




# Unsupervised Classification of Congenital Inner Ear Malformations Using DeepDiffusion for Latent Space Representation

Paula López Diez<sup>1(✉)</sup> , Jan Margeta<sup>3,4</sup>, Khassan Diab<sup>5</sup>, François Patou<sup>2</sup>,  
and Rasmus R. Paulsen<sup>1</sup>

<sup>1</sup> DTU Compute, Technical University of Denmark, Kongens Lyngby, Denmark  
plodi@dtu.dk

<sup>2</sup> Oticon Medical, Research and Technology, Smørum, Denmark

<sup>3</sup> Oticon Medical, Research and Technology, Vallauris, France

<sup>4</sup> KardioMe, Research and Development, Nova Dubnica, Slovakia

<sup>5</sup> Tashkent International Clinic, Tashkent, Uzbekistan

**Abstract.** The identification of congenital inner ear malformations is a challenging task even for experienced clinicians. In this study, we present the first automated method for classifying congenital inner ear malformations. We generate 3D meshes of the cochlear structure in 364 normative and 107 abnormal anatomies using a segmentation model trained exclusively with normative anatomies. Given the sparsity and natural unbalance of such datasets, we use an unsupervised method for learning a feature representation of the 3D meshes using DeepDiffusion. In this approach, we use the PointNet architecture for the network-based unsupervised feature learning and combine it with the diffusion distance on a feature manifold. This unsupervised approach captures the variability of the different cochlear shapes and generates clusters in the latent space which faithfully represent the variability observed in the data. We report a mean average precision of 0.77 over the seven main pathological subgroups diagnosed by an ENT (Ear, Nose, and Throat) surgeon specialized in congenital inner ear malformations.

**Keywords:** Unsupervised · Classification · DeepDiffusion · Inner Ear

## 1 Introduction

Inner ear malformations are found in 20–30% of children with congenital hearing loss [1]. While the prevalence of bilateral congenital hearing loss is estimated to be 1.33 per 1000 live births in North America and Europe, it is much higher in sub-Saharan Africa (19 per 1,000 newborns) and South Asia (up to 24 per 1,000) [8]. Early detection of sensorineural hearing loss is crucial for appropriate intervention, such as cochlear implant therapy, which is prescribed to approximately

80,000 infants and toddlers annually worldwide [16]. Radiological examination is essential for an early diagnosis of congenital inner ear malformation, particularly when cochlear implant therapy is planned. However, detecting and classifying such malformations from standard imaging modalities is a complex task even for expert clinicians, and presents challenges during CI surgery [2]. Previous studies have proposed methods to classify congenital inner ear malformations based on explicit measurements and visual analysis of CT scans [5]. These methods are time-consuming and subject to clinician subjectivity. A suggested approach for the automated detection of inner ear malformation has relied on deep reinforcement learning trained for landmark location in normal anatomies based on an anomaly detection technique [9]. However, this method is only limited to the detection of a malformation but does not attempt to classify them.

Currently, supervised deep metric learning garners significant interest due to its exceptional efficacy in data clustering and pathology classification. Most of these approaches are fully supervised and use supervisory signals that model the training by creating tuples of labeled training data. These tuples are then used to optimize the intra-class distance of the different samples in the latent space, as has been done mostly for 2D images [15, 20, 21] and 2D representation of 3D images [4]. Several recent studies have demonstrated promising outcomes from unsupervised contrastive learning from natural images. However, their utility in the medical image domain is limited due to the high degree of inter-class similarity. Particularly in heterogeneous real clinical datasets in which the image quality and appearance can significantly impact the performance of such methods, rendering them less effective. In [22] an unsupervised strategy to learn medical visual representations by exploiting naturally occurring paired descriptive text in 2D images is proposed. Typically, in 3D images, an unsupervised low-dimensional representation is utilized for further clustering, as demonstrated in [14]. Nonetheless, such approaches are commonly developed using quite homogeneous datasets that are not representative of real-world applications and the diverse clinical settings in which they must operate.

