

Predict the Therapeutic Effect of Bevacizumab Treatment Using a Channel Attention Convolution Neural Network

Hongtao Xu*

School of Information Technology and Engineering, Guangzhou College of Commerce,
Guangzhou, Guangdong, 510000, China

* Corresponding Author Email: 201906140011@xs.gcc.edu.cn

Abstract. One of the malignancies with the highest mortality rates among women worldwide is ovarian cancer. Epithelial ovarian cancer (EOC) is the most common kind of ovarian cancer which takes ~90% of ovarian cancer patients. peritoneal serous surface papillary carcinoma (PSPC) is rare cancer whose incident rate is 7% in women. Bevacizumab has been used as a monotherapy along with chemotherapy to treat advanced EOC and PSPC. Bevacizumab has a significant effect on chemotherapy, however, due to the high cost and side effects of the bevacizumab, how to predict the therapeutic effect of Bevacizumab treatment is very important. In this paper, the author uses the proposed attention module, ECA, embedding to the ResNet, compose as ECA-Net, to predict the treatment effect of the bevacizumab in current tissue is effective or invalid through the histopathological image. As a result, the ECA-Net gained novel performance, scoring highly on several evaluation metrics. Specifically, the classification accuracy of the ECA-Net is 94.54% and the f1 score is 95.00%. Bevacizumab is pricey and has side effects, the classification model will forecast its therapeutic impact. In this situation, the experiment will assist the gynecologist in selecting the best course of therapy while also saving money.

Keywords: Ovarian, efficient channel attention, histopathological image, bevacizumab, stain normalization.

1. Introduction

Ovarian cancer is one of the cancers which has the most death rates worldwide among women. It has a low incident rate but a high death rate. The American Cancer Society estimates that in the United States of America in 2021, there will be about 21410 new cases of ovarian cancer and about 13770 deaths due to the disease [1]. Patients who present with advanced-stage, high-grade serous ovarian cancer (HGS-OvCa) account for most fatalities (about 70%)[2]. The majority type of ovarian cancer is epithelial ovarian cancer (EOC). Exceeds 90% of ovarian cancer patients are diagnosed with epithelial ovarian cancer. Peritoneal serous surface papillary carcinoma (PSPC) is an uncommon malignancy that has an incidence of 7% and is characterized by diffuse peritoneal malignancy associated with macroscopically normal-sized ovaries.

Nowadays, the patients may be cured only when they are in the primary stage. However, although more than 80% of the EOC patients responded with cytoreductive surgery and platinum-based chemotherapy, Chemo-resistance and the rate of recurrence remain high [1]. The bevacizumab is used to improve this problem, it has been combined with chemotherapy as a monotherapy for the treatment of the advanced EOC and PSPC, it has also been verified to can improve the progression-free survival to 2-4 months, and even can improve the overall survival. However, the payment of the EOC treatment is expensive and the effectiveness of the bevacizumab is uncertain, furthermore, scientists said that there are no reliable biomarkers that can be used to predict how well EOC and PSPC will respond to bevacizumab therapy.

Predicting the treatment effect is the most important thing before deciding on the treatment plan. Benefit from the development of the neural network, more and more high-efficiency neural networks were proposed. Meanwhile, scientists gradually change the focus overall to focusing on the importance, in this context, scientists proposed a novel attention mechanism, Efficient Channel Attention (ECA) which is based on the SE Attention. As early mentioned, it's hard the predict the treatment effect of the bevacizumab because of the deeply hidden feature. In this paper, the author

uses the deeply ResNet as the backbone and employs an ECA module to obtain the local feature by non-dimensional reduction and local cross-channel interaction [3]. The ECA module is inserted behind each residual block to reassign the weights. The deeper network will extract deeper features; therefore, the author selects the ResNet50 to do this classification task. To assess the model's effectiveness, the author uses the F1 score, accuracy, confusion matrix, and sensitivity to verify the model's efficiency. In Section 2, the author will firstly describe the dataset and the data preprocessing. Second, shows the installation flow of the ECA-Net. Third, introduce the background, and impact of the Efficient Channel Attention, after that, elaborate on its implementation. Finally, several methods of evaluating the model's performance used in this paper will be shown. The evaluation result and the relative discussion will be shown in Section 3.

