Prototype Learning for Explainable Regression

Linde S. Hesse, Nicola K. Dinsdale, Ana I. L. Namburete

***Abstract*— The lack of explainability limits the adoption of deep learning models in clinical practice. While methods exist to improve the understanding of such models, these are mainly saliency-based and developed for classification, 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. We demonstrate our proposed model on the task of brain age prediction on two image datasets: adult MR and fetal ultrasound. Our approach achieved state-of-the-art prediction performance while providing insight in the model's reasoning process.**

***Index Terms*— Brain Age Prediction, Brain MRI, Explainability, Fetal Ultrasound, Regression**

I. 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 behaviour and gain patients' trust [1]. It is therefore vital to develop models that are explainable and hence capable of providing insight into their reasoning process [2].

The most frequently used methods to explain a model's prediction are saliency-based, which explain the prediction of an already trained model *post-hoc* [3]. 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 and can for example resemble edge maps rather than being dependent on the trained model [4]. Furthermore, they often result in noisy saliency maps, which are hard to interpret and prone to confirmation bias [4], [5].

On the other hand, inherently explainable models provide an accurate representation of the model's decision-making process by design [2]. However, designing such models for medical imaging is challenging as there is typically a trade-off between performance and explainability.

A promising example-based explainable model for the classification of natural images was proposed by [6]. 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 [7], [8]. As *ProtoPNet* uses the categorical labels to pull (or push) samples together (or apart), this architecture cannot be directly applied for regression. Furthermore, the prototypes in *ProtoPNet* represent only *image parts*, with the intuition of discovering whether the pattern in a prototype is present in any of the image patches. This provides patch-level detail in the prediction but is less suitable for continuous regression problems where we are interested in more gradual structural changes, as opposed to class-specific patterns.

In this work, we propose *ExPeRT*: an Explainable Prototype-based model for Regression using optimal Transport. We improve upon a preliminary version of this work (published at MICCAI 2022 [9]) by incorporating metric learning to map the images to an inherently continuous representation space in which the distances between images and prototypes in *latent* space are relative to their differences in *label* space. The prediction for a new sample is then made from a weighted average of prototype labels within a given distance. As all prototypes can be visualized, this provides an intuitive explanation of the prediction for a regression task.

In contrast to *ProtoPNet* and our previous work [9], the prototypes in *ExPeRT* are latent representations of whole training images. We instead incorporate spatial detail into the

LH acknowledges the support of the UK Engineering and Physical Sciences Research Council (EPSRC) Doctoral Training Award and of the Microsoft Fellowship Award. ND is supported by an Academy of Medical Sciences Springboard Award. AN is grateful for support from the Academy of Medical Sciences under the Springboard Awards scheme (SBF005/1136), and the Bill and Melinda Gates Foundation.

Linde S. Hesse, Nicola K. Dinsdale and Ana I. L. Namburete are with the Oxford Machine Learning in NeuroImaging (OMNI) laboratory, Department of Computer Science, University of Oxford, Oxford, UK. LH is also with the Institute of Biomedical Engineering, Department of Engineering, Oxford, UK. AN is also affiliated with the Wellcome Centre for Integrative Neuroscience (WIN), Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK (email: linde.hesse@seh.ox.ac.uk; nicola.dinsdale@cs.ox.ac.uk; ana.namburete@cs.ox.ac.uk).

model's decision-making process by decomposing the image-level distances into patch-wise similarities between image patches. The patch-wise similarities are computed in latent space and then structurally matched using optimal transport (OT). The 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, which can also be referred to as the Earth Mover's Distance [10]. The resulting matches can be inspected to verify whether anatomically corresponding patches are matched together between an image and prototype, thus providing a detailed decomposition of the image-level distance.

The network is trained using a combination of two losses: A *metric loss*, which enforces the prototype-image distances to be relative to label differences, and a *consistency loss* that encourages the distances between identical image patches under an anatomically justified transformation to be small. 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 [7], [11]; during gestation the predicted brain age can be compared to true post-conceptual age to quantify fetal brain development [12], [13].

In summary, our key contributions are as follows:

1) Building on our previous work [9], we propose a novel explainable model for continuous regression and demonstrate that it obtains competitive performance for brain age prediction on two datasets: fetal US and adult MRI.
2) Our model demonstrates how optimal transport can be incorporated into a prototype-based model, providing patch-level matching scores while computing a single image-level distance.
3) A geometric consistency loss is introduced which improves the model performance and leads to more anatomically correct matching between patches of the prototype and sample image.

II. Related Work

A. 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 [8], [14]–[17], but their explanations are not always consistent or comparable between different methods [18]. Other studies have used patch- or slice-based approaches to predict local brain age [19], [20]. 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, Popescu *et al.* [21] used a U-Net to predict brain age voxelwise. The predictions were more fine-grained than previous slice- or patch-based approaches, but the reported prediction error was considerably higher than the error obtained with baseline models (average MAE ~10 years vs ~3 years). Alternatively, in [22], [23], 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 might not always be available.

