Received: 24 April 2020 | Revised: 1 August 2020 | Accepted: 21 August 2020

DOI: 10.1002/nbm.4408

RESEARCH ARTICLE

NMR IN BIOMEDICINE WILEY

Classification of parotid gland tumors by using multimodal MRI and deep learning

Yi-Ju Chang[1] | **Teng-Yi Huang**[1] | **Yi-Jui Liu**[2] | **Hsiao-Wen Chung**[3] | **Chun-Jung Juan**[4,5,6]

[1]Department of Electrical Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan

[2]Department of Automatic Control Engineering, Feng Chia University, Taichung, Taiwan

[3]Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan

[4]Department of Medical Imaging, China Medical University Hsinchu Hospital, Hsinchu, Taiwan

[5]Department of Radiology, School of Medicine, College of Medicine, China Medical University, Taichung, Taiwan

[6]Department of Medical Imaging, China Medical University Hospital, Taichung, Taiwan

Correspondence
Chun-Jung Juan, Department of Medical Imaging, China Medical University Hsinchu Hospital, 199, Sec. 1, Xinglong Rd, Zhubei City, Hsinchu County 302, Taiwan.
Email: peterjuancj@yahoo.com.tw

Funding information
Ministry of Science and Technology, Taiwan, Grant/Award Numbers: MOST-107-2314-B-011 -002 -MY3, MOST-107-2314-B-039-071, MOST-108-2314-B-039-014, MOST 107-2314-B-011-002-MY3, MOST-107-2314-B-039-071 and MOST-108-2314-B-039-014

Various MRI sequences have shown their potential to discriminate parotid gland tumors, including but not limited to T_2-weighted, postcontrast T_1-weighted, and diffusion-weighted images. In this study, we present a fully automatic system for the diagnosis of parotid gland tumors by using deep learning methods trained on multimodal MRI images. We used a two-dimensional convolution neural network, U-Net, to segment and classify parotid gland tumors. The U-Net model was trained with transfer learning, and a specific design of the batch distribution optimized the model accuracy. We also selected five combinations of MRI contrasts as the input data of the neural network and compared the classification accuracy of parotid gland tumors. The results indicated that the deep learning model with diffusion-related parameters performed better than those with structural MR images. The performance results (n = 85) of the diffusion-based model were as follows: accuracy of 0.81, 0.76, and 0.71, sensitivity of 0.83, 0.63, and 0.33, and specificity of 0.80, 0.84, and 0.87 for Warthin tumors, pleomorphic adenomas, and malignant tumors, respectively. Combining diffusion-weighted and contrast-enhanced T_1-weighted images did not improve the prediction accuracy. In summary, the proposed deep learning model could classify Warthin tumor and pleomorphic adenoma tumor but not malignant tumor.

KEYWORDS
deep learning, head and neck, MRI, parotid gland tumor, transfer learning

1 | INTRODUCTION

Parotid gland tumor (PGT) is the most common type of salivary gland tumor. Major PGT types include pleomorphic adenoma (PMA), Warthin tumor (WT), and malignant tumor (MT). Determination of the type of PGT is crucial for clinical diagnosis and subsequent treatment. Imaging

Abbreviations: 2D, two-dimensional; 3D, three-dimensional; ADC, apparent diffusion coefficient; ANTs, Advanced Normalization Tools; BraTS, Multimodal Brain Tumor Segmentation Challenge; CNN, convolution neural network; DWI, diffusion-weighted imaging; EPI, echo-planar imaging; MT, malignant tumor; NT, nontumor; PGT, parotid gland tumor; PMA, pleomorphic adenoma; WT, Warthin tumor; NPV, Negative Predictive Value; PPV, Positive Predictive Value.

Presented, in part, at the ISMRM 27th Annual Meeting & Exhibition; May 11-16, 2019; Montreal, Quebec, Canada.

[Corrections added on 12 September 2020, after first online publication: The department names in the 4th and 6th author affiliations have been corrected.]

modalities, such as MRI and computed tomography, are useful to identify the location and size of PGTs. Fine-needle aspiration biopsy is the primary method for identifying the tumor type, but its sensitivity is low (70%-80%) for recognizing malignant PGTs.[1,2]

MRI can be useful for tumor classification. For example, T_1- and T_2-weighted images clearly present the texture of tumors, including areas of normal and lesion tissues.[3] High-grade malignant salivary gland tumors are distinguished on routine MR images by ill defined borders, cystic components, low T_2 signal intensity, necrosis, and invasion of surrounding tissues. However, MRI is often unable to distinguish between benign and malignant salivary tumors.[4–6] The apparent diffusion coefficient (ADC) derived from diffusion-weighted imaging (DWI) has been shown to be associated with tumor cellularity, and MTs exhibit hyperintensity in DWI.[7,8] The ADC value of a PGT region is useful for differentiating between WTs and PMAs.[9–12] However, the mean ADC values of WTs and MTs are not significantly different.[8,10,13] The ADC has a sensitivity of 50%-60% in distinguishing MTs.[9,14] Therefore, identifying MTs through MRI remains a challenge.

Recently, deep learning methods, particularly convolution neural network (CNN)-based models, have demonstrated effectiveness in image recognition tasks. CNN methods for pixel-wise classification, also referred to as semantic segmentation, are now widely employed in computer-vision applications, such as robotics and self-driving cars.[15,16] The semantic segmentation method has also been used in MRI applications. For example, in the global competition of the Multimodal Brain Tumor Segmentation Challenge (BraTS),[17,18] researchers achieved an accuracy of more than 80% for the pixel-wise classification of brain gliomas. In addition, deep-learning-based tumor segmentation and classification have been investigated for several cancers, including breast cancer,[19,20] liver tumor,[21–23] and nasopharyngeal carcinoma.[24] We hypothesize that the deep learning method on MRI data can also help detect and distinguish PGTs. In this study, we implemented a semantic segmentation method of multimodal MRI images for the segmentation of PGTs and classification of tumor types.