2. Method

2.1. Dataset

The dataset was published by the Cancer Image Archive and uploaded from Taiwan, China. The collection includes 288 H&E stained WSIs, 162 of which are effective and 126 of which are invalid. These WSIs were collected from various tissue blocks of post-treatment specimens. [2]. A digital scanner (Leica AT Turbo, Leica, Germany) with a 20-objective lens was used to collect the ovarian cancer WSIs. The average size of the WSIs gathered for this project is 54342 x 41048 pixels, with a physical size of 27.43 x 20.66 mm [2].

The specimens were obtained from clinical surgery and used the H&E for staining. Finally, group the different tissue blocks of post-treatment specimens into two groups: effective and invalid. This dataset has a large of images split by WSIs (nearly 40000 images) and the data original is from the real environment. In this context, using this dataset to train the classification model will ensure the model's useability and make it able to use in the real environment. This model will provide a suggestion which whether add the bevacizumab to the chemotherapy process.

2.2. Preprocessing

2.2.1. Whole slide image split

As early mentioned, the dataset is constructed by whole slide images (WSIs). Apparently, such types of images can't be used in train models directly. Understanding the spatial organization of different cell types in the tumor microenvironment (TME) provides information on cancer progression, metastasis, and response to treatment [4]. To obtain more images and extract deeper features, the author divides each WSI into 20x magnification cell-level images. In this paper, the author employs a python library called OpenSlide to split the WSI into cell-level images. After the pre-processing and data augmentation. The author finally retains nearly 50000 images to train the model.

2.2.2. Stain normalization

There has a problem with the image's colors: the cell-level images have many different colors, furthermore, the number of images which has different colors is ununiform. In this context, it has a probability of overfitting or local optimal while training the model. classification accuracy can be strongly impacted by color spaces [5]. Stain normalization can lessen hue and intensity variations found in stained images and increase the precision of the classification model [6]. To avoid the mentioned problem and improve the classification model's accuracy, the author uses the stain normalization technique to uniform the color of the images.

2.3. Efficient channel attention

Efficient Channel Attention has three primary characteristics: lightweight, non-dimensionality, and local cross-channel interaction. This module aims to learn effective channel attention in a more

efficient way [3]. Each characteristic is used to deal with different problems present in the modules that have been proposed in recent years.

2.3.1. Installation

Figure 1 shows the whole process from data input to output of the classified result and the Figure 2 illustrates a part of image preprocessing including resize and stain normalization. To start with, the histopathological images change the size to 224x224 to speed up the training, for avoid overfitting, the author used data augmentation and stain standardization before the data input to the model. The result which is output by the model finally will be classified in full connection pooling (FC).