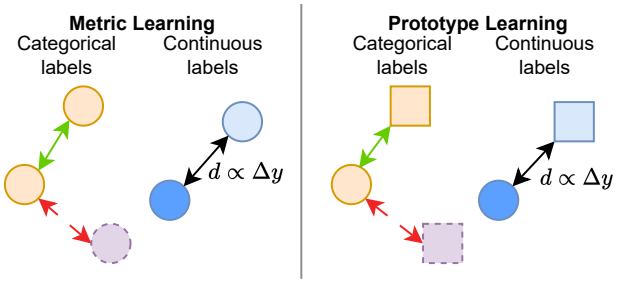

Fig. 1. Schematic overview of metric and prototype learning for categorical and continuous labels, illustrating the similarities between the two approaches. Circles represent training samples, squares prototypes, and the colours represent different classes or continuous values. The training objective for classification is minimizing the distances between samples of the same class (green arrows) while maximizing those of different classes (red arrows). For regression, the distances are trained to be relative to their label difference.

B. Prototype Learning

In prototype learning, a set of feature representations in latent space is learned during model training, the so-called *prototypes*, 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 [24]. In most early deep prototype learning approaches, the learned prototypes were unconstrained points in latent space [25], [26], making it challenging to visualise or interpret the learned prototypes. To achieve visualisation of the prototypes, [27] introduced an autoencoder to reconstruct an image from the latent representation. Alternatively, in [6] 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, enabling a more fine-grained explanation. Variations on the *ProtoPNet* architecture have also been applied to several medical imaging tasks [28]–[32], but have mainly been limited to classification tasks.

C. Deep Metric Learning

Prototype-based methods use distances in latent space to make a prediction, and can therefore be considered closely related to (deep) metric learning approaches (Fig. 1). Metric learning aims to learn a mapping to a representation space in which distances between similar samples are small and, inversely, distances between dissimilar samples are large [33]–[37]. The mapping by a neural network is typically achieved by training with contrastive or triplet losses [38]. These losses were initially developed for tasks with a clear distinction

between similar and dissimilar samples, such as classification, and were less suitable for regression tasks with continuous labels [33], [39]. Early work simply quantized the continuous differences [40], [41], but more recent extensions also adapted the triplet loss for continuous labels [37], [42]–[44]. Specifically designed for regression, [45] suggested regularizing the latent space by enforcing the feature distance between samples in a batch, computed with the Euclidean distance, to be relative to their difference in label space. Furthermore, samples were weighted by a Gaussian function so as to weigh samples close to each other more than those far apart. This was motivated by the fact that distances in the feature space should be computed along the manifold (*geodesic distances*), and that simple distance measures such as the Euclidean distance are only a good approximation for small distances. 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.

D. Optimal Transport

Optimal transport is the mathematical optimization problem computing the shortest distance (or *lowest cost*) between two distributions, given a cost matrix. This could, for example, be the optimal assignment of a set of workers to a set of tasks, given the time each worker needs for a certain task. It can be optimized efficiently when adding an entropic regularization term [46], and has been used in several deep learning architectures [47]. Most comparable to our application, OT has previously been applied to compute distances between pairs of natural images in [34]–[36]. In those works, OT was used to match the image patches with each other based on patch-level distances, resulting in improved performance and explainability.

III. 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 [34]–[36]. A schematic overview of the $ExPeRT$ architecture is shown in Fig. 2a.

A. 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 fixed at the beginning of training and are not optimized. On the other hand, 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 prototype distances of length n_p, denoted by $\boldsymbol{d}$.

B. Prototype Projection

During training, the prototypes are updated with each step and can therefore be located throughout the latent space. However, as proposed in [6], every N epochs these are replaced by the closest latent representation of an image in the training dataset, which is referred to as the *prototype projection*. We save our model checkpoints only straight after the projection, ensuring that in the saved model each prototype can be visualized using the corresponding image from the training set.

C. Distance Metric Loss

The distances between samples and prototypes need to be regularised, so that for a certain sample with ground-truth label y, the distance in latent space to a prototype k is related 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 neighbourhood of a sample representation using the Euclidean distance, which is a good approximation for small distances on the manifold.

We train our network using an adapted version of the loss proposed in [45]. Instead of regularising distance over the samples in the batch, we regularise distances between samples in the batch and each of the prototypes in the *local neighbourhood*. The neighbourhood 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:

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

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

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

where σ controls the size of the neighbourhood (i.e. the standard deviation of the Gaussian kernel), and α is a small

