European Archives of Oto-Rhino-Laryngology
https://doi.org/10.1007/s00405-022-07455-y

HEAD AND NECK

Deep learning model developed by multiparametric MRI in differential diagnosis of parotid gland tumors

Emrah Gunduz[1] · **Omer Faruk Alçin**[2] · **Ahmet Kizilay**[3] · **Ismail Okan Yildirim**[4]

Received: 29 April 2022 / Accepted: 16 May 2022

Abstract

Purpose To create a new artificial intelligence approach based on deep learning (DL) from multiparametric MRI in the differential diagnosis of common parotid tumors.

Methods Parotid tumors were classified using the InceptionResNetV2 DL model and majority voting approach with MRI images of 123 patients. The study was conducted in three stages. At stage I, the classification of the control, pleomorphic adenoma, Warthin tumor and malignant tumor (MT) groups was examined, and two approaches in which MRI sequences were given in combined and non-combined forms were established. At stage II, the classification of the benign tumor, MT and control groups was made. At stage III, patients with a tumor in the parotid gland and those with a healthy parotid gland were classified.

Results A stage I, the accuracy value for classification in the non-combined and combined approaches was 86.43% and 92.86%, respectively. This value at stage II and stage III was found respectively as 92.14% and 99.29%.

Conclusions The approach presented in this study classifies parotid tumors automatically and with high accuracy using DL models.

Keywords Artificial ıntelligence · Deep learning · Parotid tumors · Computer aided diagnosis · Head and neck cancer

Introduction

Salivary gland tumors constitute 3–12% of head and neck tumors, and 80% of salivary gland tumors originate from the parotid gland. 80% of parotid gland tumors (PGT) are benign tumors (BT), 20% are malignant tumors (MT), and the most common benign tumors are pleomorphic adenomas (PMA) and Whartin tumors (WT) [1]. The preoperative prediction of the differential diagnosis of parotid tumors strongly affects the surgical plan. Although the diagnostic strategy varies in PGT, Magnetic Resonance Imaging (MRI) and fine-needle aspiration cytology (FNAC) are frequently used [2, 3]

Multiparametric MRI is crucial for the differential diagnosis of PGT. T1-weighted contrast-enhanced (T1Wce) and T2-weighted (T2W) images present the properties of the textures of tumors, and diffusion-weighted imaging (DWI) shows tumor cellularity. Evaluating these MRI sequences together improves diagnostic accuracy [4]. It becomes more critical for the otolaryngologist to work with a radiologist experienced in the field of head and neck radiology to reach a diagnosis by evaluating the outputs of these imaging techniques correctly. Recently, many studies aiming to make computers behave like an experienced radiologist have been carried out by using artificial intelligence (AI) algorithms in medical image analysis [5]. Deep learning (DL), a subset of AI, has been widely used in recent years to create medical decision support systems [6]. These DL-based systems provide high-accuracy results and open-access diagnosis

This article was produced from the first author's doctoral thesis.

✉ Emrah Gunduz
emrah.gunduz.ctf@gmail.com

1 Department of Otorhinolaryngology Head and Neck Surgery, Malatya Training and Research Hospital, Malatya, Turkey

2 Department of Electric and Electronics Engineering Faculty of Engineering and Natural Sciences Malatya, Turgut Ozal University Malatya, Malatya, Turkey

3 Department of Otorhinolaryngology Head and Neck Surgery, Inonu University Faculty of Medicine, 44000 Malatya, Turkey

4 Department of Radiology, Inonu University Faculty of Medicine, Malatya, Turkey

Published online: 21 May 2022

support tools for physicians [7–9]. There has been a drastic increase in the volume of the otolaryngology literature describing novel applications of DL, while there are still exciting studies waiting to be developed in this area [10].

This study aimed to create a new non-invasive differential diagnosis method for PGT with a DL model using multiparametric MRI. The InceptionResNetv2 model was employed as a DL model. In this study, healthy parotid tissue called the control group (CG), and PMA, WT and MT, are considered in the differential diagnosis of PTG.

Materials and methods

This retrospective study was carried out at the Department of Otorhinolaryngology at the Faculty of Medicine at Inonu University with the approval of the Health Sciences Non-Interventional Clinical Studies Ethics Committee, with the decision numbered 2020/1038. The key objective of this study is to create a new AI approach based on deep learning from multiparametric MRI in the differential diagnosis of common parotid tumors.

Patient group and MRI protocol

One hundred and twenty-three patients with a PGT who underwent parotidectomy in our clinic between 2010 and 2020 were included the study. Patients who were operated for PGT and had the five aforementioned MRI sequences (T1Wce, T2W, b0, b1000, ADC) pre-operatively and a control group without any disease in the parotid gland were included. Patients whose pre-operative MR images had a quality level that was not standardized for radiological evaluation and did not contain these five sequences were excluded. Tumor subtypes that were rare and did not constitute a sufficient sample size were also excluded. Pleomorphic adenoma (PMA) and Whartin tumor (WT) cases were included in the study as benign tumors (BT). These were also the most frequently operated BT in our clinic. Since malignant tumors (MT) of the parotid gland are very rare, the MT were examined in a single group without dividing them into histopathological subtypes. However, these MT had malignancy findings, such as ill-defined borders, cystic components, necrosis, and invasion of surrounding tissues on MRI.

