Development of Error Passing Network for Optimizing the Prediction of VO_2 peak in Childhood Acute Leukemia Survivors

Nicolas Raymond* NICOLAS.RAYMOND2@USHERBROOKE.CA
Department of Computer Science, Université de Sherbrooke, Sherbrooke, Canada

Hakima Laribi* HAKIMA.LARIBI@USHERBROOKE.CA
Department of Computer Science, Université de Sherbrooke, Sherbrooke, Canada

Maxime Caru MCARU@PENNSTATEHEALTH.PSU.EDU
Department of Pediatrics, Division of Hematology and Oncology, Pennsylvania State Health Children's Hospital, Hershey, PA, USA.
Department of Public Health Sciences, Penn State University College of Medicine, Hershey, PA, USA

Mehdi Mitiche GM_MITICHE@ESI.DZ
Department of Computer Science, Université de Sherbrooke, Sherbrooke, Canada

Valérie Marcil VALERIE.MARCIL@UMONTREAL.CA
Research Center, Sainte Justine University Health Center, Department of Nutrition, Université de Montréeal, Montreal, Canada

Maja Krajinovic MAJA.KRAJINOVIC@UMONTREAL.CA
Research Center, Sainte Justine University Health Center, Department of Pediatrics, Université de Montréal, Montreal, Canada

Daniel Curnier DANIEL.CURNIER@UMONTREAL.CA
Research Center, Sainte Justine University Health Center, School of Kinesiology and Physical Activity Sciences, Faculty of Medicine, Université de Montréal, Montreal, Canada

Daniel Sinnett DANIEL.SINNETT@UMONTREAL.CA
Research Center, Sainte Justine University Health Center, Department of Pediatrics, Université de Montréal, Montreal, Canada

Martin Vallières MARTIN.VALLIERES@USHERBROOKE.CA
Department of Computer Science, Université de Sherbrooke, Sherbrooke, Canada
Centre de recherche du Centre hospitalier universitaire de Sherbrooke, Sherbrooke, Canada

Abstract

Approximately two-thirds of survivors of childhood acute lymphoblastic leukemia (ALL) cancer develop late adverse effects post-treatment. Prior studies explored prediction models for personalized follow-up, but none integrated the usage of neural networks to date. In this work, we propose the Error Passing Network (EPN), a graph-based method that leverages relationships between samples to propagate residuals and adjust predictions of any machine learning model. We tested our approach to estimate patients' VO_2 peak, a reliable indicator of their cardiac health. We used the EPN in conjunction with several baseline models and observed up to 12.16% improvement in the mean average percentage error compared to the last established equation predicting VO_2 peak in childhood ALL survivors. Along with this performance improvement, our final model is more efficient considering that it relies only on clinical variables that can be self-reported by patients, therefore removing the previous need of executing a resource-consuming physical test.

Data and Code Availability Software code allowing to run the experiments used to produce the results presented in this work is freely shared under the GNU General Public License v3.0 on the GitHub website at https://github.com/Rayn2402/ErrorPassingNetwork. The PETALE dataset (Marcoux et al., 2017) anal-

* These authors contributed equally.

ysed during the current study is not publicly available for confidentiality purposes. However, a randomly generated dataset with the same format as used in our experiments is publicly shared in our GitHub repository to test the code implemented for this work.

Institutional Review Board (IRB) All the analyses conducted for the study were compliant with the Declaration of Helsinki and approved by the Institutional Review Board of Sainte-Justine University Health Center. Written informed consent was obtained from study participants or parents/guardians.

1. Introduction

Childhood acute lymphoblastic leukemia (ALL) is the most frequently diagnosed type of cancer in children (Lemay et al., 2019). The 5-year relative survival rate is currently above 90% (Hunger et al., 2012). Nevertheless, approximately two thirds of childhood ALL survivors will present one or more health complications resulting from the treatment (e.g., exposure to chemotherapy, cranial radiation therapy) known as late adverse effects (LAEs) (Nathan et al., 2009). The existing follow-up measures, used in clinical settings and offered to patients during their visits to the hospital, are rather standardized for all childhood cancer survivors and not necessarily personalized for childhood ALL survivors (Hudson et al., 2021). As a result, LAEs may be underdiagnosed, and in most cases, only taken care of once they have already appeared in adulthood. In recent decades, survivorship studies have investigated associations between different factors in childhood ALL populations (e.g., received treatment, physical fitness, genetic sequence) and LAEs (Szymon et al., 2011; Wilson et al., 2015, 2018; Geneviève et al., 2024). In particular, between 2013 and 2016, 246 childhood ALL survivors participated to a series of clinical, physiological, biological and genetic evaluations as part of the PETALE study (Marcoux et al., 2017). The objective of the latter was to identify clinical, genetic, and biochemical biomarkers that are relevant to develop targeted prevention and treatment strategies reducing LAEs prevalence.

Since then, many studies have honed in the development of better personalized follow-up methods using the data acquired from the PETALE cohort (Labonté et al., 2020; England et al., 2017; Morel et al., 2018; Nadeau et al., 2019; Caubet F. et al., 2019; Caru et al., 2019). As an example, an equation based on a linear regression was specifically developed by Labonté et al. (2020) to estimate the maximal oxygen consumption (i.e., VO_2 peak) in childhood ALL survivors following a 6-minute walk test (6MWT). The VO_2 peak is an excellent predictor of cardiac health in patients with cancer and is recognized as the gold standard in exercise physiology to measure patients' cardiorespiratory fitness (Smart, 2013), which plays an important role towards the prevention of LAEs in childhood ALL survivors (Lemay et al., 2019). However, the direct measurement of the VO_2 peak, which is usually done by performing a maximal cardiopulmonary exercise test (CPET), is not an optimal solution in clinical settings due to financial and time constraints. Therefore, there is an interest in using a walking test (e.g., 6MWT) when the access to comprehensive testing is limited (e.g., CPET) (Mizrahi et al., 2020). However, even if it has been shown that using a disease-specific VO_2 peak equation from the 6MWT provides a robust tool to estimate the patient's cardiorespiratory fitness with lower costs (Mizrahi et al., 2020), there is still place for improvement considering that the 6MWT requires time and resources.

In medical contexts, simple models such as linear regression and logistic regression are often favored over more complex machine learning approaches (e.g., deep learning models) due to their ability of being easily interpreted (Lundberg and Lee, 2017; Fan et al., 2021). Additionally, because of to their modest number of parameters to optimize (i.e., reduced capacity), simple models are less inclined to overfit on small training datasets and may have higher generalization potential. Hence, these models are well adapted to clinical contexts with small cohorts of patients. However, more sophisticated model architectures (e.g., neural networks) have lately achieved better results in the prediction of clinical events using data from electronic health records (Choi et al., 2016; Ma et al., 2017). Interpretability of neural networks has also been the subject of multiple studies over the last years (Fan et al., 2021; Zhang et al., 2021). In particular, Fan et al. (2021) highlighted that the interpretation of neural networks can be facilitated from their design, with the inclusion of components with specific functionalities. For example, recent works motivated the usage of attention mechanisms within their models to help depict the decision-making process behind individual samples (Ma et al., 2017; Arık and Pfister, 2021). Other studies also

explicitly integrated graph-based architectures (i.e., graph neural networks) to leverage the importance of the similarities between patients to solve a prediction task (Lu and Uddin, 2021; Liu et al., 2020). Recent research has also harnessed these similarities to combine samples' residuals and improve interpretability and credibility of machine learning models (Papernot and McDaniel, 2018) and their performances (Yang et al., 2024).

