POLISH
JOURNAL *of* RADIOLOGY

http://www.polradiol.com

Original paper

Differentiation of carcinosarcoma from endometrial carcinoma on magnetic resonance imaging using deep learning

Tsukasa Saida[1,A,B,C,D,E,F], Kensaku Mori[1,D,E], Sodai Hoshiai[1,D,E], Masafumi Sakai[1,D,E], Aiko Urushibara[2,B,F], Toshitaka Ishiguro[1,B,D], Toyomi Satoh[3,A,E], Takahito Nakajima[1,A,E]

[1]Department of Radiology, Faculty of Medicine, University of Tsukuba, Tsukuba, Japan
[2]Department of Radiology, University of Tsukuba Hospital, Tsukuba, Japan
[3]Department of Obstetrics and Gynecology, Faculty of Medicine, University of Tsukuba, Tsukuba, Japan

Abstract

Purpose: To verify whether deep learning can be used to differentiate between carcinosarcomas (CSs) and endometrial carcinomas (ECs) using several magnetic resonance imaging (MRI) sequences.

Material and methods: This retrospective study included 52 patients with CS and 279 patients with EC. A deep-learning model that uses convolutional neural networks (CNN) was trained with 572 T2-weighted images (T2WI) from 42 patients, 488 apparent diffusion coefficient of water maps from 33 patients, and 539 fat-saturated contrast-enhanced T1-weighted images from 40 patients with CS, as well as 1612 images from 223 patients with EC for each sequence. These were tested with 9-10 images of 9-10 patients with CS and 56 images of 56 patients with EC for each sequence, respectively. Three experienced radiologists independently interpreted these test images. The sensitivity, specificity, accuracy, and area under the receiver operating characteristic curve (AUC) for each sequence were compared between the CNN models and the radiologists.

Results: The CNN model of each sequence had sensitivity 0.89-0.93, specificity 0.44-0.70, accuracy 0.83-0.89, and AUC 0.80-0.94. It also showed an equivalent or better diagnostic performance than the 3 readers (sensitivity 0.43-0.91, specificity 0.30-0.78, accuracy 0.45-0.88, and AUC 0.49-0.92). The CNN model displayed the highest diagnostic performance on T2WI (sensitivity 0.93, specificity 0.70, accuracy 0.89, and AUC 0.94).

Conclusions: Deep learning provided diagnostic performance comparable to or better than experienced radiologists when distinguishing between CS and EC on MRI.

Key words: artificial intelligence, magnetic resonance imaging, uterus, carcinosarcoma, malignant mixed Müllerian tumours.

Introduction

Uterine carcinosarcoma (CS) is a rare, biphasic tumour composed of high-grade carcinoma and sarcoma components, which account for less than 5% of all uterine tumours, while endometrial carcinoma (EC) represents approximately 95% of uterine tumours. The carcinomatous component is most often high-grade endometrioid or serous carcinomas, and uncommonly clear cell carcinoma [1].

The mesenchymal component is most commonly high-grade sarcoma not otherwise specified, but heterologous elements include rhabdomyosarcoma, chondrosarcoma, and osteosarcoma [1]. CS shows highly aggressive behaviour; at presentation, approximately 60% of patients with CS have advanced disease with extrauterine spread [2]. The prognosis of CS is also poor, with an overall 5-year survival rate of approximately 30% (stage I- II: 59%; stage III: 25%; stage IV: 9%) [3,4], in contrast to 90% in low-grade EC [5] and over 80% in stage I high grade EC [6]. Serous histology and heterologous rhabdomyoblastic differentiation have been reported to be significantly associated with worse survival [1].

Correspondence address:
Tsukasa Saida, Department of Radiology, Faculty of Medicine, University of Tsukuba, Tsukuba, Japan, e-mail: saida_sasaki_tsukasa@yahoo.co.jp

Because of sampling and biopsy errors due to the fact that carcinosarcoma contains various components, the preoperative radiologic diagnosis of CS is clinically important. This diagnosis guides optimal surgical planning and adjuvant therapy, including novel targeted immunotherapy [7]. According to previous studies investigating MRI findings of uterine CS, large exophytic (12-75%), heterogeneous endometrial masses both on T2-weighted imaging (T2WI) 71-82%) and T1-weighted imaging (T1WI) (18-50%) have been reported to be suggestive of CS [8-10]. In dynamic studies, CSs typically show areas of early and persistent marked enhancement (50-100%) [8-10]. However, despite significant differences in prognosis between CS and EC, MRI findings of CS are far from sufficient to distinguish them from EC, including the report by Bharwani *et al.* showing that 88% of CS were indistinguishable from EC [8].

Convolutional neural network (CNN) is a class of deep learning models that combine imaging filters with artificial neural networks through a series of successive linear and nonlinear layers. It is considered a promising tool for diagnostic imaging, and several CNN models for diagnostic imaging have been constructed [11-14]. The modalities studied include radiography, US, CT, and MRI. However, to the best of our knowledge, no study has built a deep learning model for diagnosing CSs using MRI; therefore, its clinical usefulness requires validation.

Here, we present CNN models for differentiating CSs from ECs on MRI, including T2WI, apparent diffusion coefficient of water (ADC) map, and contrast enhanced-T1WI (CE-T1WI), and compare their diagnostic performance with interpretations by experienced radiologists.

Material and methods

Patients

This retrospective study was approved by the institutional review board of University of Tsukuba Hospital, and the requirement for written informed consent was waived (Approval number: R02-054).