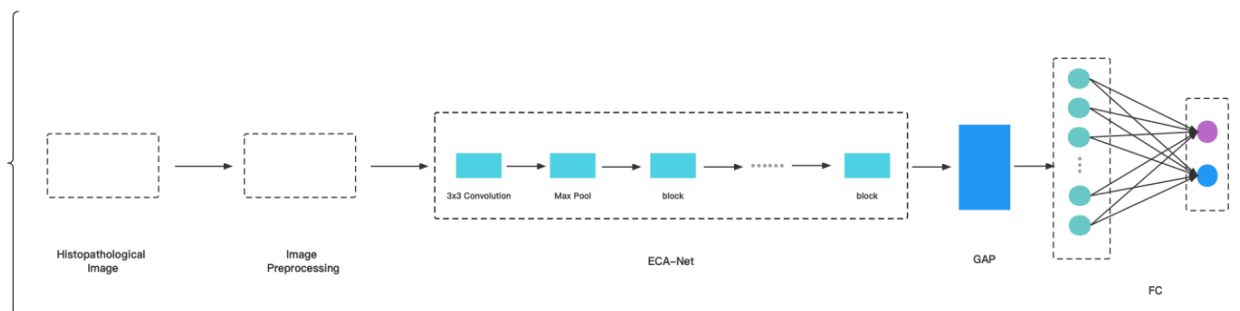


Figure 1. The overall architecture of the ECA-Net

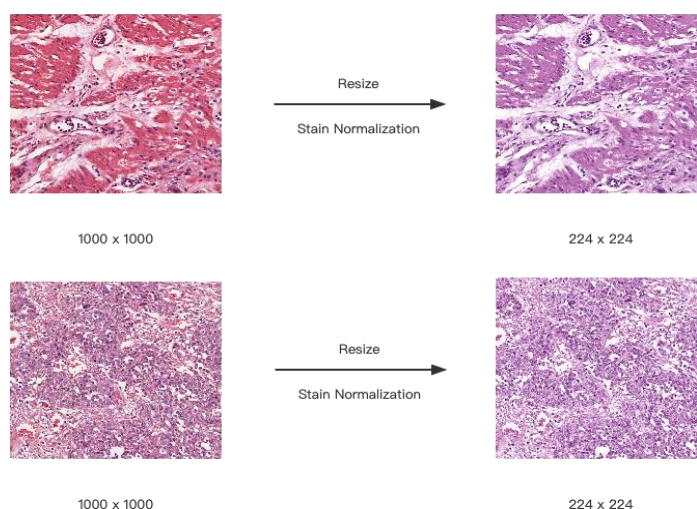


Figure 2. Images resize and stain normalization

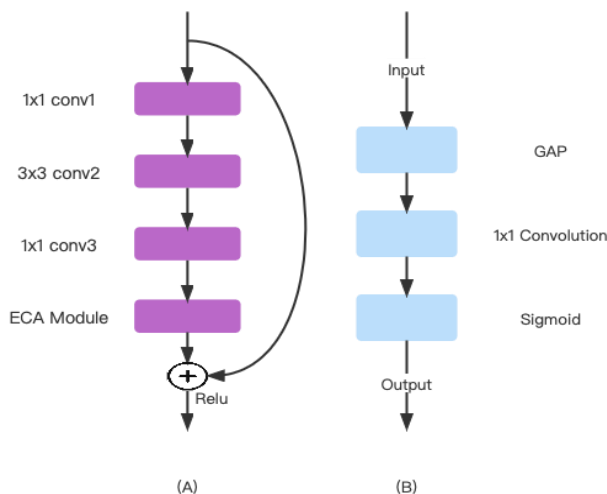


Figure 3. (A) Architecture of ECA-Net's Bottleneck (B) Architecture of efficient channel attention module

In this paper, the ResNet50 was employed as the backbone of the ECA-Net, and an ECA module was inserted into the back of each residual block's final 1D convolution. As the Figure 3 (A) shows, the data through the last 1D convolution then entrance the ECA module after BatchNormalization. In addition, the structure of Figure 3 (A) equals the block in Figure 1. Figure 3 (B) illustrates the data firstly was changed to a tensor whose height and width are all 1 by Adaptive Average Pooling. Second, the 1D convolution implements the local cross-channel interaction. Finally, the result of the sigmoid function as the output.

2.3.2. Attention mechanism

The attention mechanism has been proven that can enhance deep convolution neural networks. The main idea of the attention mechanism is from the focus overall, to focus on the importance. It can give different weights to each part of the input such as sequence and images.

2.3.3 Lightweight

Currently, there have properly two directions for the development of the attention module: feature aggregation enhancement and the combination of spatial and channel attention. In recent years, a lot of attention module has been proposed. RA-CNN [7] assumes that there are coordinated positions of the focus features. The non-Local neural network proposed a non-local mean filter operation which calculate the weighted average of all pixels in the image [8].

