Prognostic Evaluation Method of Ovarian Granulosa Cell Tumor Based on Semi-supervised Collaborative Intelligence Model

Xin Liao[1], Xin Zheng[2], Juan Zou[1], Min Feng[1], Liang Sun[1], Yan Li[3] and Kaixuan Yang[1*]

[1] *Department of Pathology, West China Second Hospital, Sichuan University, Chengdu 610041, China*
[2] *College of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China*
[3] *Department of Immunology, Cleveland Clinic, Cleveland, Ohio 44195, United States*

Abstract

Prognostic evaluation of ovarian granulosa cell tumors (GCT) is difficult because of limited case samples, differences in recurrence periods, and challenges in follow-up. Therefore, an evaluation method using artificial intelligence was proposed to provide stable prognosis on ovarian GCTs. First, data of GCT samples were preprocessed, and prognostic evaluation was conducted using a semi-supervised collaborative intelligence model. Experiments were conducted on 102 samples from real GCT cases to confirm the validity of the method. Results show that the method has superior prognostic evaluation effect on different pathological sample sets of ovarian GCTs and has practical value to evaluate the prognosis. After deleting the outliers from the pathological sample sets of ovarian GCTs, the true positive rate (TPR) and false positive rate (FPR) are improved in the intelligent model-based prognostic evaluation, and the area under its performance curve is increased from 0.741 to 0.958. Prognostic evaluation of ovarian GCTs by using semi-supervised collaborative intelligence model can be used to evaluate the prognosis and provides a solution to the difficulties in the prognostic evaluation of ovarian GCTs. This method can help clinicians precisely evaluate the recurrence risk of patients, select an optimal treatment scheme, and increase the long-term survival ratio of patients.

Keywords: Ovarian GCT, Prognostic evaluation, Artificial intelligence, Data preprocess, Semi-supervised collaborative intelligence model

1. Introduction

Ovarian granulosa cell tumor (GCT) has morbidity between 0.05 per 100,000 and 1.7 per 100,000 and represents 2%–5% of all ovary tumors and 40% of sex cord-stromal tumors [1]. Ovarian GCTs are divided into two sub-groups: adult granulosa cell tumor (AGCT) and juvenile granulosa cell tumor (JGCT), which account for 95% and 5% of GCTs, respectively [2]. Approximately 65% of AGCT cases usually occur in postmenopausal women, rarely occur before menarche. Menstrual disorder in reproductive age or irregular vaginal bleeding in postmenopausal age are the most common clinical symptoms. Approximately 80% of JGCTs usually occur in infants [3] and first present as menstrual disorder in reproductive age and irregular vaginal bleeding in postmenopausal age, followed by abdominal pain, pelvic mass, and ascite.

Ovarian GCT is a type of tumor with low-grade malignancy characterized by long-term and inconsistent recurrence time. Extra-ovarian lesions and tumor recurrence result in significantly poor prognosis. Previous studies were limited by the rarity of cases and high loss of follow-up due to low morbidity and long recurrence time [3, 4]. Existing studies on prognostic evaluation of ovarian GCTs were mainly based on statistical methods, such as regression analysis and chi-square analysis. Limited sample cases and lacking consensus on related pathological index of the prognosis resulted in difficult prognostic evaluation of ovarian GCTs. Small datasets in prognostic evaluation can be evaluated with the help of artificial intelligence technology, such as a semi-supervised collaborative intelligence model [5]. With this breakthrough, difficulties in case sample size and inconsistencies in pathological features and clinical prognosis of ovarian GCTs can be solved.

2. State of the art

Ovarian GCTs are tumors with long-term recurrence and problematic prognosis-related factors [2]. Attempts have been made to find stable and reliable prognosis-related factors as basis of clinical evaluation to help clinically treat ovarian GCTs.

Fox et al. reported recurrences within 2 years in half of GCT cases [6], whereas Schwartz et al. reported recurrences within 3 years in 14 out of 19 cases (73.6%) of ovarian GCTs [7]. However, recurrences of ovarian GCTs following initial diagnosis commonly occur after more than 5–10 years [2, 8, 9]. Sommers reported that recurrences were detected for more than 20 years in 6 cases of ovarian GCTs [10]. To date, the longest period before recurrence is 37 years [11]. Some studies reported clinical features, such as tumor recurrence, tumor spread in the pelvis, and tumor involvement of extra-ovarian organs, signal a poor prognosis of ovarian GCTs [3, 12, 13]. Concerns on the concordance

of pathological features and clinical prognosis of ovarian GCTs have been raised. Diffuse or insular pattern of tumor cells signal poor prognosis, whereas follicular pattern of tumor cells and presence of Call-Exner bodies indicate high differentiation and good prognosis [14, 15]. The number of nuclear mitosis of tumor cells, Ki-67 index, and expression of oncogenes and anti-oncogene markers, such as P53, P16, and PTEN, are related to the prognosis of ovarian GCTs [3, 16, 17]. However, no consensus was drawn on the above findings. Pathological data on ovarian GCTs are characterized by rare sample and costly acquisition of data. The difficulties are explained as follows:

(1) Pathological features and clinical prognosis of ovarian GCTs are not directly related due to limited cases and low persistence of follow-up as a result of long recurrence period. Considering this phenomenon, poor prognostic factors, which are obtained by regression analysis, have low accuracy.

