UNIVERSITÄT AUGSBURG
Universitätsbibliothek

OPEN ACCESS
OPUS AUGSBURG

Multimodal classification of parotid tumors –a proof of concept

Varvara Kondratyeva, Zoe Reinke, Marija Blažyte, Julian V. Karg, Johannes Zenk, Thomas Wendler

Angaben zur Veröffentlichung / Publication details:

Kondratyeva, Varvara, Zoe Reinke, Marija Blažyte, Julian V. Karg, Johannes Zenk, and Thomas Wendler. 2026. "Multimodal classification of parotid tumors –a proof of concept." *Current Directions in Biomedical Engineering* 12 (1): 259–62. https://doi.org/10.1515/cdbme-2026-0166.

Nutzungsbedingungen / Terms of use:

CC BY 4.0

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under these conditions:
CC-BY 4.0: Creative Commons: Namensnennung
Weitere Informationen finden Sie unter: / For more information see:
https://creativecommons.org/licenses/by/4.0/deed.de

Varvara Kondratyeva, Zoe Reinke*, Marija Blažytė, Julian Valentin Karg, Johannes Zenk, and Thomas Wendler

Multimodal Classification of Parotid Tumors – A Proof of Concept

https://doi.org/10.1515/cdbme-2026-0166

Abstract: Preoperative differentiation between benign and malignant parotid tumors is challenging but can influence treatment and outcome. In this proof-of-concept study, we investigated multimodal AI-based classification combining clinical tabular data and ultrasound images, comparing unimodal and multimodal approaches for binary benign–malignant classification. The study included 594 patients with parotid tumors (449 benign, 145 malignant) after preprocessing. The image-only classifier achieved 0.73 accuracy at image level and 0.75 after patient-level aggregation. The tabular model reached 0.88 accuracy, while the multimodal model maintained this accuracy with a slightly improved weighted F1-score (0.89) and malignant recall of 0.93. These results suggest that image-derived features provided complementary information, primarily improving sensitivity for malignant tumors.

Keywords: parotid tumors, image classification, multimodal data analysis

1 Introduction

Preoperative differentiation between benign and malignant parotid tumors remains challenging, as the criteria used for classification are often ambiguous. Accurate preoperative assessment can improve patient outcomes by avoiding unnecessary surgery, guiding treatment planning and informing patient counseling.

In routine practice, this assessment relies on clinical examination, ultrasound, and, when needed, cytology and cross-sectional imaging [1]. In this work, we investigate whether combining clinical tabular and ultrasound imaging data can improve automated benign–malignant classification.

Recent artificial intelligence studies on parotid tumor differentiation have mostly focused on single imaging modalities, including T2-weighted MRI [4], contrast-enhanced CT [6], and ultrasound-based models, incorporating selected clinical variables [8, 10]. However, explicit fusion of routine clinical tabular data with ultrasound for benign–malignant classification has not been studied. An existing multimodal approach with ultrasound has addressed differentiation between benign subtypes rather than malignancy detection [2]. This gap motivates the present study.

2 Materials and Methods

2.1 Data

A single-center 658 patient retrospective study was conducted at the University Hospital Augsburg. Patient data routinely collected for tumor type assessment was included alongside ultrasound images acquired in B-mode, Doppler, and elastography. As B-mode was the only modality available for all patients, subsequent analyses were restricted to those images.

The tabular dataset comprises clinical data with histopathologically confirmed benign or malignant parotid tumors, distributed across the following subclasses: **benign** (pleomorphic adenoma (PLA), cystadenolymphoma (Warthin tumor), other) and **malignant** (mucoepidermoid carcinoma (MUC), malignant lymphoma (LYM), other). It also includes the pre-operative binary malignancy prediction by the treating physician.

The ultrasound dataset comprised 658 patients, with images acquired as both standalone frames and video sequences (1,071 videos in total), corresponding to 5,694 benign and 2,037 malignant unprocessed images (Tab. 1). One patient with a malignant tumor had no matching tabular record and was excluded.