The inclusion criteria for the CS group were as follows: (a) pathologically proven CS by total hysterectomy and (b) a pelvic MRI in the oblique axial direction perpendicular to the long axis of the uterine body, performed at our hospital between January 2005 and December 2020. Note: Because CS is a rare tumour, cases of CS without diffusion-weighted imaging (DWI) or CE-T1WI were included. The inclusion criteria for the EC group were as follows: (a) pathologically proven EC by total hysterectomy and (b) a pelvic MRI including T2WI, DWI, and CE-T1WI in the oblique axial direction perpendicular to the long axis of the uterine body, performed at our hospital between January 2015 and December 2020. The exclusion criteria were as follows: (a) history of surgery other than caesarean section, chemotherapy, or radiation therapy of the uterus and (b) concomitant pregnancy. A flowchart for the patient selection process is shown in Figure 1.

MRI was performed using 3 T or 1.5 T equipment (Ingenia®, and Achieva®; Philips Medical Systems, Amsterdam, Netherlands). The imaging protocols for CS and EC are the same and included oblique axial T2WI, DWI with a *b* value of 0, 1000, and CE-T1WI with gadopentetate dimeglumine 5 mmol (Magnevist® or Gadovist®; Bayer, Wuppertal, Germany). The bolus intravenous contrast injection rate was 4 ml (2 mmol)/s (in case of Gadovist, diluted with saline solution and injected at 4 ml/s). Further details of these parameters are provided in Table 1.

Dataset

In order to detect the most suitable sequence for CNN to differentiate CS from EC, each sequence alone was targeted for this verification.

To create a dataset, only the slices in which the tumour was visualised were extracted based on the consensus of the 2 radiologists (T.S., A.U.) with reference to the pathology
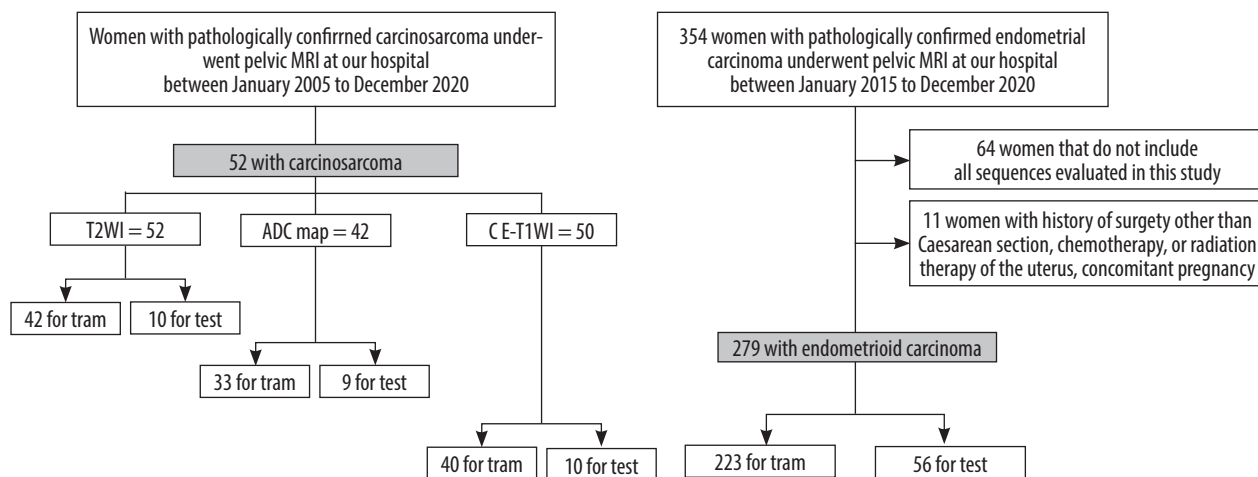

Figure 1. Flowchart for the patient selection process

Table 1. Magnetic resonance imaging acquisition parameters

Scanner	Sequence	Type	TR/TE (ms)	FA (degree)	Slice/Gap (mm)	FOV (mm)	Matrix
Ingenia® 3.0T	T2WI	2D-TSE	4955-6752/100-110	90	3-5/0.3-0.5	280	704 × 704
	DWI	EPI	6500-7500/77-79	90	3-5/0.3-0.5	280	224 × 224
	CE-T1WI	3D-GRE SPIR	4/2	10	3.3/1.6	280	576 × 576
Achiva® 1.5T	T2WI	2D-TSE	1400-6013/100-110	90	3-5/0.3-0.5	280	512 × 512-704 × 704
	DWI	EPI	3963-7500/70-78	90	3-5/0.3-0.5	280	224 × 224-256 × 256
	CE-T1WI	3D-GRE SPIR	5/2	15	2/1	280	320 × 320-576 × 576

CE-T1WI – contrast-enhanced fat-saturated T1-weighted imaging, DWI – diffusion-weighted imaging, EPI – echo planar imaging, FA – flip angle, FOV – field of view, GRE SPIR – gradient echo spectral pre-saturation with inversion recovery, T2WI – T2-weighted imaging, TE – echo time, TR – repetition time, TSE – turbo-spin echo

report. The same cross-section was extracted for all the sequences. A total of 331 patients were randomly assigned to the training and testing groups with the same ratio of CS to EC and a training-to-test ratio of approximately 5 : 1.

In the training phase of the CS group, 572 images of T2WI from 42 patients, 488 images of ADC map from 33 patients, and 539 images of CE-T1WI from 40 patients were used. In the training phase of the EC group, 1621 images of each sequence from 223 patients were used.

In the test phase, only 1 central image of the stacks was automatically extracted for each patient, and 66 images (10 images from 10 patients with CS and 56 images of 56 patients with EC) were used on T2WI and CE-T1WI, and 65 images (9 images from 9 patients with CS and 56 images from 56 patients with EC) were used on the ADC map.

In this study, the device used was unable to handle Digital Imaging and Communications in Medicine (DICOM) images, so the DICOM images were converted to Joint Photographic Experts Group (JPEG) images using the viewing software Centricity Universal Viewer (GE Healthcare, Chicago, IL, USA). Subsequently, the margins containing the patient information were automatically cropped and resized to 240 × 240 pixels using XnConvert (Gougelet Pierre-Emmanuel, Reims, France).

