Prototype Learning for Explainable Brain Age Prediction

Linde S. Hesse[1] Nicola K. Dinsdale[1] Ana I.L. Namburete[1,2]
[1]OMNI Lab, Department of Computer Science, University of Oxford, UK
[2]Wellcome Centre for Integrative Neuroimaging, FMRIB, University of Oxford, UK
{linde.hesse, nicola.dinsdale, ana.namburete}@cs.ox.ac.uk

Abstract

The lack of explainability of deep learning models limits the adoption of such models in clinical practice. Prototype-based models can provide inherent explainable predictions, but these have predominantly been designed for classification tasks, despite many important tasks in medical imaging being continuous regression problems. Therefore, in this work, we present ExPeRT: an explainable prototype-based model specifically designed for regression tasks. Our proposed model makes a sample prediction from the distances to a set of learned prototypes in latent space, using a weighted mean of prototype labels. The distances in latent space are regularized to be relative to label differences, and each of the prototypes can be visualized as a sample from the training set. The image-level distances are further constructed from patch-level distances, in which the patches of both images are structurally matched using optimal transport. This thus provides an example-based explanation with patch-level detail at inference time. We demonstrate our proposed model for brain age prediction on two imaging datasets: adult MR and fetal ultrasound. Our approach achieved state-of-the-art prediction performance while providing insight into the model's reasoning process.

1. Introduction

Deep learning models are typically considered to be black boxes, meaning that it is not possible to understand how a model's prediction was made. This severely limits the adoption of such methods in clinical practice, as the decision-making process needs to be transparent to understand model behavior and gain patients' trust [15]. It is thus vital to develop models that are explainable and, hence, capable of providing insight into their reasoning process [42].

The most frequently used methods to explain a model's prediction are saliency-based, which explain the prediction of a trained model *post-hoc* [49]. Saliency methods show the importance of each pixel in the input image with regard to the model's prediction. However, these explanations are not always a faithful representation of the original prediction, often resembling edge maps rather than being dependent on the trained model [1]. Furthermore, they often result in noisy saliency maps, which are hard to interpret and prone to confirmation bias [1, 2]. On the other hand, inherently explainable models reflect the model's decision-making process by design [42]. However, designing such models for medical imaging is challenging as there is typically a trade-off between performance and explainability.

An example-based explainable model for the classification of natural images was proposed by [11]. Their model architecture (*ProtoPNet*) learned a set of embeddings in latent space, referred to as *prototypes*, and used the distances to each of these prototypes to classify a new sample. Each prototype was assigned a class label, and could thus be considered as a representative example (in latent space) for that class. The prototypes in the final model were enforced to equate representations of actual training images, which made it possible to visualize the prototypes. In contrast to *post-hoc* methods, this type of architecture is thus **inherently explainable**, as the final prediction is directly generated from the prototype distances. However, many important medical image tasks are continuous regression problems, such as brain age prediction [14, 50]. As *ProtoPNet* uses the categorical labels to pull (or push) samples together (or apart), this architecture cannot be directly applied for regression.

In this work, we propose *ExPeRT*: an Explainable Prototype-based model for Regression using optimal Transport. We incorporate metric learning to map the images to an inherently continuous representation space in which the distances between images and prototypes in *latent* space are proportional to their differences in *label* space. This improves upon our earlier work for ordinal regression [20], in which the latent space was not truly continuous, but regularised with an attraction loss that worked on prototypes within a certain set label range.

The prediction for a new sample is made from a weighted average of prototype labels within a given distance. As all prototypes can be visualized, this provides an intuitive ex-

planation of the prediction for regression tasks. Unlike previous approaches [11, 20], *ExPeRT*'s prototypes are latent representations of whole training images. This is motivated by the fact that the signal of interest in continuous regression problems is often a gradual structural change, rather than class-specific patterns well represented in a single image patch.

We instead incorporate spatial detail into the model's decision-making process by decomposing each image-level latent distance into patch-wise similarities between image patches. The patch-wise similarities are computed in latent space and then structurally matched using optimal transport (OT). OT finds an optimal matching matrix that contains the *soft assignment* scores between the image patches of the sample and prototype. This matrix is then used to compute a single image-level distance (the Earth Mover's Distance [41]). The resulting matches can be inspected to verify whether anatomically corresponding patches are matched together between an image and prototype, providing a detailed spatial decomposition of the image-level distance.

The network is trained using a combination of a *metric loss*, which forces the prototype-image distances to be proportional to label differences, and a *consistency loss* that minimizes the distances between identical image patches under an anatomically justified transformation. Our network and all loss elements are fully differentiable and, thus, the network can be trained end-to-end.

We demonstrate our approach on the task of brain age prediction for both adult magnetic resonance (MR) and fetal ultrasound (US) images. Brain age prediction is an important medical imaging task: for adult MRI the difference between true and predicted age is a potential biomarker for disease [12, 14]; during gestation the predicted brain age can be compared to true post-conceptual age to quantify fetal brain development [17, 35]. We demonstrate that our architecture obtains state-of-the-art performance on both datasets while providing insight into the model's prediction process.

2. Related Work

Explainable Brain Age Prediction: Several studies have attempted to introduce explainability into brain age prediction models, predominantly for adult MRI. Saliency methods have been used to explain brain age predictions [9, 21, 28, 30, 50], but their explanations are not always consistent or comparable between different methods [16]. Other studies have used patch- or slice-based approaches to predict local brain age [3, 8]. While these methods provide a more detailed prediction, separate networks have to be trained for each slice or patch, increasing the computational overhead. Using only a single model, [40] used a U-Net to predict brain age voxelwise. The predictions were more fine-grained than slice- or patch-based approaches, but the reported prediction error was considerably higher than the error obtained with baseline models (MAE ~10 years vs ~3 years). Alternatively, in [5, 6], generative models were used to demonstrate the changes that would be expected in the image for different ages. However, training generative models is challenging and typically requires large amounts of training data which is often unavailable.

Prototype Learning: *Prototypes* are a set of feature representations in latent space learned during model training, which can be considered representative examples for a certain label. Subsequently, inference for a new sample is performed using the distances to the learned prototypes in latent space, for example by using nearest-neighbour classification [37]. In most early deep prototype learning approaches, the learned prototypes were unconstrained points in latent space [7, 23], making it challenging to visualise or interpret the learned prototypes. To achieve visualisation of the prototypes, [29] introduced an autoencoder to reconstruct an image from the latent representation. Alternatively, in [11] it was proposed to constrain the prototypes in the final model to be feature representations of real training images, which could easily be visualized. Further, the prototypes represented patches of training images, as opposed to whole images, creating a more fine-grained explanation. Variations on the *ProtoPNet* architecture have also been applied to several classification [4, 24, 33, 44, 45] and ordinal regression [20] tasks in medical imaging.

