Ovarian cancer (OC) is the leading cause of gynecologic cancer deaths and high-grade serous ovarian cancer (HGSOC) is the most common and most lethal histological type [1,2]. The lethality of HGSOC mainly comes from its high risk of recurrence [3,4]. Consequently, preoperative identification of recurrence in patients with HGSOC is important since it guides the personalized treatment and surveillance planning, such as selecting the agent of chemotherapy [5]. To this end, prognostic biomarkers related to recurrence of HGSOC are needed. Current studies indicated that clinical characteristics such as stage of International Federation of Gynecology and Obstetrics (FIGO), and preoperative serum cancer antigen (CA-125) were associated with recurrence of HGSOC [5,6,7]. However, the clinical biomarkers are invasive and provide only limited information about tumor due to the spatial and temporal pathologic heterogeneity of tumor [8,9].

Benefiting from the ability to noninvasively visualize a cancer's appearance on a macroscopic level, medical imaging demonstrated strong prognostic value [10–12]. Computed tomography (CT), as a...
Deep learning (DL) [17] as an artificial intelligence method has shown promising results in detecting valuable features from medical imaging [18–20]. The computational units in DL are defined as layers that are integrated together to mine the intrinsic characteristics of images [21]. Through a hierarchical neural network structure and convolutional operation, DL extracts the intrinsic characteristics of tumor that showed promising prognostic value [22]. Further study demonstrated that DL feature showed superior performance in comparison to hand-crafted image feature [23].

In this study, we explore an unsupervised DL method to extract the prognostic biomarkers of HGSOC from preoperative CT images, aiming at providing a non-invasive individualized recurrence prediction model in HGSOC. We hypothesize that the features learned by the unsupervised DL method can describe the tumor state thoroughly and reflect the intrinsic characteristics of tumor. Therefore, the DL feature contains much prognostic information on HGSOC. Moreover, the DL method can extract prognostic biomarkers of tumor requiring only tumor image data, which provides a new prognostic analysis method that enables us to utilize a large amount of data without follow-up.

Materials and methods

Patients

The institutional review board approval was granted for this retrospective study, and the requirement for informed consent was waived. We studied 245 patients who were pathologically confirmed to have primary HGSOC from the West China Second University Hospital of Sichuan University (WCSUH-SU, n = 200, between February 2010 and September 2015) and Henan Provincial People’s Hospital (HPPH, n = 45, between May 2012 and October 2016). The contrast-enhanced CT scanning was acquired for all the patients at diagnosis time, and the venous phase image was used for this study (CT scanning parameters in Supplementary Methods 1). Supplementary Methods 2 and Supplementary Fig. S1 provided detailed inclusion and exclusion criteria along with the recruitment pathway.

All the patients underwent primary debulking surgery [3,24], and were followed up every 2–4 months for the first two years, every 3–6 months from the third year, and annually from the 5th year [25]. The date of last follow-up was August 23, 2017 for the WCSUH-SU dataset and March 23, 2018 for the HPPH dataset. The endpoint of this study was recurrence, which was diagnosed combining clinical symptoms, rising CA-125 levels, and radiological findings. The time between clinical remission and first recurrence is defined as recurrence-free survival (RFS) [6,7]. The preoperative clinical characteristics were collected from the Institutional Picture Archiving and Communication System (PACS), including age, preoperative CA-125, FIGO stage, tumor location and maximum tumor diameter (Table 1).

In the WCSUH-SU dataset, patients without follow-up or follow-up time less than 3 years were included into the feature-learning cohort; while patients with 3+ year’s complete follow-up information were allocated into a primary cohort and a validation cohort according to the diagnosis time. The first 49 patients who underwent preoperative CT examination between 2010 and 2013 comprised the primary cohort, while the other 49 patients who underwent preoperative CT examination between 2014 and 2015 comprised the validation cohort 1. In the HPPH dataset, all the 45 consecutive patients were used for independent external validation (validation cohort 2). The median follow-up time was 32.83 months for the primary cohort, 31.07 months for the validation cohort 1, and 17.83 months for the validation cohort 2.