However, most of the attention modules have high complexity so the modules only can insert into a single or a few convolution blocks. ECA is a plug-and-play lightweight module. To implement this function, the ECA uses efficient convolution. The most usually used efficient convolution is group convolutions, and depth-wise separable convolutions. Although those efficient convolutions contain fewer parameters (i.e., reduce the module's complexity), they show little effectiveness in channel modules. To solve this problem, ECA uses a 1D convolution with adaptive kernel size k to replace the FC layer in the attention module [3].

2.3.4. Non-Dimensionality

SE-Net is the first proposed effective mechanism to learn channel attention [3]. It has many variants such as SE-Var2, SE-Var3, SE-GC, etc. Assuming the result of the convolution block as $X \in \mathbb{R}^{W \times H \times C}$, where W is the input's width, H is height, and C is the number of the channels, in this context, the weight of the channel attention can be expressed as:

$$\omega = \sigma(f\{W_1, W_2\})(g(X)) \quad (1)$$

Where σ Eq. (1) is sigmoid function, and $g(X)$ represent the channel-wise global average mean pooling (GAP) and it also can represent by the equation:

$$g(X) = \frac{1}{WH} \sum_{i=1, k=1}^{W, H} X_{i,k} \quad (2)$$

Let $y = (X)$, and $f\{W_1, W_2\}$ can be change as:

$$f\{W_1, W_2\}(y) = w_2 \text{ReLU}(w_1 y) \quad (3)$$

To avoid the high complexity of the module, in Eq. (3) the W_1 and W_2 are set as $C \times (\frac{C}{r})$ and $(\frac{C}{r}) \times C$, respectively. Such operation can reduce the complexity of the module, reduce the dimensionality, and take fewer parameters. However, while reducing the dimensionality, the direct correspondence between channel and weight has been broken. In general phenomenon, channel and weight are the direct correspondence but after Eq. (3), the channel's feature maps are reflected in a low-dimensionality space at the first and then mapping them back. This process changes the direction correspondence into indirection. The non-dimensionality has been proven that helpful to learn efficient channel attention [3].

To solve this problem, ECA learns the channel weight by using a fast 1D convolution with adaptive kernel size, the 1D convolution's input is the output of the global average mean pooling (GAP), where

the adaptive kernel size is used to determine the coverage local cross-channel interactions. This proposed operation achieves faster process speed and better performance with reducing dimensionality. Meanwhile, this operation also keeps the direct correspondence between channel and weight [9].

2.3.5. Local cross-channel interaction

SE-Net has two main methods of capturing local cross-channel interaction: SE-Var2 uses the diagonal matrix, SE-Var3 uses the full matrix, and the SE-GC is the combination of the SE-Var2 and the SE-Var3, which uses the block diagonal matrix.

Given the aggregate feature as $y \in \mathbb{R}^C$, the channel weight ω can be expressed as:

$$\omega = \sigma(W_y) \quad (4)$$

Where W is the $C \times C$ parameter matrix [3]. The SE-Var2 and SE-Var3's parameter matrix can be show as:

$$SEVar2 = \begin{bmatrix} \omega^{1,1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \omega^{C,C} \end{bmatrix} \quad (5)$$

$$SEVar3 = \begin{bmatrix} \omega^{1,1} & \dots & \omega^{1,C} \\ \vdots & \ddots & \vdots \\ \omega^{C,1} & \dots & \omega^{C,C} \end{bmatrix} \quad (6)$$

