









ORIGINAL ARTICLE

Open Access



# Deep learning detects retropharyngeal edema on MRI in patients with acute neck infections

Oona Rainio<sup>1\*</sup> , Heidi Huhtanen<sup>2</sup> , Jari-Pekka Vierula<sup>2</sup> , Janne Nurminen<sup>2</sup> , Jaakko Heikkinen<sup>2</sup> , Mikko Nyman<sup>2</sup> , Riku Klén<sup>1</sup>  and Jussi Hirvonen<sup>1,2</sup> 

## Abstract

**Background** In acute neck infections, magnetic resonance imaging (MRI) shows retropharyngeal edema (RPE), which is a prognostic imaging biomarker for a severe course of illness. This study aimed to develop a deep learning-based algorithm for the automated detection of RPE.

**Methods** We developed a deep neural network consisting of two parts using axial T2-weighted water-only Dixon MRI images from 479 patients with acute neck infections annotated by radiologists at both slice and patient levels. First, a convolutional neural network (CNN) classified individual slices; second, an algorithm classified patients based on a stack of slices. Model performance was compared with the radiologists' assessment as a reference standard. Accuracy, sensitivity, specificity, and area under receiver operating characteristic curve (AUROC) were calculated. The proposed CNN was compared with InceptionV3, and the patient-level classification algorithm was compared with traditional machine learning models.

**Results** Of the 479 patients, 244 (51%) were positive and 235 (49%) negative for RPE. Our model achieved accuracy, sensitivity, specificity, and AUROC of 94.6%, 83.3%, 96.2%, and 94.1% at the slice level, and 87.4%, 86.5%, 88.2%, and 94.8% at the patient level, respectively. The proposed CNN was faster than InceptionV3 but equally accurate. Our patient classification algorithm outperformed traditional machine learning models.

**Conclusion** A deep learning model, based on weakly annotated data and computationally manageable training, achieved high accuracy for automatically detecting RPE on MRI in patients with acute neck infections.

**Relevance statement** Our automated method for detecting relevant MRI findings was efficiently trained and might be easily deployed in practice to study clinical applicability. This approach might improve early detection of patients at high risk for a severe course of acute neck infections.

## Key Points

- Deep learning automatically detected retropharyngeal edema on MRI in acute neck infections.
- Areas under the receiver operating characteristic curve were 94.1% at the slice level and 94.8% at the patient level.
- The proposed convolutional neural network was lightweight and required only weakly annotated data.

**Keywords** Artificial intelligence, Magnetic resonance imaging, Neural networks (computer), Respiratory tract infections, Retropharyngeal abscess

\*Correspondence:

Oona Rainio  
ormrai@utu.fi

<sup>1</sup>Turku PET Centre, University of Turku and Turku University Hospital, Turku, Finland

<sup>2</sup>Department of Radiology, University of Turku and Turku University Hospital, Turku, Finland



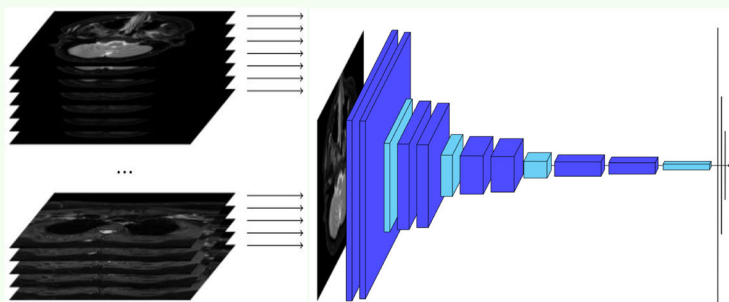
© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## Graphical Abstract

## Deep learning detects retropharyngeal edema on MRI in patients with acute neck infections


 EUROPEAN SOCIETY OF RADIOLOGY

- Deep learning automatically detected retropharyngeal edema (RPE) on MRI in acute neck infections.
- AUC was 94.1% at slice level and 94.8% at patient level.
- The proposed convolutional neural network was lightweight and required only weakly annotated data.



Light-weight CNN detects the presence of RPE from individual MRI slices

**An automated method for detecting prognostically significant MRI findings in patients with acute neck infections might benefit the deployment of advanced imaging methods in the emergency department**



**Eur Radiol Exp (2025) Rainio O, Huhtanen H, Vierula JP et al;  
DOI: 10.1186/s41747-025-00599-6**

### Background

Acute neck infections require prompt diagnosis and intervention to prevent complications, especially when they reach the deep neck spaces [1]. Therefore, emergency imaging is critical in the early assessment of these patients [2]. Magnetic resonance imaging (MRI) has recently been validated as an effective primary diagnostic modality for imaging acute neck infections of various etiologies [3]. MRI has excellent soft tissue contrast and can offer improved diagnostic precision compared to the traditionally employed computed tomography. Previous research has demonstrated that emergency MRI is feasible in the acute setting [4, 5] and can accurately delineate abscesses and show edema patterns with high sensitivity [4, 6–9].

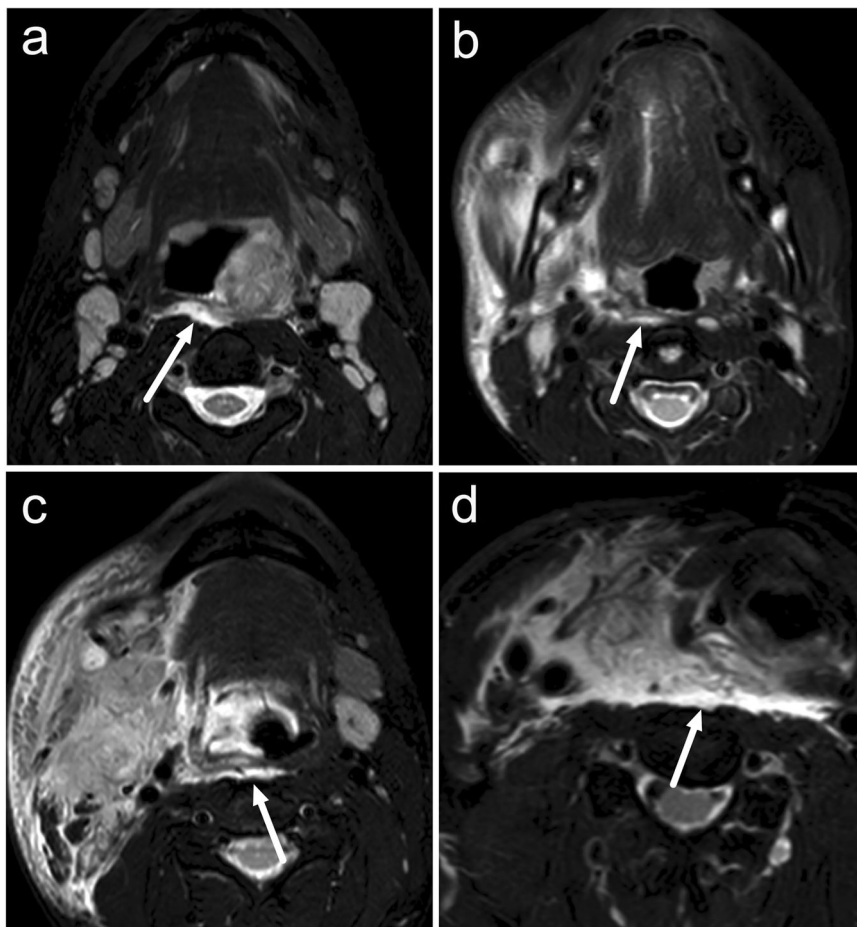
About half of the patients with an acute neck infection show a hyperintense signal in the retropharyngeal space on MRI fluid-sensitive, fat-suppressed T2-weighted imaging, called retropharyngeal edema (RPE) [6] (Fig. 1). RPE is not a drainable fluid collection but rather a reactive edema pattern suggesting a severe course of illness. A previous study found that patients with RPE are much more likely to require admission to the intensive care unit than those without, suggesting that RPE could serve as an imaging biomarker for a severe course of illness [6].