Deep Metric Learning: Prototype-based methods are closely related to metric learning approaches. Metric learning aims to learn a feature space where distances between similar samples are small and, inversely, distances between dissimilar samples are large [25, 31, 39, 51, 52]. The mapping by a neural network is typically achieved by training with contrastive or triplet losses [43]. Specifically designed for regression, [10] regularized the latent space by enforcing the feature distances between samples in a batch to be relative to their difference in label space, measured with the Euclidean distance. However, distances in feature space should be computed along the manifold (*geodesic distances*), and Euclidean distance are only a good approximation for small distances. Thus, samples were weighted by a Gaussian function so as to weigh samples close to each other more than those far apart. Inference was then done using a weighted mean of neighbouring training sample labels. However, this approach requires storing all latent training samples for inference and is dependent on larger batch sizes, which can be problematic when working with large input images due to memory constraints.

Optimal Transport (OT): OT is the mathematical optimization problem computing the shortest distance (or *lowest cost*) between two distributions, given a cost matrix. It can be optimized efficiently when an entropic regularization term [13] is added, and has been incorporated into several deep learning architectures [48]. Most relatedly, OT

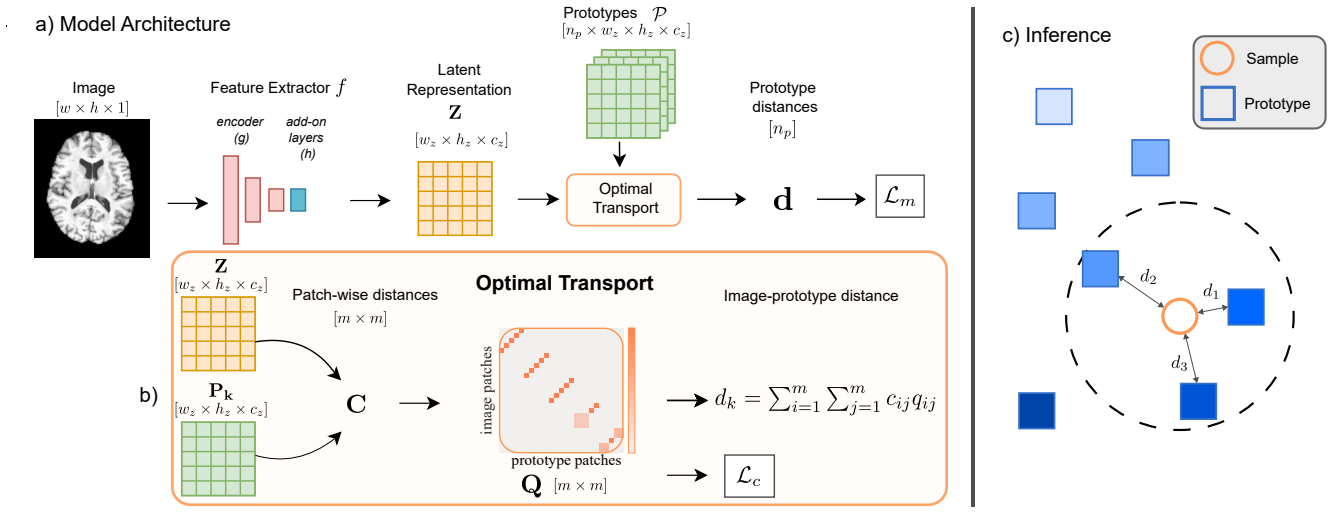

Figure 1. Schematic overview of (a) the *ExPeRT* architecture, (b) distance computation with OT, and (c) inference. The OT matching is shown for one sample and prototype, and is repeated for each prototype to obtain $\boldsymbol{d}$. The m is the number of patches per image, given by $w_z h_z$. During inference (c), a sample prediction is made using a weighted average of labels of prototypes within a certain radius.

has previously been applied to compute distances between pairs of natural images in [39, 51, 52], where OT was used to match the image patches based on patch-level distances, resulting in improved performance and explainability.

3. Methods

The proposed *ExPeRT* model aims to learn a set of prototypes in latent space, each representing a whole image from the training set. Each prototype has a continuous label (e.g. age), and sample predictions are made using a weighted mean of distances to these prototypes. To incorporate patch-level detail, the image-level distances between the image and prototypes are decomposed into patch-level distances and structurally matched using OT matching [39, 51, 52]. A schematic overview of the architecture is shown in Fig. 1a.

3.1. Network Architecture

Given an image $\mathbf{X} \in \mathbb{R}^{w \times h \times 1}$, we aim to learn a feature extractor f that can extract a latent representation of $\mathbf{X}$, denoted by $\mathbf{Z} \in \mathbb{R}^{w_z \times h_z \times c_z}$, with w_z and h_z the spatial dimensions and c_z the channel dimension. The feature extractor f is composed of a base encoder (g), and an additional block (h) with two convolutional layers with 1×1 kernels and a sigmoid as the last activation. Simultaneously, we learn a set of *prototypes*, which are essentially learned embeddings in latent space, each of the same size as $\mathbf{Z}$: $\mathcal{P} = \{\mathbf{P_i} \in \mathbb{R}^{w_z \times h_z \times c_z}, \forall i \in [1, n_p]\}$, with n_p as the number of prototypes. Each of the prototypes also has an assigned label, resulting in a vector of prototype labels, denoted by $\boldsymbol{y^{proto}}$. These prototype labels are assigned uniformly at the beginning of training within the label range present in the training dataset and are fixed during training. The prototypes themselves are considered model parameters, and can, therefore, be trained end-to-end with the feature extractor. For each sample, the network computes the distances to each of the prototypes in latent space, resulting in a vector of distances of length n_p, denoted by $\boldsymbol{d}$.

3.2. Prototype Projection

During training, the prototypes are updated with each step and can, therefore, be located throughout the latent space. However, following [11], every N epochs these are replaced by the closest latent representation of an image in the training dataset, referred to as the *prototype projection*. Checkpoints are saved only straight after the projection, ensuring that each prototype can be visualized using the corresponding image from the training set.

3.3. Distance Metric Loss

The distances between samples and prototypes are regularised, so that for a certain sample with ground-truth label y, the distance in latent space to a prototype k is proportional to its difference in label space: $d_k \propto |y_k^{proto} - y|$, with d_k and y_k^{proto} as the distance to and label of the kth prototype (i.e. the kth element of $\boldsymbol{d}$ and $\boldsymbol{y^{proto}}$), respectively.