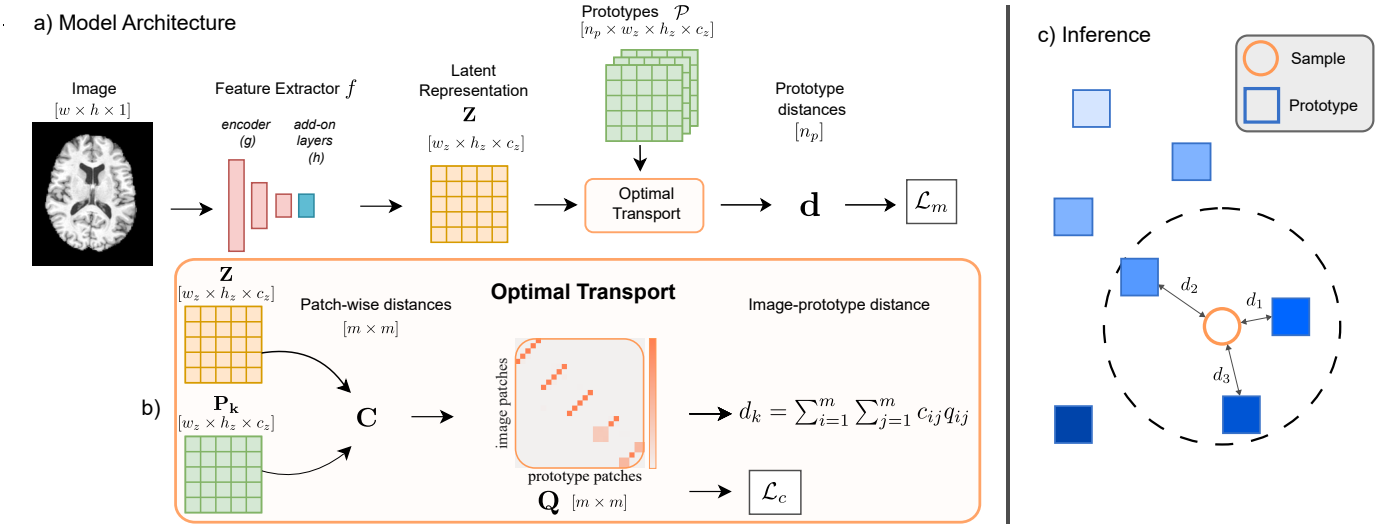

Fig. 2. Schematic overview of the *ExPeRT* architecture (a), distance computation with OT (b) and inference (c). The OT matching is shown for one sample and prototype, and is repeated for each prototype to obtain $\boldsymbol{d}$. The $\boldsymbol{m}$ is the total number of patches per sample or prototype, given by $\boldsymbol{w_z h_z}$. During inference (c), a sample prediction is made using a weighted average of labels of prototypes within a certain radius.

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 embed close together in feature space despite being dissimilar.

D. 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 that compresses the spatial dimensions, resulting in a vector of size c_z. The Euclidean distance can then be computed between these vectors [42], [45] but this discards all spatial information. To provide information about the spatial makeup of the distance between a prototype and sample, we propose to use a patch-based distance metric instead.

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 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, we use OT to obtain a matching matrix, $\mathbf{Q} \in \mathbb{R}^{m \times m}$ (more details in next section). This matrix $\mathbf{Q}$ can be considered as a *soft assignment* matrix, in which each feature vector $\boldsymbol{z_i}$ can be partly 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 kth prototype as follows:

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

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

Optimal Transport: Optimal transport 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 $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$:

$$\sum_{i=1}^{m} q_{ij} = \boldsymbol{\mu_1}, \sum_{j=1}^{m} q_{ij} = \boldsymbol{\mu_2} \tag{5}$$

This original problem is computationally expensive to solve, and so [46] introduced an *entropic regularization* to smooth the optimization problem. The resulting system can be considered as a matrix scaling problem, and can efficiently be solved using the classical Sinkhorn divergence algorithm [48], [49]. This algorithm iteratively updates the rows and columns of a matrix so that it sums up to the required marginal distributions, and is fully differentiable [49]. 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 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 as well ([35], [36]). As the

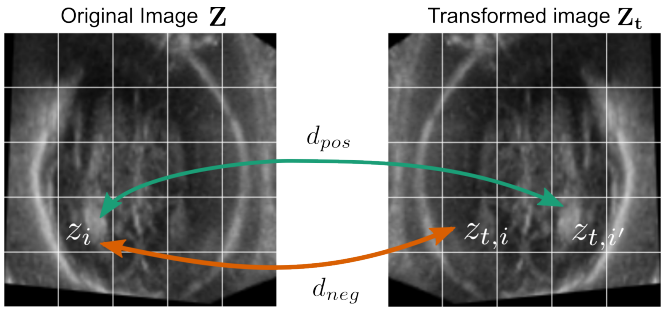

Fig. 3. Overview of the used consistency loss between a latent image representation ($\mathbf{Z}$) and the representation of a transformed image ($\mathbf{Z_t}$) for a single triplet. The grid indicates the size of the latent pixels.

OT matching is fully differentiable, we can train our whole network end-to-end.

E. Consistency Loss

The OT optimization aims to find structural matches between the image and prototype. To achieve anatomically correct matches, we also introduced a consistency loss that acts on 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 : $\boldsymbol{z_{t,i}} \in \mathbf{Z_t} = f(T(\mathbf{X}))$. The consistency loss 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, which encourages 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 [38]. 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.

F. Inference

At inference, a prediction for a new sample is made using a weighted average of the prototype labels within a certain radius r (Fig. 2b), 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 $r/3$.

G. Datasets

1) Fetal Ultrasound: For this study, 2D ultrasound images sampled from 3D volumes acquired as part of the INTERGROWTH-21st Fetal Growth Longitudinal Study [50] were used. That study aimed to describe fetal growth and neurodevelopment in a geographically diverse, healthy population of women, enrolled before 14 weeks' gestation. The volumes were acquired with a Philips US device (Philips HD-9, Philips Ultrasound, USA) using a curvilinear abdominal transducer. All gestational ages were determined based on the last menstrual period and confirmed with US crown-rump measurements that needed to agree within 7 days. We used a total of 4290 volumes between 14 and 31 gestational weeks, selected based on having sufficient ultrasound quality [12].