This edema pattern has demonstrated substantial inter-observer agreement between radiologists, suggesting favorable generalizability. However, considering the complex anatomy of the neck, variability in MRI sequences, and the subjective nature of this MRI-derived prognostic biomarker, the broader clinical applicability in the emergency setting would benefit from an automated method.

Deep learning (DL)-based artificial intelligence methods are commonly used for classification tasks in medical imaging [10] and would be particularly beneficial in an acute diagnostic setting concerning complex anatomy and advanced imaging [11]. Previous studies have successfully applied DL methods for the automated detection and segmentation of bone marrow edema on MRI in the sacroiliac [12–16], hip [17], and knee [18] joints and the spine [19], but an algorithm for detecting edema in the neck soft tissues is currently lacking. Here, we sought to develop and validate a DL algorithm for automated detection of RPE in fat-suppressed (water-only) T2-weighted MRI images from patients with acute neck infections.

### Methods

The overall study workflow is presented in Fig. 2.



**Fig. 1** Examples of retropharyngeal edema (arrows) on axial T2-weighted Dixon water-only images of four different patients (**a, b, c, d**) with acute neck infections

#### Software and hardware

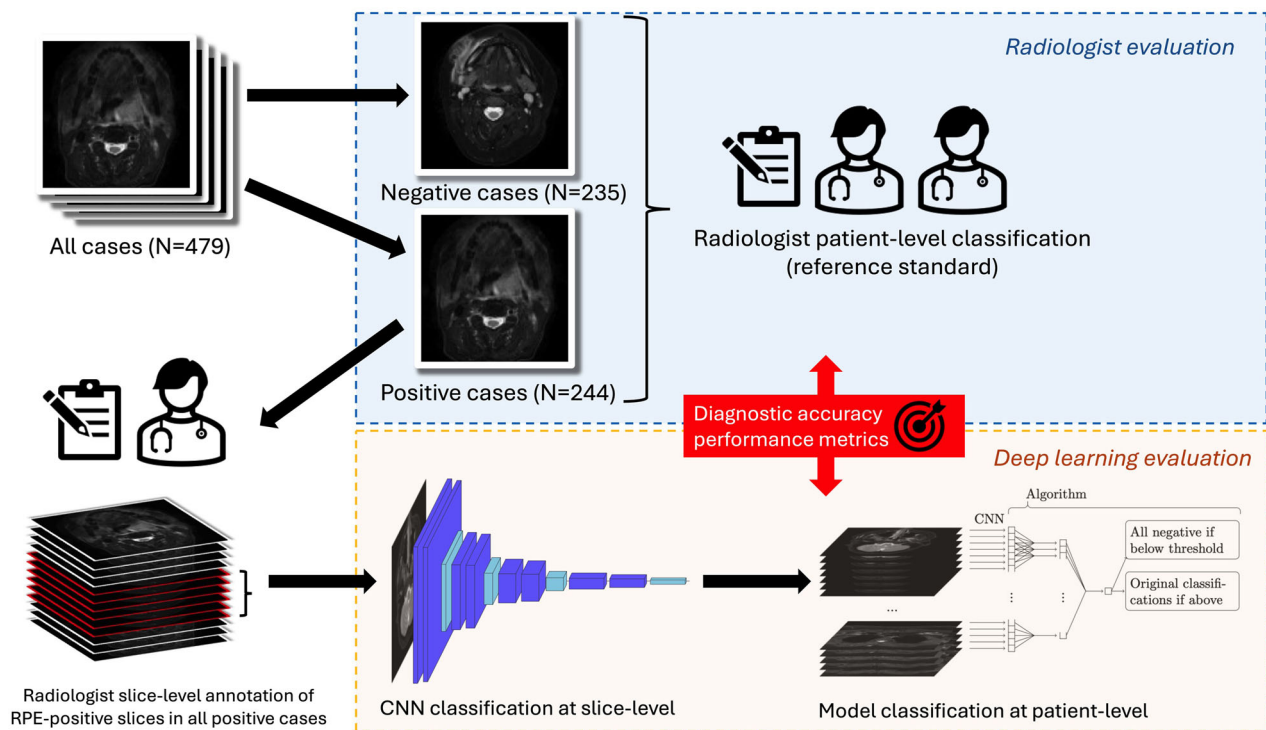
The image data was viewed and annotated with Carimas (version: 2.10) [20] and Mango (version 4.1, <https://mangoviewer.com/>). The experiments were performed with Python (version: 3.12.8) [21], using packages Keras (version: 3.7.0) [22], Tensorflow (version: 2.18.0) [23], sklearn (version: 1.5.2) [24] and SciPy (version: 1.14.1) [25]. No external computing services were used due to the sensitive nature of the patient imaging data. To obtain comparable estimates of the training time, all experiments were run on the same computer with Intel Core Ultra 5 125H processor and 16GB random-access memory.