All MRI examinations were performed at 1.5 Tesla (T) (Magnetom AERA, Siemens Healthcare, Erlangen, Germany) including diffusion weighted sequences. The scan volumes varied according to the tumor location and included the following sequences: A coronal turbo inversion recovery sequence, an axial T1-weighted (T1W) sequence (voxel size: $0.3 \times 0.3 \times 3.0$ mm, FoV read: 220 mm, slice thickness: 3.0 mm, repetition time / echo time: 580 ms/12 ms), an axial T2W sequence ($0.6 \times 0.6 \times 3.0$ mm, 220 mm, 3.0 mm, 4000 ms/84 ms) and DW-images in the axial plane ($1.3 \times 1.3 \times 4.0$ mm, 250 mm, 4.0 mm, 3000 ms/111 ms). After contrast administration (gadolinium–DTPA, 0.1 mmol/kg), T1W contrast-enhanced fat-saturated sequences were performed in the axial plane ($0.4 \times 0.4 \times 3.0$ mm, 220 mm, 3.0 mm, 533 ms/12 ms) and in the coronal plane ($0.5 \times 0.5 \times 3.0$ mm, 300 mm, 3.0 mm, 509 ms/9.6 ms). Between three and six different of the following b values (0, 50, 100, 150, 300, 500, 800, 900 and 1000 mm^2/sec) were applied in three different orthogonal directions, thus minimizing the effects of diffusion anisotropy. ADC maps were reconstructed by DWI of the different b factor, $b = 0$ and 1000 s/mm^2 on workstation. Thus, we collected data sets containing five multiparametric MRI sequences for each patient, namely, T1Wce, T2W, b0 (DWI, $b = 0$ s/mm^2), b1000 (DWI, $b = 1000$ s/mm^2), and apparent diffusion coefficient (ADC) maps [11].

Multiparametric MR images in five different sequences were used as input to the DL model. In the PGT group, the cross sections of all MR images which contained the tumor were included in the data set to increase the number of images. In the control group, healthy parotid tissue images were obtained from two cross sections of each patient's right and left parotid glands. Standardization was achieved by selecting identical cross sections in each sequence. The MR images were converted from the DICOM format to the JPEG format to use a universal image format. Both in CG and the PGT group, the region containing the parotid gland and its surroundings were cropped from these images as multiple single slices. In this rectangular manual cropping process, the borders of the rectangle were the masseter muscle in the anterior direction, the posterior belly of the digastric muscle in the posterior direction, and the medial pterygoid muscle in the medial direction. On the lateral border, the borders slightly exceeding the skin were clipped. The images were cropped by an otolaryngology consultant (first author) and a senior head and neck surgeon (third author) under the supervision of a radiologist. This way, 3495 (699 cross sections × 5 sequences) parotid gland region images were cropped from the MR images when including all patients and controls (total, $n = 173$). An example of input data for each group is shown in Fig. 1. The cropped radiological images were applied to the deep learning model without any preprocessing.

The data sets were analyzed using the SPSS 22 (Statistical Package for the Social Sciences; SPSS Inc., Chicago, IL) package program. As descriptive statistics, the qualitative variables are expressed as frequencies and percentages (n and %), while the quantitative variables Are expressed as mean ± standard deviation values. Analysis of variance (ANOVA) was used to compare the variables between groups. $p < 0.05$ was used as the statistical significance level in the analyses.