Deep learning with convolutional neural networks

Deep learning was performed on a deep station entry (UEI, Tokyo, Japan) with a GeForce RTX 2080Ti graphics processing unit (NVIDIA, Delaware, CA, USA), a Core i7-8700 central processing unit (Intel, Santa Clara, CA, USA), and deep-learning software Deep Analyzer (GHELIA, Tokyo, Japan) [15,16].

The conditions optimized based on the ablation and comparative studies of the previous research are as follows: CNN with Xception architecture [17] was used for deep learning. Xception is a novel architecture that has a similar parameter count as Inception V3, and inception modules have been replaced with depthwise separable convolutions. Therefore, the performance gains are not due to increased capacity but rather to more efficient use of model parameters. ImageNet [18] was used for pre-

training. Optimiser algorithm = Adam (learning rate = 0.0001, β1 = 0.9, β2 = 0.999, eps = le-7, decay = 0, AMS-Grad = false). The batch size was automatically selected by the Deep Analyzer to fit into the graphics processing unit (GPU) memory. Additionally, the horizontal flip, rotation (±4.5°), shearing (0.05), and zooming (0.05) were also automatically used as data augmentation techniques. The numbers of epochs used for the training were 50, 100, and 200. The validation/training ratio (validation ratio) was set at 0.1 or 0.2.

Radiologist interpretation

Three experienced radiologists (K.M., S.H., and M.S.) with 26, 12, and 8 years of experience in interpreting pelvic MRI independently reviewed the test images in a random order. Each image was evaluated by assigning confidence levels in the differentiation of CS and EC using a 6-point scale (0, definitely CS; 0.2, probably CS; 0.4, possibly CS; 0.6, possibly EC; 0.8, probably EC; and 1.0, definitely EC). They were blinded to pathological and clinical findings. A duration of a week or more was observed between each sequence interpretation. Prior to interpretation, several cases of CS and EC were presented, and the typical imaging findings of CS (large exophytic, heterogeneous endometrial masses with marked enhancement) were reconfirmed.

Statistical analysis

The age and tumour stage for each group were compared using the Mann-Whitney U test and χ^2 test.

The test dataset was used to calculate the sensitivity, specificity, and accuracy of the CS diagnoses. In radiologists, a value of 0-0.4 was treated as CS and 0.6-1.0 was treated as EC. In CNN, the classification into CS and EC groups was by output as a continuous number from 0 to 1, where 0-0.49 was considered as CS and 0.50-1.00 was considered as EC. Receiver operating characteristic (ROC) analysis was performed to assess the diagnostic performance [19]. For statistics, 95% confidence intervals (CIs) and significance difference were estimated. In addition, the inter-observer agreement for the 2 choic-

es of CS or EC was assessed using kappa (κ) statistics. The κ-statistic interpreted the agreement as follows: less than 0, no; 0-0.20, slight; 0.21-0.40, fair; 0.41-0.60, moderate; 0.61-0.80, substantial; and 0.81-1.00, almost perfect [20].

All statistical analyses were performed using SPSS software (SPSS Statistics 27.0; IBM, New York, NY, USA). The statistical significance was set at $p < 0.05$.

Results

A total of 331 women (mean age 59 years; age range 30-91 years) were evaluated across the datasets. Table 2 shows the patients' characteristics, clinical stages, and image slices of CS and EC lesions. There was no significant difference between the CS and EC groups and training and testing data regarding patient age and tumour stage. In the CS group, 7 (train 5, test 2) patients were scanned at 1.5 T. In the EC group, 13 (train 10, test 3) patients were scanned at 1.5 T. The other patients were scanned at 3 T.

For the selection of the validation ratio and epochs among the models, a model with a validation ratio of 0.2 and 100 epochs was adopted for T2WI (Figure 2), a model with a validation ratio of 0.1 and 100 epochs was adopted for the ADC map, while a model with a validation ratio of 0.2 and 200 epochs was adopted for CE-T1WI, owing to their high diagnostic performance.

Table 3 presents the results of the interpretation. Table 4 lists the diagnostic performances of the CNN models and the 3 radiologists in diagnosing CS. The ROC curves comparing the performance of the CNN models and the radiologists are shown in Figure 3; the sensitivity, specificity, accuracy, and AUC of the CNN model of each sequence in diagnosing CS were comparable to those of Reader 1 and superior to those of the other 2 readers. The CNN models and Reader 1 were able to differentiate between CS and EC, with significant differences in all sequences. And the AUC was significantly higher on T2WI and CE-T1WI for the CNN model than for Readers 2 and 3. The CNN model showed the highest diagnostic perfor-

**Table 2.** Characteristics of patients and lesions

Parameter	Training data			Test data		
	CS	EC	All	CS	EC	All
Patients, n	42	223	265	10	56	66
Age, years						
Mean (SD)	66 (11.60)	53 (11.23)	59 (12.13)	65 (8.34)	60 (14.00)	61 (13.44)
Range	39-86	30-71	30-86	50-75	31-91	31-91
Pathological type, n						
CS						
Homologous differentiation	17			3		
Heterologous differentiation	22			5		
N/A	3			2		
EC						
Endometrioid carcinoma Grade 1		129			33	
Endometrioid carcinoma Grade 2		55			12	
Endometrioid carcinoma Grade 3		21			10	
Other		18			1	
Clinical stage, n						
I	15	138		5	40	
II	4	26		0	4	
III	12	38		0	6	
IV	11	21		5	6	
Images, n/slices						
T2WI	42/572	223/1,612	265/2,184	10/10	56/56	66/66
ADC map	33/488	223/1,612	256/2,100	9/9	56/56	65/65
CE-T1WI	40/539	223/1,612	263/2,151	10/10	56/56	66/66

ADC – apparent diffusion coefficient, CE-T1WI – contrast-enhanced fat-saturated T1-weighted imaging, CS – carcinosarcoma, EC – endometrial carcinoma, SD – standard deviation, T2WI – T2-weighted imaging

mance on the T2WI with an AUC of 0.94 when comparing the sequences and interpreters.