Our objective is to develop a fully automated pipeline for the classification of inner ear malformations, utilizing a relatively large and unique dataset of such anomalies. The pipeline’s design necessitates a profound comprehension of this data type and the congenital malformations themselves. Given the CT scans in this region are complex, and the images originate from diverse sources, we employ an unsupervised approach, uniquely based on the 3D shape of the cochlear structure. We have observed that the cochlear structure can be roughly but consistently segmented by a 3D-UNET model trained exclusively on normal cochlear anatomies. We then use these segmentations and adopt an entirely unsupervised approach, meaning the deep learning model is trained from scratch on these segmentations, and the class labels are not used for training. To map these shapes to an optimal latent space representation, we utilize DeepDiffusion, which combines the diffusion distance on a feature manifold with the feature learning of the encoder.

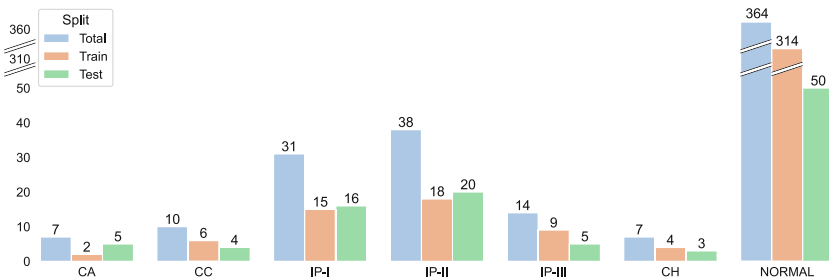
In this paper, we present the first automatic approach for the classification of congenital inner ear malformations. We use an unsupervised method to find the latent space representation of cochlear shapes, which allows for their further classification. We demonstrate that shapes from a segmentation model trained on normative cases, albeit imperfect, can be used to represent abnormalities. Moreover, our results indicate the potential for successfully applying this approach to other anatomies.

## 2 Data

Our dataset comprises a total of 485 clinical CT scans, consisting of 364 normal scans and 121 scans with various types of inner ear malformations. The distribution of inner ear scans for each type of malformation is shown in Fig. 1. We utilized the region-of-interest (ROI) extraction technique developed by [18], which involves selecting anatomical points of interest that are not part of the inner ear region to achieve a standardized and robust image orientation. To ensure consistency, all images were resampled to a spacing of 0.125 mm, and their intensities were normalized by scaling the 5<sup>th</sup> and 95<sup>th</sup> percentiles of the intensity distribution of each image to 0 and 1, respectively. Figure 1 also shows the data split used for training our model. We chose to use an approximate 50% split for abnormal cases, while the vast majority of normal cases, approximately 86%, were used for training. Other configurations were explored, including using only normal cases for training. However, it was demonstrated that while this approach may work for anomaly detection, it does not adequately categorize the different types of malformations.

## 3 Methods

### 3.1 Anatomical Representation

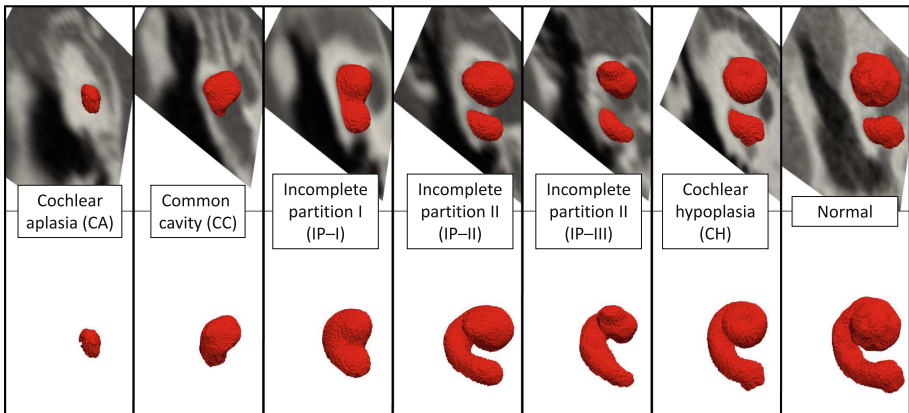


**Fig. 1.** Distribution of cases among the different classes and the split used for our approach. Cochlear aplasia (CA), common cavity (CC) incomplete partitioning type I, II, and III (IP-I, IP-II, IP-III), cochlear hypoplasia (CH), and normal.