In this work, we built on the practicability of graphs and attention mechanisms to propose a novel interpretable approach named Error Passing Network (EPN). Our model leverages similarities among patients to propagate residuals from a machine learning model and subsequently adjust its predictions. We first shown that the EPN was able to significantly improve the VO_2 peak predictions emerging from the equation already established by Labonté et al. (2020). We further realized post-hoc analyses of the attention scores calculated by the EPN to explain how predictions are refined, thereby strengthening the interpretability of the model. We finally used the EPN to develop a new model predicting the VO_2 peak based on clinical variables that can be entirely self-reported by patients. The latter, while being more efficient in terms of required resources (e.g., time, equipment) is also the most accurate prediction approach in the population of childhood ALL survivors that we studied. We believe our new findings in precision medicine will help towards efficiently and accurately monitoring cardiac health of childhood ALL survivors, and hence, help for the prevention of LAEs.

2. Materials and methods

2.1. Error Passing Network (EPN)

In this work, we propose a novel graph-based approach that leverages the power of attention (Vaswani et al., 2017) to improve predictions made by any machine learning model. The latter, which we refer to as **E**rror **P**assing **N**etwork (EPN), predicts the error that will be made by a model for a new data point (i.e., node) by calculating a weighted average of the errors made by the same model on neighboring nodes that were part of the training set. The EPN prediction is further combined with the one of the original model to get a corrected estimate. In summary, the EPN seeks to diminish the risk of making wrong predictions on new data points by integrating past errors made on similar ones in the training set.

More formally, let G be a simple undirected graph with a set of vertices V and a set of edges $E \subseteq V^2$. Let us also assume that each vertex $v_i \in V$ is associated to a feature vector $\boldsymbol{x}_i \in \mathbb{R}^D$ and a real-valued target $t_i \in \mathbb{R}$. We can represent V as the union of the disjoint sets V_{train} and V_{test}, such that V_{train} contains the nodes (i.e., vertices) for which the features and targets were used previously to train a machine learning model $f : \mathbb{R}^D \to \mathbb{R}$, while V_{test} contains the ones that were used to test f. Finally, let $N(v_i)$ be the set of training nodes sharing an edge with a node $v_i \in V$:

$$N(v_i) = \{v_j \mid v_j \in V_{train}\} \cap \{v_j \mid (v_i, v_j) \in E\}.$$

The EPN model can be used in conjunction with f to estimate the target of any node v_i as follow:

$$\hat{t_i} = \overbrace{\Big(\sum_{j|v_j \in N(v_i)} \alpha_{ij}\,(t_j - f(\boldsymbol{x}_j))\Big)}^{\mathrm{EPN}(G,\boldsymbol{x}_i)} + f(\boldsymbol{x}_i), \quad (1)$$

where

$$\alpha_{ij} = \mathrm{softmax}\left(\{e_{ij}\}_{j|v_j \in N(v_i)}\right) \quad (2)$$

$$e_{ij} = \frac{(\mathbf{W}^Q \boldsymbol{x}_i)^t \mathbf{W}^K \boldsymbol{x}_j}{\sqrt{D}}. \quad (3)$$

In this context, α_{ij} represents the attention of node v_i towards node v_j. It is determined by normalizing the values e_{ij} resulting from the dot products between the query and key projections of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ obtained with matrices $\mathbf{W}^Q$ and $\mathbf{W}^K$ respectively. This bi-directional attention mechanism refers exactly to the work of Vaswani et al. (2017). The approach presented in Equation (1) can be interpreted as a gradient boosting method where the EPN is used to estimate the *residuals* of f. However, the main element that distinguishes the current approach to broadly used boosting methods is the fact that the EPN predicts residuals using graph relationships.

2.2. Dataset

All data was taken from the PETALE study. All participants of this study were survivors with European origins who have been diagnosed for childhood ALL between 1987 and 2010 before the age of 19 and were at least 5 years post-diagnosis (see the article from Marcoux et al. (2017) for a complete list of the

Table 1: Descriptive analysis of the clinical features. The p-values, calculated using the Welch's t-test (Welch, 1947), are used to evaluate significance of the difference between biological sex means.

Feature	Unit	All survivors (n=177)	Female (n=92)	Male (n=85)	p-value
Weight	kg	67 ± 15.8	64.5 ± 16.4	69.6 ± 14.8	0.03
Age	years	22.7 ± 6.4	22.8 ± 6.5	22.7 ± 6.3	0.95
DT	years	2.1 ± 0.2	2.1 ± 0.26	2.1 ± 0.15	0.38
VO_2 peak	ml/kg/min	31.8 ± 8.3	27.0 ± 6.8	36.9 ± 6.6	$<$ **0.001**
MVLPA	min/day	27 ± 30.5	26.6 ± 33.4	27.3 ± 27.1	0.87
6MWD	m	611.8 ± 78.6	586.8 ± 69.4	637.6 ± 79.6	$<$ **0.001**
HR_{end}	bpm	149.6 ± 22.3	154 ± 20.6	145.1 ± 23.1	0.01

DT: duration of treatment; MLVPA: moderate-to-vigorous leisure physical activity; 6WMD: 6-minute walked distance; HR_{end}: heart rate at the end of the walk.

eligibility criteria). Our dataset consisted of 177 survivors who reached a valid maximal oxygen consumption while performing a cardiopulmonary exercise test (Labonté et al., 2020; Caru et al., 2021). A descriptive analysis of the aforementioned is presented in Table 1. The age, the weight and the duration of treatment (DT) of each of the participants were determined from their records. The moderate-to-vigorous leisure physical activity (MVLPA) was self-reported by surviors. The 6-min walked distance (6WMD) and the heart rate at the end of the walk (HR_{end}) were acquired during a 6-minute walk test (Labonté et al., 2020).

2.3. Experimental setup

2.3.1. Baselines

We evaluated the EPN with predictions made by different machine learning models frequently considered for regression tasks on small tabular datasets. The latter, which we refer to as baselines, are comprised of the Random Forest and Linear Regression algorithms from *scikit-learn* library (Pedregosa et al., 2011), and the XGBoost algorithm from *xgboost* library (Chen and Guestrin, 2016). We also implemented the previously established equation by Labonté et al. (2020):

$$\begin{aligned} VO_2 = &-0.236 \cdot \text{Age} - 0.094 \cdot \text{Weight} \\ &- 0.120 \cdot \text{HR}_{end} + 0.067 \cdot \text{6MWD} \\ &+ 0.065 \cdot \text{MVLPA} - 0.204 \cdot \text{DT} + 25.145 \end{aligned} \quad (4)$$

2.3.2. Training and test sets

To evaluate each of the baselines introduced previously, we performed a nested five-fold stratified cross-validation (Figure 1). We separated the dataset five times into disjoint training and test sets containing 80% and 20% of the data respectively. Each of the training set was subsequently separated into disjoint inner training and inner test sets of the same proportions. The inner sets were entirely dedicated to optimize the hyperparameters of the models for each outer training fold (see Section 2.3.3). Stratified sampling of the test sets (as well as inner test tests) was done according to the biological sex of the patients. With Table 1 providing evidence that the biological sex of childhood ALL survivors has a significant impact on their VO_2 peak values, we assumed that not preserving sexes' proportions among data splits would have affected the performances measured within each fold.

