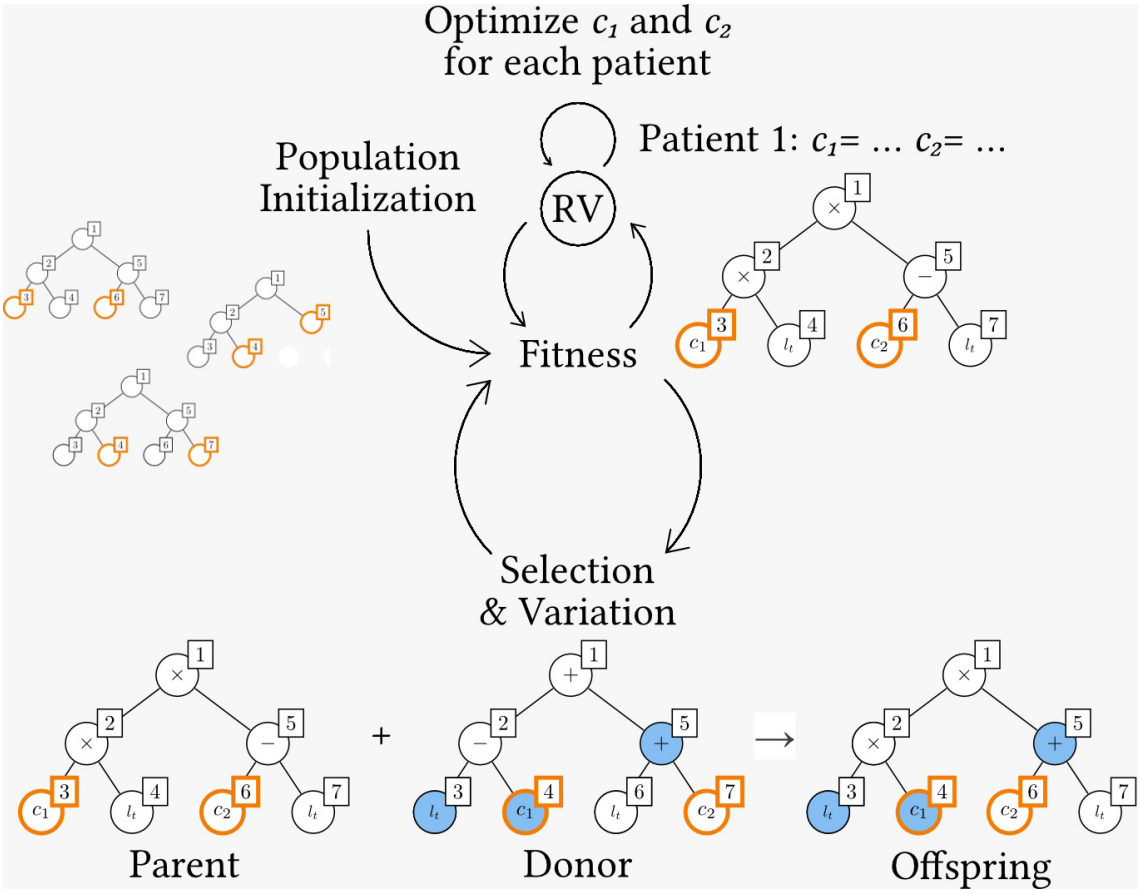

Figure 2 The Function Class learning cycle. First, we initialize the population. Then, we calculate the fitness for each function class (or individual) and tune the function class constants (in orange) to each patient. Finally, we perform variation and selection and calculate the fitness again

To evaluate the function found by the above function class algorithm, Function Class Genetic Programming Real Valued Gene-Pool Optimal Mixing Evolutionary Algorithm (FC-GP-RV-GOMEA), we compare it to six suggested functions for growth published in the literature. For the fitting of these six functions to our patient data we use RV-GOMEA with a maximum of 500,000 evaluations per function and we apply the same constraints as for FC-GP-RV GOMEA.

3.1. Feature engineering

We use the linear dimension measurement method [1] to measure the volume of the tumors. In this method, the largest diameter is measured manually by a PhD student in the X, Y, and Z dimension in 3D space based on a 3D TOF gadolinium enhanced MR scan. The tumor volume is then calculated by assuming it has an ellipsoid shape.

The measurements for the training data were performed by two PhD students. If measurements at the same point in time differed more than the previously determined smallest detectable difference (10% for carotid body and 25% for vagal body paragangliomas), consensus was reached. Otherwise, the mean was taken over these two measurements.

