Deep reinforcement learning and convolutional autoencoders for anomaly detection of congenital inner ear malformations in clinical CT images

López Diez, Paula; Sundgaard, Josefine Vilsbøll; Margeta, Jan; Diab, Khassan; Patou, François; Paulsen, Rasmus R.

Published in:
Computerized Medical Imaging and Graphics

Link to article, DOI:
10.1016/j.compmedimag.2024.102343

Publication date:
2024

Document Version
Publisher's PDF, also known as Version of record

Citation (APA):
López Diez, P., Sundgaard, J. V., Margeta, J., Diab, K., Patou, F., & Paulsen, R. R. (2024). Deep reinforcement learning and convolutional autoencoders for anomaly detection of congenital inner ear malformations in clinical CT images. *Computerized Medical Imaging and Graphics*, *113*, Article 102343. https://doi.org/10.1016/j.compmedimag.2024.102343

Computerized Medical Imaging and Graphics 113 (2024) 102343

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

Deep reinforcement learning and convolutional autoencoders for anomaly detection of congenital inner ear malformations in clinical CT images

Paula López Diez [a,*], Josefine Vilsbøll Sundgaard [a,f], Jan Margeta [c,e], Khassan Diab [d], François Patou [b], Rasmus R. Paulsen [a]

[a] Department of Applied Mathematics and Computer Science, Technical University of Denmark, Lyngby, Denmark
[b] Oticon Medical, Research & Technology group, Smørum, Denmark
[c] KardioMe, Research & Development, Nova Dubnica, Slovakia
[d] Tashkent International Clinic, Tashkent, Uzbekistan
[e] Oticon Medical, Research & Technology, Vallauris, France
[f] Novo Nordisk A/S, Denmark

ARTICLE INFO

Keywords:
Computed tomography
Anomaly detection
Deep reinforcement learning
Congenital malformation
Cochlear implant

ABSTRACT

Detection of abnormalities within the inner ear is a challenging task even for experienced clinicians. In this study, we propose an automated method for automatic abnormality detection to provide support for the diagnosis and clinical management of various otological disorders. We propose a framework for inner ear abnormality detection based on deep reinforcement learning for landmark detection which is trained uniquely in normative data. In our approach, we derive two abnormality measurements: D_{image} and U_{image}. The first measurement, D_{image}, is based on the variability of the predicted configuration of a well-defined set of landmarks in a subspace formed by the point distribution model of the location of those landmarks in normative data. We create this subspace using Procrustes shape alignment and Principal Component Analysis projection. The second measurement, U_{image}, represents the degree of hesitation of the agents when approaching the final location of the landmarks and is based on the distribution of the predicted Q-values of the model for the last ten states. Finally, we unify these measurements in a combined anomaly measurement called C_{image}. We compare our method's performance with a 3D convolutional autoencoder technique for abnormality detection using the patch-based mean squared error between the original and the generated image as a basis for classifying abnormal versus normal anatomies. We compare both approaches and show that our method, based on deep reinforcement learning, shows better detection performance for abnormal anatomies on both an artificial and a real clinical CT dataset of various inner ear malformations with an increase of 11.2% of the area under the ROC curve. Our method also shows more robustness against the heterogeneous quality of the images in our dataset.

1. Introduction

Inner ear malformation has been reported with an incidence of 20%–30% among children with congenital hearing loss (Brotto et al., 2021). The prevalence of bilateral congenital hearing loss is estimated as 1.33 per 1000 live births in North America and Europe, while in sub-Saharan Africa, the estimate is 19 per 1,000 newborns, and in South Asia up to 24 per 1,000 (Korver et al., 2017). Sensorineural hearing loss is generally detected early in countries with good access to healthcare services, which allows the prescription of interventions that mitigate the risk of abnormal social, emotional, and communicative development. These interventions include cochlear implant (CI) therapy, which is prescribed each year to about 80,000 infants and toddlers worldwide (Paludetti et al., 2012).

Radiological examination of children born with sensorineural hearing loss is key for an early diagnosis of congenital inner ear malformation. When the patient is treated with cochlear implant therapy, during the radiological examination the anatomy of the patient is evaluated to plan the surgical strategy. This surgery usually consists of drilling a precise tunnel from the surface of the scalp to the scala tympani in the cochlea where the implant is placed. The final location of the implanted electrode is critical for the patient's outcome (Chakravorti et al., 2019). Cases presenting congenital inner ear malformations raise

[*] Corresponding author.
E-mail address: plodi@dtu.dk (P. López Diez).

https://doi.org/10.1016/j.compmedimag.2024.102343
Received 21 September 2023; Received in revised form 25 January 2024; Accepted 28 January 2024
Available online 2 February 2024

many challenges during the planning and execution of CI surgery, often necessitating the surgeon to discover and adapt as the procedure progresses. Detecting and identifying such malformations from standard imaging modalities is a complex task even for expert clinicians. The detection and classification of the type of malformation is not a trivial task given the complexity of the anatomy and the great anatomical variation among malformations. Studies have defined several categories for these malformations, such as Sennaroğlu and Bajin (2017) which is one of the most popular works used for classifying congenital malformations of the inner ear. Dhanasingh et al. (2021, 2022) defined possible strategies for clinicians to detect congenital inner ear malformations based on the visual exploration of CT scans involving explicit measurements and humans' natural ability for pattern recognition. The strategy described in Dhanasingh et al. (2022) is based on visualizing the cochlea in two specific planes (oblique-coronal and mid-modiolar) and following three steps: the cochlear A and B distances defined as Escudé et al. (2006), the number of cochlear turns, and a visual analysis based on an assessment of resemblance to different objects such as the Aladdin's lamp and a side view of a dog's face. This methodology provides empirical guidelines for clinicians to detect these malformations based on hand-crafted features. As with any human-based image interpretation method, it is time-consuming and can be subject to the clinician's subjectivity during evaluation.

In López Diez et al. (2022b), we introduced the first automated approach for detecting inner ear congenital malformations. In that approach, we used a deep reinforcement learning (DRL) model trained for landmark localization exclusively in normal anatomies to derive two anomaly measurements: the first was based on the variability of the predicted configuration of the landmarks in a subspace formed by the point distribution model of the normative landmarks' location using Procrustes shape alignment and principal component analysis (PCA) projection. The second measurement was based on the distribution of the predicted Q-values of the model for the last ten states before the landmarks were localized. In the present paper, we build on our prior work and compare this approach to a 3D convolutional autoencoder approach in which the patch-based reconstruction error is used for anomaly detection in 3D images, as described by Sato et al. (2018).

This journal paper presents a significant extension of the conference work presented at MICCAI 2022 (López Diez et al., 2022b). We have extended the literature review, especially for unsupervised anomaly detection (UAD) in medical images. Furthermore, we have chosen to use the MARL architecture and not the communicative version (C-MARL), as it has proven best for this task based on our findings, and we have closely benchmarked our method against another state-of-the-art 3D-based approach for UAD in both artificially generated anomalies and clinical images of congenital inner ear malformations. This helps us to understand how current approaches tested on brain datasets might perform in different and more challenging datasets, as is the case with our rare clinical dataset. Furthermore, our approach allows us to assess how the performance generalizes from artificially generated datasets and heterogeneous real clinical scans to scans from a very specific and controlled environment.

2. Background

Recently, machine learning has enabled the development of automated medical image analysis methods that achieve great levels of performance in the detection of abnormal anatomies and other types of anatomical anomalies (Sajid et al., 2019; Wang et al., 2022). Despite their outstanding performance, these methods, which are based on supervised learning, have a major disadvantage: they require large labeled datasets that faithfully represent the spectrum of possible anomalies. These datasets are scarce, and costly to obtain, especially for rare diseases such as congenital inner ear malformation. Furthermore, it is very difficult to predict how these supervised models will behave with new unseen data. Lately, some deep learning approaches that seek to enhance UAD have been introduced. These new deep-learning-based UAD methods resemble the clinical approach to image exploration and can detect anomalies without prior knowledge about the anomalies' appearance.