To train the EPN, we created an additional validation set (as well as an inner validation set) for each of the five cross-validation splits. The latter were used to track the model's performance during training and proceed to early stopping. The creation of each of the validation sets (inner validation sets) was performed by sampling 20% points from the training sets (inner training sets) of the same split, also using the stratified sampling by biological sex.

2.3.3. Hyperparameter optimization

We optimized the hyperparameters of each model using the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) (Hansen, 2023) implementation provided by the *optuna* (Akiba et al., 2019) library. For this purpose, 500 sets of hyperparameter values were sampled sequentially from pre-defined search spaces (Figure 1). Models' hyperparameters and

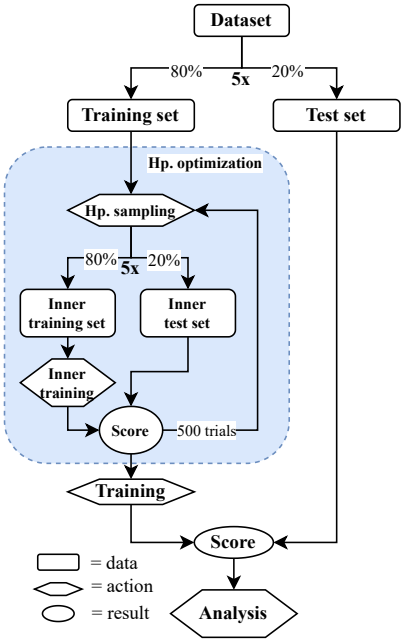

Figure 1: Nested five-fold cross validation. The nested loop (i.e., inner loop), in blue, is dedicated to hyperparameter optimization (i.e., hp. optimization). The same data splits are used for all baselines.

search spaces are provided in Appendix A. Each sampled set of hyperparameters was evaluated by training the model on each of the five inner training sets and then measuring the average of the root mean squared error (RMSE) obtained on their respective inner test sets. The set of hyperparameter values associated to the lowest RMSE was selected to train the model on the whole training set of the outer loop of the nested cross-validation process.

2.3.4. Construction of the graph

Since no underlying graph structure was associated with the dataset prior to our experiments, we decided to assume that each patient could be connected to all the others. Therefore, we represented our dataset as a fully connected graph where each node was associated to a patient. Although the usage of k nearest neighbors (k-NN) has been discussed in past studies as a means to create a graph structure when no prior connections are defined between nodes (Chen et al., 2020; Fatemi and El A., 2021; Qian et al., 2021), we did not take this approach since we hypothesized that it would have introduced a bias in our framework. In fact, with a k-NN graph, the EPN could have only learned from a pre-defined number of connections, which would have been determined by calculating a distance between every patients considering their features. Here, by using a fully connected graph, we let the EPN use its attention mechanism to determine what represents a strong or a weak connection (i.e., edge) between two patients.

2.3.5. Data imputation and transformation

For each pair of training and test sets created (as well as inner training and inner test sets pairs), we imputed the missing data in the numerical columns using the empirical means calculated from the observed data in the training set. In the overall dataset, columns with missing values were 6MWD and HR_{end}. 8 patients did not have a documented 6MWD (4.52%) and 11 patients did not have a documented HR_{end} (6.21%). Once imputed, transformation steps were applied to each pair of training and test sets. For all baselines except the Labonté et al. (2020) equation defined in Equation (4), numerical columns were reduced and centered using the empirical means and standard deviations of the observed data in the training set. The modalities of the biological sex were changed to nominal encoding (women: 0, men: 1).

2.3.6. EPN training procedure

We implemented the EPN using Pytorch (Paszke et al., 2019). The EPN is trained by batch gradient descent using a set of training samples $\mathcal{X} = \{(\boldsymbol{x}_i, t_i) \in \mathbb{R}^D \times \mathbb{R}\}_{i=1}^n$, a loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$, a pre-trained machine learning model $f : \mathbb{R}^D \to \mathbb{R}$, and a simple undirected graph G representing the connections between the training samples in $\mathcal{X}$. The elements contained in each batch represent a subset of training patients (i.e., nodes) $\mathcal{X}' \subset \mathcal{X}$ for which the EPN needs to estimate the targets using both the predictions made with the pre-trained model f and the residuals of the other patients in the training set that are not in the batch (i.e., $\mathcal{X}/\mathcal{X}'$). In Figure 2, we illustrate a training epoch with a fully connected graph of four training samples (i.e., $\mathcal{X} = \{\boldsymbol{x}_i, t_i\}_{i=1}^4$). In this example, each batch used for gradient descent is comprised of two samples (i.e., $|\mathcal{X}'| = 2$). At each step of the epoch, losses of the samples in the batch, shown in red, are averaged to update the weights of the attention mechanism of the EPN. In this work, we trained the EPN using the Adam optimizer (Kingma and Ba, 2014) with parameters $\beta_1 = 0.9, \beta_2 = 0.999$, and a batch size of 32. Batches were shuffled be-

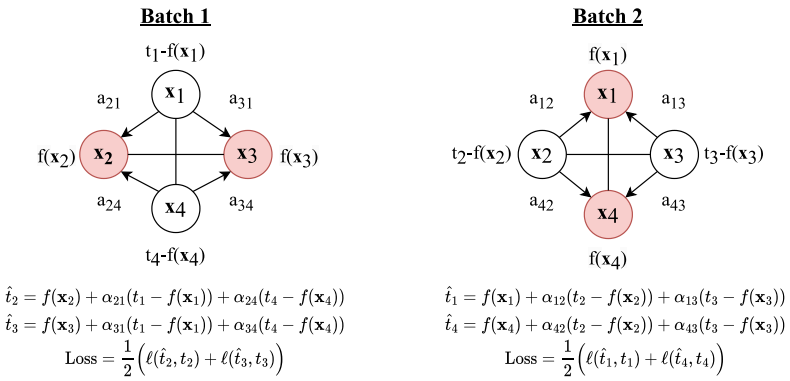

Figure 2: Error passing network (EPN) training procedure. In this example, a training epoch is executed with a fully connected simple undirected graph of four training patients (i.e., nodes) and a batch size of two. Samples in each batch are shown in red. Although the graph is undirected, arrows are used to indicate the flow of the residuals (i.e., $t_i - f(\boldsymbol{x}_i)$) from the training patients outside of the batches ($\boldsymbol{x}_i \in \mathcal{X}/\mathcal{X}^{'}$) to the ones inside the batches ($\boldsymbol{x}_i \in \mathcal{X}^{'}$).

tween each epoch. We set a maximum budget of 100 epochs and applied early stopping with a patience of 10 epochs. Mean squared error (MSE) and root mean squared error (RMSE) were used as the training loss ℓ and the early stopping metric respectively.