As Eq. (5) and Eq. (6) illustrate, the primary difference between SE-Var2 and SE-Var3 is that SE-Var3 consider the cross-channel interaction, but SE-Var2 doesn't. Scientists have already proven that the SE-Var3 gain better performance than the SE-Var2, in other words, cross-channel interaction will improve the module's effectiveness. However, SE-Var3 has a high complexity caused by plenty of parameters. Before ECA was proposed, Scientists combined the SE-Var2 and SE-Var3 and used the block diagonal matrix as the parameter matrix, named SE block with group convolution (SE-GC). The primary advantage of the SE-GC compared with SE-Var2 and SE-Var3 is has cross-channel interaction and fewer parameters at the same time. The main idea of the SE-GC is to divide the whole channel into different groups and calculate the channel weight independently. The SE-GC can be shown as:

$$W_G = \begin{bmatrix} W_G^1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & W_G^G \end{bmatrix} \quad (7)$$

As the Eq. (7), however, the problem with the SE-GC is it only calculates the channel weight of each group and ignores the relationship between channels. ECA proposed a new method of capturing the local cross-channel interaction which used the band matrix to guarantee both efficiency and effectiveness. The band matrix can be show as:

$$W_k = \begin{bmatrix} \omega^{1,1} & \dots & \omega^{1,k} & \dots & 0 & \dots & \dots & 0 \\ 0 & \omega^{2,2} & \dots & \omega^{2,k+1} & 0 & \dots & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \dots & \dots & \dots & \dots & \omega^{C,C-k+1} & \dots & \omega^{C,C} \end{bmatrix} \quad (8)$$

Where W_k , which comprises the $k \times C$ parameters, is used to calculate the channel attention weight. Only the interaction between y_i and its k neighbours is considered when calculating the weight.

$$\omega_i = \sigma(\sum_{j=1}^k w^j y_i^j), y_i^j \in \Omega_i^k \quad (9)$$

Notably, weight may be easily calculated using the fast 1D convolution can calculate weight easily. As a result, Eq. (9) can be altered to:

$$\omega = \sigma(C1D_k(y)) \quad (10)$$

As Eq. (10) illustrate, in ECA, calculating the weight of channel attention only needs to determine the coverage of local cross-channel interaction (i.e., kernel size k). ECA also proposed a method for the adaptive kernel size.

$$k = \psi(D) = \left\lceil \frac{\log_2(D)}{\gamma} + \frac{b}{\gamma} \right\rceil_{\text{odd}} \quad (11)$$

Where $|x|_{\text{odd}}$ represent the nearest odd number of x . In this paper's experiment, the author set γ and b to 2 and 1, respectively. The high-dimensional channel will achieve a greater range of interactions through Eq. (11), while low-dimensional channels will experience a lower range of interactions by utilizing a non-linear mapping [3].

2.4. Evaluation metrics

Evaluation is the indispensable part of training the model, in this paper, the author utilizes 7 evaluation metrics for validating the model's performance, the metrics including training speed, accuracy, precision, recall, sensitivity, and f1 score. Training speed is used to explore the relationship between training speed and layer depth, it is represented by the accuracy in the same number of iterations.

The confusion matrix, as the name implies, gives us a matrix as output and describes the complete performance of the model [10]. Use the sci-Kit learned library and input the true labels and the predicted labels can obtain the confusion matrix and it can be represented as Table 1.

Table 1. Confusion Matrix

True Predict	1	0
1	TP	FN
0	FP	TN

Where row means the true label and column means prediction result.

Precision represents the rate of the predicted positive samples that are positive samples, and can be calculate as:

$$P = \frac{TP}{(TP+FP)} \quad (12)$$

It is the number of correct positive results divided by the number of all relevant samples (all samples that should be identified as positive)

$$R = \frac{TP}{(TP+FN)} \quad (13)$$

F1 Score is an indicator used to measure the accuracy of dichotomous model in statistics, it can be expressed as:

$$F = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (14)$$

According to all positive data points, sensitivity is the percentage of positive data points that are accurately interpreted as positive.

$$\text{Sensitivity} = \frac{TP}{(FN+TP)} \quad (15)$$

Accuracy is the ratio of the number of correct predictions to the total number of input samples.