Historically, UAD was based on statistical models (Van Leemput et al., 2001), out-of-the-distribution techniques (Prastawa et al., 2004; Allenby et al., 2021), hand-crafted features (Martins et al., 2020), content-based retrieval or clustering (Taboada-Crispi et al., 2009). Nowadays, approaches based on isolation forest (Liu et al., 2012), like (Hariri et al., 2021; Xu et al., 2023a) for UAD, have gained significant popularity and demonstrate high performance. However, these approaches rely on the anomalies being distinct and sparse, being this last one a condition not met by the datasets utilized in this study. Furthermore, these approaches have been used only in features or lower-dimensional spaces derived from CT images as in the work presented by Hainan et al. (2019) and Welch et al. (2020). Nevertheless, they are not yet suitable for direct application to high-dimensional clinical data, such as CT or MRI images. This misalignment contradicts our goal of utilizing an approach directly applicable to the entire CT image. New deep learning techniques are being tested for UAD as the ones presented in Wang et al. (2023), Xu et al. (2023b). However, in the medical image domain, most of the UAD methods used for 3D images are based on an autoencoder approach. The underlying concept is to learn an implicit and synthesized representation of a certain type of image in normative samples and use the difference between the original image and the one produced by the generative branch of the model to estimate the probability of the given sample being an anomaly. Different versions of convolutional autoencoders (CAE) (Baur et al., 2019; Sato et al., 2018; Atlason et al., 2019; Astaraki et al., 2022) and variational autoencoders (VAE) (Chen et al., 2020; Pawlowski et al., 2018; Zimmerer et al., 2019; Astaraki et al., 2022) have been tested for UAD in medical images. These are popular strategies to tackle unsupervised anomaly segmentation by modeling the distribution of normal images. In a similar manner, different GANs-based approaches (Schlegl et al., 2019; Baur et al., 2020; Sun et al., 2020; Schlegl et al., 2017) have been used for this purpose as well. More recently, autoencoders with transformers (Pinaya et al., 2022b) and diffusion models (Pinaya et al., 2022a; Wolleb et al., 2022) have been proposed for UAD. Finally, attention-map-based approaches such as the ones presented by Silva-Rodríguez et al. (2022) and Venkataramanan et al. (2020) are also being used for UAD on medical images.

Besides the work by Sato et al. (2018) and Pinaya et al. (2022b), all the previously mentioned works have proposed 2D-based approaches, even though some are used for processing volumetric data, mainly MRI and CT scans. These 2D approaches do not exploit all the implicit information of the 3D scans, even if they are computationally more efficient. This inability to exploit all the information is problematic in UAD for complex anatomies, such as the inner ear, for which 3D spatial information is essential in correctly analyzing the internal structures, which are small and interconnected with a high degree of curvature. Despite their success, transformer-based approaches, such as Pinaya et al. (2022b), still have some weaknesses intrinsic to their autoregressive nature, as it is the fixed order of sequence elements that creates a bias to attention. This problem is more noticeable in 3D images, where even more transformers might be required to achieve good coverage of the image context given the images' higher dimensionality. Therefore, we chose to compare our proposed 3D-based UAD method with the asymmetric 3D convolutional autoencoder described by Sato et al. (2018) as we consider it the best-suited approach for direct comparison with a low demand of computational resources.

In a similar fashion, we tackle the anomaly detection problem with a parametric approach instead of a classification one, in an attempt to build implementations that move toward more interpretable results. We decided to use the landmark-based approach as an object search problem using a DRL approach trained in normative data to derive implicit information that can be used for UAD. The implicit information used is of two different types. The first type is based on the variability of the predicted configuration of the pre-specified landmarks in a subspace defined by the point distribution model of the normative location of the landmarks using Procrustes shape alignment and Principal Component Analysis projection. The second type is based on the distribution of the predicted Q-values of the model for the last ten states before the landmarks are localized as an agent hesitation measurement. Landmarks located with DRL have been used for anomaly detection by Bekkouch et al. (2022), where a method for abnormality detection in 2D X-ray images of the hip is proposed. This method is however not a UAD as it is not unsupervised because the agents are trained to localize the landmarks in abnormal cases and their prediction is then used to estimate if they fall within the healthy population expected inter-landmark relationship. This approach is therefore limited, as are all the supervised methods, by the size and representativity of the dataset used in comparison to all the possible abnormal cases.

Deep reinforcement learning consists of the use of deep learning to solve a reinforcement learning problem. Deep reinforcement learning has been applied to medical images with great success over the last years for parametric medical image analysis, optimization problems, and image classification (Zhou et al., 2021). Even though the automatic detection of the different types of inner ear malformation is, by nature, a classification problem, due to the lack of availability of representative and heterogeneous datasets that faithfully represent the full spectrum of these congenital malformations, we use a parametric approach (landmark detection) in normative data to derive implicit information that can potentially detect an anomaly in the anatomy in an unsupervised manner.

Many existing UAD approaches focus on the detection of brain cancer (Nazir et al., 2021), where cancer can appear at almost random locations in the entire brain. The goal of those approaches is to detect where the brain looks abnormal compared to a normative population. In our case, we are specifically looking at a very small anatomical region, the cochlea, that when abnormal might have a very different overall appearance compared to a normative population. Our assumption is that the configuration of a limited number of anatomical landmarks can provide the necessary information for anomaly detection. For a brain scan with a randomly placed tumor, a landmark-based UAD would require an extensive amount of anatomical landmarks and we do not believe that our DRL approach would be suitable for that task.

3. Materials and methods

3.1. Data

In this study, two different datasets have been used to evaluate the different methods: an artificial dataset and a clinical dataset.

The artificial dataset consists of 119 clinical CT scans of patients with normal inner ear anatomy. The cochlear structure presents some variability in normative patients but it is fairly consistent across this population (Demarcy et al., 2017). This dataset is composed of images from diverse CT scanners which were cropped to a standard view and orientation of the region of interest (ROI) of 32.1^3 mm^3 using the Nautilus software (Margeta et al., 2022) and their proposed orientation. These images were labeled with five anatomical points of interest for nerve characterization (López Diez et al., 2021), an example of which is shown in Fig. 1. To test our approach, we synthetically generated abnormal inner ear CT scans from the original images by removing the cochlea (simulating cochlear aplasia), thus generating corresponding pairs of normal and abnormal CT scans with the same surrounding structures. The cochlea was segmented using ITK-SNAP software (Yushkevich et al., 2006) and then replaced by Gaussian noise with mean and standard deviation estimated from the intensities of the tissue surrounding the segmentation (López Diez et al., 2022a).

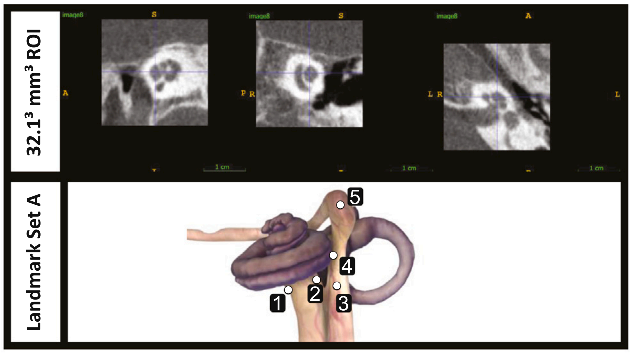


Fig. 1. Artificial dataset. Top: Example of CT scan from the artificial dataset with a ROI of 32.1^3 mm^3 and isotropic spacing of 0.2 mm. Bottom: Landmark set A: **1, 2** - Opposite sides of bony cochlear nerve canal in axial view. **3** - Facial Nerve (FN) exiting the internal acoustic canal. **4** - Closest point of FN and cochlea. **5** - Geniculate ganglion of the FN.
Source: Figure edited from Trier et al. (2008).

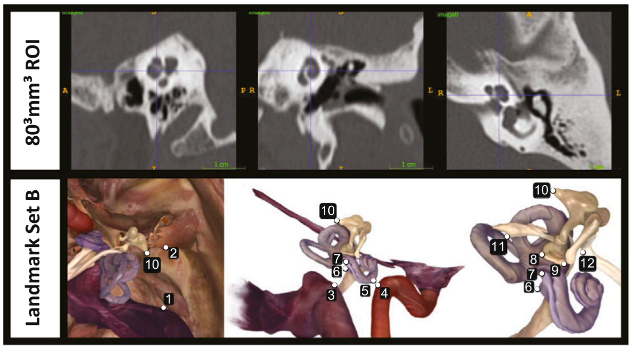


Fig. 2. Clinical dataset. Top: Example of CT image from the clinical dataset with a ROI of 80^3 mm^3 and isotropic spacing of 0.15 mm. Bottom: Landmark set B: **1** - Sigmoid sinus (closest point to the external acoustic canal). **2** - External acoustic canal (closest point to).sigmoid sinus **3** - Jugular bulb (closest point to round window). **4** - Carotid artery (closest point to basal turn of the cochlea). **5** - Basal turn (closest point to JB). **6,7** - Anterior and posterior edges of RW. **8,9** - Anterior and posterior crus of staples. **10** - Short process of incus. **11** - Pyramidal process **12**- Cochleariform process.
Source: Figure edited from Trier et al. (2008).