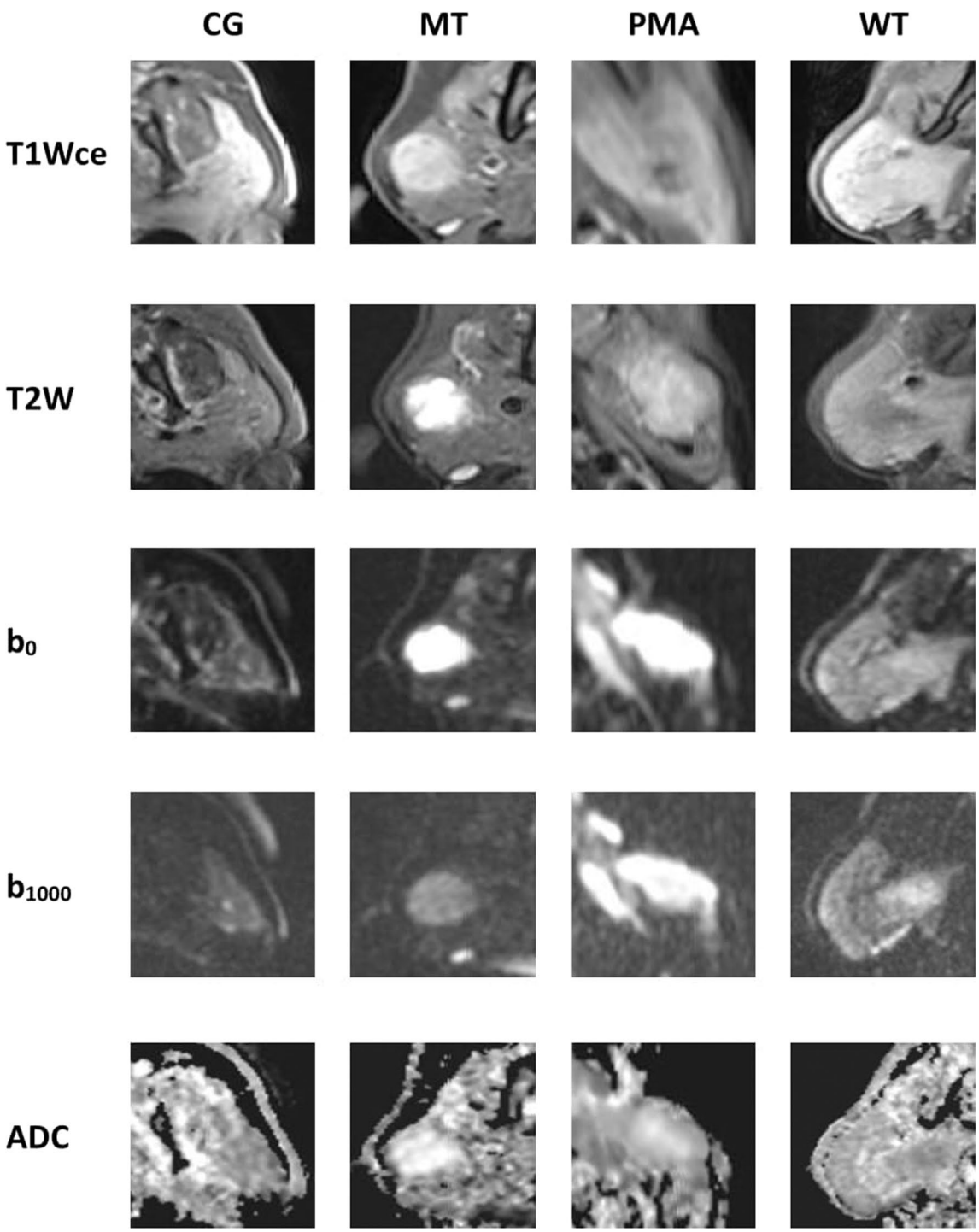

Fig. 1 Example input data set for each group. *CG* control group, *PMA* pleomorphic Adenoma, *WT* warthin tumor, *MT* malignant tumors

Deep learning, InceptionResNetv2 and conventional classifiers

Deep learning (DL) is a complicated version of machine learning. DL contains various multiple processing layers with large depths. DL can automatically learn the representations of input data thanks to multiple convolutional abstractions [12–15]. A common form of usage of DL is transfer learning. Transfer learning provides the image classification process with limited input images. In this study, the InceptionResNetv2 model was used to perform transfer learning for the PGT classification. The

InceptionResNetv2 model was trained with the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) data set. InceptionResNetv2 contains inception modules and residual connections [14, 16, 17]. The inception modules and the residual connections provide the opportunity to avoid a degradation problem of the deep structure and less training time [17, 18]. In this study, the conventional classifiers that were used included the support vector machine (SVM), k-nearest neighbor (KNN) and linear discriminant classifier (LD) approaches [18–21].

Proposed methodology

This study aimed to detect and classify PGT by using DL approaches. The InceptionResNetv2 model was used to perform transfer learning. The experimental procedures were designed in three stages: Stage I: the differential diagnosis of CG, PMA, WT and MT. Stage II: the differential diagnosis of benign tumors (PMA + WT) and malignant tumors, stage III: the differential diagnosis of CG and PGT (PMA + WT + MT).

We considered two approaches in the experiment design, the combined and non-combined approaches, in stage I. In the combined approach, we applied the five MRI sequence images together in the training and validation of the InceptionResNetv2 model. Thus, it was aimed to train the network with more input images. In the non-combined approach, each MRI sequence was considered a data set. Then, a DL model was trained for each MRI sequence, so we obtained five InceptionResNetv2 models. The input data set consisted of conventional T1Wce, T2W, diffusion-weighted b0, b1000 and ADC sequence images. The test images were evaluated for each MRI sequence and with the majority voting (MV) approach in the testing phase. MV was applied to boost the test performance of the proposed method. In this study, the MV approach was applied to the MRI sequences instead of competitive classifiers as a similar approach to a radiologist examination. MV was intended to assign a class label's highest probability in five sequence images to new test MRI data. The proposed framework and the MV process are shown for stage I with the combined approach in Fig. 2a and b respectively.

As seen in Fig. 2a that depicts stage I with the combined approach, class numbers and names should be adopted for other stages. An example can be seen in Fig. 2b, where the trained model produced three PMA results, one WT result and one MT result for the five sequences. Finally, MV assigned a PMA class to the new test data belonging a patient because of the high probability class.

The data set was randomly divided into three parts, training, validation and testing, and it was kept constant in all experimental procedures, with ratios of 70%, 10% and 20%, respectively. The InceptionResNet2 model was trained using the stochastic gradient descent optimizer method, 32 input batch size, 30 epochs, and a learning rate of 0.05 for all approaches. The deep features of InceptionResNetv2 were classified with three conventional machine learning methods as SVM, KNN and LD. The deep features were obtained from the fully connected layer of InceptionResNetv2. For comparison, the same training and testing data set was utilized in the extraction of the deep features. SVM, KNN and LD were trained with the deep features of the training data set. This procedure was applied to all stages.

Results

The proposed model was evaluated with the data set which was created with parotidectomy patients in our clinic. According to postoperative histopathological results, 46 patients had PMA, 53 patients had WT, and 24 patients had MT (PGT, $n = 123$). The subtypes of MT were high-grade mucoepidermoid carcinoma in 11 cases, acinic cell carcinoma in 5 cases, adenoid cystic carcinoma in 4 cases, squamous cell carcinoma in 2 cases and malignant lymphoma in 2 other cases. Parotid tumor subtypes and numbers according to histopathological results are shown in Table 1. Additionally, 50 patients who had a head neck MRI for other reasons and no disease in the parotid gland were included in the study as the control group (CG, $n = 50$). The PGT patients consisted of 73 men and 50 women, and their mean age was 53.9 ± 14.8 years, while the individuals in CG consisted of 31 men and 19 women, and their mean age was 56.9 ± 16.6 years.

For all stages, the numbers of images and class distributions of the training, validation and testing data are given in Table 2 according to the ratios mentioned above.

The experiments started with stage I using the combined and non-combined approaches. The test results of all classifiers for stage I with the combined and non-combined approaches are given in Table 3.

Considering the overall performance results presented in Table 3, the combined approach had the best performance metrics compared to the non-combined approach. This situation was observed during all experiments, and for simplicity, only the results of the combined experiments are included. The conventional classifiers were used only with the combined approach. The MV method increased accuracy by about 10% and 6% in the non-combined and combined approaches, respectively. The other metrics also supported the performance of the MV method. Experiments were conducted to classify the deep features with SVM, KNN and LD to show the convenience of InceptionResnetv2.