(2) Ovarian GCTs are costly and have small sample size. Therefore, the intelligent model should be able to address small sample sizes, which cannot be done by the initial model alone. The samples conforming to standards should be selected first by the intelligent model, while a prediction for the unlabeled samples is made by the initial model to improve the forecasting performance and generalization ability of the entire model.

(3) Outliers, local or global, should be avoided as the samples are iterated in the model to prevent performance degradation. Outlier detection should be performed on the data samples, and a prognostic evaluation is made on detected outlier samples not iterated in the model. Verification needs be performed of the integrity, validity, and consistency of the data samples [18].

We propose a prognostic evaluation method for ovarian GCTs. This method is suitable for small datasets and is combined with artificial intelligent technology. Here, outliers [19, 20] of datasets are deleted by data preprocessing, and prognostic evaluation on the pathologic datasets of ovarian GCTs is conducted using the semi-supervised collaborative intelligence model.

The rest of this study is organized as follows. Section 3 presents the research methods, which include the pre-processing method and the semi-supervised collaborative intelligence model for making the prognostic evaluation on the pathologic datasets of ovarian GCTs. Section 4 is the discussion of related experiments and result analysis, which is conducted on 102 samples of ovarian GCTs collected from January 2007 to April 2017. Section 5 is a summary in which related conclusions are made.

3. Methodology

3.1 Pathologic data preprocessing of ovarian GCTs
Initial pathological data have lower quality than that required by the intelligent model [18]. Therefore, preprocessing of initial data is a prerequisite for the prognostic evaluation on ovarian GCTs using the semi-supervised collaborative intelligence model.

The data preprocessing includes data cleaning based on expert rules and improved outlier detection algorithm based on density. Data cleaning verifies the uniqueness, integrity, validity, and consistency of the data according to clinical significance of the pathologic data of ovarian GCTs. Improved outlier detection algorithm based on density is proposed to simultaneously detect local and global outliers and avoid iteration of the outliers by the intelligent model [18, 21]. Density is defined as the reciprocal of average distance from k–nearest neighbor samples and an isolated point in a lower density region [22]. The algorithm flowchart is shown as Fig. 1, and its specific implementing procedure is described as follows.
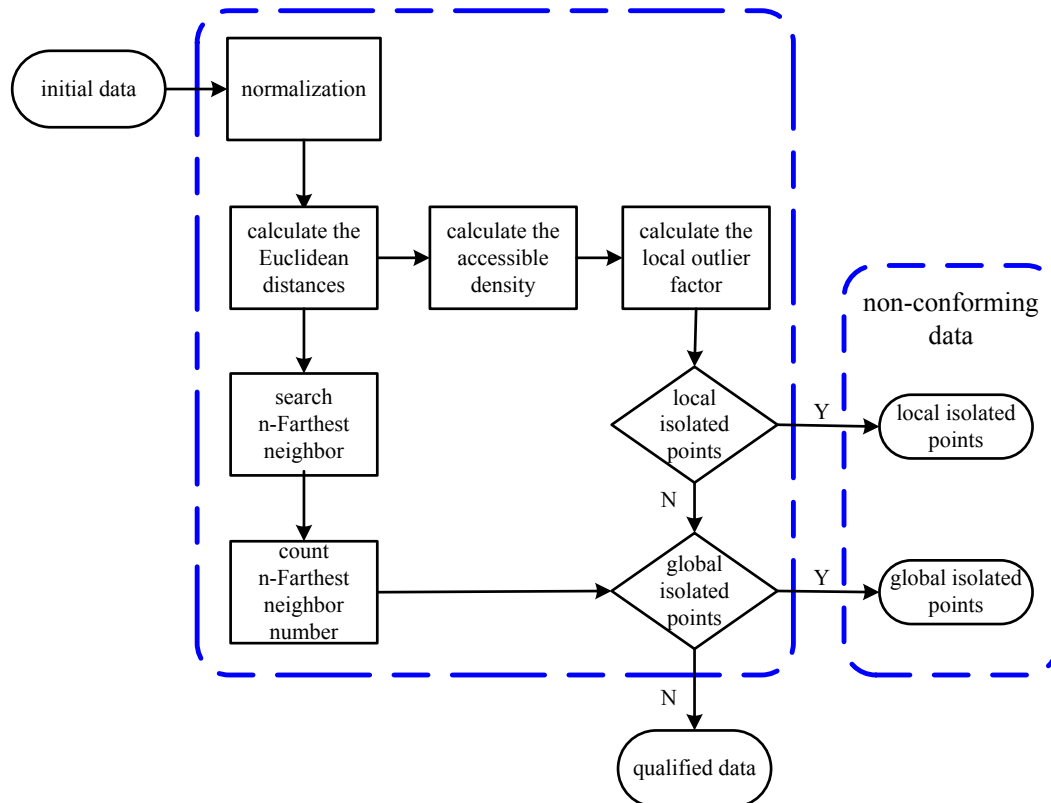

Fig. 1. Improved outlier detection algorithm based on density

Step 1) Centralization and normalization using Eq. (1) are made in samples within [0, 1.0], where max and min are the maximum and minimum values, respectively, in the column where sample x is located.

$$x^* = \frac{x - \min}{\max - \min} \tag{1}$$

Data items in large dimensions are used directly in modeling and thus result in the distortion of the model because of the difference in dimension of each pathological sample [23, 24]. Thus, normalization is needed for each item in the sample sets.