For the test data, we use the measurements collected during regular follow-up by a radiologist as a starting point. Note that the radiologist may differ over the different measurements. Not all measurements were complete, and possibly suffer from inconsistencies. To remedy this, a single radiologist completed the missing measurements, and checked the existing measurements using the MR scans.

3.2. Explainable AI approach

We propose this function class algorithm because of the intended inherent interpretability of the resulting models when compared to a model that just predicts the tumor volume of the next follow-up based on former measurements of the tumor. By giving a function based on time, we explicitly model the connection between time and volume. We think this may well result in a model that is more interpretable than, for example, using SHAP, which would result in an overview of how each volume measurement is estimated to contribute to future predictions. Furthermore, we find our approach more adequate for predicting tumor growth over time, because it also shows the relation between volume and time at points in time that are not in the data set.
The main difference in interpretability between our approach and methods like SHAP is that we create inherently interpretable models that explicitly model the relation between input and output variables, whereas SHAP attempts to visualize an estimation of this relation.

4. Preliminary results

4.1. Data

The data consists of volume measurements at 3 time points of 77 tumors. These are all used as training examples. For 10 of the 77 tumors we collected additional volume measurements, which are used as test data. For these 10 tumors, we calculated the volume from all the available MR scans acquired after the 3rd time point used for training. This resulted in 15 additional measurements: six tumors with one additional measurement, four tumors with two additional measurements, and one tumor with three additional measurements.

4.2. Performance discussion

In Figure 3, a boxplot of the Root Mean Squared Error (RMSE) of the 77 tumors for the six known functions and the newly discovered function is illustrated. In this discussion we excluded the exponential function since it can obviously not fit the data well. Figure 2 enables us to compare the quality of the new function class found and the quality of the known functions for the 77 tumors. As we can see, the function found by the implemented algorithm (FC-GP-RV GOMEA) has the best median RMSE as well as the lowest maximum RMSE. Additionally, we compare the RMSE of the known function with the newly found function using a Wilcoxon Signed Rank Test with a Bonferroni correction and an alpha of 0.05. We find that the new function is significantly better (i.e., having a smaller RMSE) than all the known functions.

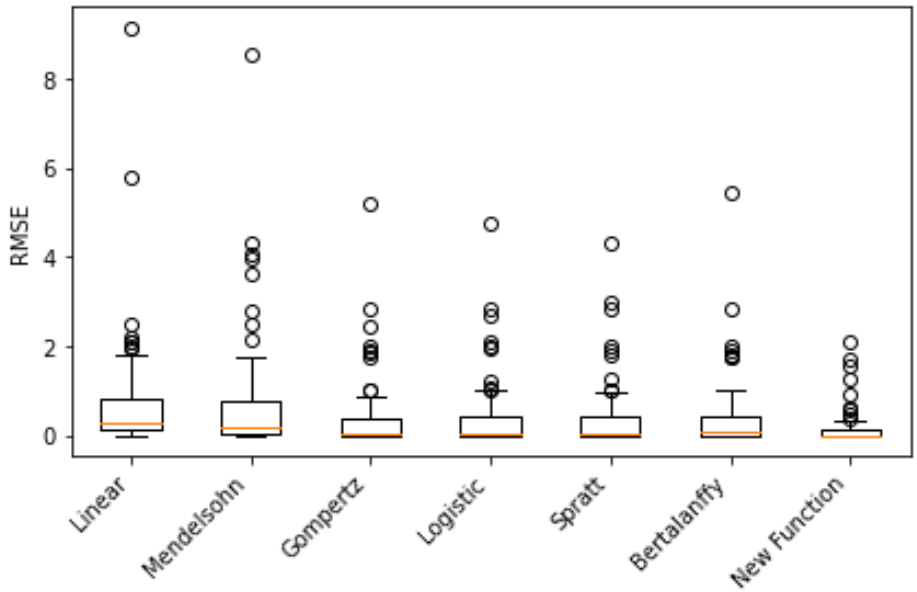

Figure 3: Box plot of the RMSE per tumor for fitting the train data. The volume is in cc. The orange bar indicates the median relative error, the upper and lower bounds of the box indicate the 75th and 25th percentile, respectively. The whiskers indicate the 0th and 100th percentile, excluding outliers, which are indicated by the circles.