Varvara Kondratyeva, Thomas Wendler, Computer-Aided Medical Procedures and Augmented Reality, Technical University of Munich, Garching bei München, Germany
***Corresponding author: Zoe Reinke,** Dept. of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Augsburg, Augsburg, Germany, e-mail: zoe.reinke@uni-a.de
Varvara Kondratyeva, Julian Valentin Karg, Thomas Wendler, Dept. of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Augsburg, Augsburg, Germany
Marija Blažytė, Johannes Zenk, Dept. of Otorhinolaryngology, University Hospital Augsburg, Augsburg, Germany
Thomas Wendler, Zoe Reinke, Digital Medicine, University Hospital Augsburg, Germany
Thomas Wendler, Bavarian Center for Cancer Research (BZKF) Augsburg, Germany
Thomas Wendler, Center for Advanced Analytics and Predictive Sciences, University of Augsburg, Germany

Open Access. © 2026 The Author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 License.

Tab. 1: Data distribution between tumor subclasses, differentiated between number of patients and images before (raw) and after preprocessing (prepr.).

Tumor subclass	$N(pat.)$ raw	$N(img.)$ raw	$N(pat.)$ $prepr.$	$N(img.)$ $prepr.$
Benign	*507*	*5,694*	*449*	*3,494*
PLA	159	1,880	140	1,166
Warthin	224	2,366	196	1,358
Other	124	1,448	113	970
Malignant	*151*	*2,037*	*145*	*1,044*
MUC	5	446	5	240
LYM	19	403	19	117
Other	127	1,188	121	687

2.2 Data preprocessing

A subset of the imaging data contained non-anatomical artifacts, including text annotations, measurement overlays, and color Doppler and elastography visualizations. Preprocessing therefore comprised image cropping, removal of Doppler and elastography frames, and text/measurement suppression via the inpainting tool IOPaint [3]; images requiring inpainting of more than 10% of the area were discarded. Tabular data preprocessing included removal of incomplete rows and one duplicated patient, yielding 594 patients (449 benign, 145 malignant; Tab. 1).

The tabular dataset was split into train/validation/test sets (80%/10%/10%), stratified by tumor subclass (Tab. 2).

Tab. 2: Data distribution between tumor subclasses, differentiated between number of patients in train/validation/test splits.

Tumor subclass	$N(pat.train)$	$N(pat.val)$	$N(pat.test)$
Benign	*358*	*46*	*45*
PLA	112	14	14
Warthin	156	20	20
Other	90	12	11
Malignant	*116*	*14*	*15*
MUC	4	0	1
LYM	15	2	2
Other	97	12	12

Images were then assigned to the split of their corresponding patient. To align the datasets and prevent patient information leakage, we matched the images to corresponding patient and used train-validation-test split by patient described above, expanding to the image level afterwards. Following that, 3,494 out of 5,694 benign and 1,044 out of 2,037 malignant images were retained (Tab. 1).

2.3 Unimodal classification: tabular classifier

Our tabular classifier is based on the following features: skin tumor history, age, preoperative facial nerve paresis, smoking status, and gender. These features are typically used in a clinical setting to predict tumor entity pre-operatively.

The test set was held out for final evaluation, and train and validation sets were combined into one set to perform training using cross-validation.

We used gradient-boosted decision trees classifier (XGBClassifier [9]) as it is fast and well-established for tabular data. For hyperparameter tuning we applied randomized search with cross-validation with 3 folds (randomizedSearchCV [5]). Following parameter search, the model was fit with best parameters on the whole training/validation set.

2.4 Unimodal classification: image classifier

We used a ResNet-18 model initialized with ImageNet-pretrained weights [7]. During training weighted sampling was used to help class imbalance. Following training, the probability threshold was fine-tuned to better fit the model uncertainty. The model predicted malignant class probabilities in range of 0.44 ± 0.18 on the validation set, so the threshold was chosen to represent the best F1-score.

The training and fine-tuning of the image classifier were done using training and validation set, with test set being held out for the final evaluation only.

2.5 Multimodal classification: combined tabular and image classifier

Rather than using hard image-level predictions (benign/malignant) in multimodal classification, we incorporated statistics of the raw predicted malignancy probabilities aggregated per patient: mean, standard deviation, minimum, maximum, and total number of images. These statistics were appended to the tabular feature set and used to train a new tabular classifier.