2 | METHODS AND MATERIALS

2.1 | The patient cohort and MRI protocol

The Institutional Review Board of Tri-Service General Hospital approved the study and waived the requirement of written informed consent for this retrospective study. Eighty-five consecutive patients with PGT (54 men and 31 women; age 49.6 ± 15.6 years) who underwent MRI examination were enrolled. Their PGTs were of the types WT (n = 27), PMA (n = 33), and MT (n = 25) according to histologic findings. All MRI examinations were performed on a 1.5 T MRI system (Signa HDx, GE Healthcare) with an eight-channel neurovascular head-and-neck array coil. Before contrast administration, the scanning protocol included T_2-weighted and DWI sequences. Acquisition parameters for T_2-weighted imaging were slice orientation axial, T_R 3150 ms, T_E 77.3 ms, number of excitations 2, and slice number 32. Moreover, the single-shot echo-planar DWI with parameters (slice orientation axial, T_R 7000 ms, T_E 72.2 ms, number of excitations 4, slice number 18, and fat saturated) were acquired with diffusion gradients b = 0 and 1000 s/mm^2 applied in each of three orthogonal directions. After contrast administration (gadolinium-DTPA, 0.1 mmol/kg), we acquired T_1-weighted images by using fat-saturated fast spin-echo with the following parameters: slice orientation axial, T_R 616.7 ms, T_E 12 ms, number of excitations 0.5, and slice number 32. Thus, we collected datasets containing four MRI contrasts for each patient, namely T_2 weighted, T_1 weighted with contrast enhancement, b_0 (DWI, b = 0 s/mm^2), and b_{1000} (DWI, b = 1000 s/mm^2).

2.2 | Data conversion and image registration

After data acquisition, we collected all DICOM image files and sorted them by identifying DICOM tags. Next, the procedure converted files into NIfTI format files with the dcm2niix software (https://github.com/rordenlab/dcm2niix). We transferred them to a workstation for further processing. A board-certified radiologist (CJJ), with more than 15 years of experience in head-and-neck MRI, manually outlined the region of the tumor on contrast-enhanced T_1-weighted images and constructed another three-dimensional (3D) volume with pixels labeled as 0, nontumor (NT; including background), 1, WT, 2, PMA, and 3, MT, according to histological records. Finally, the procedure saved the 3D volume with the tumor labels into another NIfTI file.

The subsequent step was to co-register the four volumes by using the Advanced Normalization Tools (ANTs) software package (http://stnava.github.io/ANTs/). We registered the T_2 volume to the contrast-enhanced T_1-weighted volume. Subsequently, we used deformable registration to obtain the coordinate transformation between the b_0 and T_2 volumes and used the transformation to obtain the registered b_0 and b_{1000} volumes. Using the obtained diffusion-weighted volumes, we calculated ADC maps using the equation ADC = $\ln[(SI_0/SI_{1000})]/1000$, where SI_0 and SI_{1000} are the signal intensities of the b_0 and b_{1000} volumes, respectively.[25] The ADC maps subsequently underwent median filtering with a 3 × 3 kernel. Therefore, we had six 3D volumes, including the contrast-enhanced T_1, T_2, b_0, b_{1000}, ADC, and tumor label volumes. They were presented as $\widehat{T_1}$, $\widehat{T_2}$, $\widehat{b_0}$, $\widehat{b_{1000}}$, $\widehat{\mathrm{ADC}}$ and $\widehat{\mathrm{SEG}}$, respectively. The matrix size of each volume was 256 × 256 × 32. For the upcoming deep learning procedures, we normalized the pixel values of $\widehat{T_1}$, $\widehat{T_2}$, $\widehat{b_0}$, and $\widehat{b_{1000}}$ based on the maximum intensity of each volume. Finally, we extracted two-dimensional (2D) slices from the 3D volumes for further usage in the deep learning training procedures. For