Table 5 shows the inter-observer agreement between the CNN models and the 3 radiologists. The CNN model and Reader 1 of T2WI and CE-T1WI and Reader 2 and 3 of all sequences showed a fair agreement. The other studies showed no agreement.

Figures 4-6 show the test images of this study with interpretations of the CNN model and the radiologists, including the confidence of the CNN model and the radiologists. Figure 4 is a typical CS case, where the CNN model and Reader 1 diagnosed CS on T2WI. On the other hand, only the CNN model could not diagnose CS on the ADC map, and all but Reader 2 diagnosed CS on CE-T1WI. Figure 5 presents a case of CS of a huge tumour with internal necrosis, where all but Reader 2 diagnosed CS on T2WI and ADC map, and only the CNN model diagnosed CS on CE-T1WI. Figure 6 shows a case of EC with haemorrhage in the endometrial cavity, where the CNN model diagnosed EC only on CE-T1WI.
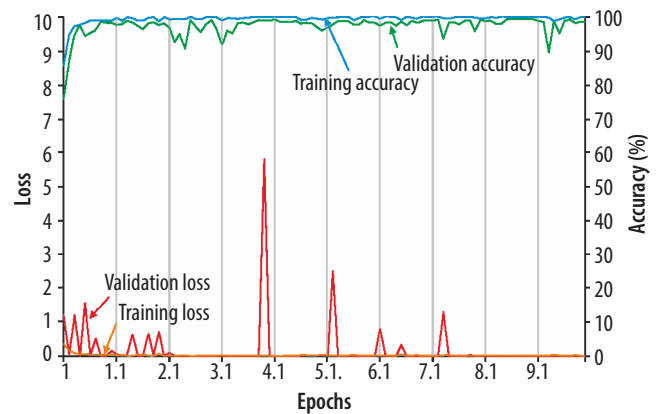
Throughput: 126 epochs/hr

● Validation acc.: 98.6%   ● Validation loss: 0.0
● Training acc.: 100.0%   ● Training loss: 4.16e-4

Figure 2. Accuracy and loss of the training data (T2-weighted imaging with validation ratio 0.2, epochs 100)

**Table 3.** Interpretation results of the CNN model and the radiologists

		CNN		Reader 1		Reader 2		Reader 3	
		CS	EC	CS	EC	CS	EC	CS	EC
T2WI	CS	7	3	7	3	3	7	6	4
	EC	4	52	9	47	13	43	31	25
ADC map	CS	4	5	7	2	3	6	5	4
	EC	6	50	7	49	13	43	32	24
CE-T1WI	CS	7	3	7	3	3	7	6	4
	EC	4	52	5	51	13	43	31	25

ADC – apparent diffusion coefficient, CE-T1WI – contrast-enhanced fat-saturated T1-weighted imaging, CNN – convolutional neural network, CS – carcinosarcoma, EC – endometrial carcinoma, T2WI – T2-weighted imaging

Table 4. Sensitivity, specificity, and AUC of the CNN model and the radiologists in diagnosing CS

Image set	Interpreter	Sensitivity	Specificity	Accuracy	AUC	*p*-value for AUC (vs. CNN model)
T2WI	CNN	0.93 (0.88-0.96)	0.70 (0.44-0.87)	0.89 (0.88-0.96)	0.94 (0.88-1.00)	–
	Reader 1	0.84 (0.79-0.87)	0.70 (0.43-0.88)	0.81 (0.74-0.87)	0.86 (0.73-0.99)	0.101
	Reader 2	0.77 (0.73-0.82)	0.30 (0.11-0.57)	0.70 (0.64-0.78)	0.52 (0.32-0.72)	0.001*
	Reader 3	0.45 (0.40-0.49)	0.60 (0.33-0.82)	0.47 (0.39-0.54)	0.52 (0.33-0.70)	0.001*
ADC map	CNN	0.89 (0.85-0.93)	0.44 (0.21-0.69)	0.83 (0.76-0.90)	0.80 (0.65-0.95)	–
	Reader 1	0.88 (0.83-0.90)	0.78 (0.49-0.93)	0.86 (0.83-0.90)	0.91 (0.83-0.98)	0.225
	Reader 2	0.77 (0.74-0.81)	0.33 (0.13-0.62)	0.71 (0.65-0.79)	0.53 (0.31-0.74)	0.063
	Reader 3	0.43 (0.38-0.47)	0.56 (0.28-0.80)	0.45 (0.38-0.47)	0.49 (0.26-0.72)	0.051
CE-T1WI	CNN	0.93 (0.88-0.96)	0.70 (0.44-0.87)	0.89 (0.82-0.95)	0.89 (0.77-1.00)	–
	Reader 1	0.91 (0.86-0.94)	0.70 (0.44-0.88)	0.88 (0.80-0.93)	0.92 (0.85-0.99)	0.532
	Reader 2	0.77 (0.73-0.82)	0.30 (0.11-0.57)	0.70 (0.64-0.78)	0.56 (0.38-0.75)	0.001*
	Reader 3	0.45 (0.40-0.49)	0.60 (0.33-0.82)	0.47 (0.39-0.54)	0.55 (0.38-0.73)	0.001*

