# Using machine learning on MRI radiomics to diagnose parotid tumours before comparing performance with radiologists.

Samy Ammari[1,2†], Arnaud Quillent[3†], Víctor Elvira[4], François Bidault[1,2], Gabriel C.T.E. Garcia[2], Dana M. Hartl[5], Corinne Balleyguier[1,2], Nathalie Lassau[1,2], Émilie Chouzenoux[3*]

**Abstract**

The parotid glands are the largest of the major salivary glands and can harbour both benign and malignant tumours. Preoperative work-up relies on MR images and fine needle aspiration biopsy, but these diagnostic tools have low sensitivity and specificity, often leading to surgery for diagnostic purposes. Machine learning methods along with radiomic features are widely used in the field of medical imaging to help radiologists make a diagnosis. The aim of this paper is to develop an algorithm based on image characteristics to automatically predict the type of parotid gland tumours. We then compare this algorithm to the diagnoses of junior and senior radiologists in order to evaluate its utility in routine practice. To create the algorithm, we enrolled a cohort of 134 patients treated for benign or malignant parotid tumours. Using radiomics extracted from the MR images of their parotid tumours, we train a random forest model to classify data into corresponding histopathological subtypes. On the test set, we obtain a 0.72 accuracy, a 0.86 specificity and a 0.72 sensitivity over all histopathological subtypes, with an average AUC of 0.838. When considering the discrimination between benign and malignant tumours, the algorithm results in a 0.76 accuracy and a 0.769 AUC, both on test set. Moreover, the clinical experiment shows that our model helps to improve diagnostic abilities of junior radiologists as their sensitivity and accuracy raised by 6 % when using our proposed method. This algorithm may be useful for training of physicians. Radiomics with a machine learning algorithm may help improve discrimination between benign and malignant parotid tumours, decreasing the need for diagnostic surgery. Further studies are warranted to validate our algorithm for routine use.

**Keywords**

Radiomics — Machine learning — Parotid glands — AI benefit analysis

[1] *Biomaps, UMR1281 INSERM, CEA, CNRS, Université Paris-Saclay, 94805 Villejuif, France*
[2] *Department of Imaging, Gustave Roussy Cancer Campus, Université Paris Saclay, 94805 Villejuif, France*
[3] *Centre de Vision Numérique, OPIS, CentraleSupélec, Inria, Université Paris-Saclay, 91190 Gif-sur-Yvette, France*
[4] *School of Mathematics, University of Edinburgh, Edinburgh EH9 3FD, United Kingdom*
[5] *Department of Otolaryngology Head and Neck Surgery, Gustave Roussy Cancer Campus, Université Paris Saclay, 94805 Villejuif, France*
*__Corresponding author__: emilie.chouzenoux@inria.fr
† These authors contributed equally to this work

This preprint has not undergone peer review or any post-submission improvements or corrections. The Version of Record of this article is published in Journal of Imaging Informatics in Medicine, and is available online at https://dx.doi.org/10.1007/s10278-024-01255-y.

## Introduction

Parotid gland tumours are benign in 70 to 80 % of cases [1, 2]. Preoperative work-up relies on MRI and fine needle aspiration cytology (FNAC), but these diagnostic tools lack sensitivity, often leading to surgery for diagnostic purposes. The sensitivity of FNAC for parotid tumours using the recommended Milan system ranges from 75 to 90 % [3]. Because of the heterogeneity of parotid gland tumours and their possible location in the deep lobe, FNAC is not always sufficient to come up with a correct diagnosis [4, 5]. Moreover, diagnostic parotid surgery carries a high risk of morbidity due to the fact that the facial nerve passes through the gland and dissection of the nerve can lead to physical and aesthetical discomfort, as well as temporary or permanent facial paralysis.

With the development of computer science, new automatic data processing techniques and machine learning can now be utilised to achieve numerous tasks such as image classification and segmentation. Usage of machine learning is increasingly

growing in the healthcare field due to the outstanding results it provides [6, 7].

Radiomics are a field of radio-diagnostics which first appeared in the 2010s. It consists in computing several statistics from a segmented area of a medical image such as its shape, volume, texture, intensity, or other high-order moments [8, 9]. Radiomics aim at extracting patterns in the data that would be invisible to the human eye. Combined with machine learning approaches, radiomics have been used by many researchers to automatically analyse MR and CT images [10, 11].

The goal of our study is to introduce a machine learning algorithm for the discrimination of parotid gland tumours into their respective histopathological subtypes as defined by the World Health Organization (WHO) [12]. Previous works were conducted on this matter, highlighting the efficiency of such kind of approach [13, 14]. However, they only used at most two MRI sequences, whereas we decide to sample four to observe whether it improves automatic diagnosis performances. Moreover, we perform a comparative analysis of our proposed algorithm with radiologists in order to evaluate benefits of machine learning algorithms combined to relevant experts' knowledge.

To conduct our research, we create a new dataset coming from a cancer treatment centre. As gathered from an expert hospital, we believe those data encompass very germane information about parotid cancers. On top of that, as all patients coming to the centre have previously been examined in community hospitals, we have at our disposal large collections of images that are not sampled from regular screening, thus resulting in a greater proportion of parotid cancer positives than observed in the whole population.

This research article is organised as follows. section 1, Materials and methods introduces our data sampling strategy, the subsequent radiomics extraction, the development of the machine learning algorithm and finally the implemented protocol for comparison with radiologists. Our findings are then presented in section 2, Results and further analysed in section 3, Discussion.

# 1. Materials and methods

## 1.1 Data acquisition

**Patient selection** The cohort includes patients that underwent an MRI examination for parotid tumours at Gustave Roussy Cancer Campus between 2012 and 2021. All the patients subsequently underwent surgery, with histopathologic examination performed in the same centre. We categorise the histopathological subtypes into pleomorphic adenomas (benign), Warthin's tumour (benign) and carcinomas (malignant). Tumours that do not belong to the previous categories are reported as 'Other type'. Descriptive statistics are shown in Table 1. As the patients were hospitalised in an institution specialised in cancer treatment, malignant tumours are over-represented in the studied cohort compared to their frequency in the worldwide population.

This research project was approved by the appropriate institutional board following the General Data Protection Regulation (GDPR) and was declared to the competent national administrative bodiess, namely the Health Data Hub and the CNIL. All patients were informed by post and provided their consent to the use of their data within the scope of this study.

**MRI protocol** Acquisition was performed on three different devices over the study period: a 1.5 T Optima MR450w, a 3 T Discovery MR750w and a 3 T Signa Premier (GE HealthCare, Milwaukee, USA). Corresponding MRI parameters are reported in Table 2. Patients were placed in a supine position, with a head and neck antenna. T1-weighted (T1w), contrast-enhanced T1-weighted (T1ce), T2-weighted (T2w) and diffusion-weighted (DWI) sequences were acquired for each patient. Gadoterate meglumine (Dotarem, Guerbet, Villepinte, France) was used as the contrast agent in all cases.

**Data cleaning** To ensure robustness of our machine learning approach, we rely on several data cleaning steps. First, as lymphomas are under-represented, they were removed from the dataset: they were too few for the machine learning algorithm to learn to classify them. Tumours that are described as 'Other type' were too heterogeneous, as expected, to be gathered into one meaningfully consistent class, and too few to be broken down into subgroups. Hence, they were also removed from the dataset. From the 134 patients of the whole cohort, only 111 were kept for the study, after this procedure (see Figure 1).