$$A = \frac{(TP+TN)}{(TP+FN+FP+TN)} \quad (16)$$

3. Result and Discussion

In this experiment, the author employs three traditional image classification models to compare the performance with the ECA-Net and implement two different layer depth ECA-Net. every model uses the same hyperparameters to make sure the facticity of the experiment. To avoid overfitting, the author used a series of methods for fine-tuning including warm-up, gradient accumulation, Xavier init, no bias decay and dynamic learning rate, etc. Specifically, the author set the initial learning rate as 0.1984 and formulates a schedule after many experiments to adjust the learning rate and let the learning rate descend step by step. The warmup technique let the first learning rate smaller than the settings learning rate and increases slowly. The batch size is set as 64 and accumulates gradient 8 iterations, so the actual batch size is 512. The weight decay is set to 0.0001 and only the FC, 2D convolution and 1D convolution will decay, the momentum is 0.9. Notably, the dataset is randomly shuffled to prevent the local optimum.

In addition, ECA-Net uses the ResNet50 as the backbone so employs the ResNet and SE-Net which let the ResNet as the backbone to compare the model's performance will have more comparable. In this context, the author will compare the model's performance in three aspects: compare accuracy, training speed, and the relationship of model depth. Finally, a series of the ECA-Net's evaluations will be on show. All experiments are implemented and trained on the same computer with NVIDIA Geforce 3060 12G GPU using PyTorch deep learning framework.

3.1. Compare the classification accuracy

In this experiment, the author uses the SE attention module embed into the backbone and trains two traditional image classification models (ResNet and Densenet) to compare the performance of ECA-Net. Table 2 lists the classification accuracy and the training speed of ECA-Net50, ECA-Net18, ResNet50, SE-Net50, and Densenet. Table 2 compares the different values between five models in two aspects: the best accuracy and the training speed, where Speed means the accuracy at the current iteration (75th times) and the accuracy is obtained by validation accuracy.

According to Table 2, the ECA-Net50 achieves 94.54% classification accuracy and the highest training speed compared with the other models. The classification accuracy of the ECA-Net18 has approximately identical to that of the ECA-Net50, however, it is slower by 3%. The SE-Net50 and ResNet50 had classification accuracy of 73.75% and 88.82, respectively. The classification accuracy of ECA-Net50 gains 6% performance improvements compared with its backbone (ResNet50) and have better accuracy than the SE-Net50 by 21%. Unexpectedly, SE-Net was the worst performer whose classification accuracy and training speed are all the lowest in Table 2. The SE-Net50 has the lowest accuracy. As early mentioned, the SE-Net destroy the direct relationship between channel and weight, this is one of the differences between the ECA module and the SE module, the relationship between channels and weight. The indirect correspondence between channels and weights may be the reason for the SE-Net's worst performance. The ECA module makes the relationship back directly and gains the novel performance. In this context, the SE-Net is not suitable for this experiment because the model's accuracy even lower than its backbone.

According to the classification accuracy of the Densenet, it gains the accuracy of 94.34, which shows a better performance compared with the SE-Net and the ResNet, however, though the ECA-Net has a lighter layer depth, the classification accuracy is still better than the Densenet121. The accuracy between the ECA-Net50 and ECA-Net18 is practically the same. Consequently, using the ECA module in conjunction with ResNet will improve classification performance and it is demonstrated that the model is valid when used to classify medical images. Overall, the ECA-Net has gained advanced classification accuracy compared with the other model.