ADC – apparent diffusion coefficient, AUC – area under the receiver operating characteristic curve, CE-T1WI – contrast-enhanced fat-saturated T1-weighted imaging, CI – confidence interval, CNN – convolutional neural network, T2WI – T2-weighted imaging. Data in parentheses are 95% confidence interval. *$p < 0.05$
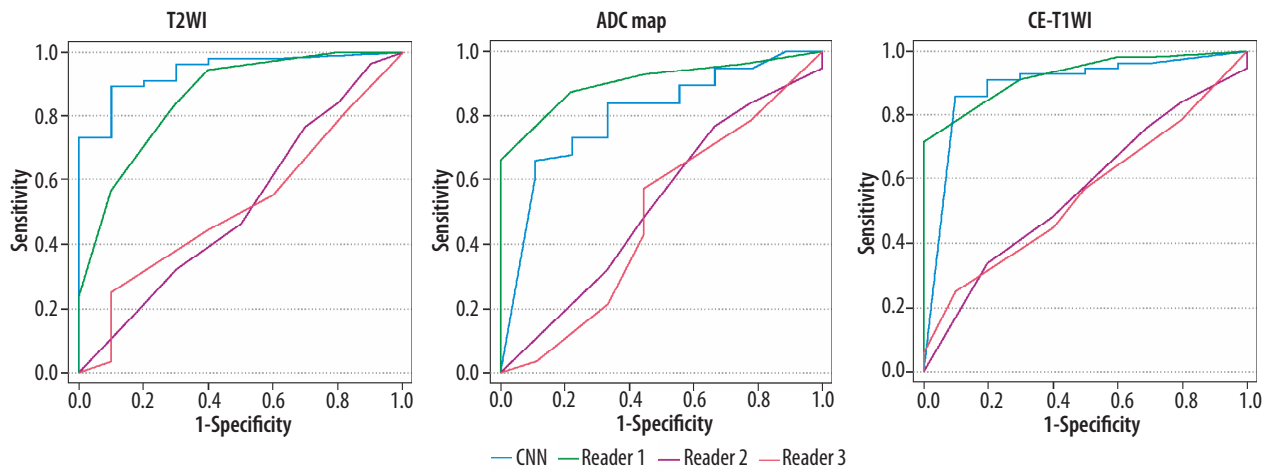
ADC – apparent diffusion coefficient, CE-T1WI – contrast-enhanced fat-saturated T1-weighted imaging, T2WI – T2-weighted imaging

Figure 3. Receiver operating characteristic (ROC) curves for the performance of the convolutional neural network (CNN) model compared to radiologists' performance

**Table 5.** Inter-observer agreement between the CNN model and the radiologists

		Reader 1	Reader 2	Reader 3
T2WI	**CNN**	0.59	0.25	0.05
	Reader 2	0.16	–	–
	Reader 3	0.01	0.40	–
ADC map	**CNN**	0.09	0.05	0.07
	Reader 2	0.04	–	–
	Reader 3	0.01	0.40	–
CE-T1WI	**CNN**	0.53	0.03	0.05
	Reader 2	0.10	–	–
	Reader 3	0.13	0.40	–

ADC – apparent diffusion coefficient, CE-T1WI – contrast-enhanced fat-saturated T1-weighted imaging, CNN – convolutional neural network, T2WI – T2-weighted imaging

Discussion

This study presented CNN models to distinguish between CS and EC using several MRI sequences, which demonstrated almost equal diagnostic performance compared to the most experienced radiologist, Reader 1, and better diagnostic performance compared to Readers 2 and 3.

Few studies have used deep learning in the field of uterine tumour MRI. Urushibara et al. developed a CNN that can differentiate between cervical cancer and non-cancerous lesions on single sagittal T2WI and showed high diagnostic performance [16]. A study by Chen et al. [21] and Don et al. [22] used T2WI, T2WI, and CE-T1WI, respectively, to evaluate the myometrial infiltration of ECs using CNN and showed comparable diagnostic performance compared to radiologists. To the best of our knowledge, this
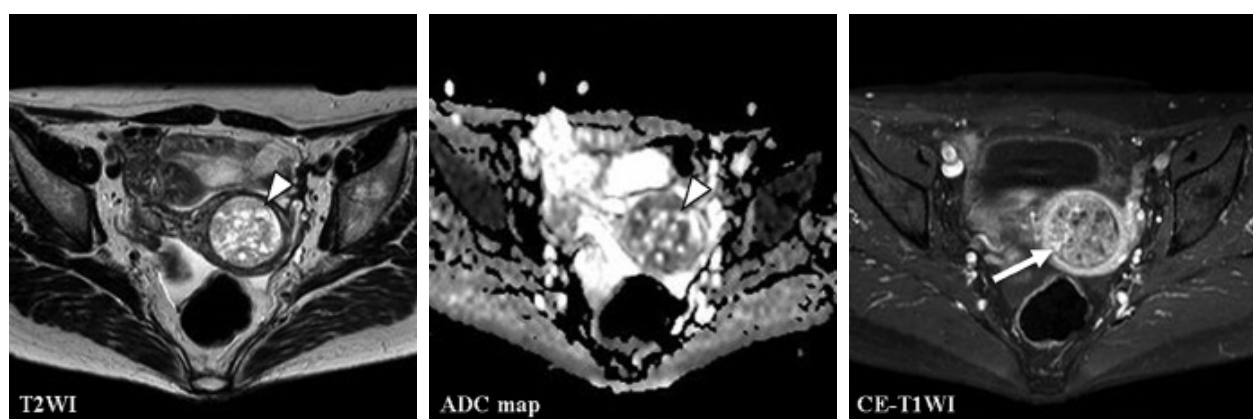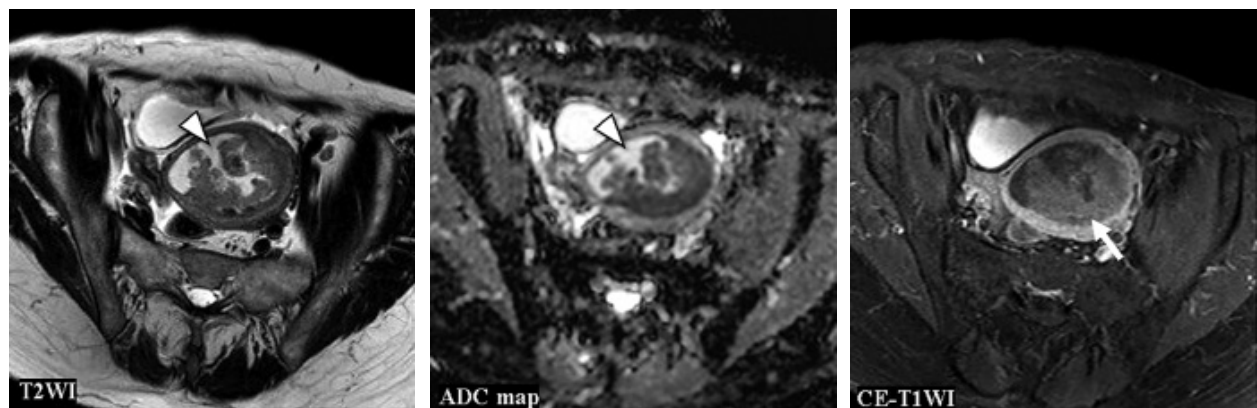