However, as the latent space exists on a high-dimensional manifold, computing the feature distances on the manifold between prototype and sample, the geodesic distance, is non-trivial. In this work, we approximate the geodesic distance in the local neighborhood of a sample representation using the Euclidean distance, which is a good approximation for small distances on the manifold [47].

Building on [10], our metric loss regularises distances between samples in the batch and each of the prototypes in

the *local neighborhood*. The neighborhood is determined by differences in labels between prototypes and samples, as opposed to the feature distance. The total loss for a single sample with ground-truth label y is thus given by:

$$\mathcal{L}_m(\boldsymbol{d}, \boldsymbol{y^{proto}}, y) = \sum_{k=1}^{n_p}(|s \cdot d_k - (|y_k^{proto} - y|)|)w_k^{train} \tag{1}$$

where s is a learnable parameter scaling the feature distances to the label differences and w_k^{train} a weight of the k^{th} prototype. w_k^{train} weights the prototype with a Gaussian function based on the label differences, defined by:

$$w_k^{train} = e^{-\frac{|y_k^{proto} - y|}{2\sigma^2}} + \alpha \tag{2}$$

where σ controls the size of the neighborhood (i.e. the standard deviation of the Gaussian kernel). α is a small number to prevent the latent embeddings from tangling: without it, samples and prototypes with a large label difference have a negligible effect on each other and could, therefore, be embedded close together in feature space.

Patch-based Distance Metric In order to train the network with the distance metric loss, distances need to be computed between the sample and each of the prototypes, both of size $h_z \times w_z \times c_z$. A common approach is to use an average pooling operator to compress the spatial dimensions, resulting in a vector of size c_z. The Euclidean distance can then be computed between these vectors [10, 53], however, this discards all spatial information. To provide information about the spatial makeup of the distance between a prototype and sample, we propose to instead use a patch-based distance metric.

The latent representation $\mathbf{Z}$ and a single prototype $\mathbf{P_i}$ can both be considered as sets of m feature vectors: $\{\boldsymbol{z_1}, \boldsymbol{z_2}, ...\boldsymbol{z_m}\}$ and $\{\boldsymbol{p_1}, \boldsymbol{p_2}, ...\boldsymbol{p_m}\}$, with $m = h_z w_z$, and each vector is of size c_z. The proposed distance metric computes the Euclidean distances between both sets of feature vectors, $d(\boldsymbol{z_i}, \boldsymbol{p_j}) \forall i, j \in \{1, ..., m\}$, resulting in a cost matrix $\mathbf{C} \in \mathbb{R}^{m \times m}$. As not all pairwise distances should contribute equally to the image-level distance (i.e. the distance between a patch containing skull and one containing the ventricles is not very meaningful), we use OT to obtain a matching matrix, $\mathbf{Q} \in \mathbb{R}^{m \times m}$ (see below). This matrix $\mathbf{Q}$ can be considered as a *soft assignment* matrix, in which each feature vector $\boldsymbol{z_i}$ can be partially matched to more than one feature vector $\boldsymbol{p_j}$. These matching and cost matrices can subsequently be used to compute the image-level distance between the sample and the *k*th prototype as follows:

$$d_k = \sum_{i=1}^{m}\sum_{j=1}^{m} c_{ij} q_{ij} \tag{3}$$

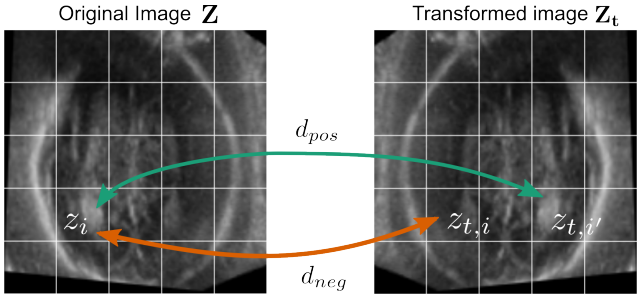

Figure 2. The used consistency loss between a latent image representation ($\mathbf{Z}$) and the representation of a transformed image ($\mathbf{Z_t}$) for a single triplet. The grid indicates the latent pixel size.

where c_{ij} and q_{ij} are the elements of $\mathbf{C}$ and $\mathbf{Q}$, respectively.

Optimal Transport (OT) OT aims to find the matching matrix (or *optimal flow*) that results in the minimal distance (or *lowest cost*) between two distributions. For discrete distributions, given the cost or distance between individual elements, this can be formalized as:

$$\min_{\mathbf{Q}} \sum_{i=1}^{m}\sum_{j=1}^{m} c_{ij} q_{ij} \tag{4}$$

which is constrained by the fact that the matching matrix $\mathbf{Q}$ needs to sum up to the initial marginal distributions given by $\boldsymbol{w_1}$ and $\boldsymbol{w_2}$:

$$\sum_{i=1}^{m} q_{ij} = \boldsymbol{w_1}, \sum_{j=1}^{m} q_{ij} = \boldsymbol{w_2} \tag{5}$$

This problem is computationally expensive to solve, and so [13] introduced *entropic regularization* to smooth the optimization problem. The resulting system can efficiently be solved using the classical Sinkhorn divergence algorithm [26, 46]. Entropic regularization transforms the optimization of Eq. 4 into the following optimization problem:

$$\min_{\mathbf{Q}} \sum_{i=1}^{m}\sum_{j=1}^{m} c_{ij} q_{ij} + \epsilon H(\mathbf{Q}) \tag{6}$$

where $H(\mathbf{Q})$ denotes the entropy function, given by: $H(\mathbf{Q}) = -\sum_{j=i}^{m}\sum_{j=1}^{m} q_{ij} log(q_{ij})$, and ϵ the strength of the entropy regularization. After computing the matching matrix, the minimal distance between the two distributions can simply be computed with Eq. 3.

The initial marginal distributions used in the optimization can be considered as importance scores or weights for each of the feature vectors $\boldsymbol{z}_i$ and $\boldsymbol{p}_j$. We set both distributions to uniform, but other options could be considered [39, 52]. As the OT matching is fully differentiable [26], we can train our whole network end-to-end.