The experiments conducted in stage II included the control group (CG), benign tumor (BT) (PMA + WT) group and

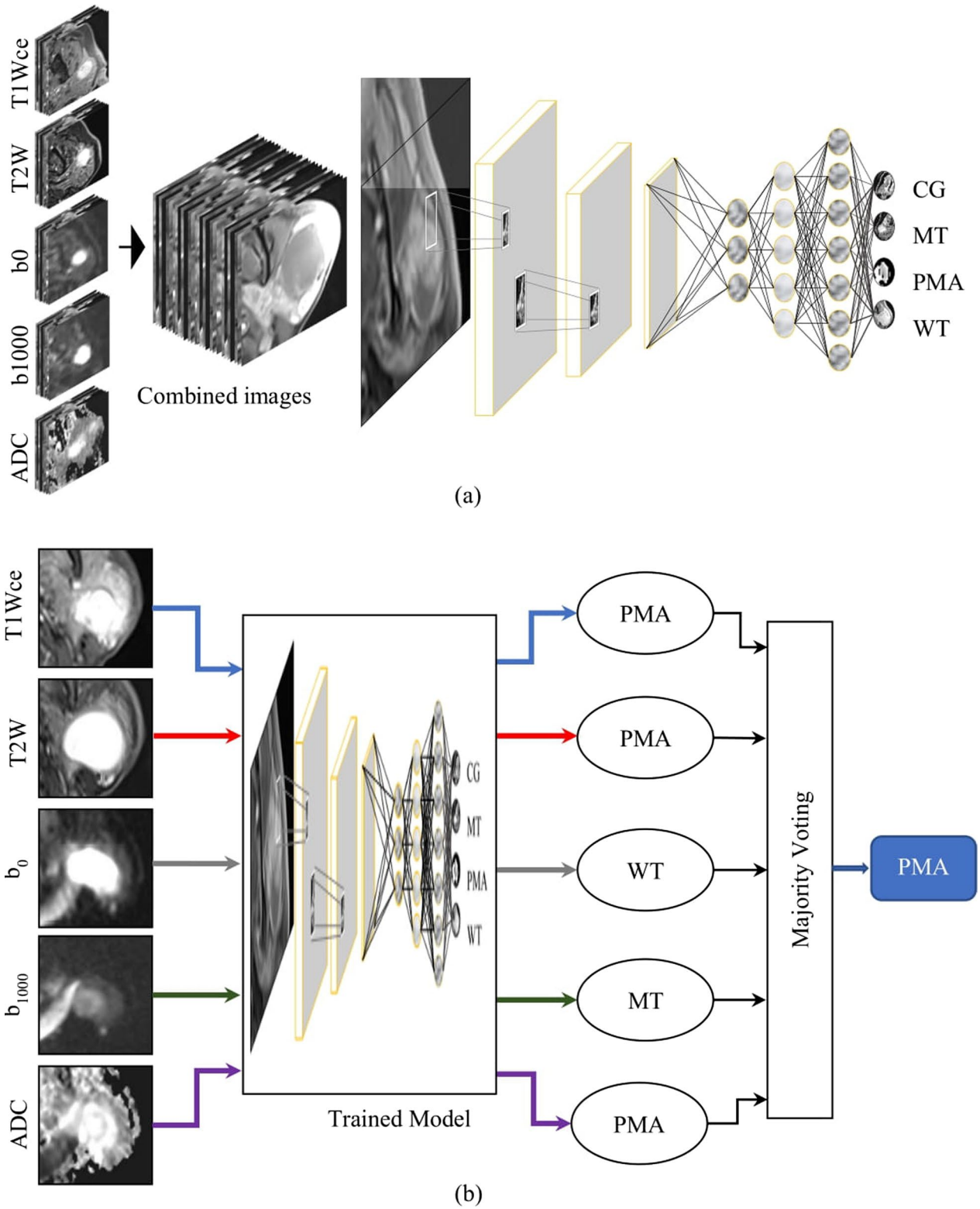

Fig. 2 Proposed methods. **a** The combined approach for stage I, **b** the majority voting process. *CG* control group, *PMA* pleomorphic adenoma, *WT* warthin tumor, *MT* malignant tumors

Table 1 Subtypes of parotid tumors

Distribution of cases according to tumor histopathology	Number of cases (*n*)
Pleomorphic adenoma	46
Warthin tumor	53
High-grade mucoepidermoid carcinoma	11
Acinic cell carcinoma	5
Adenoid cystic carcinoma	4
Squamous cell carcinoma	2
Malignant lymphoma	2
Total cases	123

malignant tumor (MT) group. The results of all classifiers are given in Table 4 for each sequence and MV.

As shown in Table 4, the diffusion b0 sequence produced the best results among the sequences. It was followed by the ADC sequence which had the second-best performance. MV achieved an accuracy improvement of about 3%. This improvement was also seen in the other metrics. The SVM method provided the best performance in the other classifiers.

Experiments were conducted to determine whether a parotid tissue was healthy or had a tumor (PGT). Table 5 presents the results of stage III for each sequence and MV.

One can see in Table 5 that the diffusion b0 sequence achieved the best performance, with metrics found as 97.86% accuracy, 100% sensitivity, and 97.06% specificity. The proposed MV approach provided a performance increase in this stage. The SVM, KNN and LD methods achieved excellent discrimination for stage III. Although Inception-Resnetv2 with MV had acceptable success levels, it lagged behind the SVM, KNN and LD models.

ROC curves were plotted to show the performance of the InceptionResnetv2 and SVM models with MV in each class for stage I, stage II and stage III. These ROC curves are given in Fig. 3.

In Fig. 3, it can be seen that the models achieved excellent CG discrimination in all stages. In stage I, the PMA and WT discrimination rates were good. The performance of the MT class lagged behind the other classes. In stage II, the SVM model provided a good discrimination performance in three classes compared to InceptionResNetv2. In stage III, SVM had excellent classification performance. Moreover, the InceptionResNetv2 classifier offered exemplary performance in diagnosing PGT.