To avoid leakage, image-derived features for the train split were generated via stratified 5-fold cross-predictions; validation and test features came from the main unimodal image classifier directly.

3 Results

3.1 Tabular classifier

Although the model achieved a good discriminative performance (Accuracy 0.88, Tab. 3), classification of malignant tumors was less robust than classification of benign tumors, with a recall of 0.80 for malignant tumors and a recall of 0.91 for benign tumors, indicating that malignant cases remained more challenging to detect reliably.

Tab. 3: Classification report on the test set for tabular classifier.

Class	Precision	Recall	F1-score	Support
Benign	0.93	0.91	0.92	45
Malignant	0.75	0.80	0.77	15
Accuracy			0.88	60
Macro avg	0.84	0.86	0.85	60
Weighted avg	0.89	0.88	0.88	60

Malignant tumor distribution is dominated by the subclass "Other" (Tab. 1), which often uses skin tumor history features for predictions in clinical settings. This can also be observed when looking at the feature relevance (Tab. 4).

Tab. 4: Features ranked by normalized gain for tabular classifier.

Feature	Normalized gain (%)
Skin tumor history	66.33
Preoperative facial nerve paresis	13.57
Smoking status	10.93
Age	6.55
Gender	2.62

3.2 Image classifier

The image-level classifier achieved an accuracy of 0.73 and weighted F1-score of 0.74, performing better on benign than malignant images. Malignant precision was notably lower (0.48), indicating a tendency toward false-positive malignant predictions (Tab. 5).

Patient-level aggregation via mean malignancy probability with a validation-tuned threshold improved overall performance slightly: accuracy increased to 0.75 and weighted F1-score to 0.76. Malignant recall improved substantially (from 0.66 to 0.80), though precision remained low. Benign precision improved at the cost of a slight recall decrease (Tab. 6).

Tab. 5: Classification report for image classifier, image-level

Class	Precision	Recall	F1-score	Support
Benign	0.87	0.75	0.81	247
Malignant	0.48	0.66	0.55	85
Accuracy			0.73	332
Macro avg	0.67	0.71	0.68	332
Weighted avg	0.77	0.73	0.74	332

Tab. 6: Classification report for patient-level aggregation of image-level predictions.

Class	Precision	Recall	F1-score	Support
Benign	0.92	0.73	0.81	45
Malignant	0.50	0.80	0.62	15
Accuracy			0.75	60
Macro avg	0.71	0.77	0.72	60
Weighted avg	0.81	0.75	0.76	60

3.3 Combined classifier

The combined model showed strong test performance, achieving an accuracy of 0.88 and a weighted F1-score of 0.89 (Tab. 7). Compared with the image-only approaches, the combined model clearly improved overall performance. Compared with the tabular classifier, accuracy remained the same, while the weighted F1-score increased slightly. Performance was excellent for benign cases, with a precision of 0.97 and recall of 0.87. For malignant cases, recall was high (0.93), while precision was lower (0.70), indicating that most malignant cases were detected, but some benign cases were incorrectly classified as malignant.

Tab. 7: Classification report on the test set for combined model.

Class	Precision	Recall	F1-score	Support
Benign	0.97	0.87	0.92	45
Malignant	0.70	0.93	0.80	15
Accuracy			0.88	60
Macro avg	0.84	0.90	0.86	60
Weighted avg	0.91	0.88	0.89	60

Feature importance analysis showed that skin tumor history remained the strongest individual predictor (33.40% normalized gain), though lower than in the unimodal tabular classifier (66.33%; Tab. 4). Aggregated image-based features collectively contributed 35.87% of normalized gain, with mean predicted probability ranking second overall, suggesting that image-derived information added relevant predictive value beyond the clinical variables (Tab. 8).

Tab. 8: Features ranked by normalized gain for combined model.