## 1.2 Image analysis

**Image pre-processing** As MR images are very heterogeneous between patients and acquisition devices, a pre-processing step of harmonisation is often performed in order to compensate the high variability of the data. In many radiomics studies, a first normalisation is applied at pixel-level on the images [15, 16]. However, as some patients in our cohort were examined with 3 T MRI devices and others with a 1.5 T machine with lower resolution, normalising the images might have resulted in a loss of information. Consequently, no normalisation was performed in our study.

**Tumour segmentation** Parotid gland tumours were manually segmented using the software Olea Sphere version 3.0.18 (Olea Medical, La Ciotat, France). All patients of the cohort present a single parotid gland tumour except for one who is affected by two tumours which were both used for this study. Each parotid gland tumour was segmented on the four different MRI sequences (see section 1.1, MRI protocol). Studied MR images are 3-dimensional, but we choose to segment 2D regions of

**Table 1.** Statistics about the studied cohort.

| **Characteristics** | **Value** |
|---|---|
| **Age** | **Years** |
| Mean | 63.7 |
| Median | 62 |
| Minimum | 18 |
| Maximum | 92 |
| **Sex** | **n (%)** |
| Male | 76 (57) |
| Female | 58 (43) |
| **MRI device power** | **n (%)** |
| 1.5 T | 36 (27) |
| 3 T | 98 (73) |
| **Tumour malignity** | **n (%)** |
| Benign | 73 (54) |
| Malignant | 61 (46) |
| **Tumour histopathological subtype** | **n (%)** |
| Pleomorphic adenoma | 36 (27) |
| Warthin's tumour | 27 (20) |
| Lymphoma | 4 (3) |
| Carcinoma | 48 (36) |
| Other | 19 (14) |

interest (ROI) using the axial slice with the largest tumour surface area. Within this slice, the entire tumour region is segmented, including its core and eventual necrosis. Figure 2 presents an example of such segmentation. Output ROIs were double-checked by an experimented radiologist.

**Features extraction** Radiomic features are computed from all four considered sequences using Olea Sphere software with the following settings: 1 $mm^3$ voxel resampling, 64 gray level bins to compute the histogram. A total of 108 radiomics per sequence are extracted and complied with the IBSI standard [17]. They are sampled as follows: 16 shape-based features, 19 first-order features, 23 gray level co-occurrence matrix (GLCM) features, 16 gray level run length matrix (GLRLM) features, 15 gray level size zone matrix (GLSZM) features, 5 neighbouring gray tone difference matrix (NGTDM) features and 14 gray level dependence matrix (GLDM) features. Appendix A summarises the computed features. Hence, a single tumour is represented by $4 \times 108 = 432$ real numbers.

Radiomics are also subject to high heterogeneity. ComBat normalisation [18, 19] can be used to reduce variability across

| **Device** | **Weighting** | **Sequence** | **TR** | **TE** | **Slice thickness** |
|---|---|---|---|---|---|
| Optima MR450w, 1.5 T | Pre-contrast T1 | Fast spin echo | 532 ms | Minimal | 3 mm |
| Installed in 2016, 70 cm tunnel, | Post-contrast T1 | Fast spin echo | 532 ms | Minimal | 3 mm |
| 32 channels, 40 cm z-axis FOV, | T2 | Fast spin echo | 9,430 ms | 102 ms | 5 mm |
| gradient 40 mT/m, SR 200 T/m/s | DWI | Spin echo | 2,310 ms | Minimal | 4 mm |
| Discovery MR750w, 3 T | Pre-contrast T1 | Fast spin echo | 680 ms | 10 ms | 3 mm |
| Installed in 2012, 70 cm tunnel, | Post-contrast T1 | Fast spin echo | 700 ms | 11 ms | 3 mm |
| 32 channels, 50 cm z-axis FOV, | T2 | Fast spin echo | 9,700 ms | 90 ms | 3 mm |
| gradient 44 mT/m, SR 200 T/m/s | DWI | Spin echo | 7,100 ms | 73 ms | 4 mm |
| Signa Premier, 3 T | Pre-contrast T1 | Fast spin echo | 500 ms | Minimal | 3 mm |
| Installed in 2021, 70 cm tunnel, | Post-contrast T1 | Fast spin echo | 500 ms | Minimal | 3 mm |
| 32 channels, 50 cm z-axis FOV, | T2 | PROPELLER | 8,831 ms | 120 ms | 3 mm |
| gradient 80 mT/m, SR 200 T/m/s | DWI | Spin echo | 3,114 ms | 60 ms | 4 mm |

**Table 2.** MRI devices parameters.

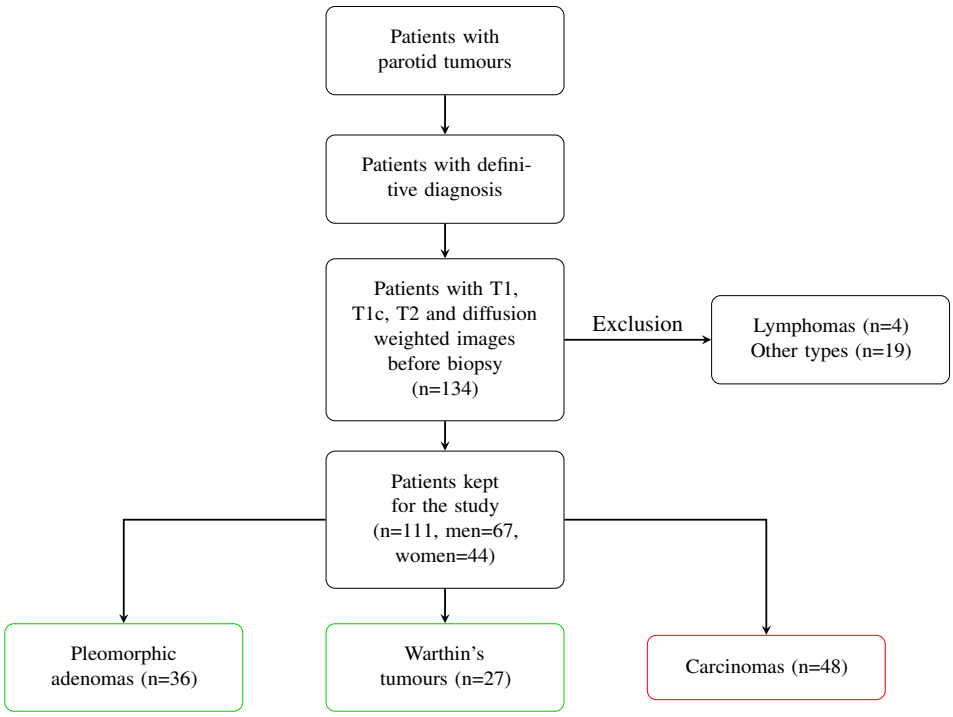


**Figure 1.** Flowchart showing how relevant data were chosen. The red box indicates malignant tumours, the green ones correspond to benign tumours.

devices, but several safety checks performed on our data suggests no significant improvement when applying this method to our problem. Here again, we do not use any pixel-level normalisation technique.