Our second dataset, the clinical dataset, consists of 300 anatomically normal CT scans from heterogeneous sources and 122 CT scans of inner ears that present different types of inner ear malformations. The ROI extraction for this dataset was done using the methodology described by Radutoiu et al. (2022) using anatomical points of interest that were not involved in the anatomy of interest in order to allow for a standardized and robust image orientation regardless of the appearance of the inner ear region. A greater ROI of 80^3 mm^3 was selected for this dataset in order to contain all the anatomical points of interest for CI therapy. For this dataset, as shown in Fig. 2, twelve anatomically relevant landmarks were carefully designed and annotated in a randomly selected subset of 160 CT scans of anatomically normal cases in collaboration with our clinical partner, an ENT surgeon specialized in CI therapy in abnormal anatomies.

3.2. DRL for landmark localization

Reinforcement learning is a computational approach for learning an optimal policy by interacting with the environment E. An agent observes its current state, s, and chooses an action, a, from its set of possible actions, A, and the environment returns a reward, r, which characterizes the quality of the action chosen. For landmark localization in 3D images, the problem is defined as the environment, E, being

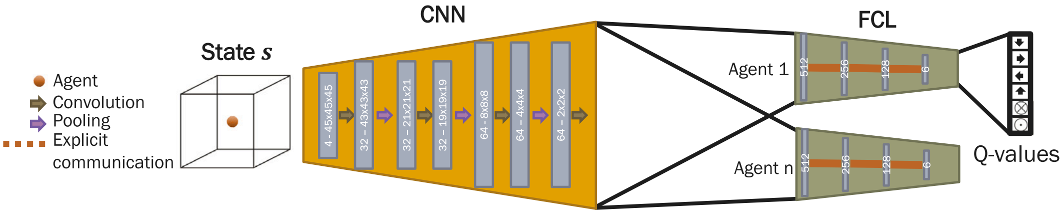


Fig. 3. Diagram of the multi-agent reinforcement learning architecture used. The input is a patch centered in the current agent's position (state). In yellow we show the convolutional neural network which extracts the relevant features of a certain patch. Those features are then passed to each corresponding agent which consists of a set of fully connected layers (in green) that map those features to the estimated expected reward (Q-values) of each of the possible actions (up, right, left, down, forward, or backward). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the 3D image and the agent is a physical location within the image. The state, s, is a patch of the image centered in the agent's location and the action set, A, is the movement in one of the six Cartesian directions (up, down, left, right, forward, and backward). The reward, r, is defined as the difference between the distance to the target landmark's location after the last action and in the previous state, meaning it is a positive value if the agent is moving closer and a negative one if it has moved away from the target landmark's location. The agent's goal is to learn an optimal policy that maximizes not only the immediate reward but also the subsequent future rewards.

The expected reward of taking a certain action given a state is defined as the Q-value. In deep reinforcement learning, this Q-value is estimated using a Deep-Q-Network which takes the current state as its input, and outputs the Q-value associated with each possible action. The architecture of the Deep-Q-Network used for landmark location resembles a typical image classification architecture. We use a MARL (Vlontzos et al., 2019) architecture which includes a set of fully connected layers for each agent. An illustration of the architecture used can be seen in Fig. 3. The agents all share the same convolutional layers (implicit communication), meaning the feature extractor is common to all the agents, but each agent has its own set of fully connected layers. While explicit communication was introduced in the C-MARL model proposed by Leroy et al. (2020), we do not include explicit communication, as we found in López Diez et al. (2022b) that it is not beneficial for anomaly detection.

We employ a multi-scale approach in which the artificial agent is trained not only to distinguish the target within the anatomy but also to learn and follow an optimal navigation path to the target landmark location in the 3D image, as introduced by Ghesu et al. (2019). The agent's search starts at the coarsest scale level with a global context and continues across three different scales, capturing increased levels of detail when transitioning to finer scales. This resembles the human approach to landmark detection in medical images, starting with a big field of view and localizing the region of interest where the clinician zooms in and continues looking for the specific features of that specific landmark.

3.2.1. *PCA shape distance method*

The landmark locations defined in Figs. 1 and 2 present a consistent spacial configuration among patients with a normal inner ear anatomy. The assumption is that when the anatomy does not resemble the anatomical appearance and configuration the agents have been trained with, the final location will deviate significantly.

To evaluate the configuration of the predicted landmark locations, we use a point distribution model (PDM) following the approach presented by Cootes et al. (1995). A full set of landmark locations in an image is denoted as a shape with a point correspondence across all shapes in the training data, and this correspondence is known. Firstly, the alignment between all the shapes from normative data is derived using Procrustes analysis (Gower, 1975). By using this transformation, we obtain a PDM that represents shape variability within the ROI for normal anatomies and is invariant to size and orientation. We can thus derive the mean shape $\bar{x}$ from this model, followed by a PCA of the shape variation (Cootes et al., 1995).

From this analysis, we obtain the matrix, Φ, which is a set of the principal components describing the variability of the shape in the healthy dataset. Based on this, a new shape, x', can be defined as: $x' = \bar{x} + \Phi b$, where the vector b defines the weights controlling the modes of shape variation and Φ contains the first t principal components, which we defined as the lower possible t such that 90% of the shape variability is contained in the Φ matrix. For the artificial dataset, shown in Fig. 1, it was found that $t = 6$ was enough, while for the clinical dataset, shown in Fig. 2, $t = 11$ was found sufficient.

When a new set of landmark locations is predicted, the new shape, x', can be aligned to the mean shape, $\bar{x}$, and be approximated by the PDM model by projecting the residuals from the average shape into principal component space: $b = \Phi^T(x' - \bar{x})$.

The vector b describes the shape coordinates in the PCA space. In this space, we evaluate the distance between the different shapes predicted by the model. We then compute the Euclidean distance between the projected shapes as

$$d_{ji} = \|b_i - b_j\|_2 \tag{1}$$

which quantifies the variation of all the different shapes predicted for a certain image, where i and j represent two different shapes within the same image. Finally, we compute the standard deviation of this distribution of distance values for a certain image in the following manner

$$D_{\text{image}} = \sqrt{\frac{\sum |d - \bar{d}|^2}{n}} \tag{2}$$

where n is the number of different distances within one images. D_{image} measures the level of agreement among the multiple predictions computed in the PCA space defined by normative shapes. A sketch of this approach is shown in Fig. 4(I).

3.2.2. *Q-value history distribution method*

It is our assumption that the expected rewards (Q-values) predicted during the landmark location search process, in which the agent is navigating the image, could represent the degree of confidence (or, defined anthropomorphically, of hesitation) an agent has about the final landmark location. This hesitation measurement should be highly correlated with the anatomical appearance, meaning it could be used to detect anomalies in the anatomy where the landmarks are localized.

Given a certain state of the agent resembles the normal anatomical configuration of such a region, we expect that the Q-values will present a uniform distribution as the agent should not expect a high reward for moving in a certain direction. On the other hand, when the anatomy of the current state does not resemble what the agent is looking for, the Q-values should be less uniformly distributed, pushing the agent to move away from the current location. We define a measurement of the variability within the distribution of the predicted Q-values of the action set in the last stages prior to the final landmark location. To compute this hesitation or uncertainty measurement, we collect the buffer of predicted Q-values of the last 10 states of the agent, which