Development of the deep learning network

We proposed a novel DL network to extract the intrinsic characteristics of HGSOC from preoperative CT images (Feature learning in Fig. 1). The proposed DL network requires only the CT scanning of HGSOC without follow-up information, which is defined as unsupervised learning. The main computational processes of DL are convolution, pooling, activation and batch normalization, which are presented in the Supplementary Methods 3. Specifically, the DL network used a convolutional auto-encoder structure [26] that includes an encoder network and a decoder network. The encoder network includes an initial convolutional layer with 24 filters and four subsequent convolutional layers with 16 filters. To accelerate training and avoid covariate shift in the network, a batch normalization layer is inserted between two adjacent convolutional layers. In addition, we used average pooling between convolutional layers to eliminate redundant features. Finally, the encoder network transforms a tumor image into a 16-dimensional mineable feature vector. We refer to this vector as DL feature. Meanwhile, the decoder network uses the DL feature to reconstruct the original tumor image, aiming at evaluating the information capacity included in the DL feature. The decoder network is similar to the encoder network except that it upsamples image by deconvolutional layer. If the DL feature extracts the intrinsic characteristics of tumor, the decoder network should be able to reconstruct the original tumor image from the DL feature. To achieve this goal, we trained the DL network in 102 patients from the feature-learning cohort. A radiologist (5+ years’ experience, Y. Rong) located tumor areas in all the CT slices from the 102 patients using a rectangle bounding box (region of interest, ROI), resulting in 8917 tumor images to train the DL network. The network training is an iterative process, which optimizes the network iteratively until it extracts the intrinsic characteristics of HGSOC (details in Supplementary Methods 4).

Deep learning feature extraction

When the DL network is well trained in the feature-learning cohort, we applied it to transform CT image slices of HGSOC into 16-dimensional DL feature. First, we selected the ROI of tumor for each patient according to the following rule: the ROI should cover the primary tumor area in ovaries. If multiple tumor areas are observed in ovaries, multiple ROIs will be selected (Y. Rong and Y. Bai selected the ROIs in the WCSUH-SU dataset and the HPPH dataset respectively, Supplementary Fig. S2 illustrated the ROIs selected by the radiologists). Afterward, the tumor image was standardized by z-score normalization and scaled to 64 × 64 voxel size, and fed into the DL network. The output of the last convolutional layer in the encoder network was extracted as DL feature [22], which was 16-dimensional. Since the tumor image included multiple 2-dimensional slices, we averaged features from all image slices to acquire the DL feature for the patient.

Reurrence analysis

To evaluate the prognostic value of the DL feature, we used a multivariate Cox-PH regression to build the association between the DL feature and recurrence of HGSOC (recurrence analysis in Fig. 1). We trained the Cox-PH model using data (DL feature, recurrence time and status) from the primary cohort, and then validated
its performance in the two independent validation cohorts. The Cox-PH model using the DL feature was defined as the DL-CPH model in this study.

For each patient, the Cox-PH model predicts a hazard score indicating the individual recurrence risk. This hazard score was used for recurrence-free survival prediction and to stratify patients into high- and low-risk groups concerning recurrence. In addition, The Cox-PH can predict the recurrence probability of patient in a specific time point. In this study, we used the Cox-PH model to predict the 3-year recurrence probability.

Comparing deep learning feature with clinical characteristics

Since clinical characteristics were used as prognostic biomarkers in HGSOC [6,7], we compared the prognostic value of the DL feature to the clinical information. We built a clinical model involving age, FIGO stage, preoperative CA-125, tumor location and tumor diameter as features, and Cox-PH regression for recurrence prediction.

Since clinical characteristics and CT imaging reflect HGSOC from different perspectives, we therefore explore the combination of these two information. Among the five clinical characteristics, we used backward step-wise selection with the likelihood ratio test to select clinical predictors, which employed Akaike information criterion as the stopping rule [27]. Afterward, we built a Cox-PH model combining the DL-predicted hazard score and clinical predictor, and defined it as combined model.

Statistical analysis

We used Harrell’s concordance-index (C-Index) to measure the concordance between the DL-predicted recurrence risk and the actual recurrence time. A C-index score around 0.70 indicates a good model, whereas a score around 0.50 means random results without predictive performance [22]. When assessing the 3-year recurrence prediction, we used area under the receiver operating characteristic curve (AUC) and accuracy to evaluate the discriminatory performance of the DL-CPH model. Moreover, calibration
curves accompanied by the Hosmer–Lemeshow test were plotted to assess the DL-CPH model, where a non-significant statistic implied that the DL-CPH model was perfectly calibrated and close to the perfect model [28].

