Ovarian Tumor Classification Using Convolutional Neural Networks and Machine Learning

Erica Liu (1724576)
Electrical Engineering
Signal Processing Systems
q.liu@student.tue.nl

I. Introduction

This is typically a hard problem for radiologists to classify ovarian tumor types using medical images. Generally, three types need to be distinguished: benign, borderline and malignant. Borderline tumors are also called low malignant tumors and the incidence is low [1]. Since the treatment of various tumors is different, it is significant for patients' treatment to categorize tumors in an early stage. This study focuses on classifying malignant tumors and benign tumors.

An ovarian tumor is abnormal growth on the ovaries, which is a form of female genital tumor [2]. Some tumors do not affect other sites of the body and only stay in the local position; these tumors are classified as benign tumors. These benign tumors are noncancerous. However, some tumors could spread over not only local positions but also spread to distant sites, which are classified as malignant tumors. Because of the characteristic of malignant tumors, they can grow unlimited and invade other organs, which leads to cancers [3].

Previous results have suggested that if ovarian cancer is detected and treated in the early stage, the survival rate is over 90%, whereas in the advanced stage this rate is only 30% [4]. Therefore, the effective detection and classification of the tumor could help treat and reduce the mortality of ovarian cancers.

The diagnosis of ovarian cancer is similar to other cancers. A biopsy of sample tissue is required [5]. However, this kind of method is invasive and risky. Thus a non-invasive classification method based on medical imaging is needed. Medical imaging helps much in visualizing tumors, in which magnetic resonance (MRI), computed tomography (CT), and ultrasound are commonly used. Furthermore, CT scans have many benefits, such as fast scanning, clear image, lower cost, and better reproducibility than other medical imaging methods. Therefore, CT images are widely applied in classifying ovarian tumors using computer vision to help improve the performance [6].

Image classification is based on feature extraction. In previous research, deep learning and machine learning (ML) methods were developed to help the classification of ovarian tumors [6] [7]. Also, the Computer-Aided Diagnosis (CAD) systems were proposed driven by Artificial Intelligence (AI), where 2D and 3D convolutional neural network (CNN) was used for classifying input information depending on its feature extraction capability [8]. Features extraction in 3D medical images can also be implemented standardly using Pyradiomics [9].

As stated in previous research, ML classifiers performed well using extracted features [7]. However, high dimension features with a small number of samples could result in a bad classifier performance or over-fitting. The ovarian tumors dataset is usually a small dataset. Hence dimensionality reduction is needed before using ML classifiers. Moreover, Cross-Validation should be used to avoid over-fitting [10].

In summary, this study therefore explores different feature extraction methods and machine learning classifiers to help with ovarian tumors classification. The chosen dataset consists of 3D CT images of ovarian tumors from 102 patients. Neural networks and Pyradiomics are used as feature extraction methods. Additionally, various machine learning classifiers, such as support vector machine (SVM), K-Nearest Neighbor (KNN), Random Forest (RF), and Neural Networks (NN), are implemented to classify tumors according to features after dimensionality reduction of features. Finally, the evaluation metric is used to compare the results of the classification. The overall process of this study is shown in figure 1.

II. Methods

This study pre-processed the ovarian tumor dataset to the appropriate format before 3 kinds of feature extraction methods followed by dimensionality reduction or feature selection. Then SVM, KNN, RF, and NN were used to classify the benign and malignant tumor. Finally, the evaluation metric was calculated for every model for comparison.

A. *Ovarian Tumor Dataset*

The Ovarian tumor dataset contains 102 samples collected from female patients' CT scans in Catharina hospital over one year, which has 47 benign samples and 55 malignant samples. Each sample has an original CT image and a tumor mask