From the medical records of the patients, phenotypic variables age and sex were retrieved, as they are often linked to different types of tumours [1, 2]. Those two variables are concatenated to the computed radiomic features and used as input to the machine learning system.

### 1.3 Machine learning task

Our goal is to discriminate parotid gland tumours according to their histopathological subtype. This corresponds to a classification task, where each type of tumour is a different class, and inputs are the radiomic features.

**Separation into training and test sets** The data are divided into a training group and a test group, composed of respectively 86 and 25 patients. Our machine learning model is developed using the training set and its performance assessed on the test set. The training group comprises data acquired from both 1.5 T Optima MR450w and 3 T MR750w whereas the test group, which was sampled after the new 3 T Signa Premier was put into operation, comprises data from the three devices. This was motivated by [13] which uses different MRI devices for training and validation, but keeps the same magnetic field strength for both. Such a set-up allows us to verify the generalisation to new MRI systems of similar power while avoiding significant bias. Class distribution of the whole dataset is preserved in a stratified manner while creating both training and test sets: the frequency of appearance of each histopathological subtype is kept.

**Model building** Random forests are more robust than other machine learning models when dealing with high-dimensional feature sets, and are also more polyvalent [20, 21]. Moreover, this family of algorithms is quite often used to classify radiomics [10, 22, 23] and tends to be less inclined to overfitting [24], a property that could help the algorithm to generalise to MR images from the new device we considered in the test set. Hence, we decide to use this kind of model in our work. In order to compensate the imbalanced class repartition, classes are given different weights to increase the influence of the least populated ones.

**Model training** A five-fold cross-validation is employed to estimate model performances while keeping an 80:20 ratio between training and test samples. This method consists in subsampling the training dataset into five folds. A model with fixed parameters is then used to predict the classes associated with four of them, and further tested on the fold that was kept aside. The performance metrics are subsequently averaged along those groups, effectively smoothing the high variability of

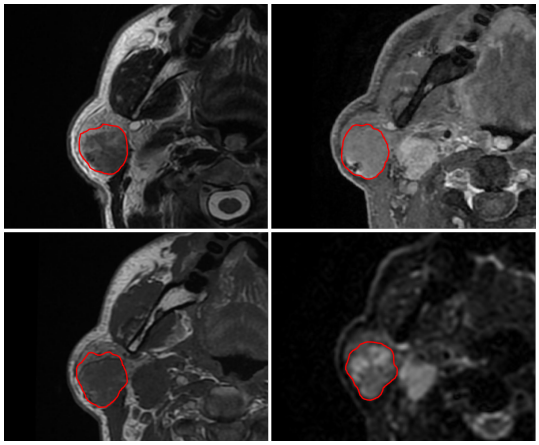

**Figure 2.** Examples of segmentation on the four acquired sequences: T1w, T1ce, DWI and T2w (clockwise).

data. To avoid overfitting, Z-score standardisation and parameters tuning are performed inside the cross-validation loop, as recommended in [25]. The random forest parameters we analysed are maximum depth of the trees, splitting criterion and associated splitting method, number of features inputted to the trees, number of samples in a leaf and number of samples to reach before splitting a node. Classes weights were also fine-tuned. All combinations of selected parameters values are used in order to find the one most relevant to our problem. Except for the former, all of them are binary-based metrics. We keep track of five metrics to assess the performance throughout our parameters search procedure: balanced accuracy (takes into account class imbalance), accuracy, sensitivity, specificity and area under the ROC curve (AUC). Definitions of those metrics are given in Appendix B, following naming conventions used by Scikit-learn[1].

**Feature selection** Feature selection is applied inside the cross-validation loop performed on the training set to compensate for the large number of features acquired and the relatively small number of patients selected ('curse of dimensionality'). Radiomics are often strongly correlated due to the high number of tweakable parameters used for their computation. They are very generic and can be used on various problems, but some of these features may be redundant for certain applications and not for others [26]. A selection of a feature subset used to train the algorithm was therefore required to avoid overfitting [27]. An F-test-based filter method was preferred because of its ability to sort out variables that are independent of the target while keeping a low computational cost [27, 28]. The best features are the ones with lowest p-value. As part of our approach, we test several values for the number of features to retain using the previously mentioned cross-validation procedure. The set of the best features is then used to subsample the test set and evaluate the metrics.

### 1.4 Comparison to the analysis of radiologists

Once we have the best performing model, we are able to compare the machine learning algorithm to radiologists' analyses of the same image sets. First, nine patients were randomly sampled from the test set, keeping the class distribution intact: 3 pleomorphic adenomas, 2 Warthin's tumours and 4 carcinomas. Seven in-training radiologists ('juniors') that just attended a lesson on parotid gland tumours and two professionals ('seniors' with 5 and 15 years of experience respectively) participated in the study. For each patient whose data were used in the experiment, all nine radiologists were shown the same four MRI sequences as the ones used as basis to extract our radiomic features (i.e., T1w, T2w, T1ce and DWI sequences). Following this procedure ensured that both radiologists and machine learning model were using the same input information to diagnose the patients. The images were displayed to physicians on a same screen, using a single software with fixed settings for every viewing. Directly after making their diagnosis, the radiologists were shown the algorithm's prediction and were asked whether they would change their mind given the machine's diagnosis. The true performance of the algorithm was never disclosed during the experiment.

## 2. Results

### 2.1 Machine learning algorithm

**Type classification** Among all the combinations of parameters tested, the best one reached an accuracy of 0.720 and a mean AUC of 0.838 on test set when considering all classes. The same model presented an OvR accuracy of 0.840 for both pleomorphic adenomas and Warthin's tumours. For carcinomas, this score was 0.760. Metrics results are available in Table 3.

[1] https://scikit-learn.org

The best model hyperparameters are as follows: 22 as maximum depth of trees, 3 as minimum number of samples in leaves, and 43 estimators. On average, the machine learning classifier has a worse sensitivity than specificity, which means that it is less able to tell a tumour belongs to a specific category than to tell it does not. Table 4 is the corresponding confusion matrix. The approach we used relies on random forests, which are not too CPU consuming on modern computers and hence could be deployed easily. For instance, our model trains in 39.5 ± 0.2 ms and predicts in 5.6 ± 0.1 ms on a computer equipped with 32 GB of RAM and an Intel Core i7-10610U (durations were averaged over 1000 runs).

| Class | Balanced accuracy | Accuracy | Specificity | Sensitivity | OvR AUC |
|---|---|---|---|---|---|
| All classes | 0.702 | 0.720 | 0.860 | 0.720 | 0.838 |
| Pleomorphic adenomas | 0.724 | 0.840 | 0.947 | 0.500 | 0.860 |
| Warthin's tumours | 0.845 | 0.840 | 0.833 | 0.857 | 0.889 |
| Carcinomas | 0.760 | 0.760 | 0.769 | 0.750 | 0.769 |

**Table 3.** Results on test set for several metrics computed across all classes and by class. Metrics definitions are available in appendix B. Overall OvR AUC is the mean of each class score.

| | | Predicted diagnosis | | |
|---|---|---|---|---|
| | | Pleomorphic adenomas | Warthin's tumours | Carcinomas |
| Ground truth | Pleomorphic adenomas | 3 | 1 | 2 |
| | Warthin's tumours | 0 | 6 | 1 |
| | Carcinomas | 1 | 2 | 9 |