Table 2. The classification accuracy and the training speed of the five models

Acc & Speed Models	ECA-Net50	ECA-Net18	SE-Net50	ResNet50	Densenet121
Accuracy	94.54	94.66	73.75	88.82	94.34
Speed (75 Epoch)	90.40	87.79	66.39	76.14	86.39

3.2. Relationship between the training speed and layer depth

The classification accuracy result shows the novel performance of the ECA-Net. To further explore the performance of the ECA-Net, the author compares the training speed between models by recording the accuracy at the current iteration (75 iterations). The classification accuracy of ECA-Net50, ECA-Net18, SE-Net, ResNet50, and Densenet is 90.40, 87.78, 66.39, 76.14, and 86.39, respectively. In the ECA-Net, training speed of ECA-Net50 is faster than the ECA-Net18 by 3%, it means that, the deeper layer model gains faster training speed, so the relationship between training speed and layer depth is a positive correlation. In other words, the training speed and layer depth increase at the same time in a certain range.

Let ECA-Net compares with the other models, although the classification accuracy of the Densenet is close to the ECA-Net, the training speed of ECA-Net still has an enormous advantage. The SE-Net gains the lowest training speed, when data through the SE module, it will be calculated by a 2D convolution, this will reduce the training speed of the model. But the ECA module employs the fast 1D convolution to speed up training and archive a better performance. In general, ECA-Net shows its novel performance, it used the comparatively light layer depth to gain the fastest training speed and best classification accuracy.

3.3. Precision, Recall, F1 Score, Accuracy, Sensitivity and Confusion Matrix

After the evaluation of classification and training speed, the typical evaluation is necessary, here, the author leverage the precision, recall, f1 score, accuracy, sensitivity, and confusion matrix to further evaluate the performance of the ECA-Net. Figure 4 shows the confusion matrix of the ECA-Net50 and ECA-Net18. Specifically, the TP, FP, FN, and TN of the ECA-Net50 are 0.95, 0.06, 0.05, and 0.94, respectively. the same values of the ECA-Net18 are 0.93, 0.04, 0.07, and 0.96, respectively. Table 3 shows a series of evaluations of the ECA-Net50 and ECA-Net18 including accuracy, precision, recall, sensitivity, and f1 score. The ECA-Net gains 94.59%, 95.00%, 95.00%, 95.01%, and 95.00%, respectively. And the ECA-Net18 gains 94.79%, 95.00%, 95.00%, 93.35%, and 95.00%, respectively. As a result, incorporating the ECA module within the ResNet will increase training speed while also extracting deeper features to increase accuracy.

Table 3. The typical evaluation of the ECA-Net18 and the ECA-Net50

Models Metrics	Accuracy	Precision	Recall	Sensitively	F1 Score
ECA-Net50	94.59	95.00	95.00	95.01	95.00
ECA-Net18	94.79	95.00	95.00	93.35	95.00