3. Results

3.1. Improving previous work with the EPN

In this work, we first evaluated the EPN model using the equation established by Labonté et al. (2020) (Equation (4)). For this experiment, we allowed the EPN to use the same variables as the equation to determine the attention scores between patients. It can be seen from Table 2 that using the EPN on top of the linear model significantly decreased the mean absolute error (MAE) and the mean absolute percentage error (MAPE) recorded on the test sets of the five-fold stratified cross-validation splits.

More precisely, enabling the correction of the predictions by sharing residuals of neighboring training patients resulted in VO_2 peak estimations that were on average 10.66 percentage points closer the the real values. Figure 3(a) supports this result by showing a comparison of the predictions and the targets generated on the five test sets. Furthermore, Figure 3(b) illustrates the distribution of the residuals (i.e., targets - predictions), calculated on the same test sets. We observe from the latter figure that the combination of the EPN with previous work (Labonté et al., 2020) led to a distribution of residuals that is more centered towards zero ($\mu = 0.14$ ml/kg/min) than that of the equation on its own ($\mu = -6.07$ ml/kg/min), the latter showing a general pattern of overestimation of the VO_2 peak values. Although the Spearman rank correlation (Spearman R) measured was higher with the EPN approach (Table 2), statistical tests revealed these changes to be non significant. Therefore, we verified that the last linear solution (Labonté et al., 2020) is still able to generate estimations that preserve the order of magnitude seen in the targets, even though it is less accurate than when adjusted with the EPN.

Table 2: Comparison of Labonté et al. (2020) equation, with and without the EPN. The median p-value over the five test folds is considered to determine the significance of the difference between the two models. Each p-value is obtained through bootstrapping (MacKinnon, 2009; Martínez-Camblor and Corral, 2012) with 100,000 repetitions.

	Labonté	Labonté + EPN
MAE	7.01 ± 0.81	4.59 ± 0.29*
MAPE	$26.36\% \pm 3.32$	$15.70\% \pm 1.80$*
Spearman R	0.69 ± 0.07	0.70 ± 0.07

*Significant difference ($p < 0.05$).
MAE: Mean Absolute Error; MAPE: Mean Absolute Percentage Error.

Next, post-hoc analyses of the attention mechanism of a trained EPN can provide important insights about the relationships between data points, here,

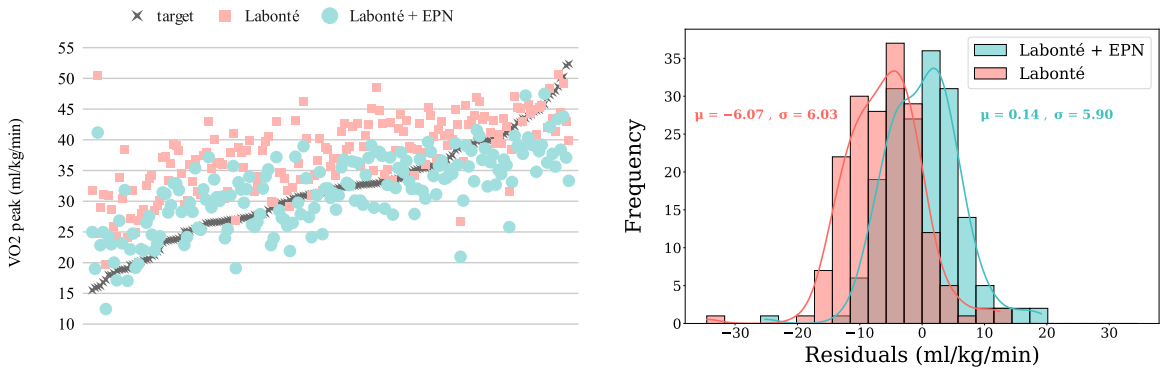

Figure 3: Impact of applying the EPN to Labonté et al. (2020) model (Equation (4)). (a) Predicted values of the equation, with and without the EPN, are compared to the VO_2 peak target values. Values predicted with the EPN are closer to the ground truth. (b) The distributions of the difference between the targets and the predicted values (i.e., the residuals) are illustrated with histograms for both the equation, and the equation with the EPN. On its own, Equation (4) generally overestimates the VO_2 peak with an average residual of -6.07 ml/kg/min , whereas, combined to the EPN, leads to a narrower distribution with a mean closer to zero.

childhood ALL survivors. In Figure 4(*a*), we presented attention scores between training patients and test patients of one of the five folds of cross-validation. We observe that some patients focus their attention on a small number of neighbors, suggesting a similarity between them, whereas others distribute their attention more uniformly across all patients. This indicates that some patients derive greater benefits from the attention mechanism than others. As a proof of concept, we selected the single patient P_{test} for which the correction effect of the EPN led to the highest improvement in VO_2 peak estimation and sought to identify its most influential neighbors of the training set.

In Figure 4(*b*), we illustrated the three highest attention scores observed between the selected test patient and the other childhood ALL survivors from the associated training set. We can see that the EPN weights most of the attention towards a single patient. In Figure 4(*c*), we shared the profile of the patient holding 82.6% of the attention and calculated its correction effect (i.e., Attn × Residual) on the test patient. We can observe that the two patients have similar age and weight, and that the value of 48.5 ml/kg/min predicted by Labonté et al. (2020) equation was adjusted to 38.1 ml/kg/min by the EPN, which is closer to the target of 32.9 ml/kg/min. We further see that the most influential training patient caused an adjustment of -12.3 ml/kg/min (0.826×-13.1), and hence, that the remaining training patients contributed as a whole to an increase of 1.9 ml/kg/min. Overall, the EPN architecture allows capturing higher-order, non-linear relationships among variables that leads to assign a high attention score to a patient for whom Labonté et al. (2020) equation (Equation (4)) predicted a VO_2 peak value with a close residual (Figure 4(*c*)). In Appendix B, we detail additional experiment results demonstrating the effectiveness of the attention mechanism in the EPN.

3.1.1. Analyzing the behavior of the EPN with linear regression models

Following the experiment conducted on the last established equation of VO_2 peak for childhood ALL survivors, we analyzed the general behavior of the EPN with linear regression models, the family of algorithm in which Labonté et al. (2020) work belongs. Hence, considering any model of the form $\boldsymbol{w}^t\boldsymbol{x} + \beta$ with $\boldsymbol{w}, \boldsymbol{x} \in \mathbb{R}^{D\times 1}$ and $\beta \in \mathbb{R}$, we started back from the definition of the EPN (Equation (1)) and de-