Step 2) Sample sets are normalized according to Eq. (2) by calculating Euclidean distances between each sample point and the other sample points.

$$d(x_i, y_j) = \sqrt{(x_{i1} - y_{j1})^2 + ... + (x_{im} - y_{jm})^2} \tag{2}$$

where X_i and Y_j are the two samples of normalized pathological sample sets, and *m* is the dimensional data items recorded in the samples.

Step 3) Euclidean distances between X_i and other samples are sorted in ascending order, where k points, k samples nearest to X_i, are found and added into k-neighborhood Ω_{ik}, largest k-distance $k\text{-}d$ (X_i) of X_i is obtained, and number of the votes for the k-neighborhood of the corresponding sample plus 1, namely:

$$k\text{-}d \ (X_i) = \text{Max}(d(X_i, Z_j)), \qquad (Z_j \in \Omega_{ik}) \tag{3}$$

The last n samples are found and added into the n-Farthest neighbor Ω_{in} of X_i, and the number of the votes for each sample in the farthest neighbor plus 1, namely:

$$Tic\,\text{ket}(P_i) = Tic\,\text{ket}(P_i) + 1, \qquad (P_n \in \Omega_{in}) \tag{4}$$

Step 4) Accessible density $\rho(X_i)$ of each sample X_i is calculated. Accessible distance $\text{Re}\,ach(Z_j)$ of each point Z_j in the k-neighborhood is computed according to Eq. (5), where the k-distance of Z_j is the longest Euclidean distance $d(X_i, Z_j)$ between Z_j and X_i.

$$\text{Re}\,ach(Z_j) = \text{Max}(k\text{-}d(Z_j), d(X_i, Z_j)), \qquad Z_j \in \Omega_{ik} \tag{5}$$

The accessible density $\rho(X_i)$ of X_i is calculated according to Eq. (6) and is the product of k and the reciprocal of the sum of the accessible distance $\text{Re}\,ach(Z_j)$ of each point Z_j in the k-neighborhood.

$$\rho(X_i) = {k}\Big/{\sum_{Z_j \in \Omega_{ik}} \text{Re}\,ach(Z_j)}, \qquad Z_j \in \Omega_{ik} \tag{6}$$

Step 5) local outlier factor $LOF(X_i)$ of X_i is calculated according to Eq. (7) and is the average of the ratios between accessible density $\rho(Z_j)$ X_i of each point in the k-neighborhood of X_i and accessible density $\rho(X_i)$ of X_i.

$$LOF(X_i) = {\sum_{Z_j \in \Omega_{ik}} \frac{\rho(Z_j)}{\rho(X_i)}}\Big/{k}, \qquad Z_j \in \Omega_{ik} \tag{7}$$

Step 6) The local isolated points are selected. If local outlier factor of a sample is greater than the threshold *VALUE*, then the outliers are determined as the local isolated points.

Step 7) The global isolated points are selected. Samples are found in the number of the votes for the farthest neighbor greater than the threshold *TICKET* and the number of the votes for the nearest neighborhood smaller than the threshold NK.

3.2 Prognostic evaluation of ovarian GCTs based on the semi-supervised collaborative intelligence model
An initial model is trained from labeled samples by semi-supervised learning, and unlabeled data are predicted by auto-mark strategy. This step improves the generalization ability of the model to learn and obtain from a small amount of labeled samples by using efficient information implied in unlabeled data [25, 26]. Co-training is important in semi-supervised learning [27, 28]. Zhou et al. proposed a Tri-Training model that is based on Co-training and eliminates the constraint that two redundant views are required to present in the sample sets [5]. In the present study, prognostic evaluation of pathological samples is conducted using Tri-Training model.

Co-training is achieved by the Tri-Training model using three base classifiers (decision tree C4.5). First, random and reset sampling is performed in labeled sample sets, and three single-labeled sample subsets are obtained to train the three base classifiers. Then, an iterative tuning of collaborative model is conducted by selecting unlabeled samples meeting requirements as supplementary sample sets. Extended training sample sets of each base classifier are obtained by combining classifiers through several iterations of the unlabeled samples [5]. The base classifiers are retrained using the labeled sample sets and extended training sample sets. During iteration, training of Tri-Training model is completed when the base classifiers are not retrained. During iterations, combination classifiers and current base classifiers are crossed and combined. The iterative training flow of the collaborative intelligent model of Tri-Training is shown in Fig. 2. The specific steps are illustrated as follows.

Step 1) Three single training sample subsets $L1$, $L2$, and $L3$ are built in the labeled sample sets by Bootstrap Resampling [29] to train the base classifiers $h1$, $h2$, and $h3$, respectively.