Discussion

The development of AI models has opened new scenarios owing to the possibility of non-invasively assessing features of medical images that are not evaluated by physicians [5]. In our approaches, using a combination of conventional and diffusion-weighted MRI as input data and the majority voting (MV) approach in test data was similar to the evaluation of an MR image in clinical practice. We designed three

Table 2 Information about data set

Stages	Training data				Validation data				Test data each sequence			
Stage I (combined)												
X	299 × 299 × 3 × 2445				299 × 299 × 3 × 350				299 × 299 × 3 × 140			
Y	CG	MT	PMA	WT	CG	MT	PMA	WT	CG	MT	PMA	WT
	720	380	595	750	90	60	105	95	38	21	42	39
Stage I (each sequence)												
X	299 × 299 × 3 × 489				299 × 299 × 3 × 70				299 × 299 × 3 × 140			
Y	CG	MT	PMA	WT	CG	MT	PMA	WT	CG	MT	PMA	WT
	144	76	119	150	18	12	21	19	38	21	42	39
Stage II												
X	299 × 299 × 3 × 2445				299 × 299 × 3 × 350				299 × 299 × 3 × 140			
Y	BT	CG	MT		BT	CG	MT		BT	CG	MT	
	1355	710	380		200	95	55		79	39	22	
Stage III												
X	299 × 299 × 3 × 2445				299 × 299 × 3 × 350				299 × 299 × 3 × 140			
Y	CG		PGT		CG		PGT		CG		PGT	
	720		1725		90		260		38		102	

X represent size and number of images (height × weight × 3 symbolize color layer × image numbers), *Y* symbolize class names and its distributions

CG control group, *PMA* pleomorphic adenoma, *WT* warthin tumor, *MT* malignant tumors, *BT* benign tumors, *PGT* parotid gland tumor

Table 3 Test results of all classifiers with combined and non-combined for stage I (CG, PMA, WT and MT classification)

	Non-combined					Combined				
	Acc (%)	Sens (%)	Spec (%)	PPV (%)	NPV (%)	Acc (%)	Sens (%)	Spec (%)	PPV (%)	NPV (%)
Each sequence DL (InceptionResNetv2)										
T1Wce	81.43	78.90	93.77	79.72	92.49	83.57	79.92	94.30	84.38	94.68
T2W	87.86	86.60	95.95	86.57	95.95	90.71	90.18	96.77	91.51	96.85
b_0	78.57	76.49	92.75	77.58	92.49	84.29	82.35	94.74	82.52	94.80
b_{1000}	72.14	70.36	90.58	72.53	91.27	85.00	84.83	95.00	83.82	94.99
ADC	76.43	72.88	91.92	76.73	92.16	86.43	82.55	95.41	84.80	95.67
MV										
DL (Inception-ResNetv2)	**86.43**	**83.24**	**95.35**	**86.52**	**95.00**	**92.86**	**90.34**	**97.51**	**93.48**	**97.78**
SVM	The matrics just calculated for DL (InceptionResNetv2) model for non-combined approach					93.57	90.98	97.73	94.93	98.03
KNN						92.14	89.74	97.26	92.79	97.51
LD						91.43	89.65	97.16	90.32	97.23

Bold denotes that the DL model created by the combined approach provides improvement in all statistical performance measures compared to the non-combined approach

T1Wce T1-weighted contrast enhanced, *T2W* T2-weighted, b_0, b_{1000}, *ADC* diffusion weighted images, *ADC* apparent diffusion coefficient, *DL* deep learning *MV* majority voting approach, *SVM* support vector machine, *KNN* k-nearest neighbor, *LD* linear discriminant, *Acc* accuracy, *Sens* sensitivity, *Spec* specificity, *PPV* positive predictive value, *NPV* negative predictive value, *CG* control group, *PMA* pleomorphic adenoma, *WT* warthin tumor, *MT* malignant tumors

Table 4 Test results of all classifiers for stage II (CG, BT and MT classification)

	Acc (%)	Sens (%)	Spec (%)	PPV (%)	NPV (%)
Each sequence DL (InceptionResNetv2)					
T1Wce	85.00	76.72	89.97	84.06	92.01
T2W	85.71	79.10	91.11	85.22	92.16
b_0	89.29	81.87	92.72	90.72	94.82
b_{1000}	83.57	77.18	91.08	79.52	91.18
ADC	88.57	80.79	92.27	87.40	94.24
MV					
DL (Inception-ResNetv2)	92.14	83.33	93.99	95.93	97.15
SVM	93.57	87.46	95.35	94.68	97.24
KNN	92.14	85.52	94.73	91.58	96.38
LD	92.86	85.94	94.97	95.27	97.02

T1Wce T1-weighted contrast enhanced, *T2W* T2-weighted, b_0, b_{1000}, *ADC* diffusion weighted images, *ADC* apparent diffusion coefficient, *DL* deep learning, *MV* majority voting approach, *SVM* support vector machine, *KNN* k-nearest neighbor, *LD* linear discriminant, *Acc* accuracy, *Sens* sensitivity, *Spec* specificity, *PPV* positive predictive value, *NPV* negative predictive value, *CG* control group, *BT* benign tumors, *MT* malignant tumors

Table 5 Test results of all classifiers for stage III (CG and PGT classification)

	Acc (%)	Sens (%)	Spec (%)	PPV (%)	NPV (%)
Each sequence DL (InceptionResNetv2)					
T1Wce	85.00	100.00	79.41	64.41	100.00
T2W	95.71	100.00	94.12	86.36	100.00
b_0	97.86	100.00	97.06	92.68	100.00
b_{1000}	92.86	100.00	90.20	79.17	100.00
ADC	94.29	100.00	92.16	82.61	100.00
MV					
DL (Inception-ResNetv2)	99.29	100.00	99.02	97.44	100.00
SVM	100.00	100.00	100.00	100.00	100.00
KNN	100.00	100.00	100.00	100.00	100.00
LD	100.00	100.00	100.00	100.00	100.00