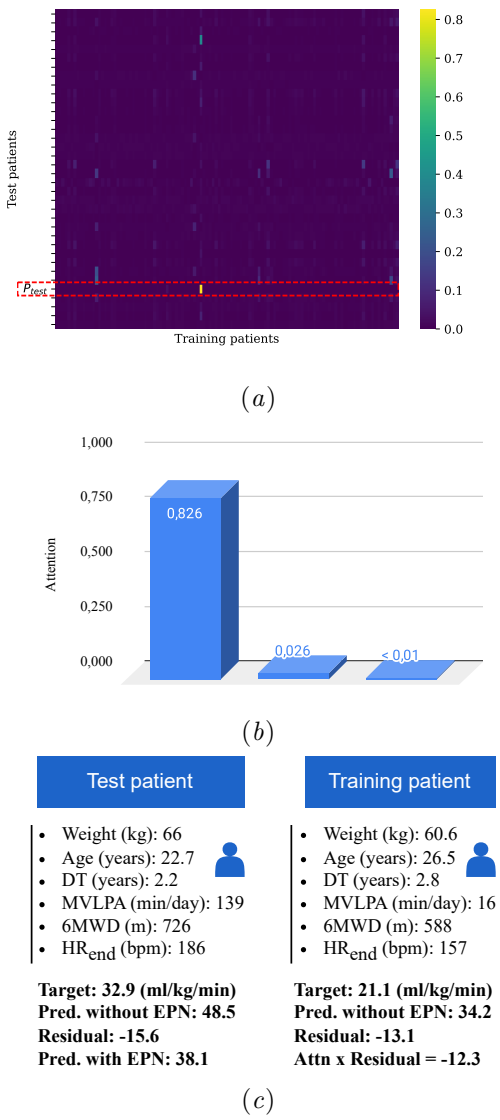

Figure 4: Post-hoc analyses of the attention scores of a trained EPN. (a) Attention scores between training patients and test patients of one fold of the five folds of cross-validation. (b) Three highest attention scores attributed by the test patient P_{test} for whom the EPN had the most significant correction effect in estimating VO_2 peak. Among all the training patients, a single one holds 82.6% of the attention. (c) Comparison of the profiles of P_{test} and the training patient that was assigned the highest attention score. Patients explicitly share similarities regarding their ages, their weights, and their residuals.

ducted that

$$\begin{aligned}\mathrm{EPN}(G, \boldsymbol{x}_i) &= \sum_{j|v_j \in N(v_i)} \alpha_{ij} \left(t_j - \left(\boldsymbol{w}^t \boldsymbol{x}_j + \beta\right)\right) \\ &= \sum_{j|v_j \in N(v_i)} \alpha_{ij} \left(t_j - \boldsymbol{w}^t \boldsymbol{x}_j\right) - \alpha_{ij}\beta \\ &= \left(\sum_{j|v_j \in N(v_i)} \alpha_{ij} \left(t_j - \boldsymbol{w}^t \boldsymbol{x}_j\right)\right) - \beta\end{aligned}$$

since the attention scores associated to as single data point sum to one. This led us again to find from Equation (1) that applying the EPN to a linear regression model removes its bias and replaces it for a personalized correction factor determined from the connections between a sample and its graph neighbors. More precisely, we have that:

$$\begin{aligned}\hat{t}_i &= \mathrm{EPN}(G, \boldsymbol{x}_i) + \boldsymbol{w}^t \boldsymbol{x}_i + \beta \\ &= \left(\sum_{j|v_j \in N(v_i)} \alpha_{ij} \left(t_j - \boldsymbol{w}^t \boldsymbol{x}_j\right)\right) + \boldsymbol{w}^t \boldsymbol{x}_i \\ &= \tau(G, \boldsymbol{x}_i) + \boldsymbol{w}^t \boldsymbol{x}_i\end{aligned}$$

where $\tau(G, \boldsymbol{x}_i)$ is the population-based correction factor. Hence, combining the EPN with a linear regression model results in a solution that offers a great compromise between flexibility and simplicity in a prediction task.

3.2. Using the EPN to predict VO_2 peak with self-reported variables

In this section, we excluded the variables recorded from the walking test (6WMD and HR_{end}), and added the patient's biological sex to predict the VO_2 peak. *The goal was to evaluate a more efficient strategy which depends only on clinical variables that can be self-reported by the patient.* We assessed three baseline models: linear regression, random forest and XGBoost, which we then combined with the EPN. Table 3 shows that all three baseline models outperformed the results obtained in Table 2 using the equation of Labonté et al. (2020), both before and after combination with the EPN. In Appendix C, we detail the performances of all baselines when combining biological sex and walking variables with the rest of the predictors. Specifically, in Table 3, the inclusion of biological sex led to lower MAEs and MAPEs along with a higher Spearman R scores compared to the best model using Labonté et al. (2020) predictors,

demonstrating the importance of biological sex in estimating the VO_2 peak. These findings align with the observed p-value in Table 1 and are consistent with patterns observed in the general population (Santisteban et al., 2022).

Next, adjusting predictions with the EPN decreased the MAPE of all three baseline models and MAE of tree-based models (Random Forest and XGboost). Nonetheless, it only slightly increased the Spearman R score of the linear regression (Table 3). This shows that the EPN was able to improve the accuracy of predictions by diminishing overestimation or underestimation of VO_2 peak, but did not enhance their ranks in correlation to the targets. Furthermore, statistical tests did not reveal a significant improvement. We believe that further experiments should consider a larger set of participants for a more accurate measurement of the benefits of the EPN with higher statistical power.

Overall, among our experiments, XGBoost used in conjunction with the EPN achieved the best performance. This final model relies solely on clinical variables that can be self-reported by the patient, eliminating the need for a walking test that requires time, financial resources and energy from the patient.

4. Discussion

Over the years, efforts have been pursued towards the development of better personalized follow-up methods for childhood ALL survivors using data from the PETALE study (Labonté et al., 2020; England et al., 2017; Morel et al., 2018; Nadeau et al., 2019; Caubet F. et al., 2019; Caru et al., 2019). Other recent works have presented interesting results associated to the use of neural networks in prediction tasks related to clinical contexts (Choi et al., 2016; Ma et al., 2017). However, until now, solutions employing neural networks have not been explored to monitor cardiac health in childhood ALL survivors. In our work, we developed a novel graph-based approach called Error Passing Network (EPN) and evaluated it in the context of VO_2 peak prediction. Our proposed solution leverages relationships between patients to improve predictions made by any machine learning model, by propagating the residuals observed on training samples. In addition to contributing to better precision medicine, our method constitutes a promising avenue for the development of artificial intelligence in clinical settings with small patient cohorts.

Along with its opportunities of involvement in healthcare, the EPN also proposes promising advances in the realm of graph-based machine learning. Compared to commonly employed GNNs (Veličković et al., 2018; Kipf and Welling, 2017), the EPN proposes a single-layered architecture that not only uses nodes' features, but also integrates their targets within the message passing procedure. Most importantly, while doing so, the EPN exploits targets in an inductive rather than transductive manner. More precisely, since it is trained to only consider target information coming from the training samples, it can be further applied to predict values for unseen nodes (i.e., unseen childhood ALL survivors). This is an important advantage over recently proposed graph-based method (Huang et al., 2020), which integrates test samples during residual propagation. Additionally, the EPN offers more flexibility than the latter approach since it propagates the residuals based on importance scores generated from a learned similarity kernel rather than edges' weights determined solely from the graph topology. Hence, it is free from the need of having a pre-defined graph structure with the dataset, as we demonstrated in this work. Although other similarity-based approaches such as k-NN do not rely on any graph structure, they come with the drawback of manually selecting or engineering an optimal similarity metric. Hence, inductively learning a kernel enables avoiding this task.