Our aim is to find a parametrized shape that is representative of the anatomy of the patient. We decided to focus on the cochlear structure as it is the main

structure of interest when trying to identify a malformation in the inner ear. To obtain a 3D segmentation of this structure we use the 3D-UNET [19] presented in [10] which has been trained exclusively in normal anatomies (130 images from diverse imaging equipment) and built using MONAI [12]. Even though no abnormal anatomies have been used for training, given the high contrast between the soft tissue of the cochlear structure and the bony structure that surrounds it, the model still performs quite well to segment the abnormal cases. This can be seen in Fig. 2 where an example of each of the types of malformations used in this study and an anatomically normal case are shown. The largest connected component of the segmentation has been selected to generate the final 3D meshes.

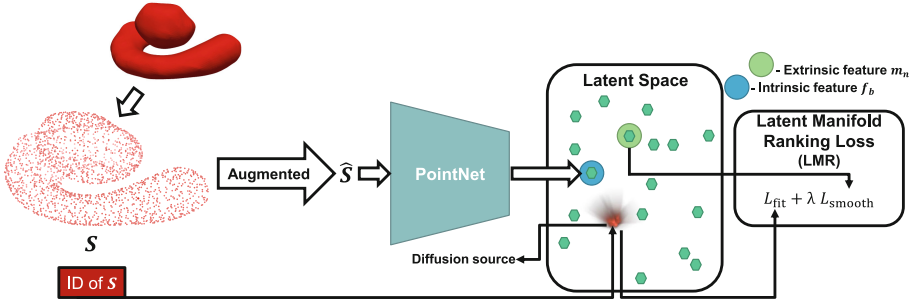
An overview of our pipeline is presented in Fig. 3. Each 3D mesh obtained from a CT image is transformed into a 1024 point cloud using the Ohbuchi method [13]. Each shape is then normalized by centering its origin in its center of gravity and enclosing the shape within a unit sphere, resulting in the point cloud representation of the shape  $S$ . Before the shape  $S$  is fed to the encoder, the shape is augmented into shape  $\hat{S}$  with a probability of 0.8. This augmentation consists of a random rotation with  $U(-5^\circ, 5^\circ)$ , an anisotropic scaling sampled from  $U(0.8, 1)$ , and a shearing and translation in each axes sampled from  $U(-0.2, 0.2)$  for both actions.



**Fig. 2.** Representative 3D-UNET segmentation meshes from each type of cochlear anatomy used in this study. Top row shows the 3D mesh with the original CT scan image; the bottom row shows exclusively the 3D mesh.

### 3.2 Deep Diffusion Algorithm

The DeepDiffusion (DD) algorithm [7] incorporates the manifold ranking [23] technique, which uses similarity diffusion on the manifold graph to learn a distance metric among the samples. The DD algorithm optimizes both the feature extraction and the embeddings produced by the encoder, which results in



**Fig. 3.** Sketch of the DeepDiffusion used for latent space representation of the cochlear 3D meshes. The pointcloud extracted from the mesh is fed to the PointNet encoder which generates the corresponding latent feature which is optimized by minimizing the LMR loss so the encoder and the latent feature manifold are optimized for the comparison of data samples.

salient features in a continuous and smooth latent space. In this latent space, the Euclidean distance among the latent features approximates the diffusion distance on the latent feature manifold. The crux behind this algorithm is the latent manifold ranking loss (LMR) which is computed using both intrinsic and extrinsic features. The LMR consists of a fitting term,  $L_{\text{fit}}$ , a smoothing term,  $L_{\text{smooth}}$ , and a balancing term,  $\lambda$ .

$$\text{LMR} = \arg \min_{M, \theta} L_{\text{fit}} \pm \lambda L_{\text{smooth}} \quad (1)$$

Where  $\theta$  characterizes the encoder and  $M \in \mathbb{R}^{N \times P}$  represents the latent feature manifold formed by the training samples, where  $N$  is the number of data samples and  $P$  is the output dimensions of the encoder. The extrinsic feature  $f$  is defined as the output of the encoder and has dimension  $P$ .  $M$  is initialized by stacking together the embeddings of the first forward pass through the encoder which has been randomly initialized as this has been shown to perform better than randomly initializing the weights of  $M$  itself as shown in [7].