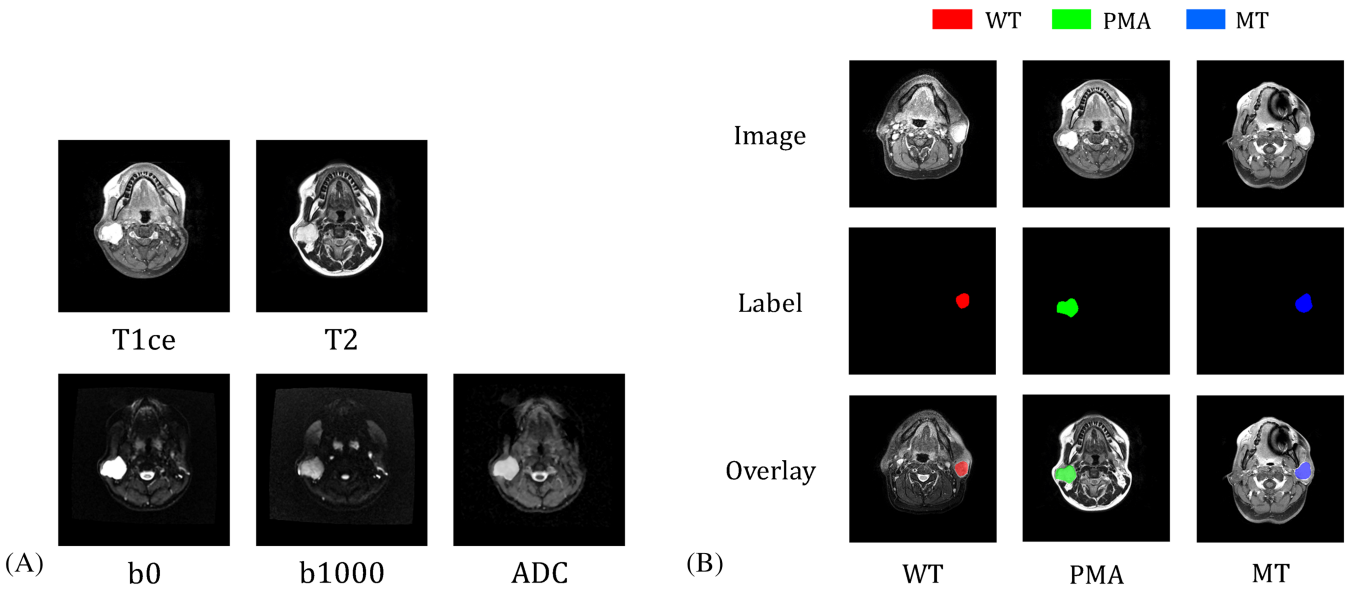


FIGURE 1 A, An example stack consisting of five types of MRI image. From left to right and top to bottom, they are contrast-enhanced T_1, T_2, b_0, b_{1000}, and ADC images. The ADC map was reconstructed from b_0 and b_{1000} images. We had 2726 stacks from 85 patients. B, Manually drawn regions of interest of the three types of tumor: NT areas are indicated in black, WT in red, PMA in green, and MT in blue. In the stored file, we used numbers 0 to 3 for NT, WT, PMA, and MT pixels, respectively

every slice, we merged $\widehat{T_1}$, $\widehat{T_2}$, $\widehat{b_0}$, $\widehat{b_{1000}}$, and $\widehat{\mathrm{ADC}}$ data into a 256 × 256 × 5 matrix. This matrix was termed a five-channel "stack." We collected 2726 stacks from the 85 patients, including 463 stacks covering PGTs and 2263 stacks without PGTs. Figure 1A shows an example stack of the five modalities of MRI. Figure 1B presents the examples of the manually outlined region of the three tumor types (red, WT, green, PMA, and blue, MT). Because of the restriction of the input layer of the implemented neural network, which will be discussed later, the input stack size was fixed to a four-channel stack (256 × 256 × 4). To compare the relation between classification accuracy and MR contrasts, we generated the following types of four-channel stack: sT2 combining four identical images ($\widehat{T_2}$, $\widehat{T_2}$, $\widehat{T_2}$, and $\widehat{T_2}$), sT1 combining ($\widehat{T_1}$, $\widehat{T_1}$, $\widehat{T_1}$, and $\widehat{T_1}$), sT1T2 combining ($\widehat{T_1}$, $\widehat{T_1}$, $\widehat{T_2}$, and $\widehat{T_2}$), sDWI consisting of (zeros, $\widehat{b_0}$, $\widehat{b_{0100}}$, and $\widehat{\mathrm{ADC}}$), sALL consisting of ($\widehat{T_1}$, $\widehat{T_2}$, $\widehat{b_0}$, and $\widehat{b_{1000}}$), and sALL2 consisting of ($\widehat{T_1}$, $\widehat{T_2}$, $\widehat{b_{1000}}$,and $\widehat{\mathrm{ADC}}$).

2.3 | Deep learning: U-Net and transfer learning

We used 2D U-Net for pixel-wise tumor classification.[26] The network consisted of encoding and decoding paths with convolutional blocks. Each of the blocks consisted of a 3 × 3 convolution layer followed by a rectified linear unit and a dropout layer. In the encoding path, the output of each block was down-sampled with a max-pooling operation with a stride of 2. In the decoding path, the input of each block was concatenated with the corresponding feature maps obtained in the encoding path, and the output of each block was up-sampled using a transpose convolution. The final output layer of the U-Net was connected to a multiclass softmax classifier.

To initialize U-Net, we used transfer learning, which refers to transferring network weights from a pretrained model to another model. In general, the pretrained models are trained with immense numbers of datasets. With the same architecture of the deep learning network, the weights of the pretrained model can be utilized as the initial values of the weights for the new model. Because the weights are linked to the process of extracting and filtering features, most deep learning machines are specialized in a particular field or task. Therefore, we adapted a method that won the third prize in BraTS 2017 to produce the pretrained model.[27] In that model, a four-channel input layer was implemented, and the U-Net was pretrained with four types of brain MR image (ie T_2, FLAIR, T_1, and contrast-enhanced T_1) and three tumor labels (1, necrotic and nonenhancing tumor, 2, peritumoral edema, and 3, gadolinium-enhanced tumor). The training set of BraTs 2017 included 285 patients with gliomas. The matrix size of the BraTs 2017 dataset was 240 × 240 × 155. We extracted 155 2D slices from each dataset and interpolated their matrix size into 256 × 256. The total number of images and steps for training the BraTS model were 155 × 285 = 44 175 and 600 000, respectively.

After constructing the pretrained model, we transferred its weights to initialize the training procedure in our current study for classifying PGTs. The training parameters of U-Net were the following: optimizer, Adam, batch size, 6 or 8, loss, cross-entropy, beta of L2 regularization,

10^{-7}, and training steps, 10 000. We employed image augmentation methods, including random image up-down and left-right flipping, rotation (ranging from 15° to −15°), and contrast adjustments (ranging from 0.6 to 1.4), in the training procedures to enhance the variability of the images. The U-Net was implemented with the TensorFlow framework (v1.8) under the Python (v3.6) environment and was trained on a home-built workstation with a 1080 GPU (1080 Ti, Nvidia Corporation, Santa Clara, CA, USA).