All 3D volumes were firstly aligned to the same coordinate system and scaled to the average brain size at 30 GW using an automated alignment method [51]. 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 using fixed coordinates, resulting in a total set of 4290 2D images of size 160×160 with an isotropic pixel size of 0.6 mm. For the experiments, this set was split into test (20%) and train/validation (80%).

2) Adult MRI: We also experimented on T1 MRI images from the IXI dataset[1], which contains MR images from healthy subjects between the age of 19 and 86 years. The images were acquired across three imaging sites in London, UK, on different imaging systems. The volumes were preprocessed using the FSL Anat pipeline[2], the key stages of which are skull stripping, bias field correction and linear registration to the 1mm MNI template. 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 and the corresponding brain atrophy is well established. This resulted in a total of 561 MR images, each of size 160×192, which was also split into test (20%) and train/validation set (80%).

H. Implementation

We implemented our proposed method with Pytorch Lightning, using PyTorch 1.13 and Python 3.10. The code is available at: (*will be released on acceptance*).

All experiments were performed on a single Nvidia A10 with 24 GB RAM and we used an implementation of the Sinkhorn algorithm in logarithmic space to avoid instabilities

[1] https://brain-development.org/ixi-dataset/
[2] https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/fsl_anat

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 was used to tune all hyperparameters, of which an overview is given in Section IV-C. The reported test set performance is obtained by applying the best fold (on the validation set) to the test set. All experiments were run 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. The prototypes were projected to the closest sample in the training set each 25 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. Prototype labels were assigned uniformly at the beginning of training within the label range present in the training dataset.

Due to shadowing artefacts of the fetal skull, usually only one of the two brain hemispheres is clearly visible in fetal brain US images [52]. As it is desired to match the two visible hemispheres with each other, the consistency loss for this dataset was implemented with horizontal flips along the midline of the brain as a geometric transformation (see Fig. 3). The weight of the consistency loss β, was set to 10, and the margin, γ to 0.1. For the adult MR images, however, both hemispheres are visible and so a consistency loss is not required. This was confirmed by cross-validation, and therefore this loss element was set to 0. During inference, only prototypes within a certain radius r of a sample are considered when generating the final prediction. Therefore, it can occur that for a certain sample, no prototypes are within this area, resulting in no prediction. In order to report the performance, we used nearest-neighbour predictions for these samples and assigned the label of the closest prototype to the sample.

IV. Results and Discussion

The quantitative prediction results for both US and MRI are shown in Table I and in Fig. 4a and c. We compared our $ExPeRT$ architecture to two baseline models: *ResNet Baseline* – a vanilla ResNet 18 model (pretrained on ImageNet) trained with mean squared error loss on our task, and *AvgPool Baseline* – a network with the same architecture as our proposed model but computing distances not with patch-based OT but by first pooling the feature representations into a single 1D vector and computing Euclidean distances between these. 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, the performance improvement did not pass a paired t-test ($p = 0.15$). Our results are in line with previously reported errors for fetal brain age prediction [8], [12], and when considering the large age range of the IXI dataset (in the weighted MAE), also in range with reported errors for adult brain MRI [53]. Specifically, we want to note that achieving increased prediction performance is not the main aim of this work. Rather, we sought to show that we can create a more explainable model without compromising on prediction performance.

Figures 4b and c 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 neighbouring prototypes are shown next to the sample image. The weight of each prototype is inferred from its distance using a Gaussian function and shown on the right. The colour of each point indicates the prototype label, which is also shown with a coloured 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. 5. For each sample-prototype pair, the OT (*soft assignment*) matrix is given in the last column of each panel, indicating which patches are most similar between the prototype and sample. The reconstructed prototype shows the OT matching more intuitively, in which for each image patch the prototype patch with the highest matching score is shown. In the US dataset, the two visible hemispheres are correctly matched (Fig. 5.1a, b), illustrating the advantage of using OT matching in our network.

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 alternate way of making a neural network more explainable. Furthermore, while in this study we have used uniform initial distributions in Eq. 5 (i.e. all patches get the same weight), this could in further work be replaced by other options such as cross-correlation between the patches [36] to generate importance scores for each of the patches, further improving the explainability of the model.

A. Fetal Ultrasound (IG)

As evidenced by the ablation study presented in Table I, both the consistency loss and the add-on layers improve the prediction performance for the US dataset, demonstrating the need for both components in the model. The effect of the consistency loss on the US images is further illustrated in Fig. 5a-c, showing sample-prototype pairs for the full model trained with consistency loss (Fig. 5a and b) and for the model without consistency loss (Fig. 5c). It can be seen that for the pair in 5a the image and prototype both have a visible right hemisphere (with the Sylvian Fissure annotated), whereas in 5b and c the visible hemisphere of the prototype is opposite to that of the image. The full model correctly identifies this and matches the patches in the visible hemisphere with each other. This can also be seen in the reconstructed prototype, in which the sides are flipped. The same pattern of matching between hemispheres was found in all test set volumes indicating that our model correctly matches the patches of the visible hemispheres. Important to note here is that no annotations

TABLE I

QUANTITATIVE RESULTS FOR US AND MRI ON THE TEST SET. THE VALUES BETWEEN BRACKETS INDICATE THE STANDARD DEVIATIONS. THE WEIGHTED MAE NORMALIZES FOR THE AGE RANGE OF THE DATASET BY DIVIDING BY IT (MAE/AGE RANGE) [53]. THE ABLATIONS SHOW THE RESULTS WITHOUT CONSISTENCY LOSS AND WITHOUT ADD-ON LAYERS IN THE FEATURE EXTRACTOR.