3.4. Consistency Loss

The OT optimization aims to find structural matches between the image and prototype. To achieve anatomically correct matches, we also introduce a consistency loss. It considers the distances between feature vectors in the latent representation of an image $\boldsymbol{z_i} \in \mathbf{Z} = f(\mathbf{X})$ and the feature vectors of the same image under an anatomically justified transformation $T : \boldsymbol{z_{t,i}} \in \mathbf{Z_t} = f(T(\mathbf{X}))$ and encourages distances of the same image content, $d(\boldsymbol{z_i}, \boldsymbol{z_{t,i'}})$ with $i' = T(i)$, to be small, while encouraging distances between patches of the same spatial location, $d(\boldsymbol{z_i}, \boldsymbol{z_{t,i}})$, to be large. This resembles a contrastive learning problem where certain feature vectors are pulled together (positive pairs) whereas others are pushed apart (negative pairs). Therefore, our contrastive loss is formulated as a triplet loss, encouraging the distance between an anchor and a positive vector to be larger than the distance between an anchor and a negative vector by a certain margin γ [43]. To create triplets, we used all vectors $\boldsymbol{z_i}$ as anchors, and the respective vectors in $\boldsymbol{z_t}$ as positive ($\boldsymbol{z_{t,i'}}$) and negative ($\boldsymbol{z_{t,i}}$) samples, excluding any triplets for which the positive and negative sample were the same. The Euclidean distance was then computed from each anchor to the positive sample, d_{pos}, and the negative sample, d_{neg}. The total consistency loss per image computes the sum over all triplets, and can be described by:

$$\mathcal{L}_c = \sum_{i=1}^{m} max(d(\boldsymbol{z_i}, \boldsymbol{z_{t,i'}}) - d(\boldsymbol{z_i}, \boldsymbol{z_{t,i}}) + \gamma, 0)[\boldsymbol{z_{t,i'}} \neq \boldsymbol{z_{t,i}}] \tag{7}$$

For training, we used a combination of the distance metric loss (Eq. 1) and consistency loss as: $\mathcal{L}_{total} = \mathcal{L}_m + \beta\mathcal{L}_c$, where β is the weight of the consistency loss.

3.5. Inference

At inference, the prediction for a new sample is made using a weighted average of the prototype labels within a certain radius r (Fig. 1b), given by:

$$\hat{y} = \frac{\sum_{k=1}^{n_p} w_k^{test} y_k^{proto}}{\sum_{k=1}^{n_p} w_k^{test}}, \quad w_k^{test} = \begin{cases} e^{-\frac{s \cdot d_k}{2(r/3)^2}}, & \text{if } s \cdot d_k \leq r \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

in which the weight of each prototype is thus determined by a Gaussian with standard deviation $r/3$.

4. Experiments

4.1. Datasets

Fetal Ultrasound: We used 2D ultrasound images sampled from 3D volumes acquired as part of the INTERGROWTH-21st Fetal Growth Longitudinal Study [36] to predict gestational age. We used a total of 4290 volumes between 14 and 31 gestational weeks, selected based on having sufficient ultrasound quality [35]. All 3D volumes were aligned to the same coordinate system and scaled to the average brain size at 30 GW using an automated alignment method [34]. Scaling was performed to enforce the network to learn patterns of structural development rather than only the volumetric size. After alignment, the 2D trans-ventricular plane was extracted from each of the 3D volumes, resulting in a total set of 4290 2D images of size 160×160.

Adult MRI: T1 MRI images from the IXI dataset [22] were used for brain age prediction, with ages between 19 and 86. The volumes were preprocessed using the FSL Anat pipeline [18]. Only subjects that completed the pipeline successfully were included. For each subject, an axial plane containing the ventricles was selected, as the increase in ventricle size with ageing is well established. This resulted in 561 MR images, each of size 160×192. More details for both datasets are given in the supplementary material.

4.2. Implementation

We implemented our proposed method with Pytorch Lightning, using PyTorch 1.13 and Python 3.10. All code is publicly available at: `github.com/lindehesse/ExPeRT_Code`. Experiments were performed on an A10 GPU with 24 GB RAM and we used an implementation of the Sinkhorn algorithm in logarithmic space to avoid instabilities with training. The Sinkhorn algorithm was run for a maximum of 25 iterations, and the weight of the entropic regularization was set to 0.1 (ϵ in Eq. 6). We used a ResNet-18 as the base encoder, g, pre-trained on ImageNet.

Five-fold cross-validation (train/test 80%/20%) was used to tune all hyperparameters, with the reported test performance being the average across folds. Networks were trained for 300 epochs for the US data and 150 for the MRI, with a learning rate of 5×10^{-4} and 1×10^{-4}, respectively. During training, the images were augmented using small random affine transformations (rotation up to 15 degrees, translation +-6 pixels, scaling between 0.95-1.05). The prototypes were projected to the closest sample in the training set each 25 epochs, starting from 75 epochs. Unless otherwise reported, for both datasets α and σ in Eq. 2 were set to 0.05 and 1, respectively, and r in Eq. 8 to 3; the number of channels in the additional layer block, h, and prototypes, c_z, to 512, and, the number of prototypes, n_p, to 100. The latent space dimensions w_z and h_z were both 5 for the US dataset, and 5 and 6, respectively, for the MRI dataset. During inference, only prototypes within a certain radius r of a sample are considered when generating the final prediction, and it is possible to have no prototypes are within this area, resulting in no prediction. In order to report performance, we assigned the label of the closest prototype to the sample.

Due to shadowing artefacts of the fetal skull, usually only one of the two brain hemispheres is clearly visible in

fetal brain US images [32]. To match patches in the visible hemispheres with each other, the consistency loss for US was implemented with horizontal flips along the midline of the brain as a geometric transformation (see Fig. 2). The weight of the consistency loss β, was set to 10, and the margin, γ to 0.1. For the MR images, the same consistency loss was applied as for the US images.

4.3. Results

Baseline Comparisons The quantitative prediction results for both US and MRI are shown in Table 1 and in Fig. 3a and c. We compared our $ExPeRT$ architecture to several non-interpretable age prediction baseline models: the SFCN model + KL loss [38], winner of the most recent PAC MR brain age prediction challenge; a ResNet-18 + CORAL loss, which showed state-of-the-art age prediction performance for fetal brain age prediction in [27]; a vanilla ResNet-18 + MSE loss; and a SFCN + MSE loss, the architecture proposed in [38] but formed as a regression task (as opposed to binning the ages in classes [38]). For both datasets the performance of our method slightly outperforms the baselines, showing that the common assumption that increased explainability results in a decrease in prediction performance does not hold. However, achieving increased prediction performance is not the main aim of this study: rather, we aimed to create a more explainable model without compromising on prediction performance.