2.4 | Cross-validation, prediction, and performance assessment

We distributed the 2726 stacks into three groups by using stratified random sampling to conduct a threefold cross-validation of the U-Net model.[26] All the stacks of one patient were dispatched into the same group, and every group had a proportional allocation of the three tumor types. The number of patients in the three groups was (WT 9, PMA 8, WT 8), (9, 8, 8), and (9, 8, 9), respectively. We performed eight trials of random sampling and U-Net training. After the training stage, we built U-Net models for the pixel-wise classification of 2D MR images. Because the PGT datasets were 3D volumes with a matrix size of 256 × 256 × 32, we split each 3D volume into 32 images, obtained the model inferences of the images, and then merged them back into a 3D volume. This volume is termed the predicted $\widehat{\mathrm{SEG}}$. The matrix size of the predicted $\widehat{\mathrm{SEG}}$ was the same as the input volume (ie 256 × 256 × 32), and pixel values presented classification results (ie 0 for background, 1 for WT, 2 for PMA, and 3 for MT). Subsequently, for each predicted $\widehat{\mathrm{SEG}}$, we counted the pixel numbers of three tumor types in it, identified the tumor type with the largest pixel number, and unified the tumor category for all slices. Finally, we stored the predicted $\widehat{\mathrm{SEG}}$ in a NIfTI file format. We then evaluated segmentation results by using Dice coefficients and calculated the accuracy of tumor classification results.

2.5 | Comparing training schemes: transfer learning and input batches

During the preliminary investigation stage, we evaluated four training schemes to determine a scheme that optimized the tumor classification performance. In Scheme 1, the input batch size was six, and transfer learning was not applied. The training procedure randomly selected six stacks from the training stacks containing all three PGTs. Scheme 2 was the same as Scheme 1 but with transfer learning. Scheme 3 was the same as Scheme 2 except that the input batch was not a random mix of three tumor types but comprised exactly two WT, two PMA, and two MT stacks. In Scheme 4, transfer learning was applied, and all stacks, including both tumor and NT stacks, were used to train the U-Net model. The batch size was eight stacks—two WT, two PMA, two MT, and two NT stacks. To select the optimized training scheme, we trained the U-Net model by using the sDWI stacks and the four schemes, and the training scheme that yielded the best results was then chosen to evaluate the segmentation and classification performance.

3 | RESULTS

3.1 | Visualization of training procedures

Figure 2 displays intermediate results obtained during the training process of sALL. The first column presents all input images, including $\widehat{T_1}$, $\widehat{T_2}$, $\widehat{b_0}$, and $\widehat{b_k}$. The following three columns demonstrated the outputs of different convolutional layers. The features were extracted layer by layer in the encoding path, and the output of the fifth layer was found to have the lowest spatial resolution. Next, an up-sampling (decoding) path was started. The output of the seventh layer demonstrated a higher spatial resolution than that of the fifth layer. The network generated softmax values for each class of each pixel. The value presented the probability distribution of the four categories (ie NT, WT, PMA, and MT) for each pixel.

3.2 | Training scheme comparison

Figure 3 compares Schemes 1 and 2. Figure 3A shows the input DWI stack. This patient had one WT in the right parotid gland. Figure 3B presents predicted $\widehat{\mathrm{SEG}}$ after different training steps for Scheme 1 (upper row) and Scheme 2 (lower row), respectively. The tumor was clearly outlined after 5000 training steps for Scheme 1 and 1000 for Scheme 2. After 5000 steps, the Dice coefficient of the Scheme 2 result was considerably higher than that of Scheme 1. This suggested that applying transfer learning in Scheme 2 not only accelerated the convergence of U-Net optimization but also improved the prediction accuracy. Table 1 presents the group analysis of the four training schemes for the recognition of PGTs. Scheme 1 exhibited the poorest prediction performance; Scheme 4 displayed the best prediction results, with accuracy values for WT, PMA, and MT being 0.81, 0.76, and 0.71, respectively.