Step 2) Extended training samples are selected from the unlabeled sample sets by making a crossover and combination of the three base classifiers.
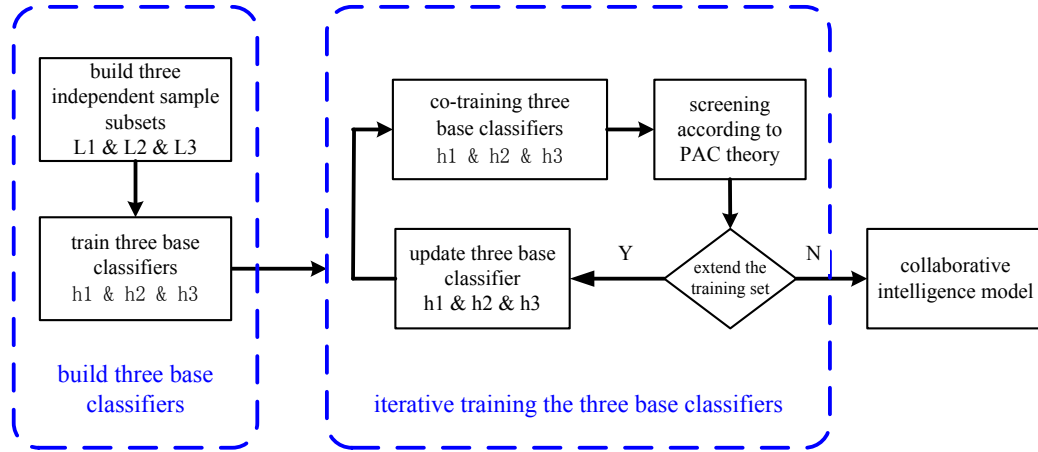
Fig. 2. Co-training flow of tri-training model

First, classification error rate e_i of the labeled sample sets is recorded for the current combination classifier, for example, $h1 \& h2$. When e_i satisfies the requirement of Eq. (8), extended training samples are selected from the unlabeled sample sets.

$$e_i < e_i', \; e_i = \frac{z - z'}{z} \qquad (8)$$

where, e_i' is the classification error rate for the last iteration with initial value of 0.5. Extended training sample set d is selected by traversing data of the unlabeled sample set U and the unlabeled sample x being classified by the base of the combined classifiers. When there is conformity in classification results, attribute x and classification result y are added into the preparatory extended set $L[3]$ of the remaining base classifier $h3$. When the number of the current candidate training samples is greater than that of the last added extended samples, flag bit bUpdate of the base classifiers is updated, and $L[3]$ is selected by Eq. (9) where, $|L^t|$ is the t iteration, which is the sample number of the preparatory extended set iterated currently.

$$e_1^t \left| L^t \right| < e_1^{t-1} \left| L^{t-1} \right| \qquad (9)$$

If the sample number of the current and last iterated extended training set satisfies the requirement of Eq. (9), then the whole extended training set $L[3]$ is retained. Otherwise, the current extended training set $L[3]$ is randomly sampled according to Eq. (10) by obtaining the extended sample set S adjusted with sample number $|S|$.

$$|S| = \left\lceil \frac{e_1^{t-1} \left| L^{t-1} \right|}{e_1^t} - 1 \right\rceil \qquad (10)$$

Using many automatic labeled samples in initial iteration may influence the learning performance when the training is performed by the base classifiers using a small amount of labeled samples, and the distribution of the sample set is not completely learned. According to the verification from

Goldman and Zhou, determining whether or not the extended sample set can improve classification performance of the classifiers using Eq. (9) is required. Thus, the initial training set $L3 = L3 + L[3]$ is directly extended. Otherwise, preparatory extended set $L[3]$ is sampled according to Eq. (10) to obtain the extended sample set S, and $L[3]$ is adjusted to achieve $L3 = L3 + S$.

Here, the training sample sets of h_1 and h_2 are extended by using the respective combination classifiers $h_2 \& h_3$ and $h_1 \& h_3$ to cross-train the co-training model.

Step 3) The supplementary sample sets are added into the three respective base classifiers in the current iteration. The current and extended sample sets are combined, and the corresponding base classifier is trained using the combined sample sets. The base classifier is set to update the flag bit bChanged to TRUE.

Step 4) If the updated flag bit bChanged of the base classifier is TRUE, then Step 2 continues to make a cross-training. Otherwise, training of the collaborative classifier stops.

4 Simulation result analysis

Experiments were conducted on 102 collected cases of ovarian GCTs to verify the validity of data preprocessing and prognostic evaluation of ovarian GCTs based on semi-supervised collaborative intelligence model.