	Ablations		Ultrasound		MRI	
	Consistency ($\mathcal{L}_c$)	Add-on layers (h)	MAE (days)	Weighted MAE	MAE (years)	Weighted MAE
ResNet Baseline	-	-	4.04 (3.16)	0.035	6.29 (4.79)	0.094
AvgPool Baseline	-	-	4.09 (3.24)	0.035	6.41 (5.47)	0.096
ExPeRT (no $\mathcal{L}_c$ or h)	×	×	4.16 (3.50)	0.036	6.37 (5.11)	0.095
ExPeRT (no $\mathcal{L}_c$)	×	✓	3.96 (3.19)	0.034	6.19 (4.95)	0.092
ExPeRT (no h)	✓	×	4.08 (3.53)	0.035	-	-
ExPeRT	✓	✓	**3.93** (3.12)	0.034	-	-

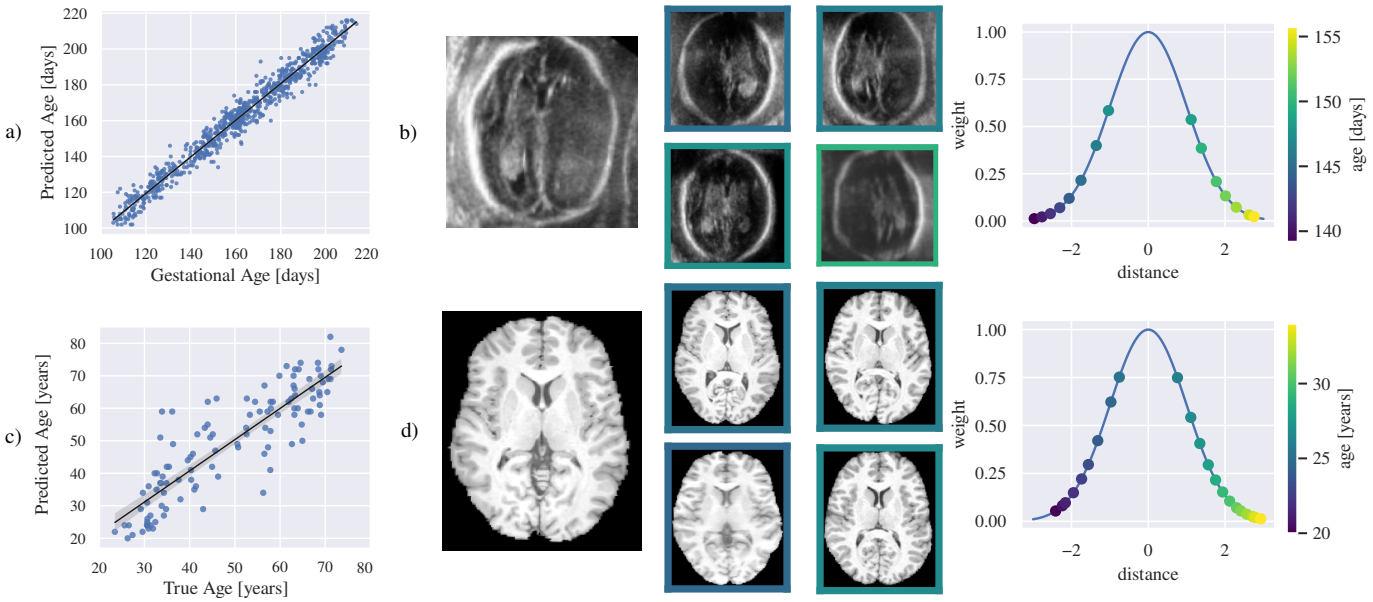

Fig. 4. Predicted versus ground-truth age for fetal US (a) and adult MRI (c), as well as an example prediction for US (b) and MRI (d). In b and d, the sample is shown on the left, and the four closest prototypes are shown in the middle. The right-most graphs indicate the weight of each of the prototypes in the final prediction. Prototypes with a label below the predicted value were plotted with negative distances for visualisation 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.

for the visible hemispheres were used during training, as the consistency loss is unsupervised.

On the other hand, in Fig. 5c it is evident that for the model trained without consistency loss, no matching occurs between the two visible hemispheres. Instead, the same spatial location in both the image and prototype are matched, as shown by the diagonal matching matrix. This thus 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. In future work, other transformations could be considered, based on the geometric variation present in the dataset.

B. Adult MRI (IXI)

In brain MR images, both hemispheres are visible, and therefore there is no need to enforce inter-hemispheric patch similarity with a consistency loss. As such, this loss was not applied to the MRI dataset (resulting in the last two rows being empty in Table I). From the ablation of the add-on layers, it is evident that these, in line with the US findings, improve the prediction performance.