Every training sample has its unique identification number ( $\text{ID}_b$ ) which is used to specify a diffusion source  $y_b$  that is consistent throughout the training procedure.  $L_{\text{fit}}$  constrains the ranking vector  $r_b$  to being close to the diffusion source  $y_b$ , which is defined as the vector containing one-hot encoding of  $\text{ID}_b$ . The ranking vector is defined as  $r_b = \text{softmax}(f_b M^T)$  and represents the probabilistic similarities between the feature  $f_b$  and all the intrinsic features contained in  $M$ . The fitting term is therefore defined as

$$L_{\text{fit}} = \sum_b \text{CrossEntropy}(r_b, y_b)$$

its minimization results in all the extrinsic features being embedded farther away from each other as they are being pulled toward their respective and unique

diffusion source vectors. The smoothing term is defined as

$$L_{\text{smooth}} = \sum_b \sum_n w_{bn} \text{Dissimilarity}(r_b, t_n) \quad (2)$$

where the dissimilarity operator is the Jensen-Shannon divergence [6] and  $t_n = \text{softmax}(m_n M^T)$  being  $m_n$  the  $n^{\text{th}}$  row of the matrix  $M$  so that  $t_n$  contains the ranking score of the intrinsic feature  $m_n$  to all the intrinsic features.  $w_{bn}$  indicates the similarity between the extrinsic feature  $f_b$  and the neighboring intrinsic feature  $m_n$  and it is defined as:

$$w_{bn} = \begin{cases} f_b m_n^T, & m_n \in \text{kNN}(f_b) \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Minimizing  $L_{\text{smooth}}$  pulls extrinsic features and their neighboring intrinsic features together which implies that an extrinsic feature is more likely to be projected onto the surface of the latent feature manifold of the intrinsic features when  $L_{\text{smooth}}$  is smaller.

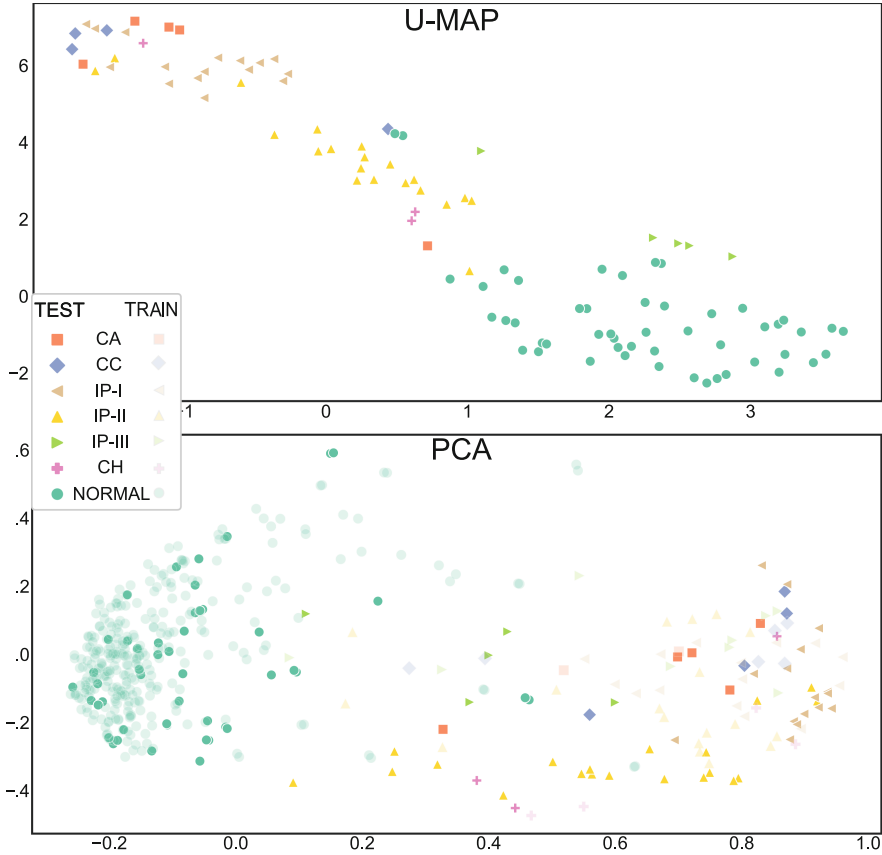
### 3.3 Implementation

For our encoder, we use the PointNet [3] architecture which takes 1024 3D points as input, applies input and feature transformations, and then aggregates point features by max pooling to a feature of dimensionality 1024 which is then compressed into dimensionality 254 with two sets of fully connected layers. The network has been trained by using mini-batch (of size 8) gradient descent using the Adam optimizer with a learning rate of  $10^{-8}$  and ReLU as the activation function. The DD algorithm is implemented in PyTorch [17] and the code used for this study is available at <https://github.com/paulalopez10/Deep-Diffusion-Unsupervised-Classification-3D-Mesh>. The models are trained on an NVIDIA GeForce RTX 3070 Laptop GPU with 8GB VRAM. The different hyper-parameters related to the approach have been explored and it has been empirically found for this specific configuration  $\lambda = 0.6$  and  $k = 10$  produce the best results that will be analyzed in the following section.