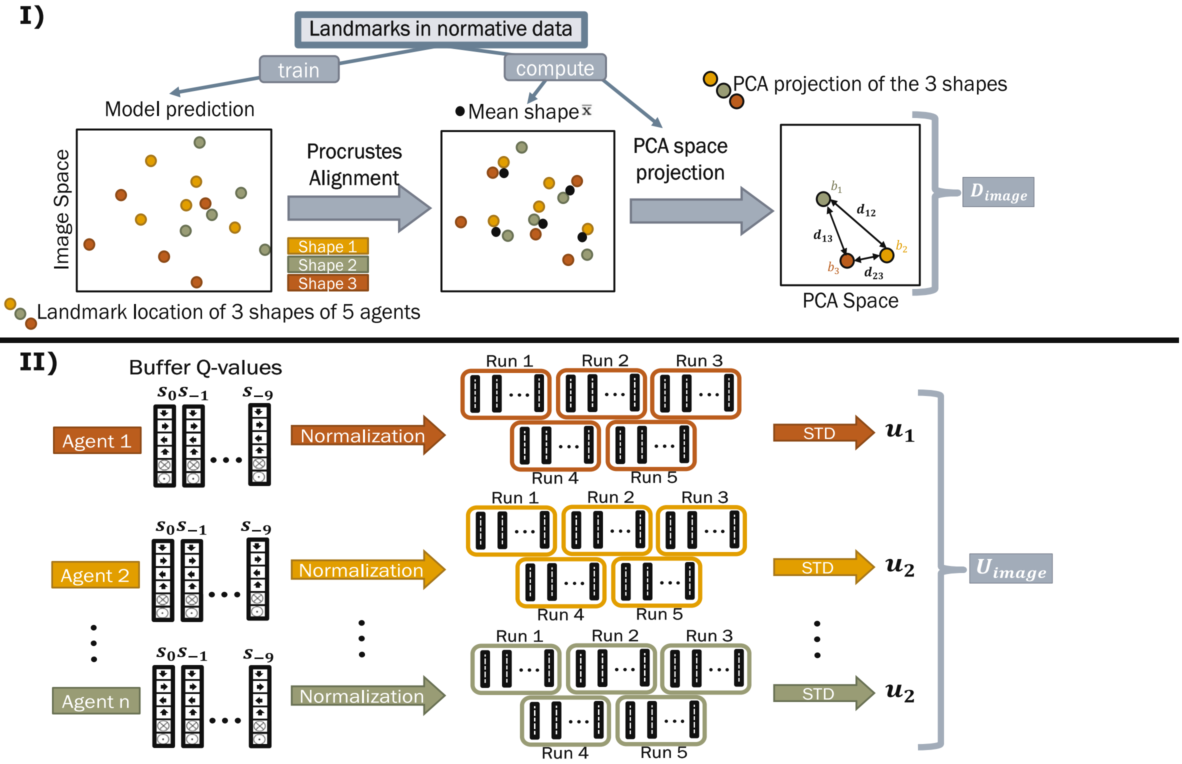


Fig. 4. (I) Diagram of the PCA shape distance method and how D_{image} is computed. (II) Diagram of the Q-value history distribution method and how U_{image} is computed.

have empirically been found sufficient to define the later stages of the landmark search procedure. The normalization of the Q-values is done using the last 10 states of an agent, these values are divided by the biggest Q-value of that agent in that run and the standard deviation is computed using those normalized values over all the runs, which is the uncertainty measurement of that landmark, u_n. These uncertainty measurements are then joined together by computing the norm of the vector containing the u_n of all the landmarks in that image into a single value per image as follows

$$U_{\text{image}} = \sqrt{\sum_n u_n^2} \tag{3}$$

We check the uniformity of the Q-values distribution for each landmark independently and not over all the landmarks in one image because different landmarks have specific anatomical relevance. An overview of this method is shown in Fig. 4(II).

3.2.3. Combined anomaly measure

To evaluate whether the two proposed methods could complement each other, the joint performance is also taken into account. Due to the magnitude of the measurements being different from one another, we introduce a weighting factor to obtain a more representative combination of both methods. This weighting factor is computed by estimating the median of both D_{image} and U_{image} over all the training images in order to determine the intrinsic magnitude difference between both measurements. The weighting factor is then defined as

$$w = \frac{\text{median}(D_{\text{training}})}{\text{median}(U_{\text{training}})} \tag{4}$$

The combined measure, which can be used to analyze the joint performance, is therefore defined as:

$$C_{\text{image}} = \sqrt{D_{\text{image}}^2 + (wU_{\text{image}})^2} \tag{5}$$

to analyze the joint performance.

3.3. 3D convolutional autoencoder

An autoencoder is an unsupervised learning algorithm that learns an identity mapping of the input by minimizing the loss function between the input and its reconstructed output. It is based on both an encoding and a decoding phase. In the encoding phase the original image, $I \in \mathcal{R}^D$, is compressed into a feature vector, $y \in \mathcal{R}^d$, that can be reconstructed back to the original space $\hat{I} \sim I$, given $D \gg d$ in the decoding phase. Autoencoders are very well suited for different tasks such as anomaly detection but also for simplifying the process of feature engineering in machine learning studies, as well as for dimensionality reduction, denoising data, generative modeling, and even pretraining deep learning neural networks (Lopez Pinaya et al., 2020).

Convolutional autoencoders (CAEs) (Masci et al., 2011) are based on the same principles but use deep convolutional layers to perform the dimensionality reduction. The local connectivity of convolutional layers enables the CAE to extract local and hierarchical features capturing the global feature of the input by combining the local features. These local connections require less computational cost than full connections. Pooling layers are used to reduce the input size and to add robustness to shift and position variance. 3D-CAE is an extended CAE composed of 3D convolution and pooling layers, applicable to volumetric data (Arai et al., 2018). We use the asymmetric architecture proposed by Sato et al. (2018), which they employed for anomaly detection in emergency head CT volumes. The architecture consists of a contracting path (3D-CNN) and a reconstructive path (3D-deCNN). Details about the architecture are shown in Fig. 5.

We consider the reconstruction error as the squared difference in intensity between input and output $\mathcal{E} = (I - \hat{I})^2$. The Mean Squared Error (MSE) is used as the loss function to train the network.

3.3.1. Abnormality measurement

Given that the CAE has only been trained on anatomically normal images, it is assumed that the model will learn how to efficiently synthesize such images. This implies that some possible implicit patterns

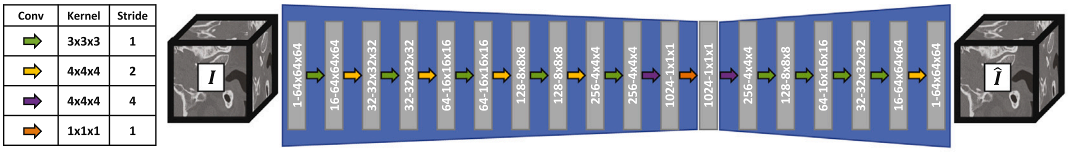


Fig. 5. Architecture of the asymmetrical 3D convolutional autoencoder used for anomaly detection where the different convolutional layers are characterized. I is the input image which is a CT scan of the inner ear and $\hat{I}$ is its generated version after the image has been synthesized into the feature vector of size 1024.

that are consistent throughout the whole dataset might not be a specific part of the encoding, given that they are always present in a similar way. This is the case for the cochlear structure, whose anatomical structure is complex yet very consistent across normative subjects. A higher reconstruction error is expected for images that are anatomically different to the ones in the training set. The chosen error measurement for the CAE model is the patch-based MSE, as used in the study by Sato et al. (2018). The abnormality measurement is defined as the MSE of a patch and all the values for the patches of a certain image are concatenated in a vector denoted a_{image}. The abnormality measurement of a certain case is thus defined as the maximum abnormality measurements over all the patches in an image, defined as

$$A_{\text{image}} = \max\{a_{\text{image}}\} \tag{6}$$

3.4. Experiments

The 119 anatomically normal cases of the artificial dataset were randomly split into: 87 cases for training, 10 for validation, and 22 for testing. The test set comprises those 22 cases and their corresponding artificial anomalies generated with the transformation explained in Section 3.1, resulting in 44 images for evaluation. The clinical dataset was randomly split into a training set of 160 (anatomically normal) images, a validation set of 18 (anatomically normal) images, and 244 images for testing (122 anatomically normal and 122 with congenital malformations). Both the MARL models for landmark localization and the 3D-CAE model were trained using the same split of the data. The best-performing model on the validation set was chosen as the final model for evaluation.

All the models were trained end-to-end on a Titan X 12 GB GPU. The MARL models were trained with one agent per landmark, meaning five agents for the artificial dataset and 12 for the clinical dataset. The models were evaluated over five inferences in order to compute the corresponding anomaly metrics introduced in Eqs. (2), (3) and (5). The approximate training time for the two MARL models used was five days. In the training process, the final state is reached when the distance to the landmark is ≤ 1 voxel using the ϵ-greedy search strategy (Watkins, 1989). During inference, the agent's oscillation is used to finalize the search. The forgetting factor γ is set to 0.9 as this has been empirically found to be the best-performing option. We used a multi-scale approach with three isotropic resolutions: 0.5, 0.250, and finally 0.125 mm and a frame history of 4 observations.

The 3D-CAE model software was developed using PyTorch (Paszke et al., 2019) building on top of the MONAI software (MONAI-Consortium, 2022). With a batch size of 8 and a patch size of 128 × 128 × 128 voxels. We used a learning rate of $5.10{-}5$ and AdamW optimizer for training.

4. Results

Processing results for the artificial dataset are shown in Fig. 6 and the ones for the clinical dataset in Fig. 7. In Figs. 6(a) and 7(a) we have included the Receiver Operating Characteristic (ROC) curve which represents the trade-off between true positive rate (TPR), also known as sensitivity, and the false positive rate, which is equivalent to 1-specificity, for all the possible thresholds of the binary classification. We also include the maximum accuracy metric and the Area under the Curve (AuC) as an overall summary of anomaly detection performance that can be observed in Table 1.