To assess whether the DL-CPH model would improve patient outcomes, we used decision curve analysis to examine clinical consequences based on threshold probability, from which the net benefit could be derived [15,29].

When assessing the clinical characteristics between the primary and validation cohorts, the independent samples t test was adopted to evaluate the significance of the mean value on age and preoperative CA-125. The chi-squared test was used to assess the difference of categorical variables. In all statistical tests, p-values smaller than 0.05 were considered significant. All the statistical analyses were conducted with R software (version 3.0.1). The Cox-PH model was implemented by lifelines package in Python 2.7. The DL network was implemented by Keras 2.1.5. We made the DL network and the DL-CPH model of this study available at http://www.radiomics.net.cn/post/111.

Results

In the primary cohort, the DL-CPH model achieved good performance on RFS prediction (Table 2, C-Index = 0.717, [95% confidence interval (CI): 0.683–0.755], Hazard Ratio (HR) = 2.711, [95% CI: 2.503–2.919]). In the two independent validation cohorts, the predictive performance of the DL-CPH model was further confirmed (C-Index = 0.713 [95% CI: 0.681–0.750] in the validation cohort 1; C-Index = 0.694, [95% CI: 0.658–0.730] in the validation cohort 2).

Importantly, the C-Index of the DL-CPH model was higher than the clinical model (C-Index = 0.448, [95%CI: 0.402–0.492] in the validation cohort 1; C-Index = 0.631, [95%CI: 0.588–0.674] in the validation cohort 2).

In addition, the strong association between the DL-predicted hazard score and the RFS was further demonstrated by the Kaplan–Meier analysis in Fig. 2. We used the median hazard score of the primary cohort as cut-off value to split patients into high- and low-risk groups [10,13]. Significant discrimination between the RFS of the two groups was observed in the three cohorts (p < 0.0001 in the primary cohort; p = 0.0038 in the validation cohort 1; p = 0.0164 in the validation cohort 2, log-rank test).

To further characterize the association between the DL feature and recurrence, we used the DL-CPH model to predict 3-year recurrence probability for patients. Table 2 and Fig. 3a indicated that the DL-CPH model achieved an AUC of 0.833 (95% CI: 0.792–0.874) in the primary cohort and AUC = 0.772 (95% CI: 0.721–0.820) in the validation cohort 1. The DL-predicted probability also showed significant difference between patients who relapsed less than three years and longer than three years (p < 0.0001 in the primary cohort; p = 0.0010 in the validation cohort 1, Fig. 3b). Good calibration in Supplementary Fig. S3a indicated that the DL-CPH model did not systematically under-predict or over-predict the 3-year recurrence probability because the Hosmer–Lemeshow test yielded a non-significant statistic to the perfect model (p = 0.475 and p = 0.404 in the primary cohort and validation cohort 1). The decision curve in Supplementary Fig. S3b showed that if the threshold probability of a patient or doctor is bigger than 30%, using the DL-CPH model to predict 3-year recurrence added more

<table>
<thead>
<tr>
<th>Models</th>
<th>Cohorts</th>
<th>C-Index (95% CI)</th>
<th>AUC (95% CI)</th>
<th>ACC (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical Model</td>
<td>Primary</td>
<td>0.680 (0.642, 0.717)</td>
<td>0.774 (0.727, 0.826)</td>
<td>0.735 (0.689, 0.784)</td>
</tr>
<tr>
<td></td>
<td>Validation 1</td>
<td>0.448 (0.402, 0.492)</td>
<td>0.443 (0.381, 0.506)</td>
<td>0.449 (0.396, 0.503)</td>
</tr>
<tr>
<td></td>
<td>Validation 2</td>
<td>0.631 (0.588, 0.674)</td>
<td>0.400 (0.268, 0.536)</td>
<td>0.541 (0.480, 0.598)</td>
</tr>
<tr>
<td>DL-CPH Model</td>
<td>Primary</td>
<td>0.717 (0.683, 0.755)</td>
<td>0.833 (0.792, 0.874)</td>
<td>0.776 (0.733, 0.820)</td>
</tr>
<tr>
<td></td>
<td>Validation 1</td>
<td>0.713 (0.681, 0.750)</td>
<td>0.772 (0.721, 0.820)</td>
<td>0.714 (0.665, 0.760)</td>
</tr>
<tr>
<td></td>
<td>Validation 2</td>
<td>0.694 (0.658, 0.730)</td>
<td>0.825 (0.765, 0.893)</td>
<td>0.730 (0.678, 0.786)</td>
</tr>
<tr>
<td>Combined Model</td>
<td>Primary</td>
<td>0.738 (0.705, 0.773)</td>
<td>0.865 (0.830, 0.903)</td>
<td>0.796 (0.757, 0.840)</td>
</tr>
<tr>
<td></td>
<td>Validation 1</td>
<td>0.700 (0.659, 0.742)</td>
<td>0.760 (0.709, 0.809)</td>
<td>0.694 (0.647, 0.741)</td>
</tr>
<tr>
<td></td>
<td>Validation 2</td>
<td>0.729 (0.696, 0.761)</td>
<td>0.762 (0.693, 0.831)</td>
<td>0.703 (0.646, 0.759)</td>
</tr>
</tbody>
</table>