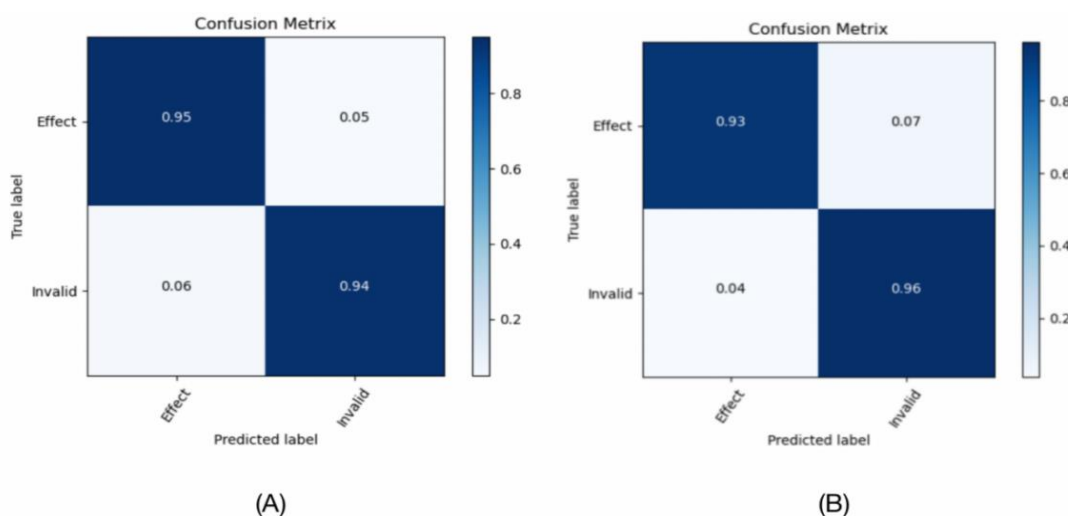


Figure 4. Confusion matrix: (A) the ECA-Net50 and (B) the ECA-Net18

Although most results in Table 3 between ECA-Net50 and ECA-Net18 are the same, there still has fewer different results, those results will play an important part in the best model selection. The

Sensitivity of the ECA-Net50 is higher than the ECA-Net18 by 1.74%. It means that the ECA-Net50 judges more Effect as Effect. Figure 4 illustrates the ECA-Net50's rate of classifying the correct between two labels is closer than the ECA-Net18, thus the ECA-Net50 is more stable than the ECA-Net18.

4. Conclusion

In this paper, the author firstly introduces the knowledge of ovarian cancer and the bevacizumab's function, on this basis, also illustrates the problem of predicting the therapeutic effect of EOC and PSPC on bevacizumab treatment: there are no effective biomarkers to classify effective or invalid. To extract the deeply features, the author uses ECA-Net as the classifier which let ResNet as the backbone and inserts the ECA module into the backbone. Finally, the ECA-Net obtained a high score in a series of evaluation metrics compared with three different models. In conclusion, the ECA-Net shows its advanced performance in classifying the treatment effect of bevacizumab. In the actual world, this experiment makes sense because the classifier aids the gynecologist in making a quick and cost-effective treatment decision. Thus, avoiding the situation where the patient spends a high price, but bevacizumab is ineffective, additionally, it also saves the cost of the hospital.

References

- [1] Wang, C. W., Chang, C. C., Khalil, M. A., Lin, Y. J., Liou, Y. A., Hsu, P. C., Lee, Y. C., Wang, C.-H., Chao, and T.K., "Histopathological whole slide image dataset for classification of treatment effectiveness to ovarian cancer," *Scientific Data* 9(1) (2022).
- [2] Integrated genomic analyses of ovarian carcinoma, "Integrated genomic analyses of ovarian carcinoma.," *Nature* **474**(7353), 609–615 (2011).
- [3] Fu, J., Zheng, H., and Mei, T., "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).
- [4] Lee, K., Lockhart, J. H., Xie, M., Chaudhary, R., Slebos, R. J., Flores, E. R., Chung, C. H., and Tan, A. C., "Deep learning of histopathology images at the single cell level," *Frontiers in Artificial Intelligence* **4** (2021).
- [5] Gowda, S. N., and Yuan, C., "Colornet: Investigating the importance of color spaces for image classification," *Computer Vision – ACCV 2018*, 581–596 (2019).
- [6] Anghel, A., Stanisavljevic, M., Andani, S., Papandreou, N., Rüschoff, J. H., Wild, P., Gabrani, M., and Pozidis, H., "A high-performance system for robust stain normalization of whole-slide images in histopathology," *Frontiers in Medicine* **6** (2019).
- [7] Wang, X., Girshick, R., Gupta, A., and He, K., "Non-local neural networks." *Proceedings of the IEEE conference on computer vision and pattern recognition.* (2018).
- [8] Yan, R., Ren, F., Wang, Z., Wang, L., Zhang, T., Liu, Y., and Zhang, F., "Breast cancer histopathological image classification using deep convolutional neural network", *Methods*, (2020).
- [9] Mishra, A., "Metrics to evaluate your machine learning algorithm," *Medium*, 28 May 2020, <<https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234>> (5 July 2022).
- [10] He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., and Li, M., "Bag of tricks for image classification with Convolutional Neural Networks," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019).