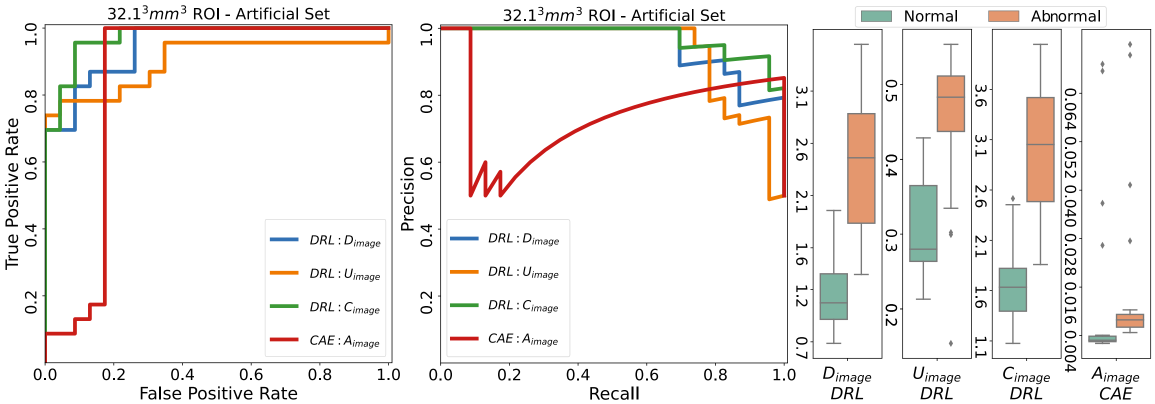
In Figs. 6(b) and 7(b) the precision–recall curve of the anomaly detection is represented. Even though we have created a perfectly balanced test set for both experiments, it is relevant to get a better overview of the classifier performance. These curves represent the relationship between recall (TPR) and precision which measures the fraction of examples classified as anomalies that are truly anomalies. We have also computed the maximum f1-score for each of the methods which is shown in Table 1.

Finally, in Figs. 6(c) and 7(c), boxplots of the distribution of the different anomaly measurements of each method are represented for the artificial and the clinical dataset respectively.

Fig. 8 shows an original image, I_{original}, from the test set and its corresponding reconstructed version, $\hat{I}_{\text{original}}$, followed by its artificially generated abnormal version, $I_{\text{artificial}}$, and its reconstruction, $\hat{I}_{\text{artificial}}$. Furthermore, we also display the reconstruction error, $\mathcal{E}$, for both the original, $\mathcal{E}_{original}$, and the artificial image, $\mathcal{E}_{artificial}$. It can be observed in the reconstructed artificial image, $\hat{I}_{\text{artificial}}$, that even though the cochlea had been artificially removed from the input image, $I_{\text{artificial}}$, it generates a normal cochlear shape, very similar to the one shown in the original image from the corresponding pair, I_{original}. This shows that the model has indeed learned the implicit representation of the normal cochlea and always reconstructs an image with normal anatomy. The resemblance between both outputs is very clear and we can understand the artificial image will have a higher reconstruction error, especially in the cochlear region which could be used to segment or indicate the region of the image that presents the anomaly.

However, is it important to note that the reconstructed images in Fig. 8 present a more smooth overall appearance with less noise than the input images. As previously mentioned, we collected both of our datasets from different clinics meaning the images come from different CT scanners and present different image quality levels, thus the datasets are quite heterogeneous. When visually analyzing the results in the artificial dataset, one notes that the more noisy scans present a greater and generalized reconstruction error due to the denoising effects of the autoencoder. This behavior is the main reason behind the results shown in Fig. 6, where a better result was expected for the 3D autoencoder given the smaller ROI and the artificially introduced anomaly that is more extreme than most of the real clinical cases. In addition, the fact that the artificial dataset presents corresponding pairs of scans with and without the malformation for evaluation allows for an analysis of the pair-wise behavior, where we also observe that the A_{image} measurement is always greater in the corrupted image but a global threshold is not successful for classification. This can be observed in Fig. 6(c) where the corresponding outliers that present a greater reconstruction error in the normal anatomies have their corresponding pairs for the abnormal cases which have a slightly greater value of A_{image}. Therefore setting a general threshold for binary classification is quite challenging for a heterogeneous dataset, but very feasible for a model implemented exclusively for a homogeneous dataset for a specific CT scanner. In Figs. 6(a) and 6(b) we can see that the CAE-based model only detects 10% of the anomalies if any false positive is tolerated.

Our DRL-based methods generally outperform the 3D-CAE approach, which is observed in the better performance curves in both Figs. 6 and 7. However, for both experiments, we see that there is a tendency

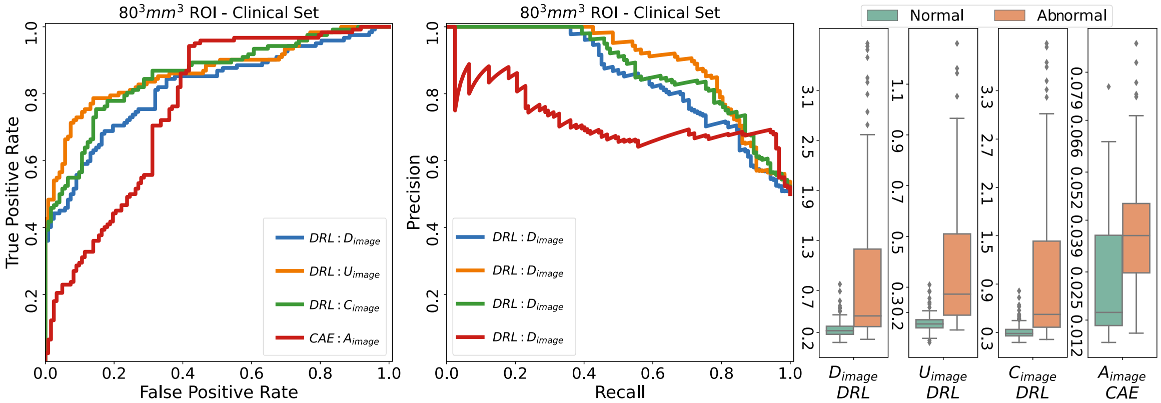


(a) Receiver Operating Characteristic (ROC) curve.

(b) Precision-Recall curve.

(c) Distribution of anomaly measurements.

Fig. 6. Evaluation in the artificial dataset. Performance metrics obtained with each of the methods based in DRL (green, blue and orange) against the performance of the 3D-CAE method (red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(a) Receiver Operating Characteristic (ROC) curve.

(b) Precision-Recall curve.

(c) Distribution of anomaly measurements.

Fig. 7. Evaluation in the clinical dataset. Performance metrics obtained with each of the methods based in DRL (green, blue and orange) against the performance of the 3D-CAE method (red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

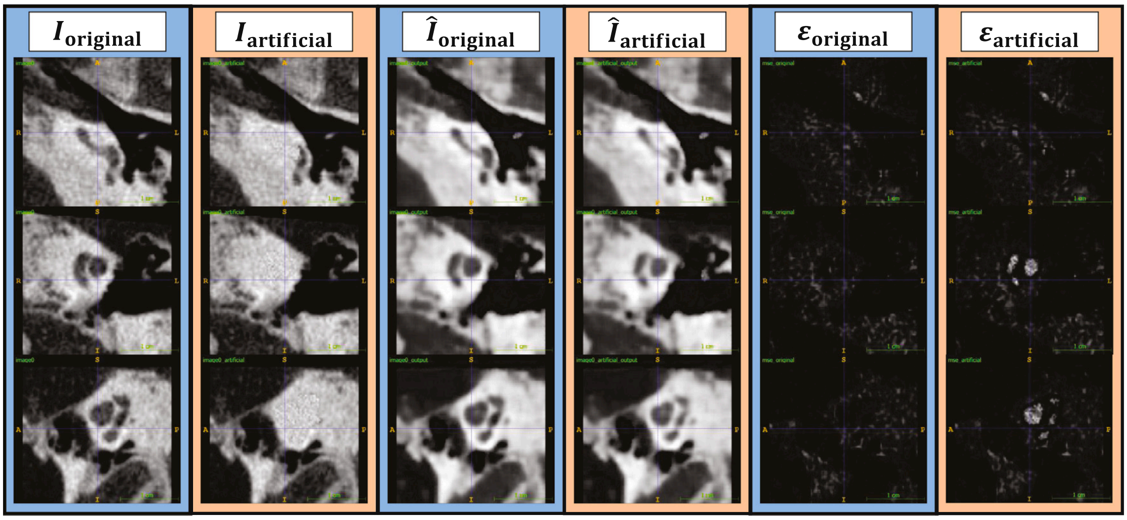


Fig. 8. 3D-CAE test examples from the artificial dataset. $I_{original}$ is the input CT clinical image and $I_{artificial}$ is its corresponding corrupted version (in-painted cochlear structure). $\hat{I}_{original}$ and $\hat{I}_{artificial}$ are the corresponding reconstructed version of the input images. $\mathcal{E}_{original}$ and $\mathcal{E}_{artificial}$ present the reconstruction error between input and output for each case.

that if false positives are not tolerated, then a DRL approach clearly outperforms the 3D-CAE, but if a high rate of normal cases detected as anomalies is tolerated (around 50% for the clinical dataset as seen in Fig. 7(a)), then the 3D-CAE has a similar or slightly better accuracy depending on the threshold and dataset. Given this, it is natural that in Fig. 7(b) we see that the 3D-CAE approach suffers a smaller drop in precision as the recall is increased, but an overall significantly lower precision for recall values from 0 to approximately 0.8.