Finally, we also compared performance to our previous work [20], an interpretable network proposed for ordinal regression (INSightR-Net). It can be seen that the performance of that model is lower than most other baselines. Furthermore, the learned prototypes did not correspond to age-discriminative image regions. Visualizations of the learned prototypes in INSightR-Net and implementation details for all baselines are given in the supplementary material.

Example Predictions Figures 3b and d show example predictions from our model. The prediction is composed of a weighted average of the labels of prototypes within a certain radius, and these neighboring prototypes are shown next to the sample image. The weight of each prototype is inferred from its distance using a Gaussian function, shown on the right. The color of each point indicates the prototype label, which is also shown with a colored border around each image. In addition to visualizing the prototypes used to make a prediction, our explanation can also decompose the computed distance between each prototype and sample into patch-level matching matrices, which are shown in Fig. 4. For each sample-prototype pair, the matching matrix is given in the last column of each panel, indicating which patches are most similar between the prototype and sample. The reshuffled prototype shows the OT matching more intuitively, in which, for each image location, the prototype patch with the highest matching score is shown. In the US dataset, the two visible hemispheres are correctly matched (Fig. 4.1a, c) by the full model, illustrating the advantage of using OT matching in our network.

Table 1. Average MAE (↓) across the five cross-validation folds on the test set, with the average standard deviations in brackets. AvgPool represents an ablation of the OT patch-based matching.

	Ablat.		US	MRI
	$\mathcal{L}_c$	h	MAE [day]	MAE [yr]
Classification				
SFCN [38]	-	-	4.14 (3.31)	7.26 (5.27)
CORAL [27]	-	-	4.14 (3.33)	5.89 (4.96)
Regression				
SFCN [38] + MSE	-	-	5.03 (3.91)	6.17 (4.30)
ResNet-18 [19]	-	-	4.10 (3.33)	5.89 (4.46)
INSightR-Net [20]	-	-	4.22 (3.41)	6.43 (5.10)
AvgPool	×	✓	4.08 (3.28)	5.80 (5.25)
ExPeRT	×	×	4.18 (3.43)	6.41 (5.33)
ExPeRT	×	✓	4.01 (3.27)	5.69 (4.59)
ExPeRT	✓	×	4.16 (3.43)	8.02 (6.41)
ExPeRT	✓	✓	**3.99** (3.17)	**5.57** (4.51)

The types of explanations we obtained are very different from more classical explainability approaches, such as saliency methods where a heat map with pixel-level importance scores is generated. We do not aim to compare directly with these kinds of methods but propose our method as an alternative way to increase the explainability of neural networks. Furthermore, while in this study uniform initial distributions were used in Eq. 5 (i.e. all patches get the same weight), this could be replaced by other options, such as cross-correlation between the patches [52], to generate importance scores for each of the patches, further improving the explainability of the model.

Ablations An ablation study on the consistency loss ($\mathcal{L}_c$) and add-on layers (h) was completed, and a further ablation study on the OT matching, *AvgPool*, where OT was replaced by pooling the feature representations into a single 1D vector and computing Euclidean distances between these. As the consistency loss works on patch distances, this loss was not applied for the *AvgPool* ablation. All ablation results are shown in Table 1, from which it is evident that for both datasets the add-on layers (h) improve the prediction performance. Furthermore, the OT patch-based matching improves performance compared to the *AvgPool* ablation.

The effect of the consistency loss ($\mathcal{L}_c$) is best observed for the US dataset in Fig. 4, which shows sample-prototype pairs for the full model trained with (4a,c) and without

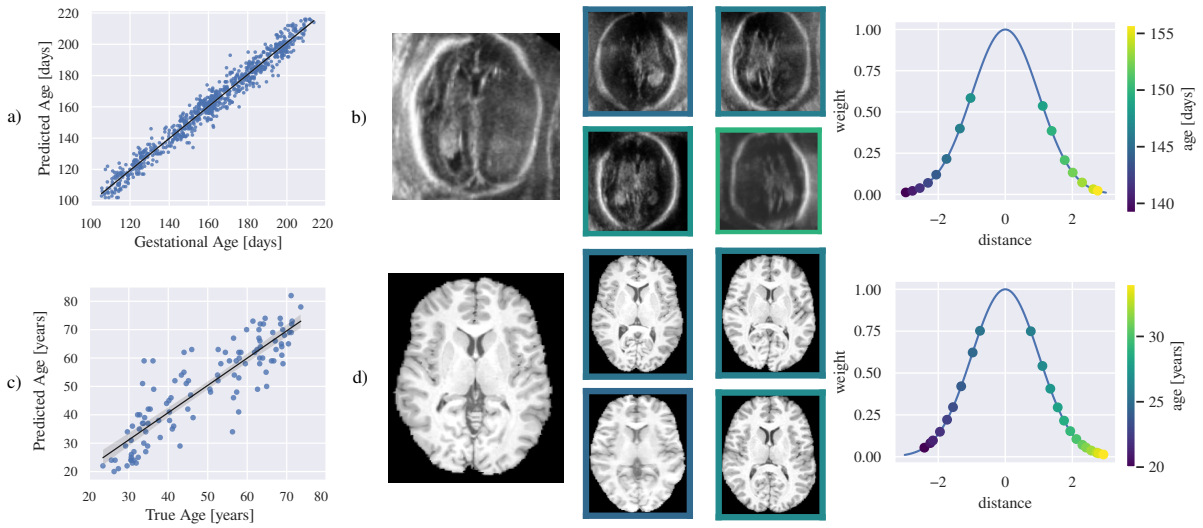

Figure 3. Predicted versus ground-truth age for fetal US (a) and adult MRI (c), as well as an example prediction for US (b) and MRI (d). In b and d, the sample is shown on the left, and the four closest prototypes are shown in the middle. The right-most graphs indicate the weight of each of the prototypes in the final prediction. Prototypes with a label below the predicted value were plotted with negative distances for visualization purposes, but unsigned distances were used in the model itself. For the US image, both the predicted and ground-truth age were 148 days and for the MRI the ground-truth age was 24 years, and the predicted age 25.6 years.