4.1 Pathological data of ovarian GCT
Cases pathologically diagnosed by West China 2nd Hospital of Sichuan University from January 2007 to April 2017 were collected. Among the 102 cases, 87 (accounting for 85.29%) are AGCTs with an onset age of 16–80 years, average age of 52.2 years. The tumor cells are mostly in follicular, insular, and trabecular patterns (Fig. 3). Fifteen cases (accounting for 14.71%) were JGCTs with an onset age of 14–54 years and average age of 35.5 years. The cells in the group of JGCTs are mainly in diffuse pattern (Fig.3(d)). In AGCTs, large amounts of Call-Exner bodies and unobvious features of luteinization of tumor cells, hemorrhage, and necrosis were found. In JGCTs, rare Call-Exner bodies and obvious features of luteinization of tumor cells, hemorrhage, and necrosis were found.
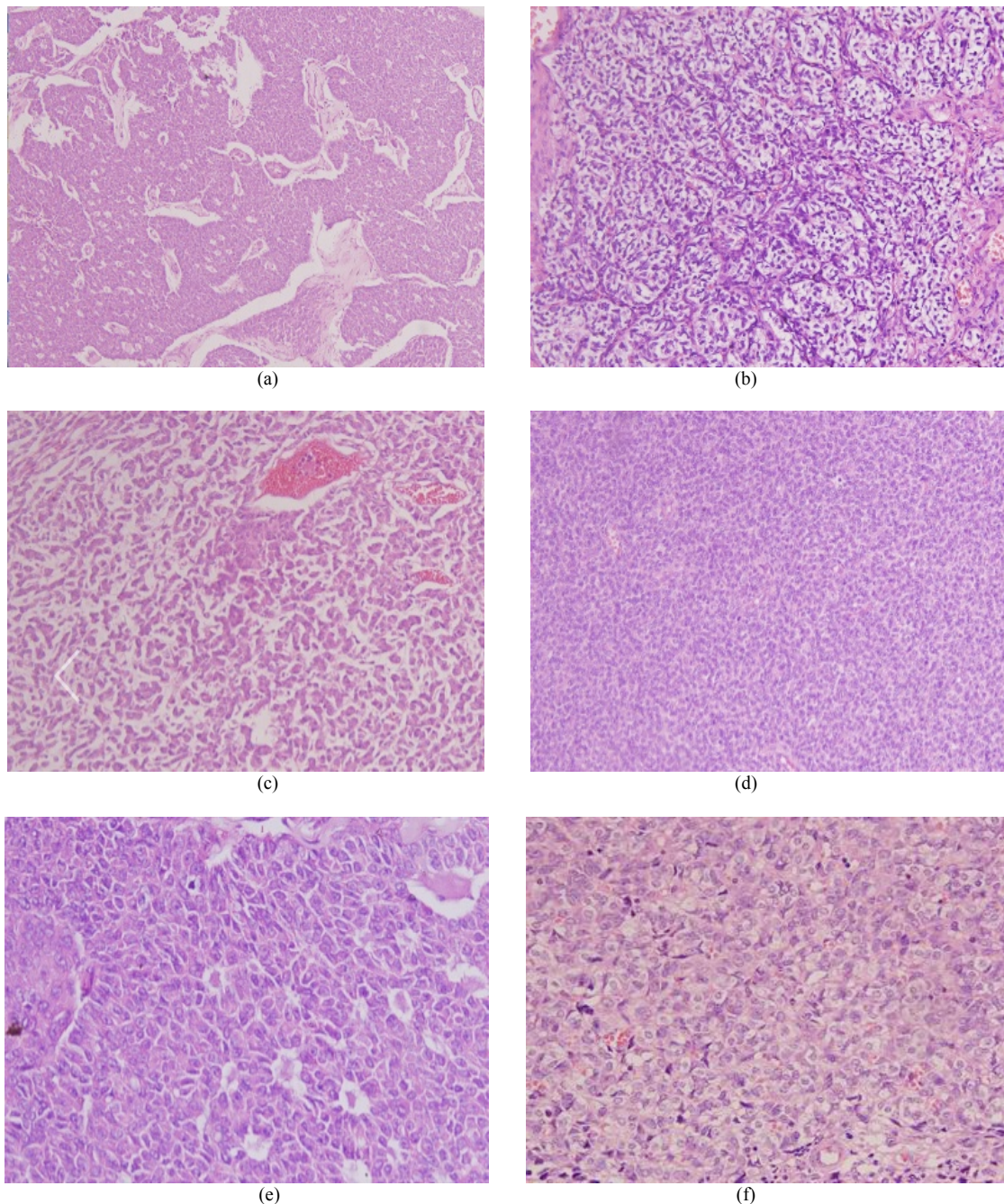
Fig. 3. Ovarian GCT images of typical fields under microscope. (a) Follicular pattern (SP×100). (b) Insular pattern (SP×100). (c) Trabecular pattern (SP×100). (d) Diffuse pattern (SP×100). (e) Call-Exner body (SP×400). (f) Luteinization of tumor cells (SP×400)

About 12 cases (accounting for 13.79%) are AGCTs with post-operation recurrence. While 15 cases (accounting for 17.24%) are AGCTs with pelvic spreading of tumor, 9 (accounting for 8.82%) were accompanied with spontaneous rupture of tumor. All above cases indicated poor prognosis of tumors [12, 17]. Statistical results in the group showed that there was no significant difference (Tab. 1) between the clinical prognosis of AGCTs and JGCTs. Thus, histopathological classification is not used in subsequent index extraction of pathological features [17].

Pathological slides were reviewed, and immunohistochemical markers such as P53, P16, PTEN, and Ki-67 were selected according to the literature [17]. Expression levels of these markers in ovarian GCTs were detected by En vision. Three immunophenotypes were confirmed as positive based on the presence of obviously tan granules in cell nucleus and were confirmed as positive cases based on the number of positive cells greater than or equal to 10%. Expression of Ki-67 was confirmed as positive based on presence of tan granules in cell nucleus and under 400 times light microscope. Sum of positive cells among 1000 cells scattered in 10 fields and the index of Ki-67 were determined based on the counting. Pathological samples of ovarian GCTs contained 15 pathological indexes and 1 prognostic state. A part of the sample data is shown in Tab. 2. Fig. 3 contains images of typical fields under the microscope.

Table 1. Comparison of clinical prognosis between AGCTs and JGCTs.

prognosis-relevant factors	AGCT	JGCT	P
	n=87	n=15	
spontaneous rupture of tumor	8 (9.20%)	1 (6.67%)	1.00
pelvic spreading of tumor	15 (17.24%)	0	0.118
post-operation recurrence	12 (13.79%)	0	0.205

.