When comparing the results in Figs. 6 and 7, a similar trend is observed for both experiments. This proves that the hypothesis tested

Table 1
Summary of evaluation metrics for the different methods for anomaly detection in the inner ear anatomy in artificial and clinical datasets.

	Artificial dataset			Clinical dataset		
	AuC	Max. f1-score	Max. acc.	AuC	Max. f1-score	Max. acc.
DRL: D_{image}	0.95	0.88	0.87	0.82	0.77	0.76
DRL: U_{image}	0.90	0.86	0.87	0.87	0.82	0.82
DRL: C_{image}	0.97	0.94	0.93	0.86	0.80	0.80
3D-CAE: A_{image}	0.85	0.92	0.91	0.76	0.80	0.76

with a sparser dataset that contains a more reduced field of view and where the images have been artificially transformed into the more severe type of malformation generalizes well to a complicated clinical dataset with different types of malformations with a greater field of view of the image.

For our DRL approach, we observe that the combined anomaly measure, C_{image}, shows improved performance on the artificial dataset as shown in Fig. 6, and a very close performance to the best-performing method on the clinical dataset as seen in Fig. 7. These results are also shown in Table 1. We consider that the combination, C_{image}, should be used as a more reliable measurement, even though in the clinical dataset we observe a small better performance from the U_{image} measurement.

It is also interesting to note how the different evaluation metrics shown in Table 1 vary for the different methods and datasets. There is a clear drop in performance between the artificial and the clinical dataset, as was expected, given the increased complexity of the task. If we look at the AuC, the difference in performance between any of the DRL methods and the 3D-CAE is very clear ($11,2\%$ improvement on average). Meanwhile, we see similar values for the maximum f1-score and accuracy, which sometimes is higher for the 3D-CAE than for some of the DRL approaches, however, the C_{image} always presents similar or superior metrics than A_{image}.

5. Discussion

Our DRL-based method outperforms the 3D-CAE mostly due to a better adaptation to heterogeneous datasets which are typical in a clinical context. In our experiments, the autoencoder presents a bigger sensitivity toward the nature and origin of the image. The search agents are more robust to the quality difference between images, even though they are trained to choose an optimal action given a certain crop of the image, the appearance is not directly correlated with the loss function, as it is in the case of the autoencoder, nor is it directly linked with the final anomaly measurement. Of course, the image's quality also affects the DRL approach's performance, as it would do for a clinician who is searching for the location of a certain number of landmarks. Images of a lower quality are still more challenging for the DRL approach because the quality of the extracted features will be affected by this, but not to the same extent as the 3D-CAE approach, as can be observed especially when analyzing the performance in the artificial set where the 3D-CAE shows a lower tolerance towards noisy scans, as explained in Section 4.

Both approaches have the potential to be used as a more interpretable anomaly detector rather than a basic classifier because both approaches contain spatial information about the original image that can be exploited. In the case of the 3D-CAE model, the reconstruction error $\mathcal{E}$ can be seen as a map of the abnormal areas indicated by a higher error, which can be highlighted to the clinician as areas of interest. For the DRL approach, the use of specific landmarks that are key for the studied anatomy provides information on the relative points of interest for each case. In both abnormality measurements D_{image} and U_{image}, the information of each agent (corresponding to one landmark) could be used to indicate which region of the image contributes more to the final measurement. This potential for interpretability allows for highlighting regions of the image that have influenced the decision. Clinicians could look into this area of interest and detect something that might have otherwise been overseen, such as an anomaly or the reason for a falsely detected anomaly, which might be, for example, an artifact in the image.

For our DRL-based approach, the normative images used for training must be annotated with all the landmarks of interest, which can be time-consuming. The 3D-CAE approach does not require pixel-level annotations but only confirms that the image contains normal anatomy. However, the 3D-CAE approach does require a more strict standardization of the input image. The DRL approach is less sensitive to scale variations or small differences in orientation given its multiscale approach. This supports the previously introduced idea that the 3D-CAE approach will be better suited for homogeneous datasets, while the DRL-based approach generalizes better for more heterogeneous datasets.

6. Conclusion

We have shown that congenital inner ear malformations in CT images can be automatically detected by training a DRL model exclusively on normative data and evaluating the output variability of its implicit information. This information contains the relative position of the predicted landmarks' location over different runs/agents in a subspace defined by the normative annotations as well as the distribution of the Q-values of the last iterations of the agents as a measurement of the uncertainty of the final location. We also compare the proposed approaches with an asymmetric 3D-CAE, which is based on a 3D-approach for volumetric data. We compare the performance between both methods and analyze the results obtained not only on artificially generated data but also in a large dataset of real clinical CT scans of patients with diverse inner ear malformations from several different clinics. The DRL approach outperforms the 3D-CAE method in both datasets, mostly because it presents a higher tolerance towards heterogeneous real clinical scans from different sources. We believe that the presented DRL approach could be readily adapted to other anatomies prone to complex anatomical anomalies. This could include, but not be limited to, congenital heart disorders or complex spine compressions.

CRediT authorship contribution statement

Paula López Diez: Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing, Funding acquisition. **Josefine Vilsbøll Sundgaard:** Methodology, Software, Writing – review & editing. **Jan Margeta:** Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing – review & editing. **Khassan Diab:** Conceptualization, Resources, Supervision. **François Patou:** Funding acquisition, Project administration, Supervision, Writing – review & editing. **Rasmus R. Paulsen:** Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Paula Lopez Diez reports financial support was provided by William Demant Foundation.

Data availability

The authors do not have permission to share data.

Acknowledgments

We would like to thank William Demant Fonden (Denmark) for financially supporting this study.

References

Allenby, M.C., Liang, E.S., Harvey, J., Woodruff, M.A., Prior, M., Winter, C.D., Alonso-Caneiro, D., 2021. Detection of clustered anomalies in single-voxel morphometry as a rapid automated method for identifying intracranial aneurysms. Comput. Med. Imaging Graph. 89, 101888. http://dx.doi.org/10.1016/j.compmedimag.2021.101888, URL https://www.sciencedirect.com/science/article/pii/S0895611121000367.

Arai, H., Chayama, Y., Iyatomi, H., Oishi, K., 2018. Significant dimension reduction of 3D brain MRI using 3D convolutional autoencoders. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. EMBC, pp. 5162–5165. http://dx.doi.org/10.1109/EMBC.2018.8513469.

Astaraki, M., Smedby, Ö., Wang, C., 2022. Prior-aware autoencoders for lung pathology segmentation. Med. Image Anal. 80, 102491. http://dx.doi.org/10.1016/j.media.2022.102491, URL https://www.sciencedirect.com/science/article/pii/S1361841522001384.

Atlason, H.E., Askell Love, M., Sigurdsson, S., Vilmundur Gudnason, M., Ellingsen, L.M., 2019. Unsupervised brain lesion segmentation from MRI using a convolutional autoencoder. In: Angelini, E.D., Landman, B.A. (Eds.), In: Medical Imaging 2019: Image Processing, vol. 10949, SPIE, International Society for Optics and Photonics, p. 109491H. http://dx.doi.org/10.1117/12.2512953, URL https://doi.org/10.1117/12.2512953.

Baur, C., Graf, R., Wiestler, B., Albarqouni, S., Navab, N., 2020. SteGANomaly: Inhibiting cyclegan steganography for unsupervised anomaly detection in brain MRI. In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (Eds.), Medical Image Computing and Computer Assisted Intervention. MICCAI 2020, Springer International Publishing, Cham, pp. 718–727. http://dx.doi.org/10.1007/978-3-030-59713-9_69.

Baur, C., Wiestler, B., Albarqouni, S., Navab, N., 2019. Deep autoencoding models for unsupervised anomaly segmentation in brain MR images. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (Eds.), Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer International Publishing, Cham, pp. 161–169. http://dx.doi.org/10.1007/978-3-030-11723-8_16.