#### Data

This study followed previously published procedures for patient selection, MRI acquisition and interpretation, and the extraction of medical and surgical information [4, 6]. We obtained study permission from the hospital district board for this retrospective cohort study. The

requirement of patient consent was waived due to the retrospective nature of the study. The inclusion criteria were: (1) emergency MRI between April 1, 2013, and August 30, 2021, for suspected neck infection scanned with the Philips Ingenia 3-T system using a dS Head-NeckSpine coil configuration; (2) MRI evidence of infection, that is, high signal on fat-suppressed T2-weighted Dixon images suggesting edema or high signal on fat-suppressed contrast-enhanced T1-weighted Dixon images suggesting abnormal tissue enhancement; (3) a final clinical diagnosis of an acute neck infection; and (4) available axial T2-weighted water-only Dixon images with diagnostic image quality as determined by a senior radiologist. The exclusion criterion was a lack of clinical and/or surgical data.

RPE was defined as an area of a hyperintense signal between the superior pharyngeal constrictor muscle anteriorly and the prevertebral muscles posteriorly on axial T2-weighted Dixon water-only images [6] (see



**Fig. 2** Study overview

Fig. 1). At the patient level, a consensus of two fellowship-trained neuroradiologists with significant experience and competence in emergency neck MRI formed the reference standard for RPE.

The final sample included 479 patients (277 males, 202 females), among whom 244 (51%) were positive and 235 (49%) negative for RPE. The mean age was 41 years (range 0–88 years).

Each MRI image contained 26–60 transaxial slices with a distance of 4–6 mm between the slices. The slices were all square-shaped images with pixel dimensions varying from  $256 \times 256$  to  $704 \times 704$ . The pixels were between  $0.342 \times 0.342 \text{ mm}^2$  and  $0.781 \times 0.781 \text{ mm}^2$  in size, so that the real size of the transaxial slice varied from  $16.0 \times 16.0 \text{ cm}^2$  to  $25.0 \times 25.0 \text{ cm}^2$ .

When training different convolutional neural network (CNN) set-ups for classification, each transaxial slice was considered as one data instance that should be classified either as positive or as negative. A fellowship-trained head and neck radiologist with significant experience in emergency neck MRI annotated the transaxial slices showing RPE in the positive patients by saving the indices of the uppermost and lowermost slices that showed RPE. By considering all the slices between these two indexes as positive and the rest as negative, we obtained 2,704 positive slices and 8,493 negative slices from the 244

positive patients. We gained an additional 10,830 negative slices from the 235 negative patients. Consequently, our total data contained 22,027 slices, 2,704 (12.3%) of which were positive.

All the slices were scaled from their original size to  $128 \times 128$  pixels. They were normalized slice-wise separately by mapping their pixel values onto the interval [0,1] with

$$x_{ij} - > (x_{ij} - \min x) / (\max x - \min x)$$

where  $\max x$  and  $\min x$  are the maximum and the minimum of the current slice and  $i, j = 1, \dots, 128$ . Due to memory restraints, the image arrays were converted from Python's default float64 type to float32.

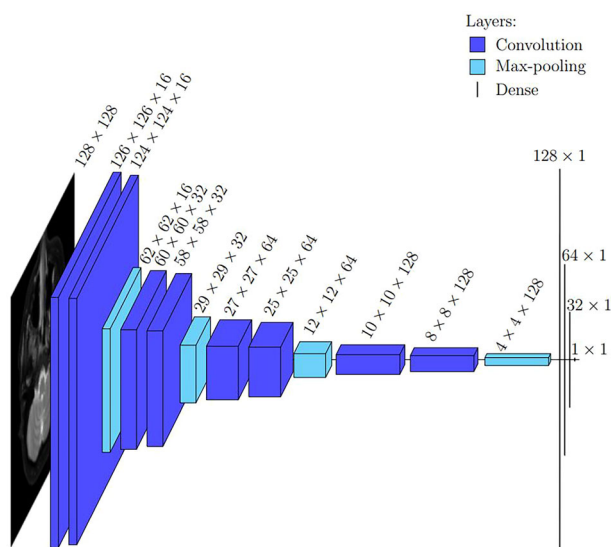
### Convolutional neural networks

The CNN of our proposed method is a modified U-Net encoder studied for head and neck cancer classification in an earlier study by Hellström et al [26]. The original U-Net introduced in 2015 by Ronneberger et al [27] was designed for medical image segmentation. It consists of two paths, a contracting one and an expanding one, the former of which is also called an encoder and the latter a decoder. The idea behind this encoder-decoder design is that the encoder first shrinks the image in size to be able

to see the context of the whole image, and the decoder returns the image to the original size so that it can focus on the details needed to perform accurate segmentation. Additionally, the original U-Net also copies the data inputs at different stages of the encoder and sends them to the decoder over several layers to prevent the loss of information during the shrinking process. Both the encoder and the decoder contain sequences of paired convolutions followed by max-pooling operations.

However, it was noted by Hellström et al [26] that the U-Net encoder can be used to create an effective light-weight CNN for classification. Namely, U-Net is able to detect relevant patterns on the input data during the encoder, and if the aim is to obtain a single class label, there is no need for the decoder. Instead, only a few dense layers are added after the encoder, as shown in Fig. 3. Given this new design worked for head and neck cancer classification, we assumed that it might be well-suited for the current classification task, too, even if our data is MRI images instead of positron emission tomography images used by Hellström et al [26].

The CNN based on the U-Net encoder was compared to InceptionV3, which is considered a state-of-the-art CNN for medical image classification [28]. Introduced by Szegedy et al [29], it is an improvement of the earlier Inception networks that utilize several differently sized kernels [30]. As InceptionV3 has three color channels, it was given three identical copies of each original grayscale image matrix stacked on top of each other as inputs. While InceptionV3 can be downloaded with weights pre-trained on ImageNet dataset to apply transfer learning, we did not use these pre-trained weights because the colored



**Fig. 3** The architecture of our convolutional neural network with all the layers and their dimensions visible

photographs of ImageNet differ from our grayscale MRI data enough to prevent a fair comparison between the proposed CNN and InceptionV3.