In Figure 4, a boxplot of the absolute error as a percentage of the total volume for the additional measurements is illustrated. By taking the error as a percentage of the total volume, we look at the error relative to the volume of the tumor, rather than the absolute error. This gives a better view of prediction accuracy when the data contains tumors with different volumes. We see here that the newly found function is quite consistent in the error percentage it makes. When interpreting these scores, one should consider that it is highly likely that there are measurement errors (with an approximate maximum of 25%) both in train and test data [2]. As well as a possible interobserver error because the measurements of the train data and of the test data are made by a different observer. Additionally, we compare the error percentages of the known function with the newfound function in a Wilcoxon Signed Rank Test with a Bonferroni correction and an alpha of 0.05. We find that there is no significant difference between the new function and all the known functions.

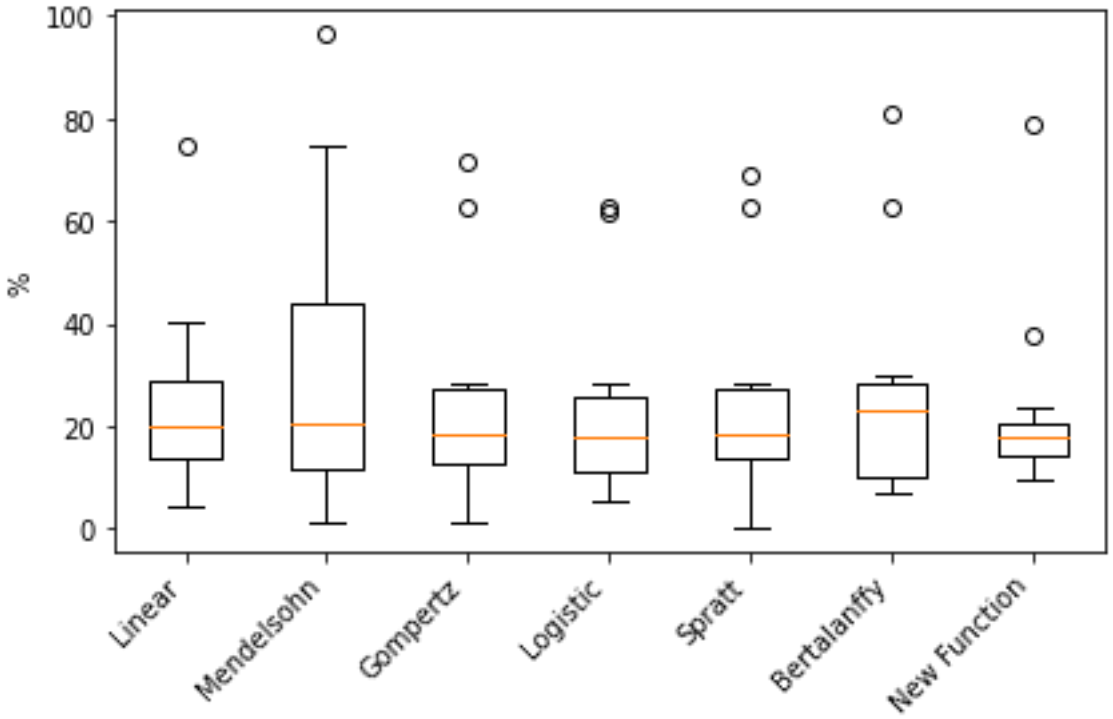

Figure 4: Box plot of the error relative to the total tumor volume (in percentages) for predicting test data volume for each measurement of a tumor. The orange bar indicates the median relative error, the upper and lower bounds of the box indicate the 75th and 25th percentile, respectively. The whiskers indicate the 0th and 100th percentile, excluding outliers, which are indicated by the circles.

4.3. Explainability discussion

The resulting model is as follows,

$$V_{T,t}(l_t) = a + \frac{b}{\frac{c}{d + l_t}},$$

where V is the tumor volume and l_t is the age of the patient in years with l_0 at conception. Further, $a, b, c,$ and d are constants that need to be optimized per tumor.

Although this function indeed expresses a certain relation between tumor volume and time, there are some improvements possible regarding interpretability.
Note that we cannot straightforwardly simplify the double division in this equation because we have used protected division. This protected division and wide variety of possible constant values make it harder to understand the relation between time and tumor volume. Future experiments will therefore include another division operator and smaller bounds for the function class constant values.

4.4. Practitioners' validation

We designed a questionnaire to have the model validated by a clinician.