Bekkouch, I.E.I., Maksudov, B., Kiselev, S., Mustafaev, T., Vrtovec, T., Ibragimov, B., 2022. Multi-landmark environment analysis with reinforcement learning for pelvic abnormality detection and quantification. Med. Image Anal. 78, 102417. http://dx.doi.org/10.1016/j.media.2022.102417, URL https://www.sciencedirect.com/science/article/pii/S1361841522000688.

Brotto, D., Sorrentino, F., Cenedese, R., Avato, I., Bovo, R., Trevisi, P., Manara, R., 2021. Genetics of inner ear malformations: A review. Audiol. Res. 11 (4), 524–536. http://dx.doi.org/10.3390/audiolres11040047, URL https://www.mdpi.com/2039-4349/11/4/47.

Chakravorti, S., Noble, J.H., Gifford, R.H., Dawant, B.M., O'Connell, B., Wang, J., Labadie, R.F., 2019. Further evidence of the relationship between cochlear implant electrode positioning and hearing outcomes. Otol. Neurotol.: Off. Publ. Am. Otol. Soc., Am. Neurotol. Soc. Eur. Acad. Otol. Neurotol. 40 (5), 617. http://dx.doi.org/10.1097/MAO.0000000000002204.

Chen, X., You, S., Tezcan, K.C., Konukoglu, E., 2020. Unsupervised lesion detection via image restoration with a normative prior. Med. Image Anal. 64, 101713. http://dx.doi.org/10.1016/j.media.2020.101713, URL https://www.sciencedirect.com/science/article/pii/S1361841520300773.

Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J., 1995. Active shape models-their training and application. Comput. Vis. Image Underst. 61 (1), 38–59. http://dx.doi.org/10.1006/cviu.1995.1004.

Demarcy, T., Vandersteen, C., Guevara, N., Raffaelli, C., Gnansia, D., Ayache, N., Delingette, H., 2017. Automated analysis of human cochlea shape variability from segmented μCT images. Comput. Med. Imaging Graph. 59, 1–12. http://dx.doi.org/10.1016/j.compmedimag.2017.04.002, URL https://www.sciencedirect.com/science/article/pii/S0895611117300332.

Dhanasingh, A., Erpenbeck, D., Assadi, M.Z., Doyle, Ú., Roland, P., Hagr, A., Rompaey, V.V., de Heyning, P.V., 2021. A novel method of identifying inner ear malformation types by pattern recognition in the mid modiolar section. Sci. Rep. 11, http://dx.doi.org/10.1038/s41598-021-00330-6.

Dhanasingh, A.E., Weiss, N.M., Erhard, V., Altamimi, F., Roland, P., Hagr, A., Rompaey, V.V., de Heyning, P.V., 2022. A novel three-step process for the identification of inner ear malformation types. Laryngosc. Investig. Otolaryngol. http://dx.doi.org/10.1002/lio2.936, URL https://onlinelibrary.wiley.com/doi/10.1002/lio2.936.

Escudé, B., James, C., Deguine, O., Cochard, N., Eter, E., Fraysse, B., 2006. The size of the cochlea and predictions of insertion depth angles for cochlear implant electrodes. Audiol. Neurotol. 11 (Suppl. 1), 27–33.

Ghesu, F.C., Georgescu, B., Zheng, Y., Grbic, S., Maier, A., Hornegger, J., Comaniciu, D., 2019. Multi-scale deep reinforcement learning for real-time 3D-landmark detection in CT scans. IEEE Trans. Pattern Anal. Mach. Intell. 41 (1), 176–189. http://dx.doi.org/10.1109/TPAMI.2017.2782687.

Gower, J.C., 1975. Generalized procrustes analysis. Psychometrika 40 (1), 33–51. http://dx.doi.org/10.1007/bf02291478.

Hainan, S., Jiang, M., Liu, Y., Lu, H., Liang, Z., 2019. The Detection of Non-Polypoid Colorectal Lesions Using the Texture Feature Extracted from Intact Colon Wall: A Pilot Study. SPIE-Intl Soc Optical Eng, p. 105. http://dx.doi.org/10.1117/12.2511823.

Hariri, S., Kind, M.C., Brunner, R.J., 2021. Extended isolation forest. IEEE Trans. Knowl. Data Eng. 33, 1479–1489. http://dx.doi.org/10.1109/TKDE.2019.2947676.

Korver, A.M., Smith, R.J., Van Camp, G., Schleiss, M.R., Bitner-Glindzicz, M.A., Lustig, L.R., Usami, S.-i., Boudewyns, A.N., 2017. Congenital hearing loss. Nat. Rev. Dis. Primers 3 (1), 1–17.

Leroy, G., Rueckert, D., Alansary, A., 2020. Communicative reinforcement learning agents for landmark detection in brain images. In: Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-Oncology. Springer, pp. 177–186. http://dx.doi.org/10.1007/978-3-030-66843-3_18.

Liu, F.T., Ting, K.M., Zhou, Z.H., 2012. Isolation-based anomaly detection. ACM Trans. Knowl. Discov. Data 6, http://dx.doi.org/10.1145/2133360.2133363.

López Diez, P., Juhl, K.A., Sundgaard, J.V., Diab, H., Margeta, J., Patou, F., Paulsen, R.R., 2022a. Deep reinforcement learning for detection of abnormal anatomies. In: Proceedings of the Northern Lights Deep Learning Workshop, vol. 3, UiT The Arctic University of Norway, http://dx.doi.org/10.7557/18.6280.

López Diez, P., Sørensen, K., Sundgaard, J.V., Diab, K., Margeta, J., Patou, F., Paulsen, R.R., 2022b. 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, Springer Nature Switzerland, Cham, pp. 697–706. http://dx.doi.org/10.1007/978-3-031-16437-8_67.

López Diez, P., Sundgaard, J.V., Patou, F., Margeta, J., Paulsen, R.R., 2021. Facial and cochlear nerves characterization using deep reinforcement learning for landmark detection. In: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention. MICCAI 2021, Springer International Publishing, Cham, pp. 519–528. http://dx.doi.org/10.1007/978-3-030-87202-1_50.

Lopez Pinaya, W.H., Vieira, S., Garcia-Dias, R., Mechelli, A., 2020. Chapter 11 - autoencoders. In: Mechelli, A., Vieira, S. (Eds.), Machine Learning. Academic Press, pp. 193–208. http://dx.doi.org/10.1016/B978-0-12-815739-8.00011-0, URL https://www.sciencedirect.com/science/article/pii/B9780128157398000110.

Margeta, J., Hussain, R., López Diez, P., Morgenstern, A., Demarcy, T., Wang, Z., Gnansia, D., Martinez Manzanera, O., Vandersteen, C., Delingette, H., Buechner, A., Lenarz, T., Patou, F., Guevara, N., 2022. A web-based automated image processing research platform for cochlear implantation-related studies. J. Clin. Med. 11 (22), http://dx.doi.org/10.3390/jcm11226640, URL https://www.mdpi.com/2077-0383/11/22/6640.

Martins, S.B., Telea, A.C., Falcão, A.X., 2020. Investigating the impact of supervoxel segmentation for unsupervised abnormal brain asymmetry detection. Comput. Med. Imaging Graph. 85, 101770. http://dx.doi.org/10.1016/j.compmedimag.2020.101770, URL https://www.sciencedirect.com/science/article/pii/S0895611120300720.

Masci, J., Meier, U., Cireşan, D., Schmidhuber, J., 2011. Stacked convolutional auto-encoders for hierarchical feature extraction. In: Honkela, T., Duch, W., Girolami, M., Kaski, S. (Eds.), Artificial Neural Networks and Machine Learning. ICANN 2011, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 52–59. http://dx.doi.org/10.1007/978-3-642-21735-7_7.

MONAI-Consortium, 2022. MONAI: Medical open network for AI. http://dx.doi.org/10.5281/zenodo.7459814.

Nazir, M., Shakil, S., Khurshid, K., 2021. Role of deep learning in brain tumor detection and classification (2015 to 2020): A review. Comput. Med. Imaging Graph. 91, 101940. http://dx.doi.org/10.1016/j.compmedimag.2021.101940, URL https://www.sciencedirect.com/science/article/pii/S0895611121000896.

Paludetti, G., Conti, G., Di Nardo, W., De Corso, E., Rolesi, R., Picciotti, P., Fetoni, A., 2012. Infant hearing loss: From diagnosis to therapy official report of XXI conference of Italian society of pediatric otorhinolaryngology. Acta Otorhinolaryngol. Italica 32 (6), 347.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., Garnett, R. (Eds.), In: Advances in Neural Information Processing Systems, vol. 32, Curran Associates, Inc., URL https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

Pawlowski, N., Lee, M.J., Rajchl, M., McDonagh, S.G., Ferrante, E., Kamnitsas, K., Cooke, S., Stevenson, S., Khetani, A., Newman, T., Zeiler, F.A., Digby, R., Coles, J.P., Rueckert, D., Menon, D.K., Newcombe, V.F.J., Glocker, B., 2018. Unsupervised lesion detection in brain CT using Bayesian convolutional autoencoders. In: 1st Conference on Medical Imaging with Deep Learning. MIDL 2018.