**Table 4.** Confusion matrix of test set: histopathological subtype classification.

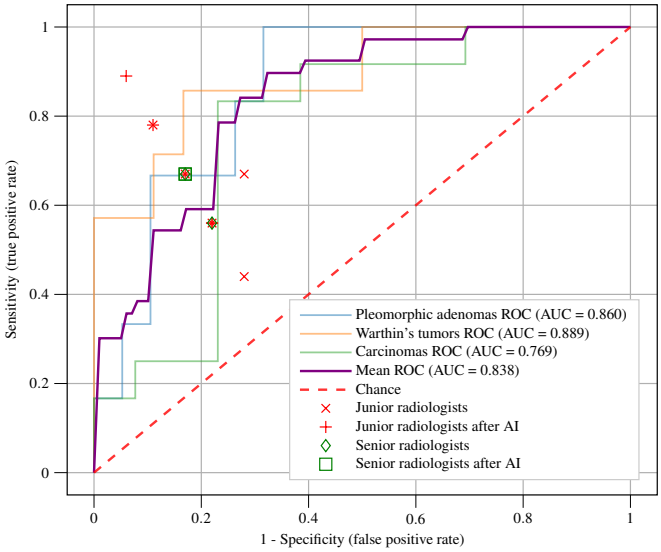


**Figure 3.** OvR ROC curves calculated from the algorithm predictions for each histopathological subtype class represented in the experiment, along with points corresponding to the metrics of the 9 radiologists. Some points are coinciding, resulting in only 6 of them being distinguishable on the graph (their corresponding overlapping symbols are displayed).

**Selected features** As previously described in section 1.3, Feature selection, we used an F-test to reduce the dimension of our feature space. This method computes p-values representing the importance of each feature in relationship to the desired output, which is the histopathological type in our case. Five relevant features were selected during the training phase, all extracted from the T2w sequence. More information can be found in Table 5. Figure 4 shows the distribution of the five selected features as box plots. We can observe that for each of these features, we can roughly discriminate the histopathological subtypes directly on the graph. This implies that those features are statistically different as per the associated class: there is a strong relationship between the selected features and the output class found by the algorithm.

| IBSI category | Feature name | F-test p-value |
|---|---|---|
| First Order | Skewness | $3.88 \times 10^{-5}$ |
| First Order | Median | $5.83 \times 10^{-5}$ |
| Gray Level Run Length Matrix | Long Run High Gray Level Emphasis | $7.21 \times 10^{-5}$ |
| First Order | Root Mean Squared | $12.07 \times 10^{-5}$ |
| Gray Level Co-occurrence Matrix | Autocorrelation | $12.32 \times 10^{-5}$ |

**Table 5.** Relevant radiomics features selected during the pre-processing step.

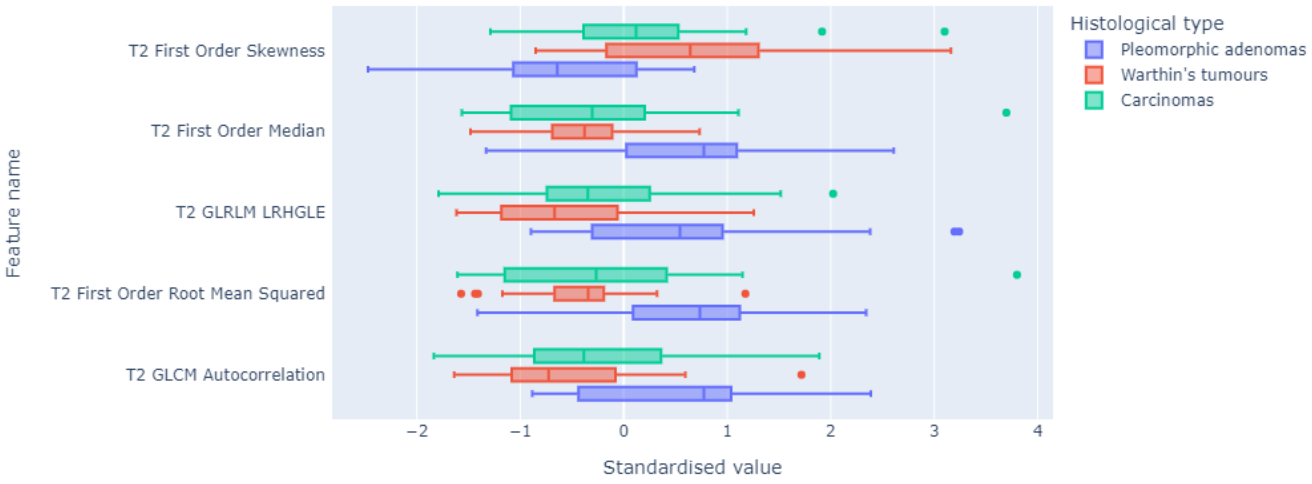


**Figure 4.** Distribution of selected radiomics features values in the training set.

### 2.2 Comparison to radiologists

**Radiologists performances** To assess the performances of the radiologists, we resort to the same metrics as described in section 1.3, Model training, allowing us to compare their results to the ones of the machine learning model. As we did not ask radiologists to evaluate their diagnoses with a probability estimate, AUC could not be computed. Table 6 and Figure 3 feature the results of the experiment. It is interesting to observe that senior radiologists seem to have worse performances than both the algorithm and the juniors. This can be explained by the fact that seniors use a lot more information than just four MRI sequences to make their diagnosis, like apparent diffusion coefficient (ADC) or patient symptoms. Radiomics cannot be extracted from such data, hence we decided not to use them in our study, since its purpose is to assess efficiency of those image statistics for tumour classification. Furthermore, experienced radiologists are used to see a wide variety of different tumours every day and thus have much more doubts than juniors that make use of their knowledge acquired in a lesson given just a few minutes before the experiment. When given access to complete patient files including other information than the four MRI sequences used for this study, radiologists' assessment of tumours is almost perfectly accurate.

**Impact of the algorithm on metrics** From Table 6, we can see that average performance of both groups improves similarly with the help of the algorithm: 8 % for balanced accuracy, 6 % for accuracy, 3 % for specificity and 6 % for sensitivity.