ADC – apparent diffusion coefficient, CE-T1WI – contrast-enhanced fat-saturated T1-weighted imaging, CNN – convolutional neural network, CS – carcinosarcoma, EC – endometrial carcinoma, T2WI – T2-weighted imaging

Figure 4. A 72-year-old woman with carcinosarcoma showing heterologous differentiation of stage IVB. T2WI: the CNN model = CS (confidence; CS = 100.0%), Reader 1 = CS (confidence; CS = 100.0%), Reader 2 = EC (confidence; CS = 0.0%), Reader 3 = EC (confidence; CS = 40.0%). ADC map: the CNN model = EC (confidence; CS = 0.2%), Reader 1 = CS (confidence; CS = 80.0%), Reader 2 = CS (confidence; CS= 60.0%), Reader 3 = CS (confidence; CS = 80.0%). CE-T1WI: the CNN model = CS (confidence; CS = 100.0%), Reader 1 = CS (confidence; CS = 100.0%), Reader 2 = EC (confidence; CS = 20.0%), Reader 3 = CS (confidence; CS = 100.0%). The tumour showed relatively heterogeneous hyperintensities both on T2WI and ADC map (arrow heads); additionally, some parts were strongly enhanced on CE-T1WI (arrow), which were typical appearances of CS. However, it seemed difficult to distinguish it from EC associated with endometrial polyp. It may also have been difficult to distinguish from adenomyosis with the ADC map alone
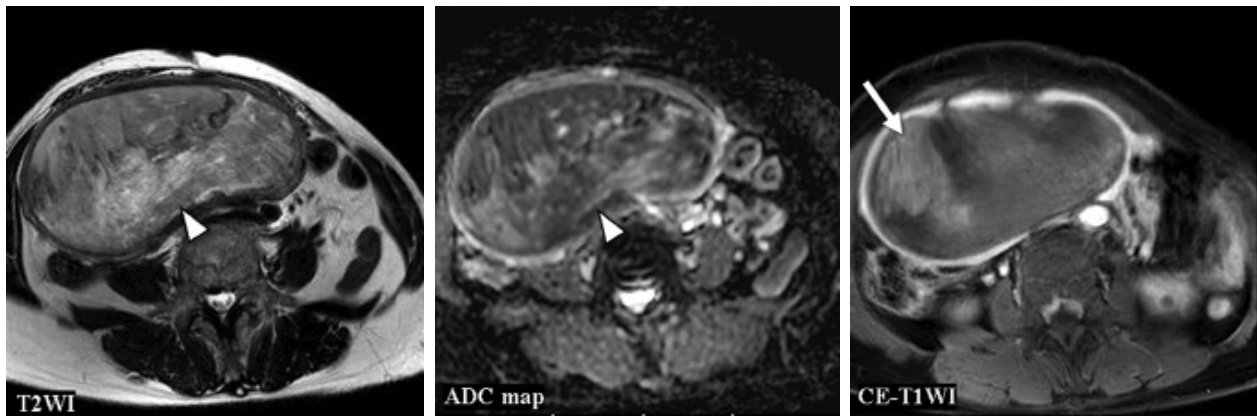
ADC – apparent diffusion coefficient, CE-T1WI – contrast-enhanced fat-saturated T1-weighted imaging, CNN – convolutional neural network, CS – carcinosarcoma, EC – endometrial carcinoma, T2WI – T2-weighted imaging

Figure 5. A 74-year-old woman with carcinosarcoma showing heterologous differentiation of stage IVB. T2WI: the CNN model = CS (confidence; CS = 100.0%), Reader 1 = CS (confidence; CS = 100.0%), Reader 2 = EC (confidence; CS = 40.0%), Reader 3 = CS (confidence; CS = 60.0%). ADC map: the CNN model = CS (confidence; CS = 69.8%), Reader 1 = CS (confidence; CS = 100.0%), Reader 2 = EC (confidence; CS = 40.0%), Reader 3 = CS (confidence; CS = 100.0%). CE-T1WI: the CNN model = CS (confidence; CS= 100.0%), Reader 1 = EC (confidence; CS = 40.0%), Reader 2 = EC (confidence; CS = 40.0%), Reader 3 = EC (confidence; CS = 40.0%). Huge tumour with exophytic growth was accompanied by extensive necrosis and showed heterogeneous signal intensities both on T2WI and ADC map (arrow heads). In CE-T1WI, the tumour enhancing effect was poorer than that in the myometrium (arrow), and no enhancing effect was observed in the necrotic area. It is presumed that all readers diagnosed EC because the tumour enhancing effect was poor on CE-T1WI

ADC – apparent diffusion coefficient, CE-T1WI – contrast-enhanced fat-saturated T1-weighted imaging, CNN – convolutional neural network, CS – carcinosarcoma, EC – endometrial carcinoma, T2WI – T2-weighted imaging