In this questionnaire, we asked several questions about three patients that had the most additional measurements. First, we provided information about the specific patient, such as the location of the tumor and the Body Mass Index (BMI). Hereafter, we provided a plot of the predicted growth curve as well as the volume measurements used as training data points. We then asked the clinician to judge the likeliness of the predicted growth curve as well as the treatment policy they would propose. After this, we again showed the patient specific information and the growth curve together with the used training data points, however, now, we also included the additional measurement(s) (i.e., the test data point(s)) in the plot. We asked the clinician to again judge the likeliness of the predicted growth curve as well as the treatment policy they would have proposed knowing these new data points. Figure 5 and 6 show examples of the plots shown to the clinician.

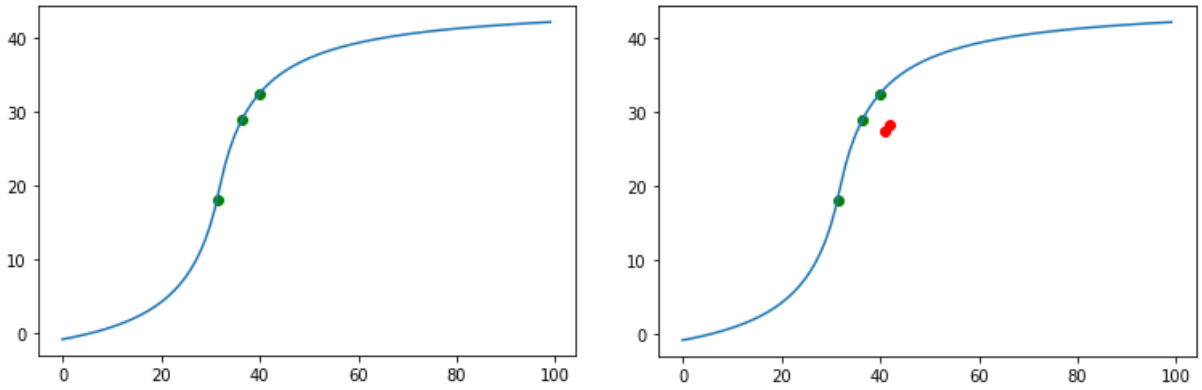

Figure 5: First example of plots of one tumor shown to the clinician in the questionnaire. In these figures, the x-axis is the age of the patient carrying the tumor, the y-axis is the volume in cc, the green points are the training points, and the red points are the test points, the blue line indicates the predicted growth curve.

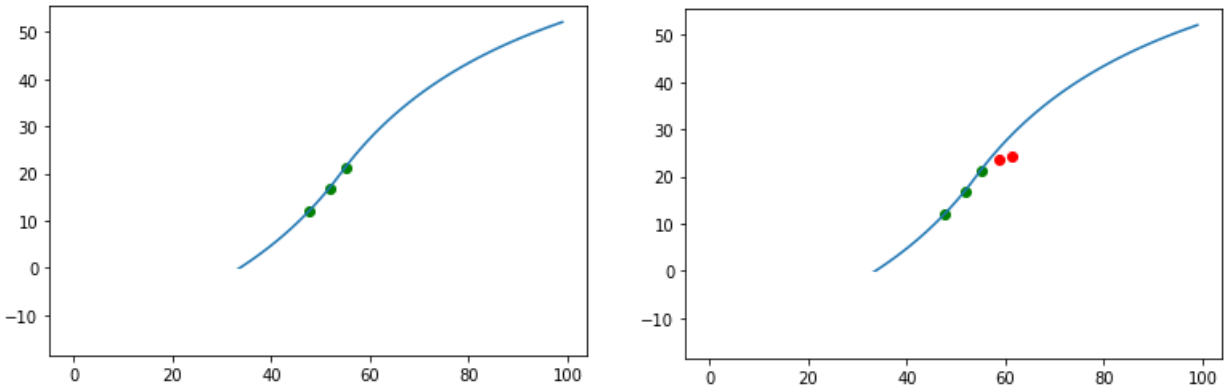

Figure 6: Second example of plots of one tumor shown to the clinician in the questionnaire. In these figures, the x-axis is the age of the patient carrying the tumor, the y-axis is the volume in cc, the green points are the training points, and the red points are the test points, the blue line indicates the predicted growth curve.

The growth curves were judged to be likely based on the illustrated curve with training data points. However, this likeliness decreased for all three patients after seeing the test data points. Based on the comments of the clinician, we point out two main reasons for this. Firstly, the prediction will likely increase in uncertainty for time points further in the future, i.e., the curve becomes less accurate over time. We observed that for two patients with more than one additional measurement, this was the case. However, in clinical practice the new data points can be taken into account once they become available. Secondly, there is a possible measurement error when using the linear dimensions method. This makes it harder to say whether the growth curve is inaccurate, or the data points are inaccurate (due to the measurement error).