Note: CI represents confidence interval. C-Index represents Harrell’s concordance index, which measures the performance of the RFS prediction. AUC represents area under the receiver operating characteristic curve, and ACC is accuracy. AUC and ACC evaluate the performance of the 3-year recurrence prediction.

Fig. 2. Kaplan–Meier’s analysis of the DL-CPH model. (a) Kaplan–Meier’s analysis in patients from the primary cohort. The vertical lines indicate censored data, and the shadow indicates the 95% confidence interval. (b) Kaplan–Meier’s analysis in patients from the validation cohort 1. (c) Kaplan–Meier’s analysis in patients from the validation cohort 2.
calibration curves were also observed. (95% CI: 0.712–0.814) in the validation cohort 2. In addition, good
0.713–0.818) in the validation cohort 1; and an AUC of 0.825
0.897) in the primary cohort; and an AUC of 0.763 (95% CI:
ity, the DL-CPH model yielded an AUC of 0.857 (95% CI: 0.815–
formance was confirmed in the validation cohort 1 (AUC = 0.772).
In the two independent validation cohorts, the DL-CPH model
yielded a higher AUC than the clinical model (AUC = 0.443, [95%
CI: 0.381–0.506] in the validation cohort 1; AUC = 0.400, [95% CI:
0.268–0.536] in the validation cohort 2) with significant difference
(p = 0.0045 and 0.0361 in the validation cohort 1 and validation
cohort 2, DeLong’s test).

Among the clinical characteristics, FIGO stage was identified as
an independent predictor for recurrence. Consequently, we
constructed a combined model integrating both clinical predictor
(FIGO stage) and the DL-predicted hazard score. The combined
model showed an improvement over the clinical model or DL-CPH
model alone (C-Index = 0.738, 95% CI: 0.705–0.773) in the pri-
mary cohort in terms of the C-Index. This result was further con-
firmed in the validation cohort 2 (C-Index = 0.729, 95% CI: 0.696–
0.761).

Since most OC were diagnosed with advanced stage (FIGO stage
III, IV), we performed a stratified analysis to evaluate the perfor-
mance of the DL-CPH model in advanced HGSOC. Similar results
were observed in Supplementary Table S1, Supplementary Fig. S4
and Supplementary Fig. S5. In the primary cohort, the DL-CPH
model achieved C-Index of 0.706 (95% CI: 0.669–0.746). This per-
fomance was confirmed in the validation cohort 1 (AUC = 0.712,
95% CI: 0.675–0.749) and validation cohort 2 (AUC = 0.687, 95%
CI: 0.649–0.723). When predicting the 3-year recurrence proba-
bility, the DL-CPH model yielded an AUC of 0.857 (95% CI: 0.815–
0.897) in the primary cohort; and an AUC of 0.763 (95% CI:
0.713–0.818) in the validation cohort 1; and an AUC of 0.825
(95% CI: 0.712–0.814) in the validation cohort 2. In addition, good
 calibration curves were also observed.