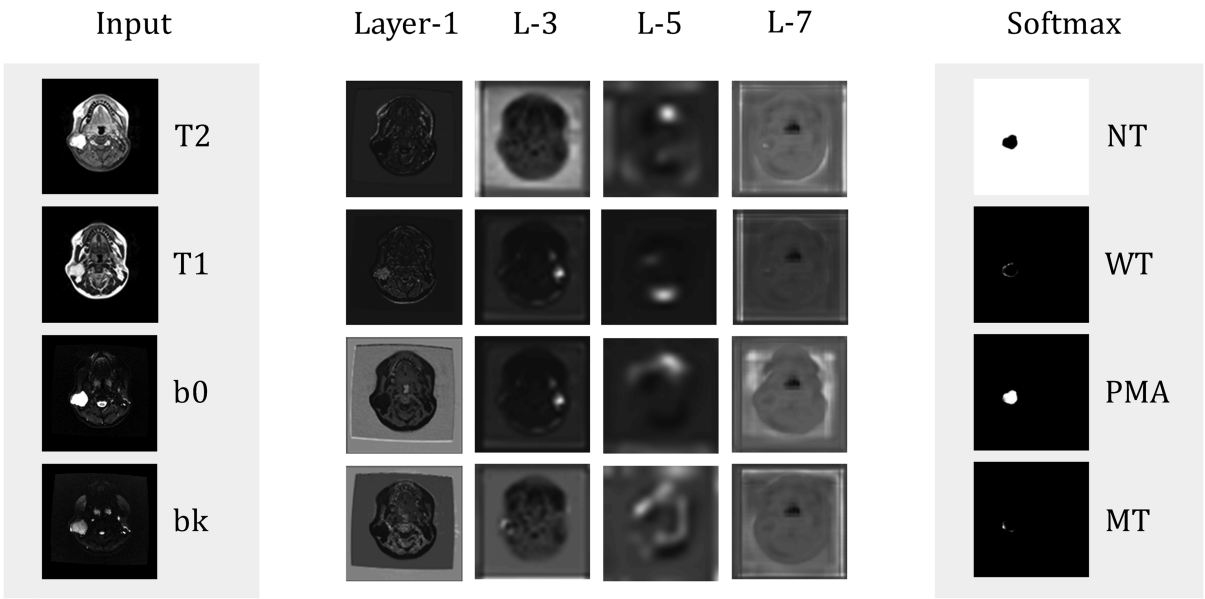
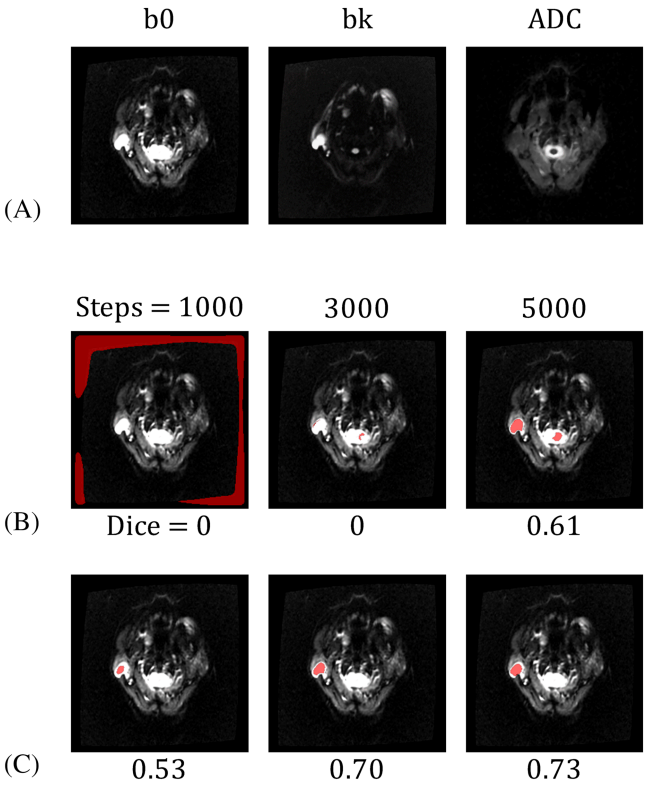


FIGURE 2 Visualization of the output images of the intermediate and softmax layers of the sALL U-Net model. Layers 1 to 5 are in the encoding path, and output images are down-sampled with max-pooling to reduce feature numbers. The pixel values of the output images of the softmax layer present the probability distribution (white for 1, black for 0) of four categories (ie NT, WT, PMA, and MT). As an example, the large region of white color in the upper-right image (the NT image from softmax output) means that these white pixels likely belong to the NT category, except for the small dark region, likely belonging to a pleomorphic tumor (cf. the PMA image from softmax output)

FIGURE 3 Demonstration of the PGT segmentation of an image containing a WT on the right side. The input is an sDWI stack (A), and WT regions are predicted using models trained with Scheme 1 (B) and Scheme 2 (C). The WT regions are presented as red pixels overlaying the ADC images. After 1000, 3000, and 5000 training steps, the WT region became progressively more accurate. The number beneath each image is the Dice coefficient between the predicted and actual tumor regions. This example demonstrates the advantage of using transfer learning in Scheme 2



3.3 | Comparing multimodal MRI

Next, we used the Scheme 4 method to obtain models with six types of stack. Each model underwent a threefold cross-validation with eight trials, with 680 (85 patients × 8 trials) predicting $\widehat{SEG}$ results for each type of stack. Figure 4 demonstrates predicted results obtained from three patients having WT, PMA, or MT. For the patients with WT or PMA, the results of the five stack types were comparable to the true label. For the

TABLE 1 Results of the methods used for the identification of PGTs

	Batch size/input data	Transfer	Accuracy of each scheme		
			WT (n = 27)	PMA (33)	MT (25)
Scheme 1	6, random (WT, PMA, MT)	×	0.73 ± 0.05	0.69 ± 0.04	0.64 ± 0.04
Scheme 2	6, random (WT, PMA, MT)	✓	0.77 ± 0.03	0.74 ± 0.01	0.69 ± 0.04
Scheme 3	6, WT × 2, PMA × 2, MT × 2	✓	0.78 ± 0.02	0.74 ± 0.03	0.67 ± 0.03
Scheme 4	8, WT × 2, PMA × 2, MT × 2, NT × 2	✓	0.81 ± 0.02	0.76 ± 0.02	0.71 ± 0.04