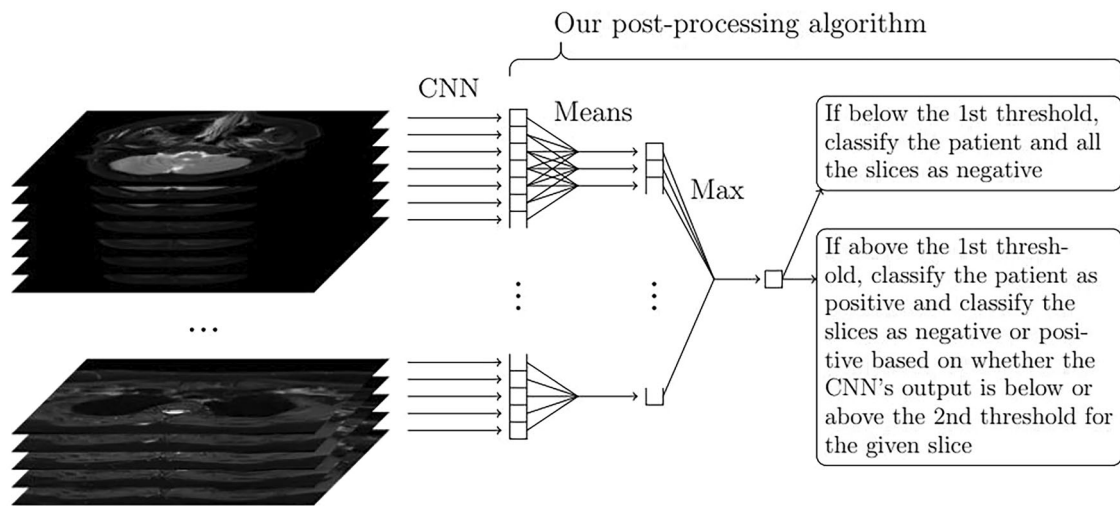
The CNN set-ups were trained with repeated five-fold cross-validation so that we first divided the data into five folds of around 20% of the total data and then each one of the original five folds was used as the test set and the other four as training data for six times, meaning 30 training iterations in total. The five-fold data split was done patient-wise, and the CNN was re-initialized after each iteration. Additionally, during each training iteration, we randomly chose a subset of 30% from the training data to create a validation set so that we could monitor the generalizability of the learning and define an early stopping criterion. The validation data was not used for any other purposes. Consequently, in each iteration, 20% of the total data was used for testing and, from the remaining 80% data, 70% (56% of the total data) was used for the actual training and 30% (24% of the total data) for validation. The loss function was the standard binary cross-entropy, which is compatible with the sigmoid activation used on the last layers of our proposed CNN and InceptionV3 implementations. Adam was chosen as an optimizer because it was noted that the convergence of the binary cross-entropy was very slow and uneven with the standard stochastic gradient descent. Based on a few initial tests, both the proposed CNN and InceptionV3 obtained an accuracy over 90% in the training set during the first two or three epochs, with very little stepwise progress afterwards. Consequently, the number of epochs was set as 15, but early stopping was used if no improvement was obtained in the validation loss during five consecutive epochs. The training time was recorded for each iteration.

#### Data augmentation

In our proposed method, the number of training images is doubled by including their reflections over the vertical axis in the training data. This type of augmentation was compared to the results obtained with no augmentation and augmentation based on clockwise rotation of 90 degrees and blurring images along filters of  $3 \times 3$  pixels. These three specific types of augmentation were originally chosen for this study because it was observed in the article [31] that a reflection in a vertical direction, a rotation of 90 degrees, and a Gaussian blur worked the best with MRI datasets.

#### Post-processing algorithm

The numeric predictions given as an output for the test set by a CNN are typically converted into binary labels by using the threshold that maximizes the Youden's index for the predictions of the training set [32]. The Youden's



**Fig. 4** The structure of our proposed method to obtain the final classifications of a single patient: First, the convolutional neural network (CNN) is used to obtain numeric slice-wise predictions, then the maximum of the mean values of five consecutive slice-wise predictions is computed, and the patient-wise classification is obtained based on whether the maximum is below or above a certain threshold. If the patient is classified as positive, the slice-wise predictions are converted into binary labels with a second threshold; otherwise, they are all classified as negative. The two thresholds are chosen by finding the thresholds that give the maximal Youden's index for the patient- and slice-wise predictions with this same method for the training set predictions. Our post-processing algorithm finds the appropriate thresholds and converts the predictions of all the test set patients into final binary labels

index is defined as the sum of sensitivity (the percentage of the positive instances classified correctly) and specificity (the percentage of the negative instances classified correctly) minus 1 [33]. After this, all the patients with at least one slice predicted as positive can be classified as positive and the rest of the patients as negative. However, this approach likely results in a very low patient-wise specificity. Namely, even if the specificity among the slices would be very high, even a single false positive prediction among all the 26–60 slices of a negative patient would cause a false positive prediction for the given patient. To avoid this issue, we need to first use the numeric predictions of the slices to classify the patients and, only after that, classify the slices in a way that all the slices of the negative patients are negative.

Consequently, in our proposed approach, we perform the post-processing scheme as in Fig. 4 with a new algorithm specifically designed by us for this purpose. As input, this algorithm receives all the numeric slice-wise predictions obtained by the given CNN set-up from both the training and the test set, ordered as vectors so that each vector contains all the slice-wise predictions of a single patient in the correct order. Additionally, the algorithm receives information about the correct classifications of the training set patients. It then goes through the vectors of the test set, lists the mean values of each set of five consecutive slice-wise predictions, and chooses the maximum of these mean values for each vector. The number five was used here because we needed both

information from at least a few adjacent slices, but still a number small enough not to lose sensitivity in the detection of RPE visible only in 2–3 slices. The patient is classified as positive or negative based on whether this maximum is above or below a threshold value, which is equal to the threshold that gives the maximal Youden's index among the predictions of the training set with this exact same approach. After this algorithm, the final predictions of the test set slices were obtained by interpreting all the slices from such patients that were classified as negative by the algorithm as negative and converting the rest of the slice-wise predictions as positive or negative according to the maximal Youden's index chosen based on the training data.

This new algorithm was compared with both the Random Forest and the Support Vector Machine (SVM) algorithms. They were trained by using similar patient-wise vectors and the correct labels of the patients within the training set, though these vectors were padded with additional zeros to obtain the same dimension for all the inputs. The predictions of the test set vectors were converted into binary labels by using the maximal Youden's index of the training set vectors.