In the qualitative results in Fig. 5e-g, it can be seen that the matching flows found by the model 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 coordinate system, this matching is expected and indicates that the model can learn the similarity between corresponding patches.

For almost all test set volumes, the sample-prototype pairs were matched correctly, but Fig. 5f shows one of the few samples where a few patches were mismatched (shown with red borders). However, even in this case most patches were correctly identified. This thus demonstrates that anatomically correct image patches are matched, even when the volumes are acquired on different MRI scanners (Fig. 5e), despite known harmonization effects [1].

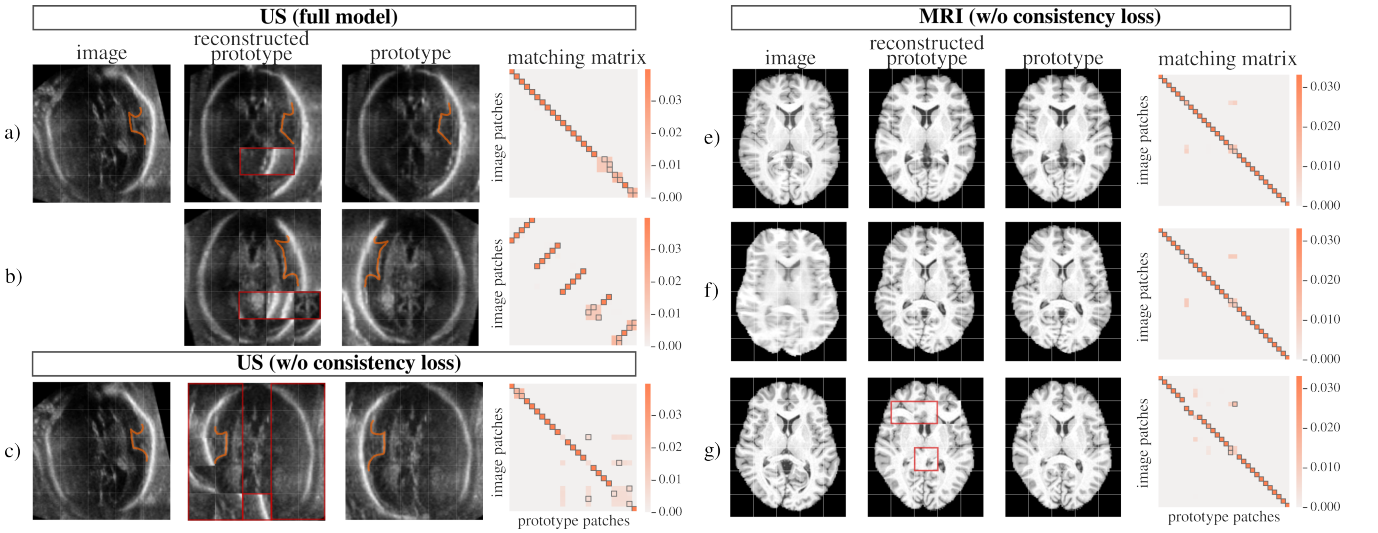

Fig. 5. Illustration of OT matching between image sample and prototypes for US (a-c) and MRI (d-f). The obtained OT matching matrix ($\mathbf{Q}$) is given in the last column of each panel, with the most similar prototype patch to each image patch given a black border. These patches are shown in order of the sample image in the reconstructed prototype, with patches that are incorrectly matched shown with red borders. For US a and b, the results are obtained from the full $\boldsymbol{ExPeRT}$ model, in which a illustrates the case where for both image and prototype a same hemisphere is visible, and b where the different hemisphere is visible. Row c is generated from the model trained without consistency loss. For MRI three examples are shown for the model without consistency loss.

C. Sensitivity to hyperparameter choice

The average validation performance across the 5 folds for each of the important hyperparameters is shown in Fig. 6 for the US dataset. For the consistency weight (Fig. 6a), the top plot shows the patch-matching accuracy, which is computed from the overlap with approximated ground-truth matching flows determined from the known visible hemisphere in each of the samples. 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, which explains the small effect for very low weights.

In Fig. 6b 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 sigma 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. 6c and d respectively. Both show improved performance when increasing the number of prototypes or channels, levelling off for higher values.

Overall, these results show that the hyperparameters introduced in our model behave expectedly. This confirms the stability of our method and shows that the results are not too sensitive to the hyperparameter selection.

V. Conclusion

In this work, we presented a novel explainable model for continuous regression tasks based on prototype learning and patch-based OT matching. We showed that our model obtained competitive prediction performance on two brain age prediction datasets; fetal US and adult MRI. On 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 in future work be applied to other continuous regression problems.

References

[1] N. K. Dinsdale, E. Bluemke, V. Sundaresan, M. Jenkinson, S. M. Smith, and A. I. Namburete, "Challenges for machine learning in clinical translation of big data imaging studies," *Neuron*, vol. 110, no. 23, pp. 3866–3881, 2022.

[2] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019.

[3] B. H. van der Velden, H. J. Kuijf, K. G. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis," *Med. Image Anal.*, vol. 79, p. 102 470, 2022.

[4] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018.

[5] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," *arXiv:1806.08049*, 2018.