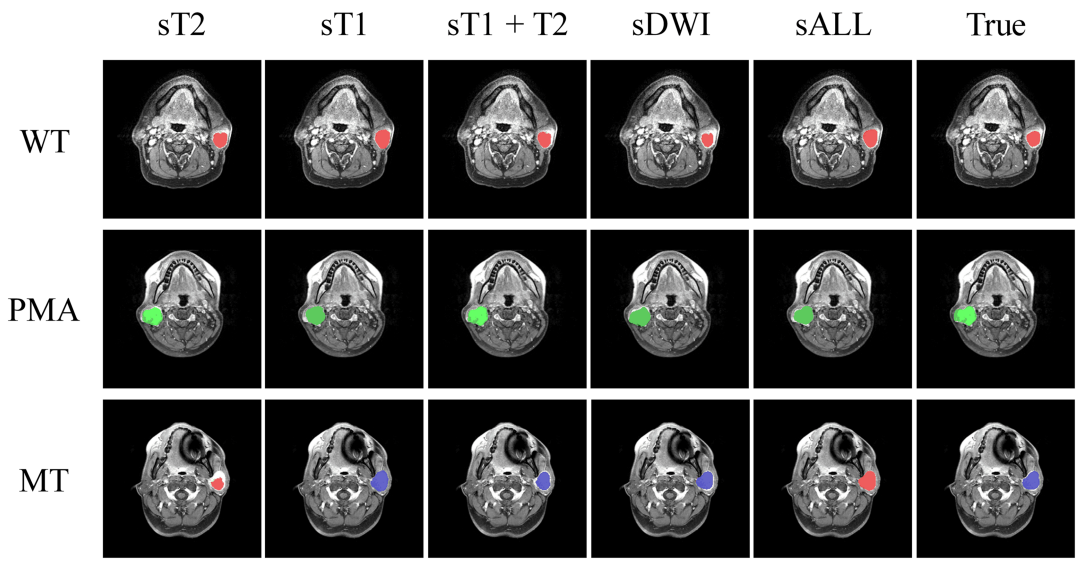


FIGURE 4 Three selected slices of the predicted PGT regions (red, WTs, green, PMAs, and blue, MTs). For WT and PMA slices, the U-Net models trained using different stacks correctly classified the tumor class. However, for the MT example, the sT2 and sALL models misidentified MTs as WTs

patient with MT, the segmentation results of all the types of stack, except sT2, were close to the correct label. However, only results obtained from sT1, sT1T2, and sDWI classify PGT into the correct category (ie MT, blue color). Table 2 presents the group statistics of the recognition of PGTs. For segmentation results, the PGT region obtained using sALL produced the highest average Dice coefficient (0.48 ± 0.01). For tumor classification, PGT types predicted using sDWI yielded the highest average accuracy. The group analysis revealed that the performance of sT2 was the poorest. Combining DWI and structural images did not improve the outcome. Among all models, classifying WTs exhibited the highest accuracy. Table 3 presents the complete results of classification using sDWI. The sensitivity of WT, PMA, and MT was 0.83, 0.67, and 0.33, respectively, suggesting that the obtained U-Net model was not sensitive to classify MTs.

TABLE 2 Results of the identification of PGTs using various combinations of MR images

Input	Dice	Accuracy of each type		
		WT (n = 27)	PMA (33)	MT (25)
sDWI	0.46 ± 0.03	0.81 ± 0.02	0.76 ± 0.02	0.71 ± 0.04
sT1	0.44 ± 0.01	0.79 ± 0.02	0.66 ± 0.03	0.68 ± 0.04
sALL	0.48 ± 0.01	0.75 ± 0.04	0.70 ± 0.03	0.62 ± 0.03
sALL2	0.40 ± 0.02	0.83 ± 0.03	0.67 ± 0.01	0.68 ± 0.02
sT1T2	0.42 ± 0.02	0.72 ± 0.04	0.65 ± 0.04	0.62 ± 0.04
sT2	0.38 ± 0.02	0.67 ± 0.05	0.63 ± 0.04	0.62 ± 0.04

TABLE 3 Classification results of sDWI

	Accuracy	Sensitivity	Specificity	PPV	NPV
WT	0.81 ± 0.02	0.83 ± 0.03	0.80 ± 0.03	0.66 ± 0.03	0.91 ± 0.01
PMA	0.76 ± 0.02	0.62 ± 0.07	0.84 ± 0.03	0.72 ± 0.02	0.78 ± 0.03
MT	0.71 ± 0.04	0.33 ± 0.07	0.87 ± 0.04	0.53 ± 0.12	0.76 ± 0.02

4 | DISCUSSION

In this study, we describe a fully automatic system for the detection and classification of PGTs. We used a 2D U-Net CNN for multiclass segmentation. To investigate a suitable training procedure, we proposed and compared four training schemes for U-Net. We found that transfer learning and manipulations of training batches gradually improved the classification accuracy. Unlike Scheme 1, we applied transfer learning in Scheme 2, which improved performance; this suggests that the network weights pretrained with 44 175 stacks in the BraTs dataset may lead the U-Net model closer to the optimal solution because filters for the detection of the brain and PGTs in convolutional layers could be similar. In Schemes 3 and 4, the training batch for the forward path of U-Net was a fixed structure. Each batch in Scheme 3 comprised two stacks of each tumor type, with class balance maintained during the training stage. Although this setup improved the classification accuracy, a nonignorable number of NT pixels were misclassified into PGTs (ie false positives). In Scheme 4, we added two stacks generated from image slices without PGTs in each batch. The performance of Scheme 4 was the best. Thus, we fixed the training procedure to that used for Scheme 4 and kept exploring the model efficiency with various combinations of MRI images to identify PGT types.

We used six types of stack as input to the U-Net model and obtained the corresponding models. The Dice coefficient results revealed that sALL and sDWI models produced better segmentation results than the other models. The sDWI model outperformed the others in accuracy. Among the stacks consisting of only structural images, the sT1 model performed better than the sT2 or sT1T2 models. As for sALL and sALL2, we assembled it with all available MR modalities, such as T_1, T_2, and DWI, and assumed that the U-Net training procedure could derive the optimal combination of all MR information. However, neither model was better than the sDWI model. This could be for two reasons: image registration and data size. For example, the four-channel sALL stack (matrix size $256 \times 256 \times 4$) was constructed from four images ($\widehat{T_1}$, $\widehat{T_2}$, $\widehat{b_0}$, and $\widehat{b_{1000}}$), with the assumption of multichannel deep learning that all the channels were aligned pixel by pixel. Although we used deformable registration to amend the misalignment between spin-echo-based structural images and echo-planar imaging (EPI)-based DWI images, residual image distortion along the phase-encoding direction in DWI images was inevitable, thereby impairing registration precision. The misregistration between channels in the sALL-based U-Net model could have reduced the accuracy of PGT recognition. We merged all MR information in the input layer and tested whether the training procedure could select dominating image channels and exclude less useful ones. In theory, if the training procedure achieves the optimal solution, the sT1T2 model should be at least comparable to the sT1 model under equal computation power. However, our limited data size (2726 stacks) could have restricted model optimization, and more information did not produce better results.