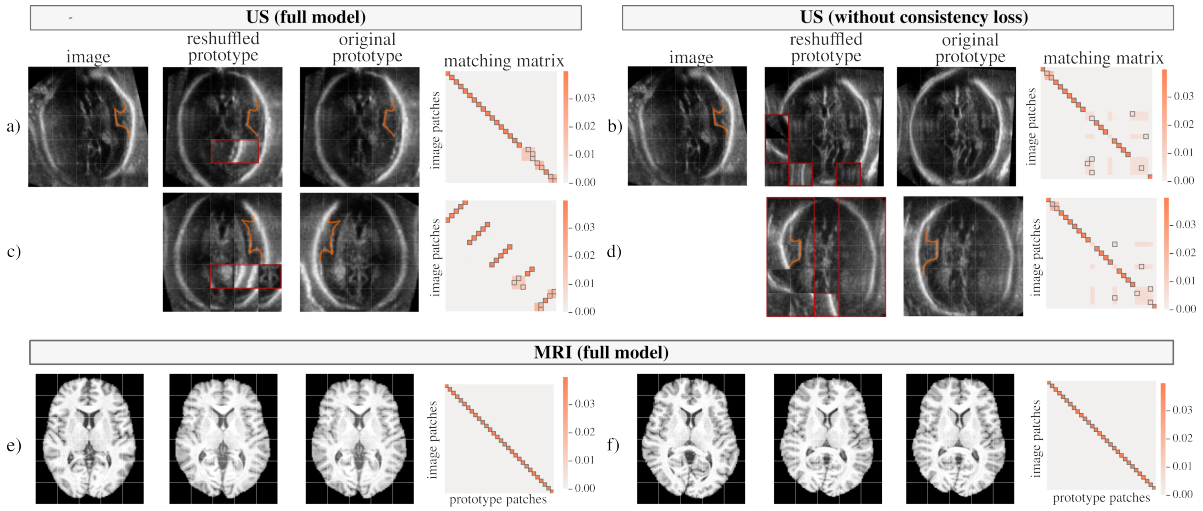

Figure 4. Illustration of OT matching between image sample and prototypes for US (a-d) and MRI (e,f). The obtained OT matching matrix ($\mathbf{Q}$) is given in the last column of each panel, with a black border indicating the most similar prototype patch to each image patch. For each image patch, this most similar prototype patch is visualized in the same location in the *reshuffled prototype*, with patches mirrored if they change side. Patches that are incorrectly matched (according to the known anatomy) have red borders. For US, the matching of the same image sample is shown with two prototypes: one with the same visible hemisphere (a, b), and one with the opposite hemisphere visible (c, d). The full model correctly matches the opposite hemispheres, but the model without consistency loss does not. The Sylvian Fissure has been drawn in orange in each of the visible hemispheres for clarity. For MRI, two examples have been shown for the full model (e,f).

(4b,d) consistency loss. It can be seen that for the pairs in Fig. 4a and b the image and prototype both have a visible right hemisphere (with the Sylvian Fissure annotated), whereas in Fig. 4c and d the visible hemisphere of the prototype is opposite to that of the image. The full model correctly identifies the visible hemispheres, matching the patches in these hemispheres with each other. This can also be seen in the reshuffled prototype, in which the sides are flipped. The same pattern of matching between hemispheres was found in all test set volumes. On the

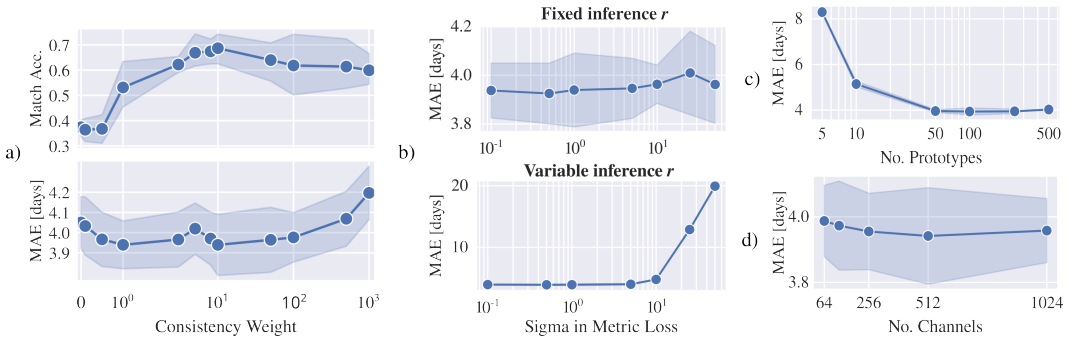

Figure 5. Overview of hyperparameter tuning for the US dataset for the (a) consistency weight, β, (b) standard deviation in metric loss, σ, (c) number of prototypes and (d) number of channels in the add-on layers and prototype representations. For σ two types of inference are shown, keeping the inference radius, r, constant (top) or varying it in line with the training sigma, $r = 3\sigma$ (bottom).

other hand, in Fig. 4d it is evident that without consistency loss, no matching occurs between the two opposite visible hemispheres. Instead, the same spatial location in both the image and prototype are matched, as shown by the diagonal matching matrix. This illustrates that the consistency loss is responsible for matching the correct hemispheres. In this work only flipping across the midline of the brain has been used as geometric transformation in the consistency loss: future work should consider other transformations, based on the geometric variation present in the dataset. It is also important to note that no hemisphere labels were used during training as the consistency loss is unsupervised.

In brain MR images, both hemispheres are visible, and there is thus less need to enforce inter-hemispheric patch similarity with a consistency loss. For this reason, the effect of this loss component for the MR images is less pronounced, with only subtle differences between the models trained with or without consistency loss. For both models, the matching matrices found (Fig. 4e,f) are mostly according to the diagonal, suggesting that each patch in the image is matched with the same spatial location in the prototype. As our images were aligned to the same template space (MNI), this matching is expected and shows that the model can learn the similarity between corresponding patches.

Sensitivity to Hyperparameter Choice The average validation performance for each of the important hyperparameters is shown in Fig. 5 for the US dataset. For the consistency weight (Fig. 5a), the top plot shows the patch-matching accuracy, which is computed from the overlap with approximated ground-truth matching matrices determined from the known visible hemisphere in each sample. It is evident that increasing the consistency weight results in an improvement of both the MAE and patch-matching accuracy up to a weight of 10, after which it starts deteriorating again. It should be noted that the consistency loss is about four orders of magnitudes smaller than the metric loss during training, hence the small effect for low weights.

In Fig. 5b the standard deviation (σ) in the metric loss (Eq. 2) is varied. The top and bottom panels show the results of the same training runs, but during inference, the radius is either fixed ($r = 3$, top panel) or adapted based on the σ during training ($r = 3\sigma$, bottom panel). The training σ has only a small effect on performance throughout the range of values tried, whereas the inference radius does considerably affect the performance, most notably at higher values. This is beneficial as the inference r can be easily adjusted after training, and can thus be optimized offline.

The number of prototypes and the number of channels in the prototype representations (and in the add-on layers) are shown in Fig. 5c and d respectively. Both show improved performance when increasing the number of prototypes or channels, leveling off for higher values.