## 4 Results

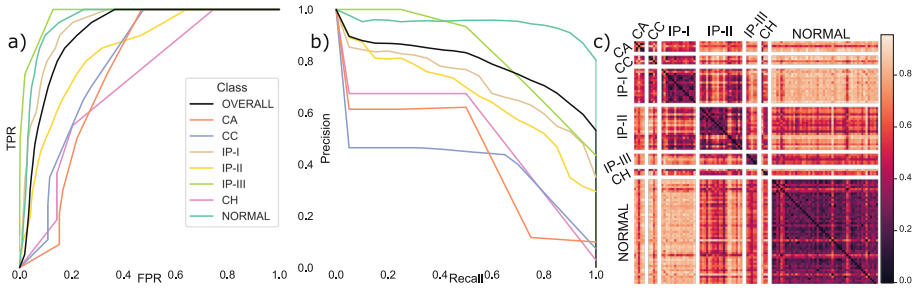
We evaluate the classification performance of our pipeline by analyzing the embeddings generated by the trained encoder. To visualize the projection of the features of the test in 2D we use the U-MAP [11], as illustrated in Fig. 4. The U-MAP visualization demonstrates the clustering of different classes in the latent space. Furthermore, it is very interesting to notice how the latent space representation displays the anatomical changes of the anatomy where the more extreme types of malformations (CA and CC) are the most distant to the normative cochlear structures. The transition between the different classes shown in the latent space properly represents the pathological variations in this anatomy.

We have also included, in Fig. 4, the projection of the features projected in the 2D-PCA space defined by the training set, where both, training and testing, sets are included to show not only the clustering in this space but also the similar distribution of the different classes in both sets within the PCA projection.



**Fig. 4. Top:** U-MAP representation of the test features where it can be observed how the different classes group together and how the anatomical variation is represented as there is a progression from the most abnormal cases towards fully normal cases. **Bottom:** Test and train features projected into the 2D PCA space defined by the training samples, where the classes are separated and show consistency between training and testing samples.

For a further analysis of the performance, we compute some evaluation metrics based on the pairwise cosine distance between samples that can be seen in Fig. 5 c). The average ROC and precision-recall curves for each of the classes can be seen in Fig. 5 a) and b). To calculate those, each test feature vector  $f_b$  is considered to be the centroid of a  $kNN(f_b)$  which consists in the  $k$  nearest



**Fig. 5.** Evaluation plots. a) Mean ROC curves for each class b) Mean Recall-Precision curve for each class c) Pairwise cosine distance between test embeddings used to evaluate the performance.

**Table 1.** Evaluation metrics reported in our experiment. ROC-Receiver operating characteristic, AUC- Area under the curve, AP-Average precision, PR - Precision-recall