Among all the models investigated in this study, the sDWI model provided the best PGT classification results. The accuracy results were 0.81, 0.76, and 0.71 for WT, PMA, and MT, respectively. The classification performance of WT is comparable to that reported in a previous study, which required sex and age information in addition to MRI images.[12] The sensitivity results were 0.83, 0.63, and 0.33 for WT, PMA, and MT, respectively. The results suggest that our current model was insensitive to MTs. This critical limitation originates from the fact that malignant PGTs in humans are related to deeper structures, such as the parapharyngeal space, adjacent muscles, and bony tissues, which are not clearly presented in MRI.[28] In addition, among WT, PMA, and MT, the ADC values of PMAs are higher than those of MTs and WTs. In previous studies on PGTs, ADC values were used to differentiate PMA or WT.[10,12,29] However, MTs with increased cellularity and WTs containing lymphoid tissues both present lower ADC values. Consequently, when only ADC values are used, an overlap of ADC values between WTs and MTs impedes the effective differentiation of MTs. Studies have suggested that fraction anisotropy values obtained using diffusion tensor imaging[30] and the wash-out pattern of dynamic contrast imaging[31] improve diagnostic accuracy. Additional studies should combine them into the deep learning framework to potentially improve sensitivity to distinguish MTs.

Compared with radiomics-based machine learning methods,[32,33] the proposed algorithm both outlines the tumor region and identifies the tumor type. It does not need region of interest drawing once the training is completed. It is therefore fully automatic and beneficial in the clinical setting. Furthermore, the procedures for the generation of image features are different in the machine and deep learning methods. CNN learns image features according to the training dataset, whereas machine learning methods use predefined features. If the training procedure of a deep learning method achieves the optimal solution with sufficient training data, the obtained image features could outperform predefined ones. However, features generated by the "black box" deep learning methods are more difficult to interpret than predefined features (eg gray-level texture features in radiomics) in machine learning models. If explainable features are desirable, machine learning algorithms may be used for this application instead of deep learning.

One study limitation is the small data size, which hampers the optimization of the large segmentation network. Although transfer learning improves the classification accuracy, deep learning models with a larger dataset are warranted. Another study limitation is the alignment of DWI, T_1-weighted, and T_2-weighted images. Although we retrospectively performed deformable registration of all images, the residual misregistration may have reduced the classification performance. Reducing EPI distortion or acquiring all images with the same type of MR sequence, such as the multishot EPI,[34] may be a solution.

In summary, we assessed the PGT classification performance of a U-Net method combined with multimodal MRI. The U-Net model based on DWI information outperformed contrast-enhanced T_1- and T_2-weighted images. Combining all available modalities did not improve accuracy. The U-Net model could simultaneously outline the tumor region and identify the tumor type. The U-Net model can be practical to use in the clinical setting to detect WTs and PMAs, but it is not sensitive for MTs.

ACKNOWLEDGEMENTS

This study was supported by the Ministry of Science and Technology, Taiwan (MOST 107-2314-B-011-002-MY3, MOST-107-2314-B-039-071, and MOST-108-2314-B-039-014). We are grateful to the National Center for High-Performance Computing for computer time and facilities.

DATA AVAILABILITY STATEMENT

The datasets generated or analyzed during the current study are available from the corresponding author on reasonable request.

CONFLICTS OF INTEREST

The authors declare that they have no financial interests or potential conflicts of interest related to the research described in this paper.

ORCID

Teng-Yi Huang https://orcid.org/0000-0002-6836-3946
Yi-Jui Liu https://orcid.org/0000-0001-5865-6836
Chun-Jung Juan https://orcid.org/0000-0001-5219-0406