Figure 6. A 63-year-old woman with endometrioid carcinoma Grade 1 of stage IVB. T2WI: the CNN model = CS (confidence; CS= 100.0%), Reader 1 = CS (confidence; CS = 60.0%), Reader 2 = EC (confidence; CS= 0.0%), Reader 3 = CS (confidence; CS = 60.0%). ADC map: the CNN model = CS (confidence; CS = 96.6%), Reader 1 = EC (confidence; CS = 40.0%), Reader 2 = EC (confidence; CS = 40.0%), Reader 3 = CS (confidence; CS = 100.0%). CE-T1WI: the CNN model = EC (confidence; CS = 0.1%), Reader 1 = CS (confidence; CS = 60.0%), Reader 2 = EC (confidence; CS = 40.0%), Reader 3 = CS (confidence; CS = 80.0%). Haemorrhage in the endometrial cavity resulted in heterogeneity (arrow heads), which may have made interpretation difficult. However, it may have been possible to make a correct diagnosis because the tumour showed a relatively homogeneous poor enhancement on CE-T1WI (arrow)

is the first study on the diagnosis of CS in deep learning using MRI.

Large exophytic heterogeneous endometrial masses have been reported to be suggestive of CS [23]. Haemorrhagic necrosis is often observed in large masses, leading to heterogeneity on T2WI and T1WI [8-10]. In dynamic studies, CS typically shows areas of early and persistent marked enhancement, similar to that of the myometrium, and/or a gradual and delayed strong enhancement, contrary to EC with consistently poor enhancement [8-10]. According to previous studies of Bharwani *et al.* with 51 CS cases [8], Tanaka *et al.* with 17 CS cases [9], and Ohguri *et al.* with 4 CS cases [10], large exophytic appearance was seen in 12% [8], 100% [9], and 75% [10], heterogeneity on T2WI was seen in 82% [8], 71% [9], and 75% [10], heterogeneity on T1WI was seen in 33% [8], 18% [9], and 50% [10], heterogeneity on CE-T1WIs was seen in 58% [8], 50% [9], and 100% [10], and delayed strong enhancement was seen in 50% [8], 81% [9], and 100% [10], repsectively. Garza *et al.* generated time-intensity curves for tumour and surrounding myometrial regions of interest in 37 patients with CS and 42 patients with EC, measuring the positive enhancement integral (PEI), maximum slope of increase (MSI), and signal enhancement ratio (SER). The threshold PEI ratio ≥ 0.67 predicted CS with 76% sensitivity, 84% specificity, and 0.83 AUC, and the

threshold SER ≤ 125 predicted CS with 90% sensitivity, 50% specificity, and 0.72 AUC; however, tumour MSI did not significantly differ between the disease groups [24]. It has been reported that there was no significant difference between the mean and minimum ADC values of CS and EC, although the mean ADC of CS was significantly higher than that of higher grade (Grade 2 and 3) EC [25], and CS demonstrated heterogenous signal intensities on DWI and ADC map reflecting complicated tissue components [26]. In summary, heterogeneous signal intensities indicating haemorrhagic necrosis, and strongly enhanced areas are predictors of uterine CSs on MRI [10]. However, the diagnostic performance of imaging features is not perfect. Especially when the tumour is small or lacks these typical diagnostic imaging features, preoperative diagnosis is not always possible and there is an oportunity for the use of deep learning to assist in interpretation.

The MRI findings of CS are not sufficient to differentiate them from EC. In addition, our CNN models consisted of only 42 CS cases, and although the images used were not cropped images of the uterus only, our CNN models showed diagnostic performances equivalent to those of experienced radiologists. The CNN model showed the highest diagnostic performance with T2WI. In contrast, the low diagnostic performance on ADC map compared to the other sequences was inconsistent with previous reports [27], and it was presumed to result from the relatively small number of training ADC maps.

The inter-observer agreement between the readers, not only between the CNN model and the readers, was generally low. This indicates that the difference in the imaging findings between CS and EC was not standardised or generalised even among experienced radiologists, and CS and EC could not be clearly classified.

Our study had several limitations. First, the number of CSs was small; however, this was the largest number of MRI studies of CSs to date. Second, each sequence was evaluated individually, which was quite different from the clinical practice wherein all sequences were referenced and comprehensively diagnosed. However, our unpublished data indicated that this case size could not be expected to improve interpretation by using a combined image set. Also, unlike radiologists, one of the CNNs' strengths may be the ability to demonstrate high diagnostic performance with only one image of one type of sequence. In addition, even when diagnosing with combined sequences by CNN, weighting each sequence will be required. Third, due to the rarity of CS, the period of adoption for CS and EC was different. Even if there is a large amount of data on one side, good results cannot be obtained. In addition, the number of images of each sequence available in CS was also different, which may have affected the results. Finally, we did not examine T1WI and dynamic studies, to avoid study complexity. Although T1WI is useful for assessing haemorrhage, it was not included in this study because the scan protocol did not include oblique axial T1WI. Moreover, we considered that the same information is included in CE-T1WI. Dynamic studies have been reported to help distinguish between CS and EC; however, we believe that delayed isolation and hyperenhancement in the late contrast phase are of utmost importance to distinguish between the 2, and that dynamic study is not essential.

The following can be considered as future improvements: For rare tumours, such as CS, it is necessary to accumulate cases at multiple centres and increase the number of images. As mentioned in the study limitations, deep-learning models trained using a combination of several sequences may show superior diagnostic performance to the current models, as reported for the CNN model of prostate cancer using fused ADC and T2WI [28]; however, additional training images may be required to achieve a higher diagnostic performance with combination images. Evaluating series images as well as learning with DICOM data can also improve diagnostic performance, and greater versatility can be achieved by using images taken with other MRI instruments. In this study, we used ImageNet [18], which comprised natural images as pre-trained data, and the diagnostic performance could be further improved if transfer learning using medical image training data [29] was used.