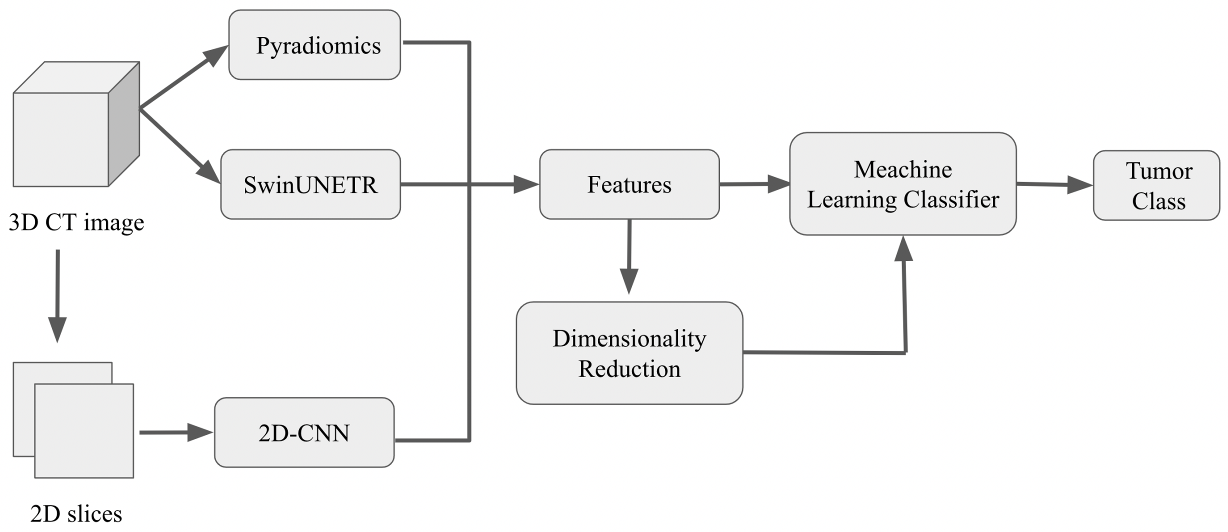

Fig. 1. Overview of the process.

after segmentation. Some samples have 2 tumors in one mask image, in which the bigger one is chosen for later process.

In every sample, the size of the original 3D CT image is the same as the size of the mask image, which is 512*512*z. Different tumor samples have different z values.

In 2D analysis, these 3D CT images were processed into 2D slices and the size of every slice is 512*512*1. To be specific, each tumor has a non-zero fragment in the projection of the z-axis. According to the middle of fragments, the mask image slices in the middle of the tumor and corresponding original image slices were selected and combined as the dataset in the 2D method. The example slices and combination of one benign tumor were demonstrated in figure 2.

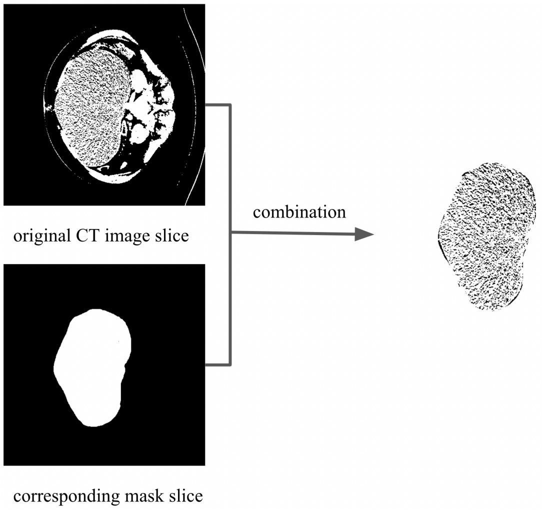

Fig. 2. Combination of a benign tumor sample.

B. *Feature Extraction*

1) Pyradiomics:

As Joost J.M.van Griethuysen et al. explored in [9], Pyradiomics were developed to calculate quantified features from medical images, e.g., lung CT images. It is an open-source python package for 2D or 3D medical data processing. Using this package, 107 radiomic features can be extracted, including 18 first-order features, 14 shape features, 24 Gray Level Co-occurrence Matrix (GLCM) features [11], 16 Gray Level Size Zone Matrix (GLSZM) features [12], 16 Gray Level Run Length Matrix (GLRLM) features [13], [14], 16 Neighboring Gray Tone Difference Matrix (NGTDM) features, 14 Gray Level Dependence Matrix (GLDM) features.

In this section, 3D CT tumor images and mask images were stored in pairs. In the python environment, feature extraction using Pyradiomics was implemented based on the original 3D CT images and corresponding mask images.

2) 2D-CNN:

In this section, the 3D CT images were processed to be 2D slices initially. Since the Ovarian tumor dataset contains original CT images and mask tumor images, this study got not only the slices from original CT images and mask images but also combines the slices for each sample. Since the slices are grayscale images, the channel of each slice is one and the size of each slice is 1*512*512 (channel*height*width). Therefore, the size of the combination slice is 2*512*512 (channel*height*width), which is also the input image size of 2D-CNN.

As suggested by Lie Zhang et al., the features extracted by the deep neural networks from the medical images are the reflection of the visual features of the tumor region [15]. In addition, pre-trained ResNet50 as a feature extractor performed well in collaboration with traditional machine learning classifiers [16]. Therefore, the deep convolutional neural network ResNet50 was used to extract features in ovarian data.

Initially, the architecture of ResNet50 is based on [17]. It consists of 5 stages and contains 50 layers. The overview architecture is shown in figure 3.

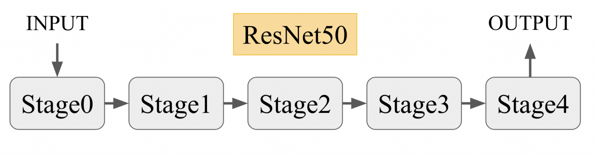

Fig. 3. Overview architecture of ResNet50.