REFERENCES

1. Gudmundsson JK, Ajan A, Abtahi J. The accuracy of fine-needle aspiration cytology for diagnosis of parotid gland masses: a clinicopathological study of 114 patients. *J Appl Oral Sci*. 2016;24(6):561-567.
2. Schmidt RL, Hall BJ, Wilson AR, Layfield LJ. A systematic review and meta-analysis of the diagnostic accuracy of fine-needle aspiration cytology for parotid gland lesions. *Am J Clin Pathol*. 2011;136(1):45-59.
3. Chen X, Wei X, Zhang Z, Yang R, Zhu Y, Jiang X. Differentiation of true-progression from pseudoprogression in glioblastoma treated with radiation therapy and concomitant temozolomide by GLCM texture analysis of conventional MRI. *Clin Imaging*. 2015;39(5):775-780.
4. Dai YL, King AD. State of the art MRI in head and neck cancer. *Clin Radiol*. 2018;73(1):45-59.
5. Abraham J. Imaging for head and neck cancer. *Surg Oncol Clin N Am*. 2015;24(3):455-471.
6. Abdel Razek AAK, Mukherji SK. State-of-the-art imaging of salivary gland tumors. *Neuroimaging Clin N Am*. 2018;28(2):303-317.
7. Kono K, Inoue Y, Nakayama K, et al. The role of diffusion-weighted imaging in patients with brain tumors. *Am J Neuroradiol*. 2001;22(6):1081-1088.
8. Abdel Razek AAK. Routine and advanced diffusion imaging modules of the salivary glands. *Neuroimaging Clin N Am*. 2018;28(2):245-254.
9. Celebi I, Mahmutoglu AS, Ucgul A, Ulusay SM, Basak T, Basak M. Quantitative diffusion-weighted magnetic resonance imaging in the evaluation of parotid gland masses: a study with histopathological correlation. *Clin Imaging*. 2013;37(2):232-238.
10. Habermann CR, Arndt C, Graessner J, et al. Diffusion-weighted echo-planar MR imaging of primary parotid gland tumors: is a prediction of different histologic subtypes possible? *Am J Neuroradiol*. 2009;30(3):591-596.
11. Habermann CR, Gossrau P, Graessner J, et al. Diffusion-weighted echo-planar MRI: a valuable tool for differentiating primary parotid gland tumors? *RöFo*. 2005;177(7):940-945.
12. Wang CW, Chu YH, Chiu DY, et al. Journal Club: the Warthin tumor score: a simple and reliable method to distinguish warthin tumors from pleomorphic adenomas and carcinomas. *Am J Roentgenol*. 2018;210(6):1330-1337.
13. Eida S, Sumi M, Sakihama N, Takahashi H, Nakamura T. Apparent diffusion coefficient mapping of salivary gland tumors: prediction of the benignancy and malignancy. *Am J Neuroradiol*. 2007;28(1):116-121.
14. Yuan Y, Tang W, Tao X. Parotid gland lesions: separate and combined diagnostic value of conventional MRI, diffusion-weighted imaging and dynamic contrast-enhanced MRI. *Br J Radiol*. 2016;89(1060):20150912.
15. Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans Pattern Anal*. 2017;39(12):2481-2495.
16. Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. *IEEE Trans Pattern Anal*. 2017;39(4):640-651.
17. Bakas S, Akbari H, Sotiras A, et al. Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci Data*. 2017;4(1):170117.
18. Menze BH, Jakab A, Bauer S, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans Med Imaging*. 2015;34(10):1993-2024.
19. Ren T, Cattell R, Duanmu H, et al. Convolutional neural network detection of axillary lymph node metastasis using standard clinical breast MRI. *Clin Breast Cancer*. 2020;20(3):e301-e308.

20. Ha R, Chang P, Mutasa S, et al. Convolutional neural network using a breast MRI tumor dataset can predict oncotype Dx recurrence score. *J Magn Reson Imaging*. 2019;49(2):518-524.
21. Wang CJ, Hamm CA, Savic LJ, et al. Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features. *Eur Radiol*. 2019;29(7):3348-3357.
22. Trivizakis E, Manikis GC, Nikiforaki K, et al. Extending 2-D convolutional neural networks to 3-D for advancing deep learning cancer classification with application to MRI liver tumor differentiation. *IEEE J Biomed Health Inform*. 2019;23(3):923-930.
23. Hamm CA, Wang CJ, Savic LJ, et al. Deep learning for liver tumor diagnosis part I: development of a convolutional neural network classifier for multi-phasic MRI. *Eur Radiol*. 2019;29(7):3338-3347.
24. Lin L, Dou Q, Jin YM, et al. Deep learning for automated contouring of primary tumor volumes by MRI for nasopharyngeal carcinoma. *Radiology*. 2019;291(3):677-686.
25. Juan CJ, Chang HC, Hsueh CJ, et al. Salivary glands: echo-planar versus PROPELLER diffusion-weighted MR imaging for assessment of ADCs. *Radiology*. 2009;253(1):144-152.
26. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. Paper presented at: International Conference on Medical Image Computing and Computer-Assisted Intervention; October 5-9, 2015, Munich, Germany, 2015.
27. Bakas S, Reyes M, Jakab A, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv*. 2018. https://arxiv.org/abs/1811.02629
28. Freling NJ, Molenaar WM, Vermey A, et al. Malignant parotid tumors: clinical use of MR imaging and histologic correlation. *Radiology*. 1992;185(3):691-696.
29. Thoeny HC, De Keyzer F, King AD. Diffusion-weighted MR imaging in the head and neck. *Radiology*. 2012;263(1):19-32.
30. Takumi K, Fukukura Y, Hakamada H, Ideue J, Kumagae Y, Yoshiura T. Value of diffusion tensor imaging in differentiating malignant from benign parotid gland tumors. *Eur J Radiol*. 2017;95:249-256.
31. Yabuuchi H, Matsuo Y, Kamitani T, et al. Parotid gland tumors: can addition of diffusion-weighted MR imaging to dynamic contrast-enhanced MR imaging improve diagnostic accuracy in characterization? *Radiology*. 2008;249(3):909-916.
32. Sheikh K, Lee SH, Cheng Z, et al. Predicting acute radiation induced xerostomia in head and neck cancer using MR and CT radiomics of parotid and submandibular glands. *Radiat Oncol*. 2019;14(1):131.
33. Pinker K, Shitano F, Sala E, et al. Background, current role, and potential applications of radiogenomics. *J Magn Reson Imaging*. 2018;47(3):604-620.
34. Chen NK, Guidon A, Chang HC, Song AW. A robust multi-shot scan strategy for high-resolution diffusion weighted MRI enabled by multiplexed sensitivity-encoding (MUSE). *NeuroImage*. 2013;72:41-47.

How to cite this article: Chang Y-J, Huang T-Y, Liu Y-J, Chung H-W, Juan C-J. Classification of parotid gland tumors by using multimodal MRI and deep learning. *NMR in Biomedicine*. 2021;34:e4408. https://doi.org/10.1002/nbm.4408