The clinician had the same proposed policy for two of the three cases before and after seeing the additional measurements. In the last case, the answer suggested that the

clinician was in doubt between 2 and 5 years for a next follow-up moment, but proposed 2 years to be on the safe side. After seeing the new measurements, the clinician decided that 5 years was the most appropriate interval for follow-up.

The clinician estimated that the clinical relevance of such a growth model is high because it can give both the patient and clinician trust in the proposed policy.

5. Conclusions

5.1. Future developments in the Use Case

In the future, we plan to make a deep-learning-based automated tumor segmentation approach based on which the tumor volume can be calculated. This would enable us to include more measurements and more patients in the growth model. Additionally, it could possibly reduce the measurement error. Finally, this model could have direct clinical relevance since measuring the tumor is a standard procedure in follow-up, and therefore it would be beneficial if measurement error and efforts that need to be made by the radiologist when measuring the tumor could be decreased.

Additionally, we plan to work on the interpretability of the growth model, for example by replacing less interpretable operators such as the protected division and setting more constraints on the possible values of the function class constants.

In consultation with the clinician, we figured that not only the expected size and the growth of the tumor is important for treatment decision making, but the location/direction of the growth plays an important role as well. Firstly, this is because growth towards some parts of the body might complicate treating the patient [7,8,9]. Secondly, growth towards specific parts of the body might cause new complaints [7]. Therefore, it might be of interest to also consider this in the project.

Lastly, we plan to implement a multi-tree multi-objective multi-modal version of the function class algorithm to get diverse sets of growth models that can be inspected by the clinicians. As mentioned before, it could be the case that there are specific groups of tumors that adhere to different function classes. Additionally, it could be the case that different function classes are possible on the same data where a clinician might be more interested in one of the classes because of their knowledge and experience. Therefore, we think the multi-tree multi-objective multi-modal approach will be useful.

5.2. Recommendations for TRUST-AI Framework

We recommend making it possible to use the multi-tree, multi-objective, multi-modal algorithm in the framework [10]. It could especially be interesting to consider implementing a custom interface for model selection that can be used alongside this algorithm.

6. References

[1] Heesterman, B. L. et al. (2019). Mathematical models for tumor growth and the reduction of overtreatment. *Journal of Neurological Surgery Part B: Skull Base*, *80*(01), 072-078.

[2] Heesterman, B. L. SDHD-related Head and Neck Paragangliomas [PhD thesis, LUMC]. https://scholarlypublications.universiteitleiden.nl/handle/1887/65453

[3] Jansen, J. C., van den Berg, R., Kuiper, A., van der Mey, A. G., Zwinderman, A. H., & Cornelisse, C. J. (2000). Estimation of growth rate in patients with head and neck paragangliomas influences the treatment proposal. *Cancer*, *88*(12), 2811-2816.

[4] Bouter, A., Alderliesten, T., Witteveen, C., & Bosman, P. A.N. (2017). Exploiting linkage information in real-valued optimization with the real-valued gene-pool optimal mixing evolutionary algorithm. In *Proceedings of the Genetic and Evolutionary Computation Conference* (pp. 705-712).

[5] Virgolin, M., Alderliesten, T., Witteveen, C., & Bosman, P. A. N. (2021). Improving model-based genetic programming for symbolic regression of small expressions. *Evolutionary Computation*, *29*(2), 211-237.

[6] La Cava, W. et al. (2021). Contemporary symbolic regression methods and their relative performance. arXiv preprint arXiv:2107.14351.

[7] Fisch, U. (1982). Infratemporal fossa approach for glomus tumors of the temporal bone. *Annals of Otology, Rhinology & Laryngology*, *91*(5), 474-479.

[8] Netterville, J. L., Jackson, C. G., Miller, F. R., Wanamaker, J. R., & Glasscock, M. E. (1998). Vagal paraganglioma: a review of 46 patients treated during a 20-year period. *Archives of Otolaryngology–Head & Neck Surgery*, *124*(10), 1133-1140.

[9] Shamblin, W. R., ReMine, W. H., Sheps, S. G., & Harrison Jr, E. G. (1971). Carotid body tumor (chemodectoma): clinicopathologic analysis of ninety cases. *The American Journal of Surgery*, *122*(6), 732-739.

[10] Sijben, E. M. C., Alderliesten, T., & Bosman, P. A.N. (2022). Multi-modal multi-objective model-based genetic programming to find multiple diverse high-quality models. In *Proceedings of the Genetic and Evolutionary Computation Conference* (pp. 440-448).