[6] C. Chen, O. Li, C. Tao, A. J. Barnett, J. Su, and C. Rudin, "This Looks Like That: Deep Learning for Interpretable Image Recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019.

[7] N. K. Dinsdale *et al.*, "Learning patterns of the ageing brain in mri using deep convolutional networks," *NeuroImage*, vol. 224, p. 117 401, 2021.

[8] M. K. Wyburd *et al.*, "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.*, Springer, 2021, pp. 242–252.

[9] L. S. Hesse and A. I. 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.*, Springer, 2022, pp. 502–511.

[10] Y. Rubner, C. Tomasi, and L. J. Guibas, "Earth mover's distance as a metric for image retrieval," *Int. J. of Comput. Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[11] J. H. Cole *et al.*, "Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker," *NeuroImage*, vol. 163, pp. 115–124, 2017.

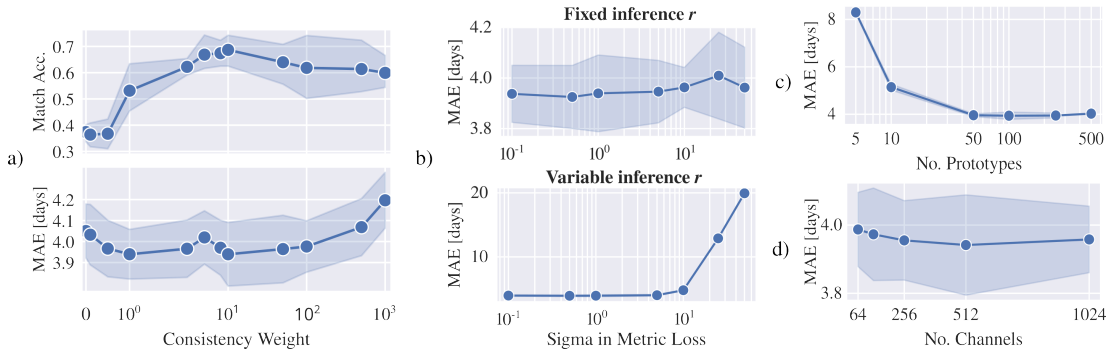

Fig. 6. Overview of hyperparameter tuning for the US dataset for the (a) consistency weight (β), (b) standard deviation in metric loss (σ), (c) number of prototypes and (d) number of channels in the add-on layers and prototype representations. For σ two types of inference are shown, keeping the inference radius, r, constant (top) or varying it in line with the training sigma, $r = 3\sigma$, (bottom).

[12] A. I. Namburete, R. V. Stebbing, B. Kemp, M. Yaqub, A. T. Papageorghiou, and J. Alison Noble, "Learning-based prediction of gestational age from ultrasound images of the fetal brain," *Med. Image Anal.*, vol. 21, no. 1, pp. 72–86, 2015.

[13] S. M. Everwijn *et al.*, "The association between flow and oxygenation and cortical development in fetuses with congenital heart defects using a brain-age prediction algorithm," *Prenatal Diagnosis*, vol. 41, no. 1, pp. 43–51, 2021.

[14] A. Lombardi *et al.*, "Explainable deep learning for personalized age prediction with brain morphology," *Front. in Neurosci.*, vol. 15, 2021.

[15] S. M. Hofmann *et al.*, "Towards the interpretability of deep learning models for multi-modal neuroimaging: Finding structural changes of the ageing brain," *NeuroImage*, vol. 261, p. 119 504, 2022.

[16] H. Li, M. Habes, D. A. Wolk, and Y. Fan, "A deep learning model for early prediction of alzheimer's disease dementia based on hippocampal magnetic resonance imaging data," *Alzheimer's & Dementia*, vol. 15, no. 8, pp. 1059–1070, 2019.

[17] M. Böhle, F. Eitel, M. Weygandt, and K. Ritter, "Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification," *Front. Aging Neurosci.*, vol. 11, 2019.

[18] F. Eitel and K. Ritter, "Testing the robustness of attribution methods for convolutional neural networks in MRI-based Alzheimer's disease classification," in *Interpretability of Mach. Intell. in Med. Image Comp. and Multimodal Learning for Clinical Decision Support*, Springer, Ed., vol. 11797, 2019, pp. 3–11.

[19] K. M. Bintsi, V. Baltatzis, A. Kolbeinsson, A. Hammers, and D. Rueckert, "Patch-Based Brain Age Estimation from MR Images," in *Mach. Learning in Clin. Neuroimag. and Radiogenomics in Neuro-oncology*, vol. 12449, 2020, pp. 98–107.

[20] P. L. Ballester *et al.*, "Predicting Brain Age at Slice Level: Convolutional Neural Networks and Consequences for Interpretability," *Front. Psychiatry*, vol. 12, 2021.

[21] S. G. Popescu, B. Glocker, D. J. Sharp, and J. H. Cole, "Local Brain-Age: A U-Net Model," *Front. in Aging Neurosci.*, vol. 13, 2021.

[22] C. F. Baumgartner, L. M. Koch, K. C. Tezcan, J. X. Ang, and E. Konukoglu, "Visual Feature Attribution Using Wasserstein GANs," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8309–8319.