	CA	CC	IP-I	IP-II	IP-III	CH	NORMAL	Overall
<b>Max Accuracy</b>	0.98	0.96	0.92	0.96	0.98	0.99	0.91	0.93
<b>Mean Accuracy</b>	0.73 $\pm 0.25$	0.77 $\pm 0.26$	0.87 $\pm 0.03$	0.77 $\pm 0.08$	0.96 $\pm 0.03$	0.69 $\pm 0.01$	0.75 $\pm 0.05$	0.78 $\pm 0.12$
<b>Max ROC-AUC</b>	0.99	0.99	0.98	0.96	0.99	0.98	0.99	0.99
<b>Mean ROC-AUC</b>	0.75 $\pm 0.25$	0.79 $\pm 0.25$	0.94 $\pm 0.04$	0.84 $\pm 0.10$	0.98 $\pm 0.02$	0.71 $\pm 0.08$	0.95 $\pm 0.09$	0.91 $\pm 0.13$
<b>Max AP</b>	0.57	0.70	0.87	0.88	0.92	0.51	0.99	0.91
<b>Mean AP</b>	0.41 $\pm 0.42$	0.38 $\pm 0.33$	0.70 $\pm 0.23$	0.65 $\pm 0.32$	0.82 $\pm 0.25$	0.50 $\pm 0.42$	0.94 $\pm 0.12$	0.77 $\pm 0.29$
<b>Max f1-score</b>	0.67	0.57	0.67	0.88	0.86	0.67	0.91	0.75
<b>Max PR-AUC</b>	0.63	0.80	0.84	0.85	0.91	0.71	0.95	0.88
<b>Mean PR-AUC</b>	0.40 $\pm 0.26$	0.37 $\pm 0.25$	0.67 $\pm 0.04$	0.62 $\pm 0.11$	0.78 $\pm 0.02$	0.48 $\pm 0.08$	0.90 $\pm 0.10$	0.74 $\pm 0.24$

features from other samples using the cosine pairwise distance shown in Fig. 5 c). We vary  $k$  until all the features from the corresponding class are within the cluster and compute the precision and false positive rate per the different recall steps, the shown results are the average among each class and overall. With the same procedure, different evaluation metrics have been obtained and are shown in Table 1. These metrics encompass the area under the curve (AUC) for the curves shown in Fig. 5 a) and b), both for the average curve and the optimal curve for each class. Furthermore, the maximum and average accuracy has been computed together with the maximum f1-score. Considering the dataset's significant class imbalance, these metrics provide a comprehensive assessment of the performance achieved. Finally, the mean average precision is also included in the table together with the optimal one for each class. The optimal or maximum value of each metric corresponds to when the optimal sample within our test features distribution is being evaluated as the centroid of its own class and the



mean values are the average over all the samples. We can observe how a bigger variance is obtained for the classes that contain a few examples as it is expected, given the nature and distribution of our dataset shown in Fig. 1.

## 5 Conclusion

We have presented the first approach for the automatic classification of congenital inner ear malformations. We show how using the 3D shape information of the cochlea obtained with a model only trained in normative anatomies is enough to classify the malformations and reduces the influence of the image's source, which is crucial in a clinical application setting.

Our method shows a mean average precision of 0.77 with a mean ROC-AUC of 0.91, indicating its effectiveness in classifying inner ear malformations. Furthermore, the representation of the different cases in the latent space shows spatial relation between classes, which is correlated with the anatomical appearance of the different malformations. These promising results pave the way towards assisting clinicians in the challenging assessment of congenital inner ear malformations potentially leading to improved patient outcome of cochlear surgery.

## References

1. Brotto, D., et al.: Genetics of inner ear malformations: a review. *Audiol. Res.* **11**(4), 524–536 (2021). <https://doi.org/10.3390/audiolres11040047>
2. Chakravorti, S., et al.: Further evidence of the relationship between cochlear implant electrode positioning and hearing outcomes. *Otol. Neurotol.* **40**(5), 617–624 (2019). <https://doi.org/10.1097/MAO.0000000000002204>
3. Charles, R.Q., Su, H., Kaichun, M., Guibas, L.J.: Pointnet: Deep learning on point sets for 3D classification and segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 77–85 (2017). <https://doi.org/10.1109/CVPR.2017.16>
4. Chen, X., Wang, W., Jiang, Y., Qian, X.: A dual-transformation with contrastive learning framework for lymph node metastasis prediction in pancreatic cancer. *Med. Image Anal.* **85**, 102753 (2023). <https://doi.org/10.1016/j.media.2023.102753>, <https://www.sciencedirect.com/science/article/pii/S1361841523000142>
5. Dhanasingh, A.E., et al.: A novel three-step process for the identification of inner ear malformation types. *Laryngoscope Investigative Otolaryngology* (2022). <https://doi.org/10.1002/lio2.936>, <https://onlinelibrary.wiley.com/doi/10.1002/lio2.936>
6. Fuglede, B., Topsoe, F.: Jensen-shannon divergence and hilbert space embedding. In: International Symposium on Information Theory, 2004. ISIT 2004. Proceedings, pp. 31– (2004). <https://doi.org/10.1109/ISIT.2004.1365067>
7. Furuya, T., Ohbuchi, R.: Deepdiffusion: unsupervised learning of retrieval-adapted representations via diffusion-based ranking on latent feature manifold. *IEEE Access* **10**, 116287–116301 (2022). <https://doi.org/10.1109/ACCESS.2022.3218909>