Overall, these results show that the hyperparameters introduced in our model behave as expected, confirming the stability of our approach and showing that the results are not overly sensitive to the hyperparameter selection.

5. Conclusion

We have presented ExPeRT, a novel explainable model for continuous regression, based on prototype learning and patch-based OT matching. Our model achieves competitive prediction performance on two brain age prediction datasets: fetal US and adult MRI. For both datasets, the patch-based distance metric was able to correctly learn the structural matches between the sample image and prototypes. Our approach is versatile and can be applied to other continuous regression problems, beyond medical imaging.

Acknowledgements

L.S.H. acknowledges the support from the EPSRC Doctoral Prize award. A.I.L.N. and N.K.D. are supported by the Bill and Melinda Gates Foundation.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Proc. Adv. Neural Inf. Process. Syst.*, 2018. 1

[2] David Alvarez-Melis and Tommi S Jaakkola. On the robustness of interpretability methods. *arXiv:1806.08049*, 2018. 1

[3] Pedro L. Ballester, Laura Tomaz da Silva, Matheus Marcon, Nathalia Bianchini Esper, Benicio N. Frey, Augusto Buchweitz, and Felipe Meneguzzi. Predicting Brain Age at Slice Level: Convolutional Neural Networks and Consequences for Interpretability. *Front. Psychiatry*, 12, 2021. 2

[4] Alina Jade Barnett, Fides Regina Schwartz, Chaofan Tao, Chaofan Chen, Yinhao Ren, Joseph Y. Lo, and Cynthia Rudin. A case-based interpretable deep learning model for classification of mass lesions in digital mammography. *Nat. Mach. Intell.*, 3(12):1061–1070, 2021. 2

[5] Cher Bass, Mariana da Silva, Carole Sudre, Logan Z J Williams, Petru-Daniel Tudosiu, Fidel Alfaro-Almagro, Sean P Fitzgibbon, Matthew F Glasser, Stephen M Smith, and Emma C Robinson. ICAM-reg: Interpretable classification and regression with feature attribution for mapping neurological phenotypes in individual scans. *IEEE Trans. Med. Imag.*, 42(4), 2023. 2

[6] Christian F Baumgartner, Lisa M Koch, Kerem Can Tezcan, Jia Xi Ang, and Ender Konukoglu. Visual Feature Attribution Using Wasserstein GANs. In *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pages 8309–8319, 2018. 2

[7] Jacob Bien and Robert Tibshirani. Prototype selection for interpretable classification. *Ann. Appl. Stat.*, 5(4):2403–2424, 2011. 2

[8] Kyriaki Margarita Bintsi, Vasileios Baltatzis, Arinbjörn Kolbeinsson, Alexander Hammers, and Daniel Rueckert. Patch-Based Brain Age Estimation from MR Images. In *Mach. Learning in Clin. Neuroimag. and Radiogenomics in Neuro-oncology*, volume 12449, pages 98–107, 2020. 2

[9] Moritz Böhle, Fabian Eitel, Martin Weygandt, and Kerstin Ritter. Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. *Front. Aging Neurosci.*, 11, 2019. 2

[10] Hanqing Chao, Jiajin Zhang, and Pingkun Yan. Regression Metric Loss: Learning a Semantic Representation Space for Medical Images. In *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, pages 427–436. Springer, 2022. 2, 3, 4

[11] Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. This Looks Like That: Deep Learning for Interpretable Image Recognition. In *Proc. Adv. Neural Inf. Process. Syst.*, 2019. 1, 2, 3

[12] James H Cole, Rudra PK Poudel, Dimosthenis Tsagkrasoulis, Matthan WA Caan, Claire Steves, Tim D Spector, and Giovanni Montana. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage*, 163:115–124, 2017. 2

[13] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Proc. Adv. Neural Inf. Process. Syst*, 2013. 2, 4

[14] Nicola K Dinsdale, Emma Bluemke, Stephen M Smith, Zobair Arya, Diego Vidaurre, Mark Jenkinson, and Ana IL Namburete. Learning patterns of the ageing brain in mri using deep convolutional networks. *NeuroImage*, 224:117401, 2021. 1, 2

[15] Nicola K. Dinsdale, Emma Bluemke, Vaanathi Sundaresan, Mark Jenkinson, Stephen M. Smith, and Ana I.L. Namburete. Challenges for machine learning in clinical translation of big data imaging studies. *Neuron*, 110(23):3866–3881, 2022. 1

[16] Fabian Eitel and Kerstin Ritter. Testing the robustness of attribution methods for convolutional neural networks in MRI-based Alzheimer's disease classification. In Springer, editor, *Interpretability of Mach. Intell. in Med. Image Comp. and Multimodal Learning for Clinical Decision Support*, volume 11797, pages 3–11, 2019. 2

[17] Sheila MP Everwijn, Ana IL Namburete, Nan van Geloven, Fenna AR Jansen, Aris T Papageorghiou, Aalbertine K Teunissen, Lieke Rozendaal, Nico Blom, Jan M van Lith, and Monique C Haak. The association between flow and oxygenation and cortical development in fetuses with congenital heart defects using a brain-age prediction algorithm. *Prenatal Diagnosis*, 41(1):43–51, 2021. 2

[18] FMRIB Software Library. FSL Anat pipeline. Available at: https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/fsl_anat. 5

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[20] Linde S. Hesse and Ana I.L. Namburete. INSightR-Net: Interpretable Neural Network for Regression Using Similarity-Based Comparisons to Prototypical Examples. In *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, pages 502–511. Springer, 2022. 1, 2, 6

[21] Simon M Hofmann, Frauke Beyer, Sebastian Lapuschkin, Ole Goltermann, Markus Loeffler, Klaus-Robert Müller, Arno Villringer, Wojciech Samek, and A Veronica Witte. Towards the interpretability of deep learning models for multimodal neuroimaging: Finding structural changes of the ageing brain. *NeuroImage*, 261:119504, 2022. 2

[22] IXI Dataset. Available at: https://brain-development.org/ixi-dataset/. 5

[23] Been Kim, Cynthia Rudin, and Julie A. Shah. The bayesian case model: A generative approach for case-based reasoning and prototype classification. In *Proc. Adv. in Neur. Inf. Process. Syst.*, volume 27, 2014. 2

[24] Eunji Kim, Siwon Kim, Minji Seo, and Sungroh Yoon. Xprotonet: diagnosis in chest radiography with global and local explanations. In *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pages 15719–15728, 2021. 2