Table 2. Sample data from part of cases of ovarian GCTs

		Sample 1		Sample 2		Sample 3	
pathological index		before	after	before	after	before	after
Call-Exner body		exist	1.0	exist	1.0	exist	1.0
nuclear mitosis		0	0.0	0	0.0	5	1.0
cell atypism		none	0.0	none	0.0	exist	1.0
hemorrhage and necrosis		none	0.0	none	0.0	exist	1.0
tumor microscopic pattern	micro-follicular	exist	1.0	exist	1.0	exist	1.0
	macro-follicular	exist	1.0	exist	1.0	exist	1.0
	insular	none	0.0	none	0.0	none	0.0
	trabecular	exist	1.0	none	0.0	exist	1.0
	ribbon	exist	1.0	none	0.0	none	0.0
	diffuse	none	0.0	exist	1.0	exist	1.0
Luteinization		none	0.0	none	0.0	exist	1.0
immunohistochemical markers	Ki-67	15%	0.15	10%	0.1	40%	0.4
	PTEN	positive	1.0	focal positive	0.33	partly positive	0.67
	EGFR	focal positive	0.33	negative	0.0	partly positive	0.67
	p53	focal positive	0.33	focal positive	0.33	partly positive	0.67
prognostic status		favorable	0	favorable	0	unfavorable	1

4.2 Analysis of Experimental Results

Detection of the integrity, validity, and conformity was performed on the initial pathological sample sets, which were normalized and standardized after deleting non-conforming data and outliers. The rules for normalization and standardization of samples were as follows.

(1) For a binary data item, presence corresponds to 1 and absence corresponds to 0 including Call-Exner body, cell atypism, hemorrhage and necrosis, tumor microscopic pattern, luteinization, and prognostic state.

(2) Multivalued data items with definite values were discretized in accordance to a required ratio including the expression rate of Ki-67, PTEN, EGFR, and p53 in immunohistochemical markers.

(3) Multivalued data items with indefinite values were truncated by setting an upper bound of value according to expert rules and discretized.

Key parameters for outlier detection are shown in Tab. 3, including k-neighborhood, outlier threshold VALUE, N-farthest distance, and number TICKET of the votes for farthest-neighbor. Set ranges of the key parameters are listed in line 2, adaptive step sizes in line 3, and optimized parameters in line 4.

Table 3. Parameter settings of data preprocess algorithm.

	K	VALUE	N	TICKET
parameter scope	[3, 6]	[1.2, 1.5]	[3, 6]	[3, 5]
adjusts step-size	1	0.1	1	1
optimized parameters	5	1.3	4	3

Prognostic evaluation on the preprocessed pathological samples of ovarian GCTs was performed using the semi-supervised collaborative intelligence model. Cross validation of their data was conducted thrice [30] and clinicopathological prognosis results were considered as gold standards of the samples. E0, E1, and E2 are detection error rates on collaborative classification model of Tri-training on the initial datasets, local isolated points, and local and global isolated points of ovarian GCTs. E1 and E2 were all deleted. During experiments, outliers data detected are evaluated using initial model and were not trained iteratively using the intelligent model. Tab. 4 lists a part of the experimental results of the prognostic evaluation using the intelligent model and parameters of the corresponding

preprocess algorithm. Probability threshold for the poor prognosis of samples was set to 50%.

From Tab. 4, E0 (18.6%) declines to E1 (mean: 9.75%, min: 6.9%) and E2 (mean: 4.84%, min: 2.0%). Experimental data show that the iterative training of the outliers, including local and global isolated points, results in a degraded prediction performance and an increased prediction error rate of the model. By deleting local isolated and global isolated points, prediction performance of the model was improved, and the prediction error rate was decreased. In the optimized parameters, a good effect was achieved using data preprocessing. Better prediction effect on the preprocessed pathological sample sets of ovarian GCTs was achieved using prognostic evaluation with E2 having an average value of 4.84% and a minimum of 2.0%. By deleting local and global isolated points from the pathological sample sets of ovarian GCTs, accuracy of the prognostic evaluation on the datasets of ovarian GCTs was greater than using the initial datasets or the datasets where only local isolated points were deleted. In clinical practice, obvious differences among cases cause large deviations in pathological data. Clinicopathological work inevitably resulted in the presence of some data errors. The adverse influence of these abnormal data on the evaluation effect of the models can be avoided by data preprocessing while ensuring evaluation accuracy and improving practical performance.

Table 4. Performance list of the prognostic evaluation of ovarian GCTs.

K	VALUE	N	TICKET	E0	E1	E2
3	1.1	4	3	18.6%	10.8%	4.9%
3	1.1	4	5	18.6%	10.8%	5.9%
3	1.1	6	3	18.6%	11.8%	6.9%
3	1.1	6	5	18.6%	9.8%	6.9%
3	1.3	4	3	18.6%	8.8%	4.9%
3	1.3	4	5	18.6%	8.8%	4.9%
3	1.3	6	3	18.6%	10.8%	4.9%
3	1.3	6	5	18.6%	12.7%	7.8%
5	1.1	4	3	18.6%	9.8%	6.9%
5	1.1	4	5	18.6%	8.8%	3.9%
5	1.1	6	3	18.6%	10.8%	5.9%
5	1.1	6	5	18.6%	9.8%	4.9%
5	1.3	4	3	18.6%	6.9%	2.0%
5	1.3	4	5	18.6%	7.8%	2.9%
5	1.3	6	3	18.6%	7.8%	2.9%
5	1.3	6	5	18.6%	9.8%	3.9%