Pinaya, W.H.L., Graham, M.S., Gray, R., da Costa, P.F., Tudosiu, P.-D., Wright, P., Mah, Y.H., MacKinnon, A.D., Teo, J.T., Jager, R., Werring, D., Rees, G., Nachev, P., Ourselin, S., Cardoso, M.J., 2022a. Fast unsupervised brain anomaly detection and segmentation with diffusion models. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (Eds.), Medical Image Computing and Computer Assisted Intervention. MICCAI 2022, Springer Nature Switzerland, Cham, pp. 705–714. http://dx.doi.org/10.1007/978-3-031-16452-1_67.

Pinaya, W.H., Tudosiu, P.D., Gray, R., Rees, G., Nachev, P., Ourselin, S., Cardoso, M.J., 2022b. Unsupervised brain imaging 3D anomaly detection and segmentation with transformers. Med. Image Anal. 79, http://dx.doi.org/10.1016/j.media.2022.102475.

Prastawa, M., Bullitt, E., Ho, S., Gerig, G., 2004. A brain tumor segmentation framework based on outlier detection. Med. Image Anal. 8 (3), 275–283. http://dx.doi.org/10.1016/j.media.2004.06.007, URL https://www.sciencedirect.com/science/article/pii/S1361841504000295. Medical Image Computing and Computer-Assisted Intervention - MICCAI 2003.

Radutoiu, A.T., Patou, F., Margeta, J., Paulsen, R.R., López Diez, P., 2022. 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. Springer Nature Switzerland, Cham, pp. 416–424. http://dx.doi.org/10.1007/978-3-031-21014-3_43.

Sajid, S., Hussain, S., Sarwar, A., 2019. Brain tumor detection and segmentation in MR images using deep learning. Arab. J. Sci. Eng. 44, 9249–9261. http://dx.doi.org/10.1007/s13369-019-03967-8.

Sato, D., Hanaoka, S., Nomura, Y., Takenaga, T., Miki, S., Yoshikawa, T., Hayashi, N., Abe, O., 2018. A primitive study on unsupervised anomaly detection with an autoencoder in emergency head CT volumes. In: Medical Imaging 2018: Computer-Aided Diagnosis, vol. 10575, SPIE, pp. 388–393. http://dx.doi.org/10.1117/12.2292276.

Schlegl, T., Seeböck, P., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U., 2019. F-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. Med. Image Anal. 54, 30–44. http://dx.doi.org/10.1016/j.media.2019.01.010, URL https://www.sciencedirect.com/science/article/pii/S1361841518302640.

Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G., 2017. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: Niethammer, M., Styner, M., Aylward, S., Zhu, H., Oguz, I., Yap, P.-T., Shen, D. (Eds.), Information Processing in Medical Imaging. Springer International Publishing, Cham, pp. 146–157. http://dx.doi.org/10.1007/978-3-319-59050-9_12.

Sennaroğlu, L., Bajin, M.D., 2017. Classification and current management of inner ear malformations. Balkan Med. J. 34, http://dx.doi.org/10.4274/balkanmedj.2017.0367.

Silva-Rodríguez, J., Naranjo, V., Dolz, J., 2022. Constrained unsupervised anomaly segmentation. Med. Image Anal. 80, http://dx.doi.org/10.1016/j.media.2022.102526.

Sun, L., Wang, J., Huang, Y., Ding, X., Greenspan, H., Paisley, J., 2020. An adversarial learning approach to medical image synthesis for lesion detection. IEEE J. Biomed. Health Inf. 24 (8), 2303–2314. http://dx.doi.org/10.1109/JBHI.2020.2964016.

Taboada-Crispi, A., Sahli, H., Orozco Monteagudo, M., Hernandez Pacheco, D., Falcon, A., 2009. Anomaly detection in medical image analysis. In: Handbook of Research on Advanced Techniques in Diagnostic Imaging and Biomedical Applications. pp. 426–446. http://dx.doi.org/10.4018/978-1-60566-314-2.ch027, chapter 2.

Trier, P., Noe, K.Ø., Sørensen, M.S., Mosegaard, J., 2008. The visible ear surgery simulator. Stud. Health Technol. Inf. 132, 523.

Van Leemput, K., Maes, F., Vandermeulen, D., Colchester, A., Suetens, P., 2001. Automated segmentation of multiple sclerosis lesions by model outlier detection. IEEE Trans. Med. Imaging 20 (8), 677–688. http://dx.doi.org/10.1109/42.938237.

Venkataramanan, S., Peng, K.C., Singh, R.V., Mahalanobis, A., 2020. Attention guided anomaly localization in images. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (Eds.), Computer Vision – ECCV 2020. Springer International Publishing, Cham, pp. 485–503. http://dx.doi.org/10.1007/978-3-030-58520-4_29.

Vlontzos, A., Alansary, A., Kamnitsas, K., Rueckert, D., Kainz, B., 2019. Multiple landmark detection using multi-agent reinforcement learning. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention. MICCAI 2019, Springer International Publishing, Cham, pp. 262–270. http://dx.doi.org/10.1007/978-3-030-32251-9_29.

Wang, S., Zeng, Y., Yu, G., Cheng, Z., Liu, X., Zhou, S., Zhu, E., Kloft, M., Yin, J., Liao, Q., 2023. E3Outlier: A self-supervised framework for unsupervised deep outlier detection. IEEE Trans. Pattern Anal. Mach. Intell. 45, 2952–2969. http://dx.doi.org/10.1109/TPAMI.2022.3188763.

Wang, S., Zhu, Y., Lee, S., Elton, D.C., Shen, T.C., Tang, Y., Peng, Y., Lu, Z., Summers, R.M., 2022. Global-local attention network with multi-task uncertainty loss for abnormal lymph node detection in MR images. Med. Image Anal. 77, 102345. http://dx.doi.org/10.1016/j.media.2021.102345, URL https://www.sciencedirect.com/science/article/pii/S136184152100390X.

Watkins, C.J.C.H., 1989. Learning from Delayed Rewards (Ph.D. thesis). King's College, Cambridge, UK.

Welch, M.L., McIntosh, C., McNiven, A., Huang, S.H., Zhang, B.B., Wee, L., Traverso, A., O'Sullivan, B., Hoebers, F., Dekker, A., Jaffray, D.A., 2020. User-controlled pipelines for feature integration and head and neck radiation therapy outcome predictions. Phys. Med. 70, 145–152. http://dx.doi.org/10.1016/j.ejmp.2020.01.027.

Wolleb, J., Bieder, F., Sandkühler, R., Cattin, P.C., 2022. Diffusion models for medical anomaly detection. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (Eds.), Medical Image Computing and Computer Assisted Intervention. MICCAI 2022, Springer Nature Switzerland, Cham, pp. 35–45. http://dx.doi.org/10.1007/978-3-031-16452-1_4.

Xu, H., Pang, G., Wang, Y., Wang, Y., 2023a. Deep isolation forest for anomaly detection. IEEE Trans. Knowl. Data Eng. http://dx.doi.org/10.1109/TKDE.2023.3270293.

Xu, H., Wang, Y., Wei, J., Jian, S., Li, Y., Liu, N., 2023b. Fascinating supervisory signals and where to find them: Deep anomaly detection with scale learning. URL http://arxiv.org/abs/2305.16114.

Yushkevich, P.A., Piven, J., Cody Hazlett, H., Gimpel Smith, R., Ho, S., Gee, J.C., Gerig, G., 2006. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. Neuroimage 31 (3), 1116–1128. http://dx.doi.org/10.1016/j.neuroimage.2006.01.015.

Zhou, S.K., Le, H.N., Luu, K., V Nguyen, H., Ayache, N., 2021. Deep reinforcement learning in medical imaging: A literature review. Med. Image Anal. 73, 102193. http://dx.doi.org/10.1016/j.media.2021.102193, URL https://www.sciencedirect.com/science/article/pii/S1361841521002395.

Zimmerer, D., Isensee, F., Petersen, J., Kohl, S., Maier-Hein, K., 2019. Unsupervised anomaly localization using variational auto-encoders. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention. MICCAI 2019, Springer International Publishing, Cham, pp. 289–297. http://dx.doi.org/10.1007/978-3-030-32251-9_32.