| Candidate | Balanced accuracy | Accuracy | Specificity | Sensitivity |
|---|---|---|---|---|
| Algorithm | 0.806 | 0.778 | 0.889 | 0.778 |
| J1 | 0.833 / 0.917 (+ 0.083) | 0.778 / 0.889 (+ 0.11) | 0.889 / 0.944 (+ 0.05) | 0.778 / 0.889 (+ 0.11) |
| J2 | 0.722 / 0.722 (+ 0.000) | 0.667 / 0.667 (+ 0.00) | 0.833 / 0.833 (+ 0.00) | 0.667 / 0.667 (+ 0.00) |
| J3 | 0.556 / 0.722 (+ 0.166) | 0.667 / 0.778 (+ 0.11) | 0.833 / 0.889 (+ 0.06) | 0.667 / 0.778 (+ 0.11) |
| J4 | 0.389 / 0.556 (+ 0.167) | 0.444 / 0.556 (+ 0.12) | 0.722 / 0.778 (+ 0.06) | 0.444 / 0.556 (+ 0.12) |
| J5 | 0.722 / 0.889 (+ 0.167) | 0.778 / 0.889 (+ 0.11) | 0.889 / 0.944 (+ 0.05) | 0.778 / 0.889 (+ 0.11) |
| J6 | 0.722 / 0.722 (+ 0.000) | 0.667 / 0.667 (+ 0.00) | 0.833 / 0.833 (+ 0.00) | 0.667 / 0.667 (+ 0.00) |
| J7 | 0.556 / 0.556 (+ 0.000) | 0.556 / 0.556 (+ 0.00) | 0.778 / 0.778 (+ 0.00) | 0.556 / 0.556 (+ 0.00) |
| **Mean** | **0.643 / 0.726 (+ 0.083)** | **0.651 / 0.714 (+ 0.063)** | **0.825 / 0.857 (+ 0.032)** | **0.651 / 0.714 (+ 0.063)** |
| S1 | 0.556 / 0.722 (0.166) | 0.556 / 0.667 (+ 0.111) | 0.778 / 0.833 (+ 0.055) | 0.556 / 0.667 (+ 0.111) |
| S2 | 0.639 / 0.639 (+ 0.000) | 0.667 / 0.667 (+ 0.000) | 0.833 / 0.833 (+ 0.000) | 0.667 / 0.667 (+ 0.000) |
| **Mean** | **0.597 / 0.681 (+ 0.084)** | **0.611 / 0.667 (+ 0.056)** | **0.806 / 0.833 (+ 0.027)** | **0.611 / 0.667 (+ 0.056)** |

**Table 6.** Performance metrics of junior and senior radiologists on 9 sub-sampled patients for the histopathological subtype classification task, first without, then with the help of our machine learning algorithm (difference between brackets). Results are given for each participant to the experiment. J1 to J7 are junior radiologists, S1 and S2 are seniors.

Table 7 shows results of the same experiment but class-wise. In this paragraph, we will consider balanced accuracy as reference metrics. On the subset, our algorithm performs best on Warthin's tumours, which is consistent with the results on the test set. Juniors benefit from the machine learning model for all histopathological subtypes. A sensitivity gain of 21.5 % is obtained on Warthin's tumours diagnosis when using the algorithm, which results in a nearly 12 % increase of corresponding balanced accuracy. Seniors profit from a rise of specificity for both Warthin's tumours and carcinomas (25 % and 10 % respectively) when trusting the model. However, they decided not to follow the machine's predictions concerning pleomorphic adenomas, resulting in a stagnancy of corresponding metrics.

**Limited model influence on radiologists** Predictions were followed by juniors in only 8 % of cases, but when they did, the algorithm was correct in 80 % of the cases. Moreover, when juniors kept their diagnosis despite the model estimations, they were mistaken 92 % of the time. As for senior radiologists, they only trusted the algorithm when they were hesitating between two diagnoses, and in 85 % of those cases, the decision was right. When seniors were mistaken in their diagnosis and the model suggested changing their mind, they never did it.

## 3. Discussion

### 3.1 Interpretations

The results confirm that an appropriate machine learning algorithm can be used to discriminate parotid gland tumours by means of basic radiomic features extraction from MR images. The deployed random forest model reaches a 0.838 AUC, a 0.860 specificity, and a 0.720 sensitivity over all considered classes on the test set. Warthin's tumours are however easier to classify than other types of tumours, as shown in Table 3.

| Histopathological subtype | Balanced accuracy | Accuracy | Specificity | Sensitivity |
|---|---|---|---|---|
| | | Algorithm | | |
| Pleomorphic adenomas | 0.834 | 0.889 | 1.000 | 0.667 |
| Warthin's tumours | 0.929 | 0.889 | 0.857 | 1.000 |
| Carcinomas | 0.775 | 0.778 | 0.800 | 0.750 |
| | | Juniors | | |
| Pleomorphic adenomas | 0.750 / 0.774 (+ 0.024) | 0.762 / 0.794 (+ 0.032) | 0.786 / 0.833 (+ 0.047) | 0.714 / 0.714 (+ 0.000) |
| Warthin's tumours | 0.745 / 0.863 (+ 0.118) | 0.841 / 0.905 (+ 0.064) | 0.918 / 0.939 (+ 0.021) | 0.571 / 0.786 (+ 0.215) |
| Carcinomas | 0.693 / 0.725 (+ 0.032) | 0.698 / 0.730 (+ 0.032) | 0.743 / 0.771 (+ 0.028) | 0.643 / 0.679 (+ 0.036) |
| | | Seniors | | |
| Pleomorphic adenomas | 0.709 / 0.709 (+ 0.000) | 0.722 / 0.722 (+ 0.000) | 0.750 / 0.750 (+ 0.000) | 0.667 / 0.667 (+ 0.000) |
| Warthin's tumours | 0.750 / 0.875 (+ 0.125) | 0.889 / 0.944 (+ 0.055) | 1.000 / 1.000 (+ 0.000) | 0.500 / 0.750 (+ 0.250) |
| Carcinomas | 0.613 / 0.663 (+ 0.050) | 0.611 / 0.667 (+ 0.056) | 0.600 / 0.700 (+ 0.100) | 0.625 / 0.625 (+ 0.000) |

**Table 7.** Performance metrics of junior and senior radiologists on 9 sub-sampled patients for the histopathological subtype classification task, first without, then with the help of our machine learning algorithm (difference between brackets). Results are given for each considered class.

The model proposed in our work rely on classical machine learning techniques, namely F-test feature selection and random forests. Besides, radiomics can be quickly computed with modern software. The combination of those characteristics allows for a fast, memory efficient diagnosis workflow that could be easily deployed on any computer while ensuring satisfactory predictive performances.

When senior radiologists are exposed to predictions made by the algorithm and have the chance to change their decision, their diagnosis accuracy ameliorates by 5.6 % on average. Accuracy is marked with an improvement of 6.3 % for junior radiologists in the same conditions. Tables 6 and 7 show that seeing the output of our model indeed helps radiologists, as it never mislead them in our experiment: when they followed the predictions of our model, the diagnosis accuracy improves. This may indicate a high degree of confidence in our algorithm from the radiologists who took part in the experiment.

Concerning the input data, most studies on parotid glands imaged with MR devices rely on T1w and T2w sequences with sometimes alterations (e.g., contrast enhancement or fat saturation) [13, 14, 29, 30]. Nevertheless, the use of DWI is rather rare, but studies such as [31] demonstrated that it could be useful for parotid gland tumours characterisation. Xia et al. in [30] observed that increasing the number of sampled sequences might output better results. Four sequences (T1w, T1ce, T2w and DWI) are given as input to our system. However, all the features automatically selected to compute the predictions of the proposed algorithm are extracted from T2w only. This nuances the statements made in those aforementioned works. It seems that the more different sequences we use, the more possible optimisation paths there are, though all sequences might not be useful to an algorithm to make reliable predictions.

Some works on parotid tumours [13, 14, 29, 30, 32] propose to gather all malignant tumours subtypes into a single class. Nonetheless, as medical treatment differs largely between the different types of malignant tumours we discarded, it was then more practical for our centre's radiologists to get precise insights about histopathological subtypes rather than broader malignity assessment.