Fig. 4 shows receiver operating characteristic (ROC) curves [31] for all sample sets used semi-supervised collaborative intelligence model. Tab.5 shows the performance statistics of the above ROC curves with the following parameters for outlier detection: nearest k - neighborhood is 3, outlier threshold VALUE is 1.3, farthest distance N is 4, and the number TICKET of the votes for the farthest-neighbor is 3. ROC curves M1, M2, and M3 have values 0.761, 0.839, and 0.958, respectively (Tab. 5). This shows that the model in the study has a prognostic evaluation value (AUC>0.70) on the all above pathological sample sets of ovarian GCTs. M3 and M2 are greater than M1 indicating that the collaborative model has a greater prognostic evaluation value on the pathological sample sets of ovarian GCTs from which outliers are deleted than on initial pathological sample sets of ovarian GCTs. Moreover, M3 is greater than M2, indicating that the collaborative model has a greater prognostic evaluation value on the pathological sample sets of ovarian GCTs from which the local and global isolated points are all deleted than on the pathological sample sets of ovarian GCTs from which only the local isolated points are deleted. In clinical practice, prognostic evaluation accuracy is directly associated with the selection of patient's postoperative treatment scheme and forms the basis of individual treatment. The specificity and sensibility of the evaluation on the samples of ovarian GCTs with deleted local and global isolated points are improved significantly by the collaborative model; thus, helping clinicians to evaluate patient's state completely and accurately, optimize the treatment scheme, decrease the risk of tumor recurrence of patients, and increase the long-term survival ratio of patients.

Based on the results, prognostic evaluation using the semi-supervised collaborative intelligence model on pathological samples of ovarian GCTs with deleted outliers achieved a better evaluation performance and a greater practical value.
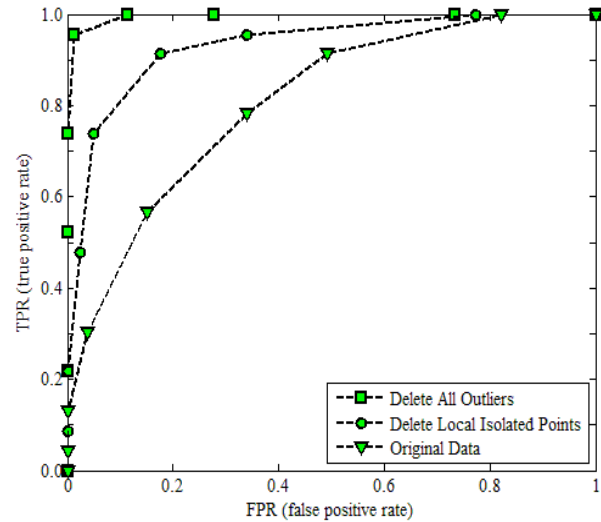

Fig. 4. ROC performance contrast among different pathological datasets of ovarian GCTs by the evaluation models.

Table. 5. Performance list of ROC curves for the prognostic evaluation model of ovarian GCTs.

Pathological sample set of ovarian GCTs	AUC	confidence interval (95%)	
		lower bound	upper bound
the initial datasets of ovarian GCTs	0.741	0.709	0.841
the datasets of ovarian GCTs from which only the local isolated points are deleted	0.859	0.779	0.889
the datasets of ovarian GCTs from which the local isolated points and the global isolated points are all deleted	0.958	0.922	0.981

5. Conclusions

Limited case samples and differences in recurrence cycles result in difficulties in the prognostic evaluation of ovarian GCTs. Here, a stable and effective prognostic evaluation method of ovarian GCTs is proposed in combination with artificial intelligent technology, which provides a way to solve difficulties in the prognostic evaluation. This can help clinicians to accurately evaluate the risk of tumor recurrence of patients, select the optimal treatment scheme, and increase the long-term survival ratio of patients. The following conclusions are made:

(1) The data cleaning based on expert rules can ensure the integrity, validity, and conformity of sample data, and improved outlier detection algorithm based on density can effectively delete local and global outliers from the pathological samples of GCTs.

(2) The semi-supervised collaborative intelligence model has a stable and effective prognostic evaluation effect on different pathological sample sets of ovarian GCTs.

(3) By deleting local and global isolated points from the pathological sample sets of ovarian GCTs, iteration of outlier samples by models is avoided, further improving the prognostic evaluation performance of the semi-supervised collaborative intelligence model.

This study explores the prognostic evaluation method of tumors by incorporating artificial intelligent technology into pathology. The proposed method overcomes the limited pathological samples of ovarian GCTs and indefinite related pathological indexes while achieving good prediction effect and great prognostic evaluation value.