Feature	Normalized gain (%)
Skin tumor history	33.40
Mean predicted probability	18.18
Preoperative facial nerve paresis	11.23
Smoking status	9.49
Minimum predicted probability	6.37
Age (years)	6.27
Standard deviation of predicted probability	4.07
Maximum predicted probability	3.77
Gender	3.74
Number of images	3.48

Physicians correctly identified 97.7% of benign and 80.0% of malignant tumors on the test set, missing 20.0% of malignant cases. The model showed higher malignant sensitivity (93.3%) but lower benign specificity (86.7%), reflecting a shift toward fewer missed malignancies at the cost of more false-positive predictions (Tab. 9).

Tab. 9: Correct classification rates (column-normalized) for physician assessment and the combined model, on the full dataset and test set.

Prediction	Physician all	Physician test	Model test
Correctly benign	0.987	0.977	0.867
Correctly malignant	0.785	0.800	0.933

4 Discussion

Incorporating image classifier outputs slightly improved overall performance: accuracy remained at 0.88 while weighted F1-score increased from 0.88 to 0.89. The main gain was in malignant recall (0.80 to 0.93), at the cost of lower malignant precision (0.75 to 0.70). Image-based features contributed 35.87% of normalized total gain, suggesting clinically relevant information can be extracted even from suboptimal retrospective imaging when combined with clinical variables.

The combined model shifted toward higher malignant sensitivity, which is desirable in a screening setting where missing malignancies is particularly costly. However, the precision drop indicates that threshold tuning, calibration, or class-balancing strategies merit further investigation. A key limitation is the use of non-standardized clinical imaging; while inpainting mitigated artifact contamination, it may introduce information loss or bias. Future work includes fusion strategy optimization and validation on larger, higher-quality cohorts.

Author Statement

Research funding: The author state no funding involved. **Conflict of interest**: Authors state no conflict of interest. **Informed consent**: The data used was retrospectively acquired and does not include patient consent. **Ethical approval**: The local institutional review board approved this study.

References

[1] Geiger JL, Ismaila N, Beadle B, Caudell JJ, Chau N, Deschler D, et al. Management of Salivary Gland Malignancy: ASCO Guideline. J Clin Oncol 2021;39:1909–1941.

[2] He Y, Zheng B, Peng W, et al. An ultrasound-based ensemble machine learning model for the preoperative classification of pleomorphic adenoma and Warthin tumor in the parotid gland. Eur Radiol 2024;34:6862–6876.

[3] IOPaint contributors. IOPaint version 1.6.0: free, open-source and self-hostable inpainting/outpainting tool [Internet]. Available from: https://www.iopaint.com/. Accessed 10 Jun 2026.

[4] Ma Q, Ren J, Ge Y, Yuan Y, Tao X. Automated segmentation and diagnosis of parotid tumors using a combined deep learning and radiomics model on T2-weighted MRI: a multivendor validation study. Cancer Imaging 2026;26:12.

[5] scikit-learn developers. sklearn.model_selection.RandomizedSearchCV. scikit-learn 1.7.2 documentation [Internet]. Available from: https://scikit-learn.org/1.7/modules/generated/sklearn.model_selection.RandomizedSearchCV.html. Accessed 10 Jun 2026.

[6] Shen XM, Mao L, Yang ZY, Chai ZK, Sun TG, Xu YC, et al. Deep learning-assisted diagnosis of parotid gland tumors by using contrast-enhanced CT imaging. Oral Dis 2023;29:3325–3336.

[7] Torchvision contributors. torchvision.models.resnet18. Torchvision main documentation, version 0.25.0 [Internet]. Available from: https://docs.pytorch.org/vision/main/models/generated/torchvision.models.resnet18.html. Accessed 12 Jun 2026.

[8] Tu CH, Wang RT, Wang BS, Kuo CE, Wang EY, Tu CT, et al. Neural network combining with clinical ultrasonography: A new approach for classification of salivary gland tumors. Head Neck 2023;45:1885–1893.

[9] XGBoost Developers. XGBoost Documentation, release 3.1.3 [Internet]. Available from: https://xgboost.readthedocs.io/en/release_3.2.0/. Accessed 10 Jun 2026.

[10] Zhang G, Zhu L, Huang R, Xu Y, Lu X, Chen Y, et al. A deep learning model for the differential diagnosis of benign and malignant salivary gland tumors based on ultrasound imaging and clinical data. Quant Imaging Med Surg 2023;13:2989–3000.