In this part, the ResNet50 was pre-trained on the ImageNet-1K. The pre-trained weights for ResNet50 were loaded except for the weights of the first layer because the input channel in this study is different from the pre-trained ResNet50. Using these weights, the ResNet50 was re-trained on the 2D ovarian dataset, where 82 images were used for training and 20 images were used for validation.

- Stage 0
  In this stage, the input images (2*512*512) were pre-processed by 64 7x7 convolutional kernels with stride=2, followed by batch normalization and ReLU activation function. MaxPooling was finally implemented in this stage. After this stage, the size of the output was 64*128*128.
- Stage 1-4
  The main part of ResNet50 can be split into several convolutional blocks and identity blocks. Convolutional (Conv) blocks can change the dimensionality of the network and identity blocks help increase the depth of the network. In stages 1-4, every stage has a Conv block. stage 1 contains 1 Conv block and 2 identity blocks, which is similar to stage 4. After stage 1, the output size was 256*128*128. Stage 2 has 3 Conv blocks and after this stage, the size of the output is 512*64*64. Stage 3 consists of 1 Conv block and 5 identity blocks, which can change the output size to 1024*32*32. Finally, The output size is 2048*16*16 after stage 4.

After these stages, the output with size 2048*16*16 was processed using AveragePooling. Moreover, the flattening out was applied and a fully connected layer was linked to categorize 2 classes.

In the ResNet50 re-training process, the batch size was assigned to 3, and the number of workers was set to 8. The learning rate was 0.0001 over the learning process. After training, the weights of neural network were updated and this model was saved as the best ResNet50 model in this study. With these updated weights, the 2D ovarian dataset was processed to 2048 features by this best model without the last fully connected layer.

3) Pre-trained SwinUNETR [18]:

SwinUNETR is a 3D CNN model based on a transformer aiming to implement the segmentation of 3D medical images. Tang et al. trained and tested the SwinUNETR on both Brain Tumor Segmentation (BraTS) dataset and Beyond the Cranial Vault (BTCV) Segmentation Challenge dataset [19]. This pre-trained model can interpret 3D data, and hence it can analyze the full volume of ovarian CT scans compared to the 2D method where only slices were used.

Since the BTCV dataset are CT images while the BraTS dataset are MRI images, this study chose to use the model trained on the BTCV dataset.

In this case, the pre-trained SwinUNETR extracted features without training this model on our ovarian dataset. Moreover, this study extracted features from not only combinations of images and masks but also original CT images.

Since this model was only pre-trained on a small dataset and no re-training process was implemented, only the better features were chosen to be processed to compare with other methods depending on these 2 sets of features.

SwinUNETR segmentation consists of encoder and decoder implementation, and the features were extracted from the last encoder step, where the output size was 768*3*3*3. 3D AveragePooling changed the output size to 768.

C. Feature Selection and Dimensionality Reduction

Many features were extracted from different extraction methods. 107 features were extracted using Pyradiomics; 2048 features were extracted using 2D-CNN and 768 features were extracted using pre-trained SwinUNETR.

In this section, those high-dimension features were reduced to help improve the later ML classifiers [10]. There are 2 methods that were chosen, one method is to compute some standardized features using projection, which is Principle Component Analysis (PCA) [20]. Another one is to select several important features and leave other features out using analysis of variance (ANOVA), which is feature selection.

1) Principle Component Analysis (PCA):

PCA aims to extract important information from all features without target categories, and this important information can be represented as several orthogonal feature variables [20].

PCA toolbox in MATLAB is the unsupervised analysis of all features, which is used to calculate PCA. In this part, PCA was implemented in MATLAB automatically when the ML classifier was being trained on the data [21].

2) Analysis of Variance (ANOVA):

There are many feature selection methods. The feature data after feature extraction is numerical and the tumor class is categorical. According to the recommendation choices of feature selection, the selection depending on known labels can be regarded as a classification predictive modeling problem, then Analysis of Variance coefficient can be used to help select important features [22].

In this part, the most k important features were selected using ANOVA F-measure via a function in Sklearn [23]. Since the best k cannot be calculated automatically, this study tested different k values to choose a better k in ANOVA to compare with other methods.

D. Classification Using Machine Learning Classifier

In this section, the naive method was initially performed to classify tumors using features without dimensionality reduction. The selected ANOVA features and the standardized PCA features were used to classify the ovarian tumors using ML classifiers. In this study, only benign tumors and malignant tumors were considered so the following 4 kinds of high-performance binary classifiers were used to train classification models separately [24]. Moreover, those ML classifiers were used in the MATLAB toolbox Classification Learner. Finally, due to the small number of samples in the dataset, the K-fold Cross-Validation was used to avoid over-fitting. Here 5-fold Cross-Validation was selected before training in Classification Learner.