[25] Sungyeon Kim, Minkyo Seo, Ivan Laptev, Minsu Cho, and Suha Kwak. Deep metric learning beyond binary supervision. In *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pages 2288–2297, 2019. 2

[26] Philip A. Knight. The Sinkhorn-Knopp algorithm: Convergence and applications. *SIAM J. Matrix Anal. Appl.*, 30(1):261–275, 2008. 4

[27] Lok Hin Lee, Elizabeth Bradburn, Rachel Craik, Mohammad Yaqub, Shane A Norris, Leila Cheikh Ismail, Eric O Ohuma, Fernando C Barros, Ann Lambert, Maria Carvalho, et al. Machine learning for accurate estimation of fetal gestational age based on ultrasound images. *NPJ Digital Medicine*, 6(1):36, 2023. 6

[28] Hongming Li, Mohamad Habes, David A Wolk, and Yong Fan. A deep learning model for early prediction of alzheimer's disease dementia based on hippocampal magnetic resonance imaging data. *Alzheimer's & Dementia*, 15(8):1059–1070, 2019. 2

[29] Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Proc. AAAI Conf. Artif. Intell.*, volume 32, pages 3530–3537. AAAI press, 2018. 2

[30] Angela Lombardi, Domenico Diacono, Nicola Amoroso, Alfonso Monaco, João Manuel RS Tavares, Roberto Bellotti, and Sabina Tangaro. Explainable deep learning for personalized age prediction with brain morphology. *Front. in Neurosci.*, 15, 2021. 2

[31] Jiwen Lu, Junlin Hu, and Jie Zhou. Deep metric learning for visual understanding: An overview of recent advances. *IEEE Signal Process. Mag.*, 34:76–84, 11 2017. 2

[32] G. Malinger, D. Paladini, K. K. Haratz, A. Monteagudo, G. L. Pilu, and I. E. Timor-Tritsch. ISUOG practice guidelines (updated): sonographic examination of the fetal central nervous system. part 1: performance of screening examination and indications for targeted neurosonography. *Ultrasound Obstet. & Gynecol.*, 56:476–484, 2020. 6

[33] Sanaz Mohammadjafari, Mucahit Cevik, Mathusan Thanabalasingam, and Ayse Basar. Using protopnet for interpretable alzheimer's disease classification. In *Canadian conf. on Artif. Intell.*, 2021. 2

[34] Felipe Moser, Ruobing Huang, Bartłomiej W Papież, Ana IL Namburete, INTERGROWTH 21st Consortium, et al. Bean: Brain extraction and alignment network for 3d fetal neurosonography. *NeuroImage*, 258:119341, 2022. 5

[35] Ana I.L. Namburete, Richard V. Stebbing, Bryn Kemp, Mohammad Yaqub, Aris T. Papageorghiou, and J. Alison Noble. Learning-based prediction of gestational age from ultrasound images of the fetal brain. *Med. Image Anal.*, 21(1):72–86, 2015. 2, 5

[36] Aris T. Papageorghiou, Eric O. Ohuma, Douglas G. Altman, Tullia Todros, Leila Cheikh Ismail, Ann Lambert, Yasmin A. Jaffer, Enrico Bertino, Michael G. Gravett, Manorama Purwar, J. Alison Noble, Ruyan Pang, Cesar G. Victora, Fernando C. Barros, Maria Carvalho, Laurent J. Salomon, Zulfiqar A. Bhutta, Stephen H. Kennedy, and José Villar. International standards for fetal growth based on serial ultrasound measurements: The Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project. *Lancet*, 384(9946):869–879, 2014. 5

[37] Nicolas Papernot and Patrick McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv:1803.04765*, 2018. 2

[38] Han Peng, Weikang Gong, Christian F Beckmann, Andrea Vedaldi, and Stephen M Smith. Accurate brain age prediction with lightweight deep neural networks. *Medical image analysis*, 68:101871, 2021. 6

[39] Hai Phan and Anh Nguyen. Deepface-emd: Re-ranking using patch-wise earth mover's distance improves out-of-distribution face identification. In *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit*, pages 20259–20269, 2022. 2, 3, 4

[40] Sebastian G Popescu, Ben Glocker, David J Sharp, and James H Cole. Local Brain-Age: A U-Net Model. *Front. in Aging Neurosci.*, 13, 2021. 2

[41] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. Earth mover's distance as a metric for image retrieval. *Int. J. of Comput. Vision*, 40(2):99–121, 2000. 2

[42] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.*, 1(5):206–215, 2019. 1

[43] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pages 815–823, 2015. 2, 5

[44] Gurmail Singh and Kin-Choong Yow. An interpretable deep learning model for covid-19 detection with chest x-ray images. *IEEE Access*, 9:85198–85208, 2021. 2

[45] Gurmail Singh and Kin-Choong Yow. These do not look like those: An interpretable deep learning model for image recognition. *IEEE Access*, 9:41482–41493, 2021. 2

[46] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pac. J. Appl. Math.*, 21(2):343–348, 1967. 4

[47] Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000. 3

[48] Luis Caicedo Torres, Luiz Manella Pereira, and M. Hadi Amini. A Survey on Optimal Transport for Machine Learning: Theory and Applications. *arXiv:2106.01963*, 2021. 2

[49] Bas H.M. van der Velden, Hugo J. Kuijf, Kenneth G.A. Gilhuijs, and Max A. Viergever. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Med. Image Anal.*, 79:102470, 2022. 1

[50] Madeleine K Wyburd, Linde S Hesse, Moska Aliasi, Mark Jenkinson, Aris T Papageorghiou, Monique C Haak, and Ana IL Namburete. Assessment of regional cortical development through fissure based gestational age estimation in 3d fetal ultrasound. In *Uncertainty for Safe Utilization of Mach. Learning in Med. Imag., and Perinatal Imag., Placental and Preterm Image Anal.*, pages 242–252. Springer, 2021. 1, 2

[51] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. DeepEMD: Differentiable earth mover's distance for few-shot learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(5):5632–5648, 2022. 2, 3

[52] Wenliang Zhao, Yongming Rao, Ziyi Wang, Jiwen Lu, and Jie Zhou. Towards interpretable deep metric learning with

structural matching. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 9887–9896, 2021. 2, 3, 4, 6

[53] Kang Zheng, Yirui Wang, Xiao Yun Zhou, Fakai Wang, Le Lu, Chihung Lin, Lingyun Huang, Guotong Xie, Jing Xiao, Chang Fu Kuo, and Shun Miao. Semi-supervised Learning for Bone Mineral Density Estimation in Hip X-Ray Images. In *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, pages 33–42. Springer, 2021. 4