**Discussion**

The DL network includes thousands of neuron paths to extract
the intrinsic characteristics of HGSOC. This self-learning structure
enables us to quantify the prognostic features of HGSOC that are
difficult to be manually defined. Through a stacked neural network
structure, the DL network manages to encode tumor into multi-
level features reflecting various characteristics of HGSOC. In
Fig. 4a, we visualized the proposed DL network [30,31]. Each layer
in the network extracted different characteristics of HGSOC, from
simple low-level features to complex high-level features. Filters
from the first layer extracted CT intensities of tumor. Afterward,
the second convolutional layer extracted tumor edge information.
When the network went deeper, the convolutional layers extracted
abstract and complex features such as more complicated edges in
Conv. 3 layer, shapes in Conv. 4 layer and a combination of multiple
features in the Conv.5 layer.

Tumors with different recurrence times can activate different
signal pathways of the DL network and finally be encoded into fea-
tures with different values. We fed two tumor images from two
patients (recurrence time = 8.53 and 31.07 months) into the DL
network, and observed different responses as shown in Fig. 4b.
The filters from the DL network had weak response on the patient
with short recurrence time and high response on the patient with
long recurrence time. This was further demonstrated in Supple-
mentary Fig. S6 that depicted the distribution of patients in the
DL feature space (reduced to 2-dimensional by principal compo-
nent analysis algorithm [32] for display convenience). This figure
indicated that patients with longer recurrence time were separated
from patients with shorter recurrence time. Furthermore, we
depicted the patients with recurrence time longer than 3 years and
shorter than 3 years, and these two classes of patients could be
divided easily in the DL feature space.

The good prognostic value of the DL feature probably comes
from the DL network design, which includes an encoder network
compressing tumor images into a lower dimensional representa-
tion (the DL feature) and a decoder network evaluating the infor-
mation capacity of the DL feature. Each layer in the encoder
network aims at using few features to include most information of
the tumor. Therefore, redundant information is eliminated, and
only features that can reflect the intrinsic characteristics of
HGSOC are preserved. We used the DL feature to reconstruct the
original tumor image in Supplementary Fig. S7. This figure indi-
cated that the tumor image reconstructed from the DL feature
was similar to the original tumor image, which demonstrated that
the DL feature included the intrinsic characteristics that are essen-
tial to represent a tumor.

In addition, we evaluated the robustness of the deep learning
features concerning the radiologist bias in selecting the ROIs of
tumor. Two radiologists annotated ROIs independently on 40
patients that were randomly selected from the whole dataset.
Afterward, we calculated the intra-class correlation coefficient
(ICC) of the deep learning features using the ROIs selected by the
two radiologists. The ICC values show that the deep learning
features are stable between two radiologists (ICC range between 0.83 to 0.98 for all the deep learning features). Since we do not require precise tumor boundary segmentation, the ROIs selected by different radiologists do not have large effects to the deep learning features.

Despite the encouraging results, our study has several limitations. First, the DL network extracted the intrinsic characteristics of HGSOC. However, the specific connection between the DL feature and genetic changes were not explored. In the future work, we can explain the DL feature in genetic level by combining the genetic profile of these patients. Second, the DL-CPH model is separate from the Cox-PH. Therefore, the integration of Cox analysis and DL architecture needs further exploration.

To conclude, this study shows that deep learning can provide new CT-based prognostic biomarkers related to the recurrence of HGSOC, which demonstrated stronger prognostic value than clinical characteristics. We also developed a non-invasive DL-CPH model to predict the recurrence of HGSOC by preoperative CT imaging, aiming at assisting individualized treatment and surveillance planning in HGSOC. Moreover, we proposed a novel method to mine the intrinsic characteristics of HGSOC by unsupervised learning. This method can take advantage of the large amount of data without the need for follow-up information.

Acknowledgements

This paper was supported by the National Key Research and Development Plan of China (2017YFA0205200, 2016YFC0103001, YS2017YFCH000397, 2017YFC1308700, 2017YFC1309100), National Natural Science Foundation of China (81227901, 81527805, 81772012, 81501549, 8172010821, 81641168), the Beijing Natural Science Foundation (7182109), Beijing Municipal Science & Technology Commission (Z171100000117023, Z161100002616022), Chinese Academy of Sciences (GJJSSTD20170004, QYZDJ-SSW-JSC005), Henan Province Scientific and Technological Cooperation Project (152106000014).

The authors would like to acknowledge the instrumental and technical support of Multi-modal biomedical imaging experimental platform, Institute of Automation, Chinese Academy of Sciences.

Conflicts of interest

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.radonc.2018.10.019.

References