Conclusions

CS is much less prevalent than EC and challenging to distinguish from EC; however, under the limited conditions of this study, deep learning provided comparable or better diagnostic performance to experienced radiologists when distinguishing between CS and EC on MRI. Although further validation is needed, deep learning has the potential to help with highly specialized MRI image interpretation.

Disclosure

The authors report no conflict of interest.

References

1. WHO Classification of Tumours Editorial Board. Female Genital Tumours WHO Classification of Tumours, 5th ed. Lyon: World Health Organization; 2020, p. 32-76.
2. Cantrell LA, Blank SV, Duska LR. Uterine carcinosarcoma: a review of the literature. Gynecol Oncol 2015; 137: 581-588.
3. Gonzalez BJ, Terstriep SA, Cliby WA, et al. The impact of multimodal therapy on survival for uterine carcinosarcomas. Gynecol Oncol 2010; 116: 419-423.
4. Callister M, Ramondetta LM, Jhingran A, et al. Malignant mixed Mullerian tumors of the uterus: analysis of patterns of failure, prog-

nostic factors, and treatment outcome. Int J Radiat Oncol Biol Phys 2004; 58: 786-796.

5. Sonoda Y, Barakat RR. Screening and the prevention of gynecologic cancer: endometrial cancer. Best Pract Res Clin Obstet Gynaecol 2006; 20: 363-377.

6. Soslow RA, Bissonnette JP, Wilton A, et al. Clinicopathologic analysis of 187 high-grade endometrial carcinomas of different histologic subtypes: similar outcomes belie distinctive biologic differences. Am J Surg Pathol 2007; 31: 979-987.

7. Kanthan R, Senger JL. Uterine carcinosarcomas (malignant mixed mullerian tumours): a review with special emphasis on the controversies in management. Obstet Gynecol Int 2011; 470795. doi: 10.1155/2011/470795.

8. Bharwani N, Newland A, Tunariu N, et al. MRI appearances of uterine malignant mixed müllerian tumors. AJR Am J Roentgenol 2010; 195: 1268-1275.

9. Tanaka YO, Tsunoda H, Minami R, Yoshikawa H, Minami M. Carcinosarcoma of the uterus: MR findings. J Magn Reson Imaging 2008; 28: 434-439.

10. Ohguri T, Aoki T, Watanabe H, et al. MRI findings including gadolinium-enhanced dynamic studies of malignant, mixed mesodermal tumors of the uterus: differentiation from endometrial carcinomas. Eur Radiol 2002; 12: 2737-2742.

11. Chartrand G, Cheng PM, Vorontsov E, et al. Deep learning: a primer for radiologists. Radiographics 2017; 37: 2113-2131.

12. Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. Z Med Phys 2019; 29: 102-127.

13. Soffer S, Ben-Cohen A, Shimon O, et al. Convolutional neural networks for radiologic images: a radiologist's guide. Radiology 2019; 290590-290606. doi: 10.1148/radiol.2018180547.

14. Shin HC, Roth HR, Gao M, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 2016; 35: 1285-1298.

15. Fujioka T, Katsuta L, Kubota K, et al. Classification of breast masses on ultrasound shear wave elastography using convolutional neural networks. Ultrason Imaging 2020; 42213-42220. doi: 10.1177/0161734620932609.

16. Urushibara A, Saida T, Mori M, et al. Diagnosing uterine cervical cancer on a single T2-weighted image: Comparison between deep learning versus radiologists. Eur J Radiol 2020; 135109471. doi: 10.1016/j.ejrad.2020.109471.

17. Chollet F. Xception: deep learning with depthwise separable convolutions. rXiv. 2017:1610.02357v3. Available at: https://arxiv.org/pdf/1610.02357.pdf.

18. Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. International Journal of Computer Vision 2015; 115: 211-252.

19. Linden A. Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristic (ROC) analysis. J Eval Clin Pract 2006; 12: 132-139.

20. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159-174.

21. Chen X, Wang Y, Shen M, et al. Deep learning for the determination of myometrial invasion depth and automatic lesion identification in endometrial cancer MR imaging: a preliminary study in a single institution. Eur Radiol 2020; 30: 4985-4994.

22. Dong HC, Dong HK, Yu MH, et al. Using deep learning with convolutional neural network approach to identify the invasion depth of endometrial cancer in myometrium using MR images: a pilot study. Int J Environ Res Public Health 2020; 175993. doi: 10.3390/ijerph17165993.

23. Shapeero LG, Hricak H. Mixed müllerian sarcoma of the uterus: MR imaging findings. Am J Roentgenol 1989; 153: 317-319.

24. Garza A, Elsherif SB, Faria SC, et al. Staging MRI of uterine malignant mixed Müllerian tumors versus endometrial carcinomas with emphasis on dynamic enhancement characteristics. Abdom Radiol (NY) 2020; 454: 1141-1154.

25. Takeuchi M, Matsuzaki K, Harada M. Carcinosarcoma of the uterus: MRI findings including diffusion-weighted imaging and MR spectroscopy. Acta Radiol 2016; 57: 1277-1284.

26. Kato H, Kanematsu M, Furui T, et al. Carcinosarcoma of the uterus: radiologic-pathologic correlations with magnetic resonance imaging including diffusion-weighted imaging. Magn Reson Imaging 2008; 26: 1446-1450.

27. Aldoj N, Lukas S, Dewey M, Penzkofer T. Semi-automatic classification of prostate cancer on multi-parametric MR imaging using a multi-channel 3D convolutional neural network. Eur Radiol 2020; 30: 1243-1253.

28. Le MH, Chen J, Wang L, et al. Automated diagnosis of prostate cancer in multi-parametric MRI based on multimodal convolutional neural networks. Phys Med Biol 2017: 626497-626514.

29. Wang X, Peng Y, Lu L, et al. ChestX-Ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. IEEE CVPR 2017: 2097-2106. Available at: https://arxiv.org/pdf/1705.02315.pdf.