#### Evaluation

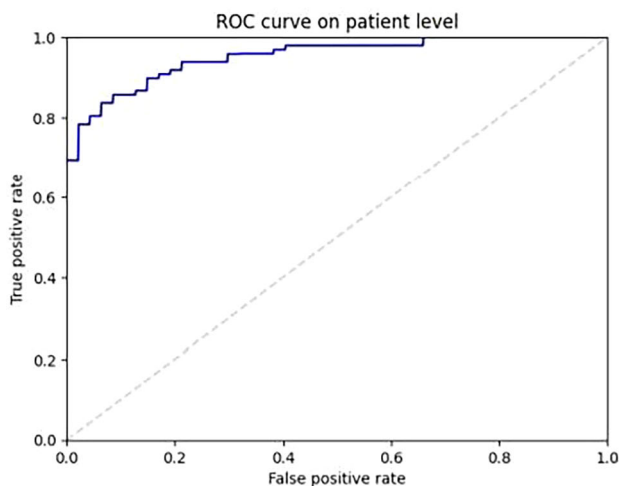
To evaluate our proposed method, we computed the values of accuracy, sensitivity, specificity, and the area under the receiver operating characteristic curve (AUROC) [32] on both patient level and slice level for the

predictions of each test set during the repeated cross-validation. The AUROC of the proposed method on the patient level was obtained by considering the curve drawn by using different thresholds to convert the patients based on the maximums for the mean values of five consecutive slice-wise predictions. Additionally, we compared the proposed CNN and vertical reflection-based augmentation to InceptionV3 and other augmentation choices by computing the same four evaluation metrics directly from the slice-wise CNN output of the test set, with no patient-wise post-processing algorithm. We also assessed the training times of the CNN setups. The comparison between our new post-processing algorithm, random forest, and SVM was done with the same four evaluation metrics on the patient level. As all the CNN setups are trained with five-fold cross-validation repeated six times, we obtained 30 different values for each method variation for the purpose of statistical testing.

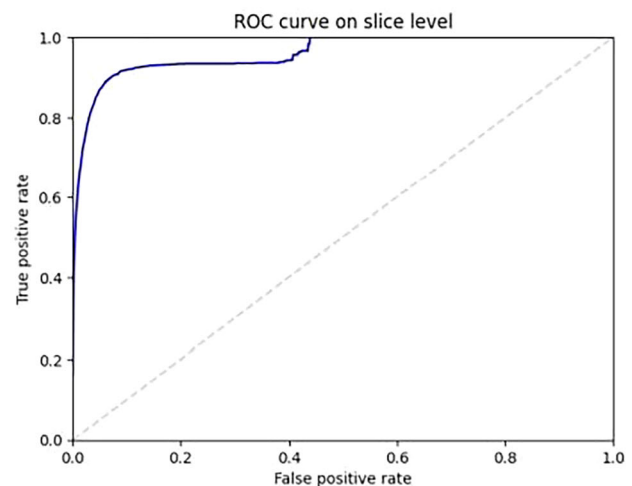
**Table 1** Evaluation metrics for the proposed CNN

Metric	Patient level	Slice level
Accuracy (%)	87.4 ± 2.8/87.0	94.6 ± 0.4/94.6
Sensitivity (%)	86.5 ± 3.2/85.7	83.3 ± 3.6/82.7
Specificity (%)	88.2 ± 4.9/87.2	96.2 ± 0.7/96.2
AUROC (%)	94.8 ± 1.2/94.9	94.1 ± 1.3/94.0

Data are given as mean ± standard deviation/median of the four evaluation metrics computed both on patient level and on slice level with our proposed method (the convolutional neural network shown in Fig. 3 with vertical reflection augmentation followed by our new algorithm) during the 30 iterations of the repeated five-fold cross-validation



(a)



(b)

**Fig. 5** The median AUROC curves on (a) patient and (b) slice level computed from the medians of the ROC curves of the test set predictions by our proposed method after each 30 training iterations

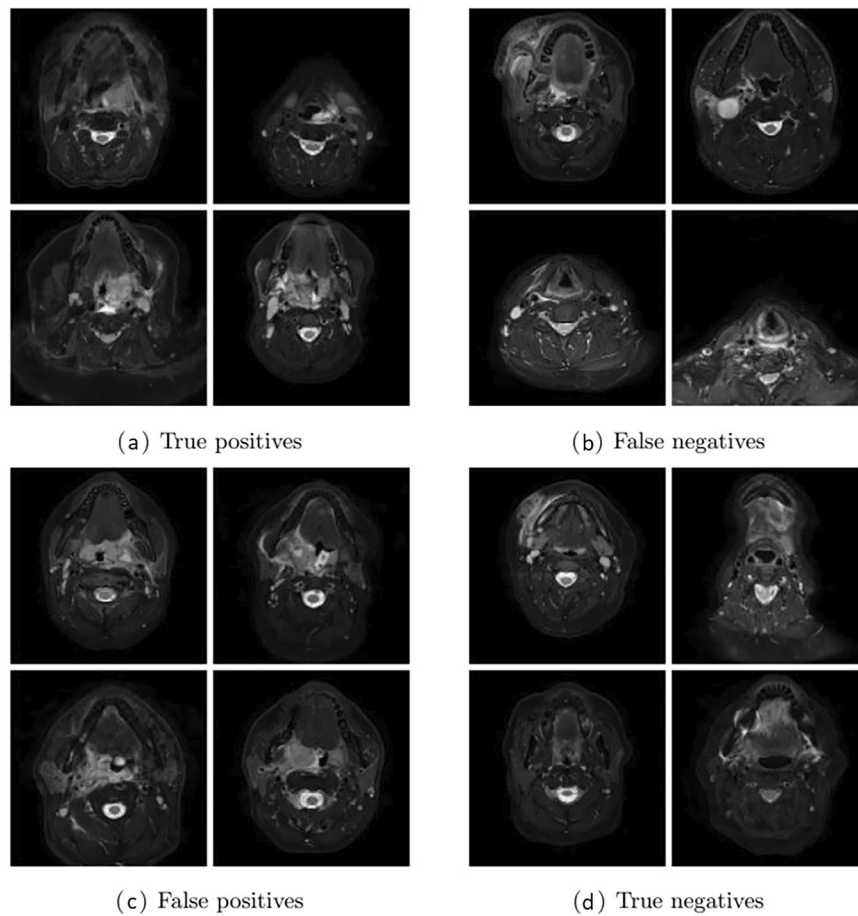
### Statistical analysis

To study whether there were significant differences between 30 different values of accuracy, sensitivity, specificity, or AUROC, we used the Wilcoxon signed-rank test. The reason for this test is that it is non-parametric and therefore not sensitive to potential outliers, and its assumption of paired observations suits our situation where the cross-validation test set varies over different iteration rounds [32]. The multiple comparisons problem was taken into account before drawing conclusions by considering only the  $p$ -values less than 0.01 or 0.001 significant instead of using the standard level of 5%.