8. Korver, A.M., et al.: Congenital hearing loss. *Nature Rev. Disease Primers* **3**(1), 1–17 (2017)
9. López Diez, P., et al.: Deep reinforcement learning for detection of inner ear abnormal anatomy in computed tomography. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) *Medical Image Computing and Computer Assisted Intervention - MICCAI 2022*, pp. 697–706. Springer Nature Switzerland, Cham (2022). [https://doi.org/10.1007/978-3-031-16437-8\\_67](https://doi.org/10.1007/978-3-031-16437-8_67)
10. Margeta, J., et al.: A web-based automated image processing research platform for cochlear implantation-related studies. *J. Clin. Med.* **11**(22) (2022). <https://doi.org/10.3390/jcm11226640>, <https://www.mdpi.com/2077-0383/11/22/6640>
11. McInnes, L., Healy, J., Saul, N., Großberger, L.: Umap: Uniform manifold approximation and projection. *J. Open Source Softw.* **3**(29), 861 (2018). <https://doi.org/10.21105/joss.00861>
12. MONAI-Consortium: Monai: Medical open network for AI (2022). <https://doi.org/10.5281/zenodo.7459814>
13. Ohbuchi, R., Minamitani, T., Takei, T.: Shape-similarity search of 3D models by using enhanced shape functions. *Int. J. Comput. Appl. Technol.* **23**(2–4), 70–85 (2005)
14. Onga, Y., Fujiyama, S., Arai, H., Chayama, Y., Iyatomi, H., Oishi, K.: Efficient feature embedding of 3d brain mri images for content-based image retrieval with deep metric learning. In: *2019 IEEE International Conference on Big Data (Big Data)*, pp. 3764–3769. IEEE (2019)
15. Pal, A., et al.: Deep metric learning for cervical image classification. *IEEE Access* **9**, 53266–53275 (2021)
16. Paludetti, G., et al.: Infant hearing loss: from diagnosis to therapy official report of xxi conference of Italian society of pediatric otorhinolaryngology. *Acta Otorhinolaryngol. Italica* **32**(6), 347 (2012)
17. Paszke, A., et al.: Pytorch: An imperative style, high-performance deep learning library. In: *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc. (2019). <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>
18. Radutoiu, A.T., Patou, F., Margeta, J., Paulsen, R.R., López Diez, P.: Accurate localization of inner ear regions of interests using deep reinforcement learning. In: Lian, C., Cao, X., Rekik, I., Xu, X., Cui, Z. (eds.) *Machine Learning in Medical Imaging*. pp. 416–424. Springer Nature Switzerland, Cham (2022). [https://doi.org/10.1007/978-3-031-21014-3\\_43](https://doi.org/10.1007/978-3-031-21014-3_43)
19. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015*, pp. 234–241. Springer International Publishing, Cham (2015)
20. Sundgaard, J.V., et al.: Deep metric learning for otitis media classification. *Med. Image Anal.* **71**, 102034 (2021). <https://doi.org/10.1016/j.media.2021.102034>, <https://www.sciencedirect.com/science/article/pii/S1361841521000803>
21. Zhang, Y., Luo, L., Dou, Q., Heng, P.A.: Triplet attention and dual-pool contrastive learning for clinic-driven multi-label medical image classification. *Med. Image Anal.* **86**, 102772 (2023). <https://doi.org/10.1016/j.media.2023.102772>, <https://www.sciencedirect.com/science/article/pii/S1361841523000336>

22. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Proceedings of the 7th Machine Learning for Healthcare Conference. Proceedings of Machine Learning Research, vol. 182, pp. 2–25. PMLR (2022). <https://proceedings.mlr.press/v182/zhang22a.html>
23. Zhou, D., Weston, J., Gretton, A., Bousquet, O., Schölkopf, B.: Ranking on data manifolds. In: Thrun, S., Saul, L., Schölkopf, B. (eds.) Advances in Neural Information Processing Systems. vol. 16. MIT Press (2003). <https://proceedings.neurips.cc/paper/2003/file/2c3ddf4bf13852db711dd1901fb517fa-Paper.pdf>