[23] C. Bass *et al.*, "ICAM-reg: Interpretable classification and regression with feature attribution for mapping neurological phenotypes in individual scans," *IEEE Trans. Med. Imag.*, vol. 42, no. 4, 2023.

[24] N. Papernot and P. McDaniel, "Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning," *arXiv:1803.04765*, 2018.

[25] B. Kim, C. Rudin, and J. A. Shah, "The bayesian case model: A generative approach for case-based reasoning and prototype classification," in *Proc. Adv. in Neur. Inf. Process. Syst.*, vol. 27, 2014.

[26] J. Bien and R. Tibshirani, "Prototype selection for interpretable classification," *Ann. Appl. Stat.*, vol. 5, no. 4, pp. 2403–2424, 2011.

[27] O. Li, H. Liu, C. Chen, and C. Rudin, "Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, AAAI press, 2018, pp. 3530–3537.

[28] E. Kim, S. Kim, M. Seo, and S. Yoon, "Xprotonet: Diagnosis in chest radiography with global and local explanations," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15 719–15 728.

[29] S. Mohammadjafari, M. Cevik, M. Thanabalasingam, and A. Basar, "Using protopnet for interpretable alzheimer's disease classification.," in *Canadian conf. on Artif. Intell.*, 2021.

[30] G. Singh and K.-C. Yow, "An interpretable deep learning model for covid-19 detection with chest x-ray images," *IEEE Access*, vol. 9, pp. 85 198–85 208, 2021.

[31] G. Singh and K.-C. Yow, "These do not look like those: An interpretable deep learning model for image recognition," *IEEE Access*, vol. 9, pp. 41 482–41 493, 2021.

[32] A. J. Barnett *et al.*, "A case-based interpretable deep learning model for classification of mass lesions in digital mammography," *Nat. Mach. Intell.*, vol. 3, no. 12, pp. 1061–1070, 2021.

[33] J. Lu, J. Hu, and J. Zhou, "Deep metric learning for visual understanding: An overview of recent advances," *IEEE Signal Process. Mag.*, vol. 34, pp. 76–84, 6 Nov. 2017.

[34] C. Zhang, Y. Cai, G. Lin, and C. Shen, "DeepEMD: Differentiable earth mover's distance for few-shot learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5632–5648, 2022.

[35] H. Phan and A. 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*, 2022, pp. 20 259–20 269.

[36] W. Zhao, Y. Rao, Z. Wang, J. Lu, and J. Zhou, "Towards interpretable deep metric learning with structural matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 9887–9896.

[37] S. Kim, M. Seo, I. Laptev, M. Cho, and S. Kwak, "Deep metric learning beyond binary supervision," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2288–2297.

[38] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 815–823.

[39] M. Kaya and H. Ş. Bilge, "Deep metric learning: A survey," *Symmetry*, vol. 11, p. 1066, 9 2019.

[40] C. L. Liu and Q. H. Chen, "Metric-based semi-supervised regression," *IEEE Access*, vol. 8, pp. 30 001–30 011, 2020.

[41] G. Mori *et al.*, "Pose embeddings: A deep architecture for learning to match human poses," *arXiv:1507.00302*, 2015.

[42] K. Zheng *et al.*, "Semi-supervised Learning for Bone Mineral Density Estimation in Hip X-Ray Images," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, Springer, 2021, pp. 33–42.

[43] X. Zhao, H. Qi, R. Luo, and L. Davis, "A weakly supervised adaptive triplet loss for deep metric learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019.

[44] W. Dai, X. Li, W. H. K. Chiu, M. D. Kuo, and K. T. Cheng, "Adaptive contrast for image regression in computer-aided disease assessment," *IEEE Trans. on Med. Imag.*, vol. 41, pp. 1255–1268, 5 May 2022.

[45] H. Chao, J. Zhang, and P. Yan, "Regression Metric Loss: Learning a Semantic Representation Space for Medical Images," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, Springer, 2022, pp. 427–436.

[46] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," *Proc. Adv. Neural Inf. Process. Syst*, 2013.

[47] L. C. Torres, L. M. Pereira, and M. H. Amini, "A Survey on Optimal Transport for Machine Learning: Theory and Applications," *arXiv:2106.01963*, 2021.

[48] R. Sinkhorn and P. Knopp, "Concerning nonnegative matrices and doubly stochastic matrices," *Pac. J. Appl. Math.*, vol. 21, no. 2, pp. 343–348, 1967.

[49] P. A. Knight, "The Sinkhorn-Knopp algorithm: Convergence and applications," *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 1, pp. 261–275, 2008.

[50] A. T. Papageorghiou *et al.*, "International standards for fetal growth based on serial ultrasound measurements: The Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project," *Lancet*, vol. 384, no. 9946, pp. 869–879, 2014.

[51] F. Moser, R. Huang, B. W. Papież, A. I. Namburete, I.-2. Consortium, *et al.*, "Bean: Brain extraction and alignment network for 3d fetal neurosonography," *NeuroImage*, vol. 258, p. 119 341, 2022.

[52] 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.*, vol. 56, pp. 476–484, 3 2020.

[53] J. H. Cole, K. Franke, and N. Cherbuin, "Quantification of the biological age of the brain using neuroimaging," *Biomarkers of human aging*, pp. 293–328, 2019.