### Results

Our proposed method is the CNN based on the U-Net encoder with vertical reflection augmentation, followed by our new algorithm, the results of which are summarized in Table 1. Due to the large number of negative slices, the method has a very high slice-wise specificity, but it is also accurate and sensitive on both the patient and slice levels. In particular, the AUROC values are very high, meaning that the method should work relatively well on different choices of thresholds used for converting numeric predictions. Figure 5 shows the median ROC curves corresponding to the AUROC values of Table 1, and Fig. 6 shows examples of the correct and the incorrect classification decisions by our proposed method.

Tables 2 and 3 show the results of the comparison between the proposed CNN and InceptionV3 and between different augmentation choices, respectively. As can be seen from Table 2, even though the proposed CNN



**Fig. 6** Examples of 16 slices of the first test set that were consistently classified either correctly or incorrectly during the 6 repetitions of the first five-fold cross-validation split, including (a) positive slices correctly classified as positive, (b) positive slices incorrectly classified as negative, (c) negative slices incorrectly classified as positive, and (d) negative slices correctly classified as negative by our proposed method. All the slices are from different patients

and InceptionV3 perform very similarly otherwise, the proposed CNN is considerably more computationally effective: The training period of the non-augmented CNN is less than 10 min, whereas training InceptionV3 takes over 4 h with the same data and hardware. Table 3 reveals that the vertical reflection produces significantly better AUROC values than the other augmentation choices, even if no augmentation method is superior in terms of all four evaluation metrics due to the trade-off between sensitivity and specificity.

Table 4 summarizes the patient-wise predictions obtained by our new algorithm compared to random forest and SVM. The predictions are all computed by using the numeric output of the proposed CNN with vertical reflection augmentation. As can be seen, our new algorithm has the highest accuracy and AUROC, and it outperforms the two other algorithms in terms of AUROC in a statistically significant way.

**Table 2** The proposed CNN versus InceptionV3

Metric	Proposed convolutional neural network	InceptionV3
Accuracy (%)	<b>91.6 ± 0.9/91.5</b>	91.6 ± 1.1/ <b>91.7</b>
Sensitivity (%)	<b>89.3 ± 2.4/89.2</b>	89.2 ± 2.6/89.1
Specificity (%)	<b>91.9 ± 1.2/91.8</b>	91.9 ± 1.5/92.1
AUROC (%)	<b>96.9 ± 0.5/97.0</b>	96.7 ± 0.7/96.5*
Time (min)	<b>9.3 ± 1.1/9.1</b>	277.7 ± 60.5/270.1***

Data are given as mean ± standard deviation/median of the four evaluation metrics for the predictions of the test set slices and the training time of a single iteration for the proposed convolutional neural network and InceptionV3 during the 30 iterations of the repeated five-fold cross-validation. No augmentation nor post-processing algorithm was used. The values of InceptionV3 that significantly differ those of the proposed convolutional neural network according to a Wilcoxon signed-rank test are denoted as follows: \* $p < 0.05$ ; \*\* $p < 0.01$ ; \*\*\* $p < 0.001$ . The better values of evaluation metrics and time are in bold

**Table 3** Impact of augmentation

Metric	Reflection	No augmentation	Rotation	Blurring
Accuracy (%)	92.8 ± 0.8/92.8	91.6 ± 0.9/91.5***	91.5 ± 1.2/91.6***	<b>92.8 ± 0.7/92.8</b>
Sensitivity (%)	89.0 ± 2.8/88.5	89.3 ± 2.4/89.2	<b>90.3 ± 2.2/90.2*</b>	81.4 ± 3.6/81.1***
Specificity (%)	93.4 ± 1.1/93.3	91.9 ± 1.2/91.8***	91.7 ± 1.6/91.8***	<b>94.4 ± 1.0/94.6**</b>
AUROC (%)	<b>97.3 ± 0.4/97.3</b>	96.9 ± 0.5/97.0***	97.1 ± 0.3/97.0**	96.3 ± 0.6/96.3***
Time (min)	22.2 ± 2.6/22.0	<b>9.3 ± 1.1/9.1***</b>	19.6 ± 2.3/19.3**	31.9 ± 3.2/32.7***

Data are given mean ± standard deviation/median of the four evaluation metrics for the predictions of the test set slices and the training time of a single iteration for the proposed convolutional neural network with different augmentation choices during the 30 iterations of the repeated five-fold cross-validation. These choices include either augmentation with vertical reflection, no augmentation, augmentation with 90-degree rotation, or augmentation with image blurring, and each augmentation type was used to double the amount of the original training data. No post-processing algorithm was used. The best values are in bold and the values that significantly differ from those of the vertical reflection augmentation according to a Wilcoxon signed-rank test are denoted by \*, \*\*, and \*\*\* as in Table 2

**Table 4** Proposed post-processing algorithm versus Random forest and SVM

Metric	Our algorithm	Random forest	Support vector machine
Accuracy (%)	<b>87.4 ± 2.8/87.0</b>	85.9 ± 2.7/86.4***	86.8 ± 3.1/86.5
Sensitivity (%)	86.5 ± 3.2/85.7	<b>87.3 ± 3.3/87.5</b>	85.7 ± 3.8/85.7
Specificity	<b>88.2 ± 4.9/87.2</b>	84.5 ± 4.5/85.1***	87.9 ± 4.3/88.3
AUROC (%)	<b>94.8 ± 1.2/94.9</b>	86.1 ± 2.6/86.4***	86.8 ± 3.1/86.5***

Data are given as mean ± standard deviation/median of the four evaluation metrics for the predicted classifications of the test set patients computed with our new algorithm, random forest, and support vector machine from the slice-wise predictions obtained with the proposed convolutional neural network with vertical reflection augmentation during the 30 iterations of the repeated five-fold cross-validation. The best values are in bold and the values that significantly differ those of our new algorithm according to a Wilcoxon signed-rank test are denoted by \*, \*\*, and \*\*\* as in Table 2

## Discussion

In emergency MRI of acute neck infections, RPE is an imaging biomarker with clinical and prognostic significance. Here, our DL-based algorithm for the automated detection of RPE from axial T2-weighted water-only Dixon images achieved high diagnostic accuracy on both slice and patient levels against the reference standard of a consensus between two neuroradiologists. Our model's high accuracy and manageable computational requirements pave the way for clinical implementation in the emergency setting.

Our proposed method consists of a classifier CNN based on the encoder of the U-Net, combined with a simple algorithm. While the CNN is two-dimensional and processes the transaxial slices separately, our new algorithm allows us to utilize the depth-wise information of the original three-dimensional MRI images. Additionally, the training of our proposed CNN takes only 3.3% of the time required by the state-of-the-art classification CNN InceptionV3 without resulting in a less accurate classification performance. While the inference times were not recorded or compared between the models, they are around a few seconds for the whole test data, and their

potential differences would, therefore, not have a practical meaning when predicting the image of a specific patient in the emergency department. Consequently, our method is both very lightweight and accurate. In addition, we used only weakly annotated data based on slice- and patient-based binary labels instead of annotating RPE specifically within positive slices.