T1Wce T1-weighted contrast enhanced, *T2W* T2-weighted, b_0, b_{1000}, *ADC* diffusion weighted images, *ADC* apparent diffusion coefficient, *DL* Deep Learning, *MV* majority voting approach, *SVM* support vector machine, *KNN* k-nearest neighbor, *LD* Linear discriminant, *Acc* Accuracy, *Sens* sensitivity, *Spec* specificity, *PPV* positive predictive value, *NPV* negative predictive value, *CG* control group, *PGT* parotid gland tumors

stages to perform comprehensive parotid gland tumor (PGT) classification. Conventional and DWI sequence images were used to construct the data set. We proposed the application of the five MRI sequences together and named this method the combined approach. The aim of the combined approach was to expand our data set and increase PGT discrimination performance. The combined method's performance was compared to those of the five InceptionResNetv2 models trained for each MRI sequence, called the non-combined approach. In testing, the five MRI sequences were evaluated

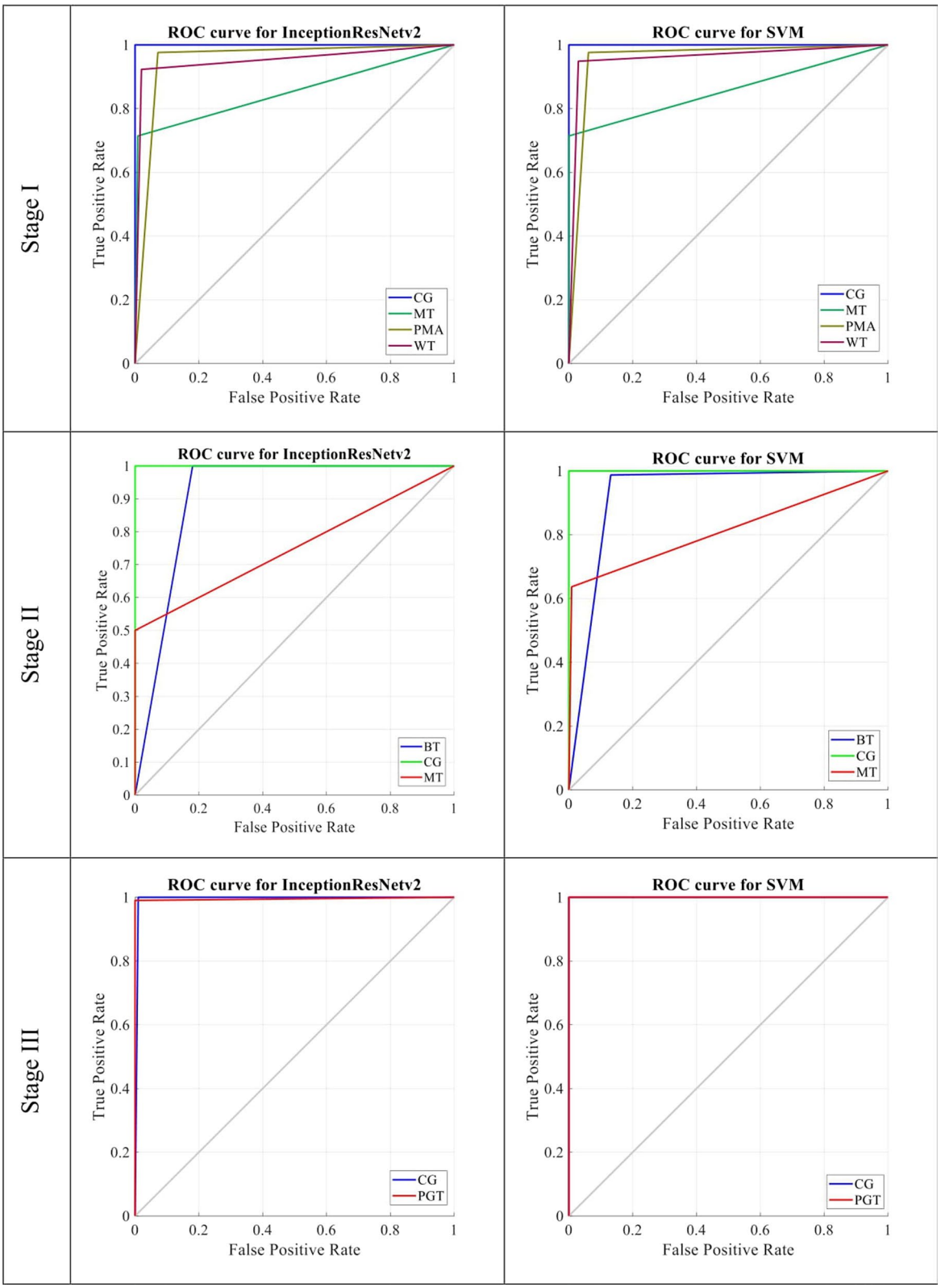

Fig. 3 ROC curves for all stages. *CG* control group, *PMA* pleomorphic adenoma, *WT* warthin tumor, *MT* malignant tumors. *SVM* support vector machine. *ROC* receiver operating characteristic

together to increase the classification performance. This new approach, inspired by radiologists, is named the MV approach. The proposed approaches were tested with a data set collected from parotidectomy patients operated in our clinic. In this study, while creating the deep learning model and completing the training, the histopathological diagnosis, which is the gold standard, was taken as the final result. The InceptionResNetv2 model performance of the hidden deep features about PGT was tested with conventional classifiers. The results of the deep features with the conventional classifiers showed that InceptionResNetv2 is capable of PGT detection.