Concerning the classification task, we make use of radiomic features, whose main role is to sum up multiscale information contained in images. Radiomics can be extracted from 2-dimension surfaces or from 3-dimension volumes. Here, we chose to compute the features on MR image slices with the largest tumour area. The use of 2D radiomics, while it might not be as efficient as 3D features [10], spares doctors a very time-consuming task as segmentation would have to be carried out on each tumour slice otherwise. As an example, the segmentation and radiomics computation from the largest 2-dimension surface of a 6,038 $mm^3$ tumour on all 4 considered sequences took us around 7 minutes, as opposed to 12 minutes using a 3D volume. Moreover, further feature analysis and selection can be manually performed as in [13, 14, 33]. Here, we preferred to embed these steps in our algorithm. This makes iterative fine-tuning of the process possible before inputting the features to our machine learning model. Those procedures can however be avoided using deep learning methods that directly extract important information from the MR images, at the cost of a loss of interpretability [29, 30, 32]. Authors of [14, 32] observed that Warthin tumours are easier to discriminate than other histopathological subtypes. Table 3 illustrates the same findings, as our machine learning model gets higher accuracy for this kind of affliction. Junior and senior radiologists as well identify Warthin's tumours more accurately than other types (see Table 7) thanks to a high specificity.

Matsuo et al. [29] proposed to gather the information of both T1w and T2w into a single image considering them as two

colour channels, thus creating pseudo-coloured pictures. This method allows to quickly compare signal intensities which is certainly a good take since Warthin's tumours are hyper-intense on T1w [34], while pleomorphic adenomas are hypo-intense on T1w and hyper-intense on T2w [35]. However, the whole process needs co-registered MR images, which were not available at the time of writing.

Regarding the trust placed in artificial intelligence (AI) by radiologists, our study is in line with previous findings. A survey conducted by the European Society of Radiology (ESR) in 2018 [36] suggested that radiologists tend to trust AI. They see it as a way to save time, allowing a significant improvement of interaction with their patients or other clinicians. As explained in section 2.2, Comparison to radiologists, junior and senior radiologists participating in our experiment followed the predictions of our algorithm in the cases where they were more uncertain, which shows confidence in the AI-based approach. Nevertheless, according to the same research work from ESR, around a half of doctors think that patients do not consider AI reliable enough to make a diagnosis without confirmation from a physician. We then believe that junior radiologists would largely benefit from the use of machine learning during their training period. Indeed, it could help them to learn new ways of making their diagnosis, for instance focusing on previously overlooked patterns in MR images, while in the end improving their overall accuracy. This would as well constitute a good way to get used to these new techniques as part of their future job. Moreover, radiologists in general could employ such algorithms as a complement to their daily diagnosis routine.

### 3.2 Limitations

In the following, we summarise the limitations of this work, highlighting ideas for future research.

First, the cohort enrolled in our research is not large enough to acquire a substantial amount of MR images. While we aspire to reduce the number of patients even more as it would mean that fewer persons are suffering from parotid tumours, insufficient data usually results in high variability and low generalisation capacities.

We could have gathered all malignant lesions in a single class, including carcinomas, lymphomas and other types we did not consider and preferred to exclude from analysis. Doing so would have included 23 images of malignant tissues, which might have increased the discrimination performances of our algorithm.

In order to characterise the sampled MR images and sum up the information contained within it, we used of radiomic features. However, the robustness of this method may be dependent on the application, as stated in [26]. This property is crucial to ensure stability of the results, but there is no consensus regarding how to assess it [37]. Thus, we cannot guarantee the prospective abilities of our algorithm.

Also, it is rather unexpected that all features automatically selected with our approach come from T2w, while radiologists tend to prefer T1ce to make their diagnosis. Gabelloni et al. [14] however got appealing results using only T2w radiomics.

Eventually, we provided an experiment to assess the impact of our algorithm on radiologists' decisions. Nonetheless, it should be noted that since we compare a machine to humans, their respective training datasets are different. Indeed, experience of the physicians can be considered as a form of pre-training, but it conveys a bias as they consequently know about the frequency of occurrence of each histopathological subtype. Moreover, radiomics given as input to our machine learning model do not encompass the same information as a radiologist's eyes. For instance, the algorithm has no data about the spatial location of the tumour within the head (e.g., deep lobe of the parotid gland, contact with surrounding nerves, etc.). Last but not least, we enrolled nine radiologists in our experiment, which is surely not enough to draw definitive conclusions or perform meaningful statistical tests.

### 3.3 Perspectives

Based upon the experience acquired while carrying out this research work, we would like to advocate some tactics to conduct future experiments on parotid gland tumours and MR radiomics.

First, radiomics normalisation techniques applied to machine learning should be further studied. Here, we tried to employ an MRI-adjusted ComBat [18], but it yielded poor standardisation results. This method should actually be adapted to suit machine learning tasks, i.e., modified so that it could be calibrated on a training set before being applied to test data. Moreover, as previously observed in [10], one should consider 3D volumes segmentation as direction of tumour growth is an indication of its histopathological subtype. Correlation is high between radiomics characteristics extracted from those volumes, but redundancy can be diminished thanks to intra-class correlation (ICC) and concordance index computations as proposed in [38]. Regarding the experiment performed with radiologists, new investigations could take advantage of a larger number of physicians involved, as well as the implementation of a method to evaluate the level of confidence into their own diagnosis. This kind of assessment could easily be interpreted as class probabilities in order to assess doubts and compute new metrics. In this work, radiologists had no input on the probability of the algorithm's predictions. Uncertainty quantification would surely help professionals to assess the quality of the AI-based diagnosis.

## Conclusion

In this research paper, we propose an easy workflow to predict the histopathological subtype of parotid gland tumours using radiomics. Thanks to the analysis of the results gave by the algorithm, we strongly believe that our method could be used as a mean of training junior radiologists. Moreover, the procedure we offer may also help physicians settle doubts when diagnosing parotid tumours.

## Author contributions

Conceptualisation, S.A., A.Q. and E.C.; methodology, A.Q., V.E. and E.C.; data collection, S.A., F.B., G.G., D.H.; data curation, S.A. and A.Q.; software, A.Q.; supervision, E.C., N.L. and C.B.; formal analysis, A.Q., V.E. and E.C.; writing – original draft, A.Q.; writing – review & editing, A.Q., V.E. and E.C.

## Acknowledgements

The authors would like to thank the patients who gave their agreement to the use of their MR images in the scope of this study. E.C. and A.Q. received funding from the European Research Council Starting Grant MAJORIS ERC-2019-STG850925. The work of V.E. is supported by ARL/ARO under grant W911NF-22-1-0235.

## References

[1] Marco Guzzo et al. “Major and Minor Salivary Gland Tumors”. In: *Critical Reviews in Oncology/Hematology* 74.2 (May 2010), pp. 134–148. ISSN: 10408428. DOI: 10.1016/j.critrevonc.2009.10.004.

[2] Ronald H. Spiro. “Salivary Neoplasms: Overview of a 35-Year Experience with 2,807 Patients”. In: *Head & Neck Surgery* 8.3 (Jan. 1986), pp. 177–184. ISSN: 01486403, 19302398. DOI: 10.1002/hed.2890080309.