We began by applying the EPN to the disease-specific VO_2 peak equation established by (Labonté et al., 2020). The VO_2 peak is the gold standard for measuring the cardiorespiratory fitness (Smart, 2013), which in turn is a key element for the prevention of LAEs such as obesity, cholesterol and depression in childhood ALL survivors (Lemay et al., 2019). The EPN ended up having a significant impact by reducing the equation's MAPE from 26.36% to 15.70% (Table 2). We further demonstrated the interpretability of the EPN by analyzing the behavior of its attention mechanism with a targeted test patient, showing how residuals of training patients are combined to adjust a VO_2 peak prediction.

Next, we developed a new VO_2 peak prediction model in childhood ALL survivors by combining the EPN with XGBoost (Chen and Guestrin, 2016). Our model achieved better performance than the last approach (Labonté et al., 2020) (MAPE of 14.20% compared to 26.36%), the previously published model seems to perform poorly due to not including the biological sex. In addition, our solution does not rely on

Table 3: Comparison of baseline methods, with and without the EPN, in the prediction of VO_2 peak. In this experiment, biological sex is included while walk variables (6WMD and HR_{end}) are excluded.

	MAE	MAPE	Spearman R
Linear regression	4.47 ± 0.85	15.02% ± 3.48	0.70 ± 0.09
Random Forest	4.35 ± 0.49	14.65% ± 1.96	**0.75 ± 0.04**
XGBoost	4.42 ± 0.37	14.78% ± 1.32	0.74 ± 0.03
Linear regression + EPN	4.49 ± 0.66	14.94% ± 2.68	0.71 ± 0.07
Random Forest + EPN	4.30 ± 0.47	14.48% ± 1.68	**0.75 ± 0.04**
XGBoost + EPN	**4.28 ± 0.39**	**14.20% ± 1.16**	0.74 ± 0.03

a walking test (e.g., 6MWT). The removal of this constraint represents a strong advantage in the context of healthcare considering that the 6MWT requires time and financial resources. In fact, our model facilitates the prediction of VO2 peak for clinicians since the variables needed by the model can be self-reported by the survivors (age, DT, biological sex, MVLPA, weight) and/or accessed directly from their medical records. Therefore, our model could be associated to an online survey that survivors would be asked to fill periodically. The resulting predictions could be further analyzed by an exercise physiologist with the support of an interface providing comparisons between the current patient and the most similar survivors for which the VO_2 peak is already known (Figure 4(c)). We acknowledge that the deployment of a model working with self-reported variables comes with challenges, considering that self-reported values can be noisy. Statistical correction techniques should be eventually investigated to account for the noise in the self-reported MVLPA and weight variables. Nonetheless, our solution still presents promising results given that it was in itself evaluated with self-reported MVLPA from the PETALE study.

We further highlighted limitations that need to be addressed in future studies. Firstly, considering that the number of edges grows exponentially with the number of nodes in a fully connected graph, our current solution with such type of graph may not be scalable with larger training sets (i.e., cohorts) with thousands of patients. Future work on larger datasets should explore the use of more efficient attention mechanisms that propose a linear complexity (Wang et al., 2020; Liu et al., 2022). Alternatively, selecting the k nearest neighbors of each patient could reduce the number of connections in the graph. However, as mentioned earlier, this approach requires careful selection of the similarity metric and number of neighbors. Secondly, only a small number of samples were available in the PETALE dataset. Therefore, the scores obtained in Section 3.2 might not be fully representative of future performance on external datasets. In addition, all survivors considered in this study were from a monocentric cohort comprised of individuals that had only European origins. Hence, our current findings may not translate to other ethnicity groups of childhood ALL survivors. Future research should focus on the evaluation and the optimization of our methods on external cohorts with a greater variety of ethnic groups. It is our responsibility to remove systemic biases from our methods and ensure that artificial intelligence solutions deployed in healthcare are accessible and accurate to all patients. For this purpose, we believe the EPN framework may have a role to play in the assessment of fairness of the algorithms by potentially helping to pinpoint whether the errors of a given algorithm are concentrated on a specific category of nodes (e.g., patients of a given race).

In conclusion, we developed the Error Passsing Network (EPN), a novel graph-based method that adjusts predictions of any machine learning model by propagating residuals between samples. We evaluated our model in the context of VO_2 peak prediction for childhood ALL survivors, and demonstrated its superiority over previously established standards. Our proposed EPN architecture is model-agnostic and could be applied to different types of medical problems.

References

T. Akiba, S. Sano, T Yanase, T. Ohta, and M. Koyama. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Confer-*

ence on Knowledge Discovery & Data Mining, page 2623–2631, 2019.

S Arık and T. Pfister. Tabnet: Attentive interpretable tabular learning. In *AAAI*, volume 35, pages 6679–6687, 2021.

M. Caru et al. Identification of genetic association between cardiorespiratory fitness and the trainability genes in childhood acute lymphoblastic leukemia survivors. *BMC Cancer*, 19(1):443, 2019.

Maxime Caru et al. Maximal cardiopulmonary exercise testing in childhood acute lymphoblastic leukemia survivors exposed to chemotherapy. *Supportive Care in Cancer*, 29:987–996, 2021.

M. Caubet F. et al. A bayesian multivariate latent t-regression model for assessing the association between corticosteroid and cranial radiation exposures and cardiometabolic complications in survivors of childhood acute lymphoblastic leukemia: a PETALE study. *BMC Medical Research Methodology*, 19(1):100, 2019.

T. Chen and C. Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 33, page 785–794, 2016.

Y. Chen, L. Wu, and M. Zaki. Iterative Deep Graph Learning for Graph Neural Networks: Better and Robust Node Embeddings. In *Advances in Neural Information Processing Systems*, volume 33, page 19314–19326, 2020.

E. Choi, MT. Bahadori, A. Schuetz, Stewart WF., and J. Sun. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. In *Proceedings of the 1st Machine Learning for Healthcare Conference*, pages 301–318, 2016.

J. England et al. Genomic determinants of long-term cardiometabolic complications in childhood acute lymphoblastic leukemia survivors. *BMC Cancer*, 17(1):751, 2017.

FL. Fan, J. Xiong, M. Li, and G. Wang. On Interpretability of Artificial Neural Networks: A Survey. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 5(6):741–760, 2021.

B. Fatemi and SM. El A., L. amd Kazemi. SLAPS: Self-Supervision Improves Structure Learning for Graph Neural Networks. In *Advances in Neural Information Processing Systems*, volume 34, page 22667–22681, 2021.

Nadeau Geneviève, Yazdanpanah Mojgan, Yazdanpanah Nahid, Forgetta Vincenzo, Girard Simon, Sinnett Daniel, Krajinovic Maja, Alos Nathalie, and Manousaki Despoina. Genetic susceptibility and late bone outcomes in childhood acute lymphoblastic leukemia survivors. *Journal of Bone and Mineral Research*, page zjad013, 2024.

Nikolaus Hansen. The CMA Evolution Strategy: A Tutorial, 2023.

Qian Huang, Horace He, Abhay Singh, Ser-Nam Lim, and Austin Benson. Combining Label Propagation and Simple Models out-performs Graph Neural Networks. In *International Conference on Learning Representations*, 2020.

MM. Hudson et al. Long-term Follow-up Care for Childhood, Adolescent, and Young Adult Cancer Survivors. *Pediatrics*, 148(3):e2021053127, 2021.