The differential diagnosis of parotid tumors is a broad and systematic approach that includes imaging, laboratory studies and pathological analysis. MRI and FNAC are the first methods in differential diagnosis [2, 3]. The ability of FNAC to determine parotid tumors is well-established, and its accuracy, sensitivity and specificity values have been reported in the ranges of 85–97%, 64–90% and 86–100%, respectively [3, 22]

Conventional and advanced MR imaging techniques come to the fore in imaging methods used to diagnose parotid tumors. Some studies have shown that a combination of conventional MRI and DWI showed higher diagnostic accuracy than conventional MRI or DWI alone [2, 4, 23]. A meta-analysis demonstrated that sensitivity and specificity values were 76% and 83% for conventional MRI, 91% and 56% for DWI and 86% and 90% for the combination of conventional MRI with DWI [2]. In our study, first, we combined all conventional MRI and DWI images in the training and validation process. In the non-combined approach, we trained models for each MRI sequence separately. We achieved better results for all statistical performance metrics in the combined approach.

Chang et al., who developed a DL model for parotid tumor classification with multimodal MRI, combined conventional and DWI MRI images as input data and reported the accuracy rates for WT, PMA and MT as 83%, 67% and 68%, respectively [9]. The best accuracy rate was obtained in WT, while in our study, PMA provided the highest accuracy as 97.62%. The results for each tumor type of the InceptionResNetv2 and SVM classifiers with the MV approach in our study are given in Table 6. In the study by Chang, while the DWI MRI combination yielded the best results, the combination of conventional MRI and DWI did not improve the results. In our study, it was seen that the best classification was made in the T2W sequence in both the combined and the non-combined approaches. It was also observed that the combination of conventional MRI and DWI increased the accuracy rate in all sequences. The accuracy rate increased by about 6.4% in the combined approach with MV.

In the DL model developed by Feng for the classification of PGT, parotid gland regions were cropped from MR images, similar to our study [24]. Only T1W and T2W images were included in their study. Patients were examined in 3 groups as benign tumors, malignant tumors and free of tumors, and a modified ResNet model was used for classification. Their model's accuracy rate was 89% for the training data set and 83% for the validation data set [24]. We had an accuracy rate of 92.14% in the classification of BT, MT and CG with the DL model and MV approach in stage II.

Table 6 Classification results of SVM and InceptionResNetv2 for SI with majority voting in the each parotid gland tumors

Tumor type	Acc (%)	Sens (%)	Spec (%)	PPV (%)	NPV (%)
DL (InceptionResNetv2)					
PMA	97.62	97.62	92.86	85.42	98.91
WT	92.31	92.31	98.02	94.74	97.06
MT	71.43	71.43	99.16	93.75	95.16
SVM					
PMA	97.62	97.62	93.88	87.23	98.92
WT	94.87	94.87	97.03	92.50	98.00
MT	71.43	71.43	100.00	100.00	95.20

T1Wce T1-weighted contrast enhanced, *T2W* T2-weighted, $b_{0,}$ $b_{1000,}$ *ADC* diffusion weighted images, *ADC* apparent diffusion coefficient, *DL* deep learning, *MV* majority voting approach, *SVM* support vector machine, *Acc* accuracy, *Sens* sensitivity, Spec specificity, *PPV* positive predictive value, *NPV* negative predictive value, *PMA* pleomorphic Adenoma, *WT* warthin tumor, *MT* malignant tumors

In another study, the authors stated that with their proposed anomaly detection (AD) and VGG16-based DL model, the classification of benign and malignant parotid tumors could also be successful in a small quantity of imbalanced data. They used non-medical images obtained from the CUReT data set to reduce the overfitting caused by the small number of images [25]. We used medical images of tumor-free parotid gland tissues to facilitate the removal of the general visual models of the PGT group and reduce overfitting. Indeed, in stage III, it was seen that the patients with PGT and CG could be distinguished with 100% accuracy with the proposed DL model. CG reduced the overfitting in stage II in which benign–malignant tumor classification was performed. While the aforementioned authors achieved 75% sensitivity and 82% specificity in the differential diagnosis of benign and malignant tumors with the VGG16-based DL model with AD, we obtained a better result with 83.33% sensitivity and 93.99% specificity in stage II with the MV approach.

Our study had several limitations. First, a limited data set was studied due to the rarity of PGT. There are many parameters to be trained in DL that require more input data than conventional methods demand. The limited input data may cause inconsistent changes, especially regarding the aspect of statistical performance criteria in evaluating test images.

Moreover, the cropping procedure was performed manually. The process of manual cropping takes approximately one minute. In the future, a fully automated pipeline could be built with a deep learning segmentation approach instead of a manual cropping process to provide standardization and time efficiency.

A point that should be kept in mind for AI studies in medicine is that DL models are supportive of clinical diagnostic methods and examination findings, and they should not be evaluated alone. Inflammatory conditions of the parotid gland or involvement with non-tumor diseases may also be included in the differential diagnosis with tumors in MRI. Therefore, the patient's symptoms and examination findings are important in clinical diagnosis. Additionally, these studies are preliminary studies of AI-based diagnosis which are currently being carried out in selected and characteristic diseases.

Conclusions

In this paper, an AI-based new approach that classifies common parotid tumors automatically and with high accuracy is presented. It was observed that the proposed DL method (Combined approach with Majority Voting) differentiated the tumor-containing parotid gland from the normal parotid gland excellently, and it made the differential diagnosis of benign parotid tumors and malignant parotid tumors very well. In the subtypes of parotid tumors, the DL model detected PMA and WT very well, but it lagged behind other methods in detecting MT. This study will provide convenience for physicians in the evaluation of MRI images of parotid masses.

Author contributions EG: conception and design or analysis and interpretation of data and writing the article. OFA: drafting of the manuscript or revising it for important intellectual content and analysis via Artificial Intelligence knowledge. AK: supervision and final approval of the version to be published. IOY: evaluation and guidance of the radiological technical details and the creation of the data set from the MRI images in the study.