References

1. Koukourakis, G. V., Kouloulias, V. E., Koukourakis, M. J., "Granulosa cell tumor of the ovary: tumor review". *Integrative Cancer Therapies*, 7(3), 2008, pp.204-215.
2. You, X. L., Yin, R. T., Li, K. M., "Clinical and pathological analysis on ovarian granulosa cell tumor". *Journal of Sichuan University (Medical science edition)*, 41(3), 2010, pp.467-470.
3. Pectasides, D., Pectasides, E. A., "Granulosa cell tumor of the ovary". *Cancer Treatment Reviews*, 34(1), 2008, pp.1-12.
4. Li, W., Wu, X., Fang, C., "Prognostic factors in adult granulosa cell tumor of the ovary". *Saudi Medical Journal*, 79(10), 1997, pp.1951-1955.
5. Zhou, Z. H., Li, M., "Tri-training: exploiting unlabeled data using three classifiers". *IEEE Transactions on Knowledge & Data Engineering*, 17(11), 2005, pp.1529-1541.
6. Fox, H., Agrawal, K., Langley, F. A., "A clinicopathologic study of 92 cases of granulosa cell tumor of the ovary with special reference to the factors influencing prognosis". *Cancer*, 35(1), 1975, pp.231-241.
7. Schwartz, P. E., Smith, J. P., "Treatment of ovarian stromal tumors". *American Journal of Obstetrics & Gynecology*, 125(3), 1976, pp.402-411.
8. Cheong, M. L., Shen, J., Huang, S. H., "Long-term survival in a patient with an advanced ovarian juvenile granulosa cell tumor with para-aortic lymph node metastasis". *Taiwanese Journal of Obstetrics & Gynecology*, 55(6), 2016, pp.907-909.
9. Majdoul, S., Tawfiq, N., Bourhaleb, Z., "Recurrence occurring ten years after the initial diagnosis of granulosa cell tumour of the ovary: about two cases and review of the literature". *Pan African Medical Journal*, 2016, pp.25-30.
10. Sommers, S. C., Gates, O. Goodof II, "Late recurrence of granulosa cell tumors: report of two cases". *Obstetrics & Gynecology*, 6(4), 1955, pp. 395-398.
11. Auranen, A., Sundstro, J., "Prognostic factors of ovarian granulosa cell tumor: a study of 35 patients and review of the literature". *International Journal of Gynecological Cancer*, 17(5), 2007, pp.1011-1018.
12. Koukourakis, G. V., Kouloulias. V. E., Koukourakis, M. J., "Granulosa cell tumor of the ovary: tumor review". *Integrative cancer therapies*, 7(3), 2008, pp.204-215.
13. Seagle, B. L., Ann, P., Butler, S., Shahabi, S., "Ovarian granulosa cell tumor: a national cancer database study". *Gynecologic Oncology*, 146(2), 2017, pp.285-291.
14. Lauszus, F. F., Petersen. A. C., Greisen. J., Jakobsen, A, "Granulosa cell tumor of the ovary: a population-based study of 37 women with stage I disease". *Gynecologic Oncology*, 81(3), 2001, pp.456-460.
15. Sakr, S., Abdulfatah, E., Thomas, S., "Granulosa cell tumors: novel predictors of recurrence in early-stage patients". *International Journal of Gynecological Pathology*, 36(3), 2017, pp.240-252.
16. Thomakos, N., Biliatis, I., Koutroumpa, I., "Prognostic factors for recurrence in early stage adult granulosa cell tumor of the ovary". *Archives of Gynecology & Obstetrics*, 294(5), 2016, pp.1-6.
17. Liao, X., Feng. M., Wang. H., "Pathologic features and prognostic factors of ovarian granulosa cell tumor". *Journal of Sichuan University (Medical science edition)*, 44(3), 2013, pp.419-423.
18. Cheng, Y. H., Liao., X., Zheng, X., "Data cleaning technology oriented health big data". *Modern Computer*, (17), 2017, pp.21-25.
19. Ekambaram, R., Fefilatyev, S., Shreve, M., "Active cleaning of label noise". *Pattern Recognition*, (51), 2016, pp.463-480.
20. Mao, Y., Long, H., Deng, R., "Study on data cleaning of healthcare big data". *China Digital Medicine*, 12(6), 2017, pp. 49-52.
21. Shi, H., Zhou. S., Xing, Y., "Average density-based outliers detection". *Journal of University of Electronic Science & Technology of China*, 36(6), 2007, pp.1285-1286.
22. Liu, B., Ma, R., Ding, J., "Density-based statistical merging algorithm for large data sets". *Journal of Software*, 26(11), 2015, pp.2820-2835.
23. Lin, Y., Wang, P., Liu, W., "Research on preprocessing methods for medical examination data". *Application Research of Computers*, (4), 2017, pp.1089-1092.
24. Beaulieu, B. K., Greene, C. S., "Semi-supervised learning of the electronic health record for phenotype stratification". *Journal of Biomedical Informatics*, 64, 2016, pp.168-178.
25. Huang, G., Song, S., Gupta, J. N., "Semi-supervised and unsupervised extreme learning machines". *IEEE Transactions on Cybernetics*, 44(12), 2014, pp.2405-2417.
26. Lu, X., Li, X., Mou, L., "Semi-supervised multitask learning for scene recognition". *IEEE Transactions on Cybernetics*, 45(9), 2017, pp.1967-1976.
27. Guo, X., Wang, W., "An improved co-training style algorithm: compatible co-training". *Journal of Nanjing University (natural sciences)*, 52(4), 2016, pp.662-671.
28. Zhan, W., Zhang, M. L., "Inductive semi-supervised multi-label learning with co-training". In: *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Halifx, NS, Canada: ACM, 2017, pp.1305-1314.
29. Joseph F., "Confidence limits on phylogenies: an approach using the bootstrap". *International journal of organic evolution*, 39(4), 1985, pp.783-791.
30. Qu S. Y., Zhang Q. J., Wang W. J., "Research of the effect of multiple cross-validation on plsda model". *Chinese Journal of Health Statistics*, 34(1), 2017, pp.15-17.
31. Fawcett T., "An introduction to roc analysis". *Pattern Recognition Letters*, 27(8), 2006, pp.861-874