SP. Hunger et al. Improved survival for children and adolescents with acute lymphoblastic leukemia between 1990 and 2005: a report from the children's oncology group. *Journal of Clinical Oncology*, 30 (14):1663–1669, 2012.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *Proceedings of the 5th International Conference on Learning Representations*, 2017.

L. Labonté et al. Developing and validating equations to predict VO_2 peak from the 6MWT in Childhood ALL Survivors. *Disability and Rehabilitation*, pages 1–8, 2020.

V. Lemay et al. Prevention of Long-term Adverse Health Outcomes With Cardiorespiratory Fitness and Physical Activity in Childhood Acute Lymphoblastic Leukemia Survivors. *Journal of Pediatric Hematology/Oncology*, 41(7):e450–e458, 2019.

Jing Liu, Zizheng Pan, Haoyu He, Jianfei Cai, and Bohan Zhuang. Ecoformer: Energy-saving attention with linear complexity. *Advances in Neural Information Processing Systems*, 35:10295–10308, 2022.

Z. Liu, X. Li, H. Peng, L. He, and PS. Yu. Heterogeneous similarity graph neural network on electronic health records. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 1196–1205. IEEE, 2020.

H. Lu and S. Uddin. A weighted patient network-based framework for predicting chronic diseases using graph neural networks. *Scientific Reports*, 11 (1):22607, 2021.

SM. Lundberg and SI Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 4768–4777, 2017.

F. Ma et al. Dipole: Diagnosis Prediction in Healthcare via Attention-based Bidirectional Recurrent Neural Networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 1903–1911, 2017.

James G MacKinnon. Bootstrap hypothesis testing. *Handbook of computational econometrics*, pages 183–213, 2009.

S. Marcoux et al. The PETALE study: Late adverse effects and biomarkers in childhood acute lymphoblastic leukemia survivors. *Pediatric Blood & Cancer*, 64(6):e26361, 2017.

Pablo Martínez-Camblor and Norberto Corral. A general bootstrap algorithm for hypothesis testing. *Journal of Statistical Planning and Inference*, 142 (2):589–600, 2012.

D. Mizrahi et al. The 6-minute walk test is a good predictor of cardiorespiratory fitness in childhood cancer survivors when access to comprehensive testing is limited. *International Journal of Cancer*, pages 847–855, 2020.

S. Morel et al. Development and relative validation of a food frequency questionnaire for French-Canadian adolescent and young adult survivors of acute lymphoblastic leukemia. *Nutrition Journal*, 17(1):45, 2018.

G. Nadeau et al. Identification of genetic variants associated with skeletal muscle function deficit in childhood acute lymphoblastic leukemia survivors. *Pharmacogenomics and Personalized Medicine*, 12: 33–45, 2019.

PC. Nathan, K. Wasilewski-Masker, and LA. Janzen. Long-term outcomes in survivors of childhood acute lymphoblastic leukemia. *Hematology/Oncology Clinics of North America*, 23(5): 1065–1082, 2009.

Nicolas Papernot and Patrick McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv preprint arXiv:1803.04765*, 2018.

A. Paszke et al. PyTorch: an imperative style, high-performance deep learning library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 8026–8037, 2019.

F. Pedregosa et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Y. Qian, P. Expert, P. Panzarasa, and M. Barahona. Geometric graphs from data to aid classification tasks with Graph Convolutional Networks. *Patterns*, 2:100237, 2021.

KJ. Santisteban, AT. Lovering, Halliwill JR., and CT. Minson. Sex differences in vo2max and the impact on endurance-exercise performance. *International Journal of Environmental Research and Public Health*, 19(9):4946, 2022.

Neil A. Smart. How do cardiorespiratory fitness improvements vary with physical training modality in heart failure patients? a quantitative guide. *Experimental & Clinical Cardiology*, 18(1):e21–e25, 2013.

Skoczen Szymon, Miroslaw Bik-Multanowski, Walentyna Balwierz, Jacek J Pietrzyk, Marcin Surmiak, Wojciech Strojny, Danuta Galicka-Latala, and Jolanta Gozdzik. Homozygosity for the rs9939609t allele of the fto gene may have protective effect on becoming overweight in survivors of childhood acute lymphoblastic leukaemia. *Journal of genetics*, 90(2):365, 2011.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz

Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph Attention Networks. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

Bernard L Welch. The generalization of 'student's'problem when several different population varlances are involved. *Biometrika*, 34(1-2):28–35, 1947.

Carmen L Wilson, Wei Liu, Jun J Yang, Guolian Kang, Rohit P Ojha, Geoffrey A Neale, Deo Kumar Srivastava, James G Gurney, Melissa M Hudson, Leslie L Robison, et al. Genetic and clinical factors associated with obesity among adult survivors of childhood cancer: a report from the st. jude lifetime cohort. *Cancer*, 121(13):2262–2270, 2015.

Carmen L Wilson, Carrie R Howell, Robyn E Partin, Lu Lu, Sue C Kaste, Daniel A Mulrooney, Ching-Hon Pui, Jennifer Q Lanctot, Deo Kumar Srivastava, Leslie L Robison, et al. Influence of fitness on health status among survivors of acute lymphoblastic leukemia. *Pediatric blood & cancer*, 65 (11):e27286, 2018.

Zitong Yang, Michal Lukasik, Vaishnavh Nagarajan, Zonglin Li, Ankit Rawat, Manzil Zaheer, Aditya K Menon, and Sanjiv Kumar. Resmem: Learn what you can and memorize the rest. *Advances in Neural Information Processing Systems*, 36, 2024.

Y. Zhang, P. Tino, A. Leonardis, and K. Tang. A Survey on Neural Network Interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742, 2021.

Appendix A. Hyperparameters

This section holds additional details related to the hyperparameters of the models and their respective search spaces (Table 4, Table 5, Table 6).

Table 4: Random forest's hyperparameters. The hyperparameters that are not mentioned were set as the default ones from version 0.24.2 of the *scikit-learn* library.

Hyperparameter	Search space
max_features	$\{\text{sqrt}, \log2\}$
max_leaf_nodes	$\{5k\}_{k=3}^{6}$
max_samples	$[0.8, 1]$
n_estimators	$\{1000 + 250k\}_{k=0}^{8}$

Table 5: XGBoost's hyperparameters. The hyperparameters that are not mentioned were set as the default ones from the scikit-learn wrapper interface of version 1.4.2 of the *xgboost* library.

Hyperparameter	Search space
max_depth	$\{1, 2, 3\}$
learning_rate	$[0.01, 0.1]$
reg_lambda	$[0.0005, 0.005]$
subsample	$[0.8, 1]$

Table 6: Error Passing Network's hyperparameters. The *weight decay* refers to the coefficient multiplying the $\mathcal{L}_2$ penalty in the mean squared error loss (MSE). The *learning rate* refers to the initial learning rate at the beginning of the training. Both parameters are given to the Adam optimizer (Kingma and Ba, 2014).