The accuracy of our DL-based algorithm was higher at the slice level than at the patient level. Specificity (the proportion of true negatives to all negatives) dropped slightly more than sensitivity from the slice level to the patient level, most likely due to the considerably higher number of true negatives at the slice level (88%) than at the patient level (49%). Nevertheless, the accuracy at the patient level is reasonably promising for clinical applications, given that the pre-test probability of RPE is not particularly low (about 50%).

Despite overall high accuracy, some misclassifications were evident (Fig. 6). RPE is defined as high signal area between the superior constrictor muscle and the pre-vertebral muscle. Some false positives appear to have been due to edema (high signal intensity) in the superior constrictor muscle itself (Fig. 6c). In contrast, some false negatives are unilateral and thin appearances of RPE (Fig. 6b) that diagnostic radiologists have nevertheless classified as positive. These false model classifications, although rare, partially highlight the difficulties associated with the subjective nature of the reference standard.

Our findings suggest some directions for future research. Our approach involved a CNN prediction of the patients' RPE status from individual axial slices. An alternative strategy would be to employ a three-dimensional CNN, although this would be computationally more burdensome. Another avenue for future work would be to use patient-level annotation and only the images as input to directly predict clinical outcomes, such as intensive care unit admissions. This would bypass the subjective reference standard of RPE, representing an operationalized imaging biomarker or a proxy for human

observers. Such direct classification might have higher accuracy for specific clinical outcomes. Still, it would inherently be limited to only one outcome at a time, unlike RPE, which is associated with multiple clinical and laboratory outcomes in acute neck infections [6] and thus perhaps more versatile.

Our algorithm might be clinically applied as an automated triage tool to prioritize neck MRI scans with clinically and prognostically significant findings, alerting the on-call radiologist. In addition, this algorithm might improve the diagnostic accuracy of radiologists who are less familiar with acute neck MRI interpretation during on-call hours. Finally, automated tools for image interpretation might be helpful for clinicians, who often view images themselves.

Particular strengths of the study include a large sample size of high-quality 3-T MRI images from patients with acute neck infections of various etiologies, corroborated by thorough clinical and surgical characterization. From the technical and generalizability perspective, a considerable strength is the high performance of the model with very manageable hardware requirements. Yet, some limitations need to be addressed. The most critical limitation is the subjective nature of the reference standard for RPE, based on the consensus of two experienced and fellowship-trained head and neck radiologists. This limitation is inherent in many studies on semiquantitative imaging biomarkers, such as bone marrow edema in sacroiliitis [12]. Next steps to further corroborate this imaging finding might be to establish the intra- and interobserver reproducibility in a large sample of annotators with varying levels of expertise and to grade RPE using automated volumetric segmentations to achieve more fine-grained correlations to patient outcomes. Related to the reference standard, perfect model performance would be as good as that of a consensus of two fellowship-trained neuroradiologists. A practical implication is that the model would benefit radiologists and clinicians who are less familiar with acute neck MRI interpretation. Yet, this is a probable setting in most emergency radiology departments, especially outside office hours. All MRI data were obtained from a single institution, excluding external site validation and limiting generalizability. We are unaware of other large published MRI datasets on acute neck infections. In general, the availability of emergency MRI is much lower than that of computed tomography, the most commonly used cross-sectional imaging method in acute neck infections. This will limit potential clinical applicability, which will be required to assess the impact on reporting times and patient outcomes. Concerning the DL algorithm, we could not produce attention maps, limiting the explainability of

our model. In any case, the high performance of the model should be useful in clinical practice.

In conclusion, we found sensitivity, specificity, and AUROC of 86.5%, 88.2%, and 94.8%, respectively, for a DL-based algorithm for automated detection of RPE, an MRI imaging biomarker of clinical and prognostic significance in patients with acute neck infections. Highly accurate automated methods for patient prognostication may improve patient outcomes and the generalizability of advanced imaging methods in the emergency setting.

#### Abbreviations

AUROC	Area under receiver operating characteristic curve
CNN	Convolutional neural network
DL	Deep learning
MRI	Magnetic resonance imaging
RPE	Retropharyngeal edema
SVM	Support vector machine

#### Acknowledgements

No LLMs (large language models) were used.

#### Author contributions

O Rainio wrote the code, ran the experiments, and reported their results. H Huhtanen, J-P Vierula, J Nurminen, J Heikkinen, M Nyman, and J Hirvonen collected the original imaging data. R Klén and J Hirvonen supervised the project. O Rainio and J Hirvonen wrote the manuscript and prepared the figures. All the authors read the manuscript and helped in editing it.

#### Funding

This study was financially supported by the Sigrid Jusélius Foundation. The first author was financially supported by the Sakari Alhopuro Foundation.

#### Data availability

Data cannot be publicly shared because of the national legislature's on the privacy of patient data. The code is available at [https://github.com/rklen/Rpe\\_project](https://github.com/rklen/Rpe_project).

#### Declarations

##### Ethical approval and consent to participate

According to the national legislation, no separate ethics committee approval is needed for retrospective studies that involve a secondary use of registry or archival data.

##### Consent for publication

According to the national legislation, no separate ethics committee approval is needed for retrospective studies that involve a secondary use of registry or archival data.

##### Competing interests

The authors declare that they have no competing interests.