[3] Sam T. H. Reerds et al. “Accuracy of Parotid Gland FNA Cytology and Reliability of the Milan System for Reporting Salivary Gland Cytopathology in Clinical Practice”. In: *Cancer Cytopathology* 129.9 (Sept. 2021), pp. 719–728. ISSN: 1934-662X, 1934-6638. DOI: 10.1002/cncy.22435.

[4] Inês Correia-Sá et al. “Fine-Needle Aspiration Cytology (FNAC): Is It Useful in Preoperative Diagnosis of Parotid Gland Lesions?” In: *Acta Chirurgica Belgica* 117.2 (Mar. 2017), pp. 110–114. ISSN: 0001-5458. DOI: 10.1080/00015458.2016.1262491.

[5] T. Tartaglione et al. “Differential Diagnosis of Parotid Gland Tumours: Which Magnetic Resonance Findings Should Be Taken in Account?” In: *Acta Otorhinolaryngologica Italica* 35.5 (Oct. 2015), pp. 314–320. ISSN: 0392-100X, 1827-675X. DOI: 10.14639/0392-100X-693.

[6] El-Sayed A. El-Dahshan et al. “Computer-Aided Diagnosis of Human Brain Tumor through MRI: A Survey and a New Algorithm”. In: *Expert Systems with Applications* 41.11 (Sept. 2014), pp. 5526–5545. ISSN: 09574174. DOI: 10.1016/j.eswa.2014.01.021.

[7] Jenna Wiens and Erica S Shenoy. “Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology”. In: *Clinical Infectious Diseases* 66.1 (Jan. 2018), pp. 149–153. ISSN: 1058-4838, 1537-6591. DOI: 10.1093/cid/cix731.

[8] Philippe Lambin et al. “Radiomics: Extracting More Information from Medical Images Using Advanced Feature Analysis”. In: *European Journal of Cancer* 48.4 (Mar. 2012), pp. 441–446. ISSN: 09598049. DOI: 10.1016/j.ejca.2011.11.036.

[9] Virendra Kumar et al. “Radiomics: The Process and the Challenges”. In: *Magnetic Resonance Imaging* 30.9 (Nov. 2012), pp. 1234–1248. ISSN: 0730725X. DOI: 10.1016/j.mri.2012.06.010.

[10] Rafael Ortiz-Ramón et al. “Classifying brain metastases by their primary site of origin using a radiomics approach based on texture analysis: a feasibility study”. In: *European Radiology*. Vol. 28. Springer-Verlag, Nov. 2018, pp. 4514–4523. DOI: 10.1007/s00330-018-5463-6.

[11] Ji Zhang et al. “Differentiating the Pathological Subtypes of Primary Lung Cancer for Patients with Brain Metastases Based on Radiomics Features from Brain CT Images”. In: *European Radiology* 31.2 (Feb. 2021), pp. 1022–1028. ISSN: 1432-1084. DOI: 10.1007/s00330-020-07183-z.

[12] Organisation mondiale de la santé and Centre international de recherche sur le cancer, eds. *WHO Classification of Head and Neck Tumours*. 4th ed. World Health Organization Classification of Tumours 9. Lyon: International agency for research on cancer, 2017. ISBN: 978-92-832-2438-9.

[13] Ying-mei Zheng et al. "MRI-Based Radiomics Nomogram for Differentiation of Benign and Malignant Lesions of the Parotid Gland". In: *European Radiology* (Nov. 2020). ISSN: 1432-1084. DOI: 10.1007/s00330-020-07483-4.

[14] Michela Gabelloni et al. "Can Magnetic Resonance Radiomics Analysis Discriminate Parotid Gland Tumors? A Pilot Study". In: *Diagnostics* 10.11 (Nov. 2020), p. 900. DOI: 10.3390/diagnostics10110900.

[15] Russell T. Shinohara et al. "Statistical Normalization Techniques for Magnetic Resonance Imaging". In: *NeuroImage: Clinical* 6 (Jan. 2014), pp. 9–19. ISSN: 2213-1582. DOI: 10.1016/j.nicl.2014.08.008.

[16] László G. Nyúl and Jayaram K. Udupa. "On Standardizing the MR Image Intensity Scale". In: *Magnetic Resonance in Medicine* 42.6 (1999), pp. 1072–1081. ISSN: 1522-2594. DOI: 10.1002/(SICI)1522-2594(199912)42:6<1072::AID-MRM11>3.0.CO;2-M.

[17] Alex Zwanenburg et al. "The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping". In: *Radiology* 295.2 (Mar. 2020), pp. 328–338. ISSN: 0033-8419. DOI: 10.1148/radiol.2020191145.

[18] Fanny Orlhac et al. "How Can We Combat Multicenter Variability in MR Radiomics? Validation of a Correction Procedure". In: *European Radiology* (Sept. 2020). ISSN: 1432-1084. DOI: 10.1007/s00330-020-07284-9.

[19] Joanne C. Beer et al. "Longitudinal ComBat: A Method for Harmonizing Longitudinal Multi-Scanner Imaging Data". In: *NeuroImage* 220 (Oct. 2020), p. 117129. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2020.117129.

[20] Rich Caruana, Nikos Karampatziakis, and Ainur Yessenalina. "An Empirical Evaluation of Supervised Learning in High Dimensions". In: *Proceedings of the 25th International Conference on Machine Learning*. ICML '08. New York, NY, USA: Association for Computing Machinery, July 2008, pp. 96–103. ISBN: 978-1-60558-205-4. DOI: 10.1145/1390156.1390169.

[21] Manuel Fernández-Delgado et al. "Do We Need Hundreds of Classifiers to Solve Real World Classification Problems?" In: *The Journal of Machine Learning Research* 15.1 (Jan. 2014), pp. 3133–3181. ISSN: 1532-4435.

[22] Helge C. Kniep et al. "Radiomics of Brain MRI: Utility in Prediction of Metastatic Tumor Type". In: *Radiology* 290.2 (Dec. 2018), pp. 479–487. ISSN: 0033-8419. DOI: 10.1148/radiol.2018180946.

[23] Hexiang Wang et al. "Preoperative MRI-Based Radiomic Machine-Learning Nomogram May Accurately Distinguish Between Benign and Malignant Soft-Tissue Lesions: A Two-Center Study". In: *Journal of Magnetic Resonance Imaging* 52.3 (2020), pp. 873–882. ISSN: 1522-2586. DOI: 10.1002/jmri.27111.

[24] Leo Breiman. "Random Forests". In: *Machine Learning* 45.1 (Oct. 2001), pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324.

[25] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Second. Springer Series in Statistics. New York: Springer-Verlag, 2009. ISBN: 978-0-387-84857-0. DOI: 10.1007/978-0-387-84858-7.

[26] Philippe Lambin et al. "Radiomics: The Bridge between Medical Imaging and Personalized Medicine". In: *Nature Reviews Clinical Oncology* 14.12 (Dec. 2017), pp. 749–762. ISSN: 1759-4782. DOI: 10.1038/nrclinonc.2017.141.

[27] Max Kuhn and Kjell Johnson. "An Introduction to Feature Selection". In: *Applied Predictive Modeling*. New York, NY: Springer New York, 2013, pp. 487–519. ISBN: 978-1-4614-6848-6 978-1-4614-6849-3. DOI: 10.1007/978-1-4614-6849-3_19.