Hyperparameter	Search space
weight decay	$[0.0005, 0.005]$
learning rate	$[0.005, 0.05]$

Appendix B. Benefits of the attention mechanism

As Table 1 suggests a significant difference in VO_2 peak between men and women, we added the biological sex variable to the EPN used in conjunction with Labonté et al. (2020) equation and analyzed the attention scores. Figure 5(*a*) shows that male subjects of the test set focus their attention on male subjects of the training set, and vice versa for female subjects. Moreover, we substituted the weights e_{ij} in Equation 3 by a dot product and a cosine similarity between the test and training samples as follows: $e_{ij} = \mathbf{x}_i^t \mathbf{x}_j$ and $e_{ij} = \frac{\mathbf{x}_i^t \mathbf{x}_j}{\|x_i\| \|x_j\|}$, in order to evaluate the effectiveness of the attention mechanism and the

learnable parameters $\boldsymbol{W}^Q$ and $\boldsymbol{W}^K$. In Figure 5(b) and Figure 5(c), we presented the weights given by each patient in the test set to each patient in the training set when using a simple dot product and a cosine similarity. Although we observe a similar overall trend to the attention scores presented in Figure 5(a), where male subjects from the testing set tend towards other male subjects from the training set, and vice versa for female subjects, Table 7 shows a drop in performances with higher MAE and MAPE, and a lower Spearman R score. This demonstrates that the computed weights with dot products and cosine similarity are less effective in combining residuals than the attention scores learned with $\boldsymbol{W}^Q$ and $\boldsymbol{W}^K$ parameters.

We also optimized the number of neighbors to be considered when implementing the EPN with a non-learnable similarity kernel (dot product and cosine similarity). We varied the proportion of training samples to consider as neighbors from 10% to 100%, and the hyperparameters tuning always ended up selecting the maximum number of neighbors accross the 5 folds of cross-validation.

Appendix C. Additional experimental results

In Table 8, we show the performances of all baseline models, with and without the EPN, when including all predictors.

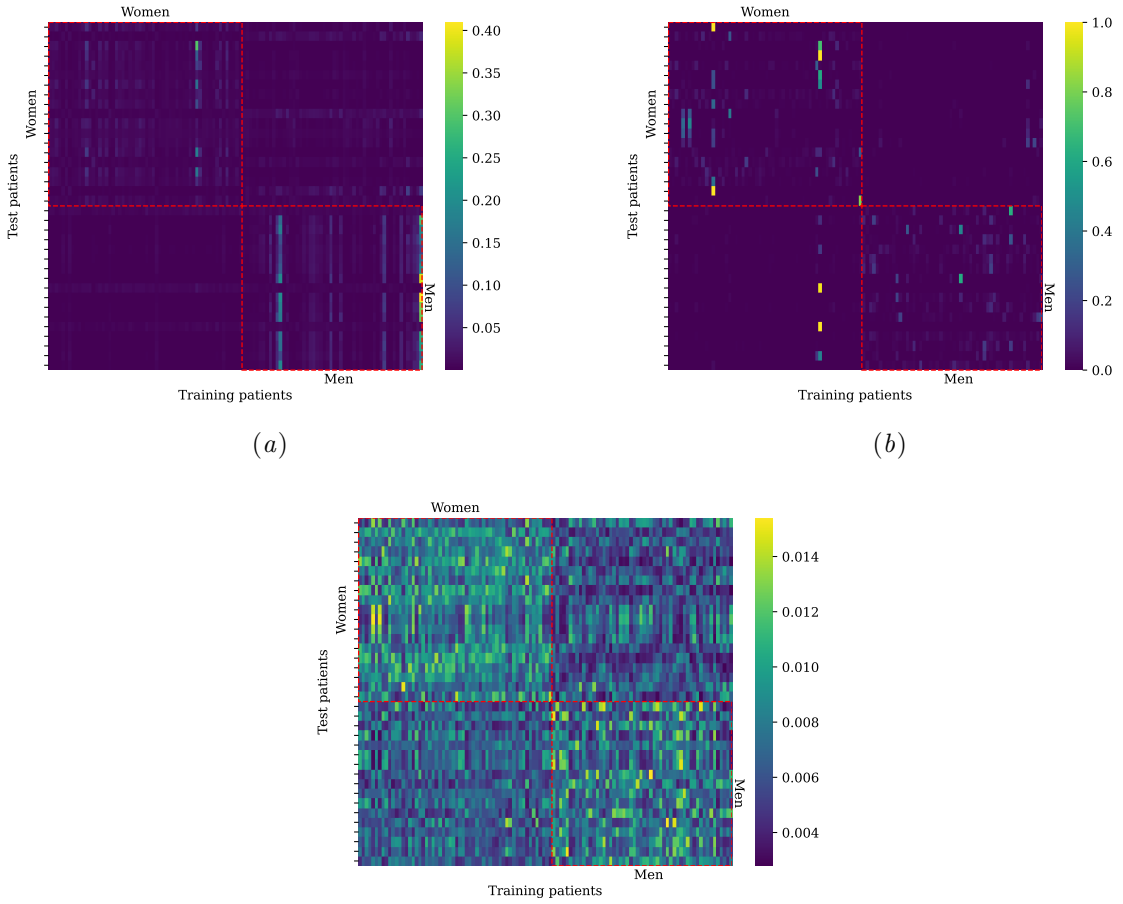

Figure 5: Post-hoc analyses of the weights assigned to training samples' residuals on one fold of the five folds of cross-validation. The weights are computed using three distinct methods. In this experiment, the EPN is used in conjunction with Labonté et al. (2020) equation and includes the biological sex in its predictors. In each set, patients are ordered according to their biological sex. Red lines distinguish between male and female subjects. (a) Residuals' weights are computed using the attention mechanism. (b) Residuals' weights are computed using a dot product between each pair of test and training samples. (c) Residuals' weights are computed using a cosine similarity between each pair of test and training samples. The color intensity scales in (a), (b) and (c) are not the same for better visibility

Table 7: Comparison of EPN using three distinct methods to compute residuals' weights: attention mechanism, dot products and cosine similarity, in the prediction of VO_2 peak. In this experiment, the EPN is used in conjunction with Labonté et al. (2020) equation and includes the biological sex in its predictors.

	MAE	MAPE	Spearman R
EPN + Attention Mechanism	4.16 ± 0.51	$14.10\% \pm 1.86$	0.76 ± 0.05
EPN + Dot Product	4.70 ± 0.40	$16.12\% \pm 2.28$	0.70 ± 0.08
EPN + Cosine Similarity	4.69 ± 0.30	$16.08\% \pm 2.01$	0.69 ± 0.07

Table 8: Comparison of baseline methods, with and without the EPN, in the prediction of VO_2 peak. In this experiment, biological sex and walk variables (6WMD and HR_{end}) are included.

	MAE	MAPE	Spearman R
Linear regression	**3.80 ± 0.5**	12.96% ± 2.19	0.78 ± 0.06
Random Forest	4.11 ± 0.34	14.00% ± 1.61	**0.80 ± 0.04**
XGBoost	4.13 ± 0.37	13.92% ± 1.6	0.78 ± 0.04
Linear regression + EPN	3.82 ± 0.46	**12.95% ± 1.93**	0.78 ± 0.06
Random Forest + EPN	4.00 ± 0.33	13.56% ± 1.27	**0.80 ± 0.05**
XGBoost + EPN	4.06 ± 0.4	13.61% ± 1.42	0.78 ± 0.04