Funding The authors received no financial support for the research, authorship, and/or publication of this article.

Data availability statement The data sets generated or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare that there is no conflict of interest regarding the publication of this article.

Ethical approval Ethical approval for this study was obtained from the local ethics committee (2020/1038).

Informed consent Informed consent was obtained from all individual participants included in the study.

References

1. Barnes LEJ, Reichart P, Sidransky D (2017) Pathology and genetics of head and neck tumors. IARC Press, Lyon
2. Liang YY, Xu F, Guo Y et al (2018) Diagnostic accuracy of magnetic resonance imaging techniques for parotid tumors, a systematic review and meta-analysis. Clin Imaging 52:36–43. https://doi.org/10.1016/j.clinimag.2018.05.026
3. Suzuki M, Kawata R, Higashino M et al (2019) Values of fine-needle aspiration cytology of parotid gland tumors: a review of 996 cases at a single institution. Head Neck 41:358–365. https://doi.org/10.1002/hed.25503
4. Tao X, Yang G, Wang P et al (2017) The value of combining conventional, diffusion-weighted and dynamic contrast-enhanced MR imaging for the diagnosis of parotid gland tumours. Dentomaxillofac Radiol 46:20160434. https://doi.org/10.1259/dmfr.20160434
5. Litjens G, Kooi T, Bejnordi BE et al (2017) A survey on deep learning in medical image analysis. Med Image Anal 42:60–88. https://doi.org/10.1016/j.media.2017.07.005
6. Angus DC (2015) Fusing randomized trials with big data: the key to self-learning health care systems? JAMA 314:767–768. https://doi.org/10.1001/jama.2015.7762
7. Menze BH, Jakab A, Bauer S et al (2015) The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans Med Imaging 34:1993–2024. https://doi.org/10.1109/TMI.2014.2377694
8. Esteva A, Kuprel B, Novoa RA et al (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542:115–118. https://doi.org/10.1038/nature21056
9. Chang YJ, Huang TY, Liu YJ et al (2021) Classification of parotid gland tumors by using multimodal MRI and deep learning. NMR Biomed 34:e4408. https://doi.org/10.1002/nbm.4408
10. Crowson MG, Ranisau J, Eskander A et al (2020) A contemporary review of machine learning in otolaryngology-head and neck surgery. Laryngoscope 130:45–51. https://doi.org/10.1002/lary.27850
11. Juan CJ, Chang HC, Hsueh CJ et al (2009) Salivary glands: echo-planar versus PROPELLER diffusion-weighted MR imaging for assessment of ADCs. Radiology 253:144–152. https://doi.org/10.1148/radiol.2531082228
12. Deepak S, Ameer PM (2009) Brain tumor classification using deep CNN features via transfer learning. Comput Biol Med 111:103345. https://doi.org/10.1016/j.compbiomed.2019.103345
13. Ismael AM, Şengür A (2021) Deep learning approaches for COVID-19 detection based on chest X-ray images. Expert Syst Appl 164:114054. https://doi.org/10.1016/j.eswa.2020.114054
14. Türkoğlu M, Hanbay D (2019) Plant disease and pest detection using deep learning-based features. Turk J Elec Eng Comp Sci. 27:1636–1651. https://dergipark.org.tr/tr/pub/tbtkelektrik/issue/45742/577521
15. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444. https://doi.org/10.1038/nature14539
16. Nguyen LD, Lin D, Lin Z et al (2018) Deep CNNs for microscopic image classification by exploiting transfer learning and feature concatenation. Paper presented at: 2018 IEEE International Symposium on Circuits and Systems (ISCAS). https://doi.org/10.1109/ISCAS.2018.8351550

17. Szegedy C, Ioffe S, Vanhoucke V et al (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. Paper presented at: Proceedings of the AAAI Conference on Artificial Intelligence. https://ojs.aaai.org/index.php/AAAI/article/view/11231
18. Siuly S, Alçin ÖF, Kabir E et al (2020) A new framework for automatic detection of patients with mild cognitive impairment using resting-state EEG signals. IEEE Trans Neural Syst Rehabil Eng 28:1966–1976. https://doi.org/10.1109/TNSRE.2020.3013429
19. Vapnik V (2000) The nature of statistical learning theory, 2nd edn. Springer, New York
20. Cover T, Hart P (1967) Nearest neighbor pattern classification. IEEE Trans Inf Theory 13:21–27. https://doi.org/10.1109/TIT.1967.1053964
21. Bekios-Calfa J, Buenaposada JM, Baumela L (2011) Revisiting linear discriminant techniques in gender recognition. IEEE Trans Pattern Anal Mach Intell 33:858–864. https://doi.org/10.1109/TPAMI.2010.208
22. Schmidt RL, Hall BJ, Wilson AR et al (2011) A systematic review and meta-analysis of the diagnostic accuracy of fine-needle aspiration cytology for parotid gland lesions. Am J Clin Pathol 136:45–59. https://doi.org/10.1309/AJCPOIE0CZNAT6SQ
23. Lechner Goyault J, Riehm S, Neuville A et al (2011) Interest of diffusion-weighted and gadolinium-enhanced dynamic MR sequences for the diagnosis of parotid gland tumors. J Neuroradiol 38:77–89. https://doi.org/10.1016/j.neurad.2009.10.005
24. Feng B, Xia X, Xu L et al (2020) Deep-learning for diagnosis of parotid gland tumor on MR images. Int J Radiat Oncol Biol Phys 108:43–44. https://doi.org/10.1016/j.ijrobp.2020.07.2155
25. Matsuo H, Nishio M, Kanda T et al (2020) Diagnostic accuracy of deep-learning with anomaly detection for a small amount of imbalanced data: discriminating malignant parotid tumors in MRI. Sci Rep 10:19388. https://doi.org/10.1038/s41598-020-76389-4