[28] Max Kuhn and Kjell Johnson. *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC Press, July 2019. ISBN: 978-1-351-60947-0.

[29] Hidetoshi Matsuo et al. "Diagnostic Accuracy of Deep-Learning with Anomaly Detection for a Small Amount of Imbalanced Data: Discriminating Malignant Parotid Tumors in MRI". In: *Scientific Reports* 10.1 (Dec. 2020), p. 19388. ISSN: 2045-2322. DOI: 10.1038/s41598-020-76389-4.

[30] Xianwu Xia et al. "Deep Learning for Differentiating Benign From Malignant Parotid Lesions on MR Images". In: *Frontiers in Oncology* 11 (June 2021), p. 632104. ISSN: 2234-943X. DOI: 10.3389/fonc.2021.632104.

[31] Hidetake Yabuuchi et al. "Characterization of Parotid Gland Tumors: Added Value of Permeability MR Imaging to DWI and DCE-MRI". In: *European Radiology* 30.12 (Dec. 2020), pp. 6402–6412. ISSN: 0938-7994, 1432-1084. DOI: 10.1007/s00330-020-07004-3.

[32] Yi-Ju Chang et al. "Classification of Parotid Gland Tumors by Using Multimodal MRI and Deep Learning". In: *NMR in Biomedicine* 34.1 (Jan. 2021). ISSN: 0952-3480, 1099-1492. DOI: 10.1002/nbm.4408.

[33] Chih-Wei Wang et al. "JOURNAL CLUB: The Warthin Tumor Score: A Simple and Reliable Method to Distinguish Warthin Tumors From Pleomorphic Adenomas and Carcinomas". In: *American Journal of Roentgenology* 210.6 (June 2018), pp. 1330–1337. ISSN: 0361-803X, 1546-3141. DOI: 10.2214/AJR.17.18492.

[34] Mitsuaki Ikeda et al. "Warthin Tumor of the Parotid Gland: Diagnostic Value of MR Imaging with Histopathologic Correlation". In: *AJNR. American journal of neuroradiology* 25.7 (Aug. 2004), pp. 1256–1262. ISSN: 0195-6108.

[35] Mika Okahara et al. "Parotid Tumors: MR Imaging with Pathological Correlation". In: *European Radiology* 13.S06 (Dec. 2003), pp. L25–L33. ISSN: 0938-7994, 1432-1084. DOI: 10.1007/s00330-003-1999-0.

[36] European Society of Radiology (ESR). "Impact of Artificial Intelligence on Radiology: A EuroAIM Survey among Members of the European Society of Radiology". In: *Insights into Imaging* 10.1 (Dec. 2019), p. 105. ISSN: 1869-4101. DOI: 10.1186/s13244-019-0798-3.

[37] Alex Zwanenburg et al. "Assessing Robustness of Radiomic Features by Image Perturbation". In: *Scientific Reports* 9.1 (Dec. 2019), p. 614. ISSN: 2045-2322. DOI: 10.1038/s41598-018-36938-4.

[38] Renee Cattell, Shenglan Chen, and Chuan Huang. "Robustness of Radiomic Features in Magnetic Resonance Imaging: Review and a Phantom Study". In: *Visual Computing for Industry, Biomedicine, and Art* 2.1 (Dec. 2019), p. 19. ISSN: 2524-4442. DOI: 10.1186/s42492-019-0025-6.

## A. Extracted radiomic features

| IBSI category | Features | Number of features |
|---|---|---|
| Shape | Volume, Surface area, Surface area to volume ratio, Sphericity, Compactness 1, Compactness 2, Spherical disproportion, Maximum 3D diameter, Maximum 2D diameter slice, Maximum 2D diameter column, Maximum 2D diameter row, Major axis, Minor axis, Least axis, Elongation, Flatness | 16 |
| First order | Energy, Total energy, Entropy, Minimum, 10th percentile, 90th percentile, Maximum, Mean, Median, Interquartile range, Range, Mean absolute deviation, Robust mean absolute deviation, Root mean squared, Standard deviation, Skewness, Kurtosis, Variance, Uniformity | 19 |
| GLCM | Autocorrelation, Joint average, Cluster prominence, Cluster shade, Cluster tendency, Contrast, Correlation, Difference Average, Difference entropy, Difference variance, Joint energy, Joint entropy, Informal measure of correlation 1, Informal measure of correlation 2, Inverse difference moment, Inverse difference moment normalized, Inverse difference, Inverse difference normalized, Inverse variance, Maximum probability, Sum average, Sum entropy, Sum of squares | 23 |
| GLRLM | Short run emphasis, Long run emphasis, Gray level non uniformity, Gray level non uniformity normalized, Run length non uniformity, Run length non uniformity normalized, Run percentage, Gray level variance, Run variance, Run entropy, Low gray level run emphasis, High gray level run emphasis, Short run low gray level emphasis, Short run high gray level emphasis, Long run low gray level emphasis, Long run high gray level emphasis | 16 |
| GLSZM | Small area emphasis, Large area emphasis, Gray level non uniformity, Gray level non uniformity normalized, Size zone non uniformity, Size zone non uniformity normalized, Zone percentage, Gray level variance, Zone variance, Zone entropy, Low gray level zone emphasis, Small area low gray level emphasis, Small area high gray level emphasis, Large area low gray level emphasis, Large area high gray level emphasis | 15 |
| NGTDM | Coarseness, Contrast, Busyness, Complexity, Strength | 5 |
| GLDM | Small dependence emphasis, Large dependence emphasis, Gray level non uniformity, Dependence non uniformity, Dependence non uniformity normalized, Gray level variance, Dependence variance, Dependence entropy, Low gray level emphasis, High gray level emphasis, Small dependence low gray level emphasis, Small dependence high gray level emphasis, Large dependence low gray level emphasis, Large dependence high gray level emphasis | 14 |

## B. Metrics definitions

Let TP, FP, TN and FN denote the true/false positives and true/false negatives.

- **Accuracy:** Defined as the ratio of samples that are correctly identified, no matter the class.

$$\frac{TP+TN}{TP+FP+TN+FN}$$

- **Specificity:** Also called true negative rate. Defined as the ratio of negatives that are correctly identified.

$$\frac{TN}{TN+FP}$$

  When not in used in a binary or OvR context, specificity is computed with overall total TN and FP (i.e., across all classes).

- **Sensitivity:** Also called true positive rate. Defined as the ratio of positives that are correctly identified.

$$\frac{TP}{TP+FN}$$

  When not in used in a binary or OvR context, sensitivity is computed with overall total TP and FN (i.e., across all classes).

- **Balanced accuracy:** Defined as the mean of sensitivity and specificity.

$$\frac{1}{2}\left(\frac{TP}{TP+FN}+\frac{TN}{TN+FP}\right)$$

  When not in used in a binary or OvR context, balanced accuracy is the average of sensitivity for each class.

- **Area under ROC curve (AUC):** The receiver operating characteristic (ROC) curve plots the TP rate against the FP rate. AUC corresponds to the area between this curve and the x-axis. An AUC greater than 0.5 means that the algorithm predicts better than random. An AUC of 1 maximizes both specificity and sensitivity.