Received: 24 January 2025 Accepted: 4 June 2025

Published online: 19 June 2025

#### References

1. Boscolo-Rizzo P, Stellin M, Muzzi E et al (2012) Deep neck infections: a study of 365 cases highlighting recommendations for management and

- treatment. *Eur Arch Otorhinolaryngol* 269:1241–1249. <https://doi.org/10.1007/s00405-011-1761-1>
2. Kamalian S, Avery L, Lev MH et al (2019) Nontraumatic head and neck emergencies. *Radiographics* 39:1808–1823. <https://doi.org/10.1148/rg.2019190159>
  3. Hirvonen J, Heikkinen J, Nyman M et al (2023) MRI of acute neck infections: evidence summary and pictorial review. *Insights Imaging* 14:5. <https://doi.org/10.1186/s13244-022-01347-9>
  4. Nurminen J, Velhonoja J, Heikkinen J et al (2021) Emergency neck MRI: feasibility and diagnostic accuracy in cases of neck infection. *Acta Radiologica* 62:735–742. <https://doi.org/10.1177/0284185120940242>
  5. Nurminen J, Heikkinen J, Happonen T et al (2022) Magnetic resonance imaging findings in pediatric neck infections—a comparison with adult patients. *Pediatr Radiol* 52:1158–1166. <https://doi.org/10.1007/s00247-021-05275-6>
  6. Heikkinen J, Nurminen J, Velhonoja J et al (2022) Clinical and prognostic significance of emergency MRI findings in neck infections. *Eur Radiol* 32:1078–1086. <https://doi.org/10.1007/s00330-021-08200-5>
  7. Heikkinen J, Nurminen J, Velhonoja J et al (2022) MRI findings in acute tonsillar infections. *AJNR Am J Neuroradiol* 43:286–291. <https://doi.org/10.3174/ajnr.A7368>
  8. Heikkinen J, Jokihaka V, Nurminen J et al (2023) MRI of odontogenic maxillofacial infections: diagnostic accuracy and reliability. *Oral Radiol* 39:364–371. <https://doi.org/10.1007/s11282-022-00646-7>
  9. Vierula JP, Nurminen J, Jussila V et al (2023) Diagnostic performance of short noncontrast biparametric 3-T MRI for tonsillar infections: comparison with a full protocol including contrast-enhanced sequences. *Eur Radiol Exp* 7:65. <https://doi.org/10.1186/s41747-023-00379-0>
  10. Montagnon E, Cerny M, Cadrin-Chênevert A et al (2020) Deep learning workflow in radiology: a primer. *Insights Imaging* 11:22. <https://doi.org/10.1186/s13244-019-0832-5>
  11. Huhtanen H, Nyman M, Mohsen T et al (2022) Automated detection of pulmonary embolism from CT-angiograms using deep learning. *BMC Med Imaging* 22:43. <https://doi.org/10.1186/s12880-022-00763-z>
  12. Bressen KK, Adams LC, Proft F et al (2022) Deep learning detects changes indicative of axial spondyloarthritis at MRI of sacroiliac joints. *Radiology* 305:655–665. <https://doi.org/10.1148/radiol.212526>
  13. Kucybała I, Tabor Z, Polak J et al (2020) The semiautomated algorithm for the detection of bone marrow oedema lesions in patients with axial spondyloarthritis. *Rheumatol Int* 40:625–633. <https://doi.org/10.1007/s00296-020-04511-w>
  14. Lin KYY, Peng C, Lee KH et al (2022) Deep learning algorithms for magnetic resonance imaging of inflammatory sacroiliitis in axial spondyloarthritis. *Rheumatology* 61:4198–4206. <https://doi.org/10.1093/rheumatology/keac059>
  15. Lee KH, Choi ST, Lee GY et al (2021) Method for diagnosing the bone marrow edema of sacroiliac joint in patients with axial spondyloarthritis using magnetic resonance image analysis based on deep learning. *Diagnostics (Basel)* 11:1156. <https://doi.org/10.3390/diagnostics11071156>
  16. Ozga J, Wyka M, Raczko A et al (2023) Performance of fully automated algorithm detecting bone marrow edema in sacroiliac joints. *J Clin Med* 12:4852. <https://doi.org/10.3390/jcm12144852>
  17. Zheng Y, Bai C, Zheng K et al (2023) Deep-learning based quantification model for hip bone marrow edema and synovitis in patients with spondyloarthritis based on magnetic resonance images. *Front Physiol* 14:1132214. <https://doi.org/10.3389/fphys.2023.1132214>
  18. Yu AS, Yang M, Lartey R et al (2024) Unsupervised segmentation of knee bone marrow edema-like lesions using conditional generative models. *Bioengineering* 11:526. <https://doi.org/10.3390/bioengineering11060526>
  19. Lin Y, Chan SCW, Chung HY et al (2024) A deep neural network for MRI spinal inflammation in axial spondyloarthritis. *Eur Spine J* 33:4125–4134. <https://doi.org/10.1007/s00586-023-08099-0>
  20. Rainio O, Han C, Teuvo J et al (2023) Carimas: an extensive medical imaging data processing tool for research. *J Digit Imaging* 36:1885–1893. <https://doi.org/10.1007/s10278-023-00812-1>
  21. Van Rossum G, Drake FL (2009) Python 3 Reference Manual ISBN: 1441412697. CreateSpace, Scotts Valley, CA
  22. Chollet F (2015) Keras <https://github.com/fchollet/keras>
  23. Abadi M, Agarwal A, Barham P et al (2015) TensorFlow: large-scale machine learning on heterogeneous systems. Mountain View, CA: TensorFlow
  24. Pedregosa F, Varoquaux G, Gramfort A (2011) Scikit-learn: machine learning in Python. *J Mach Learn Res* 12:2825–2830
  25. Virtanen P, Gommers R, Oliphant TE et al (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* 17:261–272. <https://doi.org/10.1038/s41592-019-0686-2>
  26. Hellström H, Liedes J, Rainio O et al (2023) Classification of head and neck cancer from PET images using convolutional neural networks. *Sci Rep* 13:10528. <https://doi.org/10.1038/s41598-023-37603-1>
  27. Ronneberger O, Fischer P, Brox T (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation in Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, Proceedings, Part III vol. 18. pp. 234–241
  28. Yu H, Yang LT, Zhang Q et al (2021) Convolutional neural networks for medical image analysis: state-of-the-art, comparisons, improvement and perspectives. *Neurocomputing* 444:92–110. <https://doi.org/10.1016/j.neucom.2020.04.157>
  29. Szegedy C, Vanhoucke V, Ioffe S et al (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE. pp. 2818–2826
  30. Li Z, Liu F, Yang W et al (2021) A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE Trans Neural Netw Learn Syst* 33:6999–7019. <https://doi.org/10.1109/TNNLS.2021.3084827>
  31. Rainio O, Klén R (2024) Comparison of simple augmentation transformations for a convolutional neural network classifying medical images. *SIVIP* 18:3353–3360. <https://doi.org/10.1007/s11760-024-02998-5>
  32. Rainio O, Teuvo J, Klén R (2024) Evaluation metrics and statistical tests for machine learning. *Sci Rep* 14:6086. <https://doi.org/10.1038/s41598-024-56706-x>
  33. Rainio O, Tamminen J, Venäläinen MS et al (2024) Comparison of thresholds for a convolutional neural network classifying medical images. *Int J Data Sci Anal* <https://doi.org/10.1007/s41060-024-00584-z>

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.