- Support Vector Machine (SVM).
- K-Nearest Neighbor (KNN).
- Random Forest (RF).
- Neural Network (NN).

E. *Evaluation Metric*

In classification processes, Receiver Operating Characteristic (ROC) curve is a graph that can illustrate the capability of a binary classification model when the differentiation threshold changes [25]. This ROC curve needs the true positive rate (TPR) as y axis and the false positive rate (FPR) as x axis, The ratio of true positive (TP) prediction numbers and all positive numbers is TPR, while the ratio of false positive (FP) numbers and all negative prediction numbers is FPR [25]. The calculation formula are illustrated in equation 1, 2.

$$TPR = \frac{TP}{TP + FN} \tag{1}$$

$$FPR = \frac{FP}{FP + TN} \tag{2}$$

The evaluation metric is needed to evaluate the performance of models after training on the feature data. Because of the small and imbalanced ovarian tumor dataset, the area under ROC curve (AUC) can be used as an evaluation metric in this situation [26] [27]. The computation of AUC was implemented in MATLAB Classification Learner.

III. Results

In this section, the results comparisons of different models' results are displayed in detail according to feature extraction methods. Since PCA was automatically used to reduce dimensionality in MATLAB, only the ANOVA was analyzed to choose k important features. All comparison classification results were based on 5-fold Cross-Validation.

A. *Analysis of Variance (ANOVA)*

This study set $k = 10, 20, 50, 100$ and trained the classifiers separately, resulting in different AUCs in table I. The results of different combinations of feature extraction and ML classifiers were shown. Half of the best results for every combination appeared with $k = 10$ so $k = 10$ was chosen for ANOVA in later processes to compare with other methods. Most of the best results of combination occurred in $k = 10, 20$. $k = 10$ was the best choice for pyradiomics and pre-trained SwinUNETR. In contrast , $k = 20$ was better for ResNet50. ResNet50 had a stable feature extraction capability. The average AUC was more than 0.80.

B. *Pyradiomics (10 Features)*

Using Pyradiomics, 107 features were extracted. Those features were fed into classifiers as predictor variables and the classes of ovarian tumors were fed as target variables. Also, the naive method used all features while the ANOVA method only used 10 features after selection. The PCA method was directly set before the training of the classification model. After the training processes, the AUCs of models were illustrated in MATLAB and the comparison is displayed in figure 4. The highest AUC was found in the combination of ANOVA and NN, which was 0.82. The second one appeared in the combination of ANOVA and SVM, which was 0.80.

In this case, PCA worked with uncertainty, the AUC got higher using KNN and RF, and it got lower using SVM and NN. Nevertheless, AUCs become much better after using ANOVA. The results of SVM, KNN, and RF were very close and stable. In contrast, NN suggested unstable performance in 3 dimensionality reduction methods.

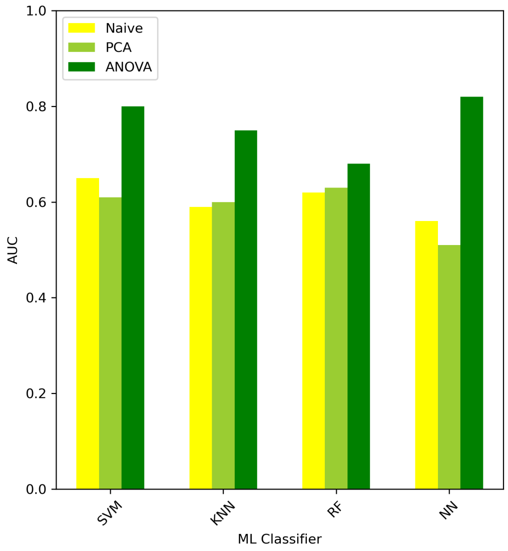

Fig. 4. AUCs for various ML classifiers in Pyradiomics.

C. *ResNet50 (10 Features)*

The training loss and validation loss in figure 5 indicated over-fitting occurring. Also, the validation accuracy kept fluctuating over the process, which was displayed in figure 6.

Therefore, the training was stopped at around ten epochs. Using the re-trained ResNet50 without the last fully connected layer, each sample in the 2D ovarian dataset was processed to 2048 features.

After the training of classifier models, their AUCs were compared in figure 7. Similarly, the combination of ANOVA and NN showed the highest AUC, and this value was 0.88. The second one was indicated in the combination of ANOVA

TABLE I
AUCs FOR VARIOUS CLASSIFIERS USING k SELECTED FEATURES.

k Features	Rad+SVM	Rad+KNN	Rad+RF	Rad+NN	2D+SVM	2D+KNN	2D+RF	2D+NN	SU+SVM	SU+KNN	SU+RF	SU+NN
10	**0.80**	**0.75**	0.68	**0.82**	0.86	0.81	0.80	**0.88**	**0.68**	**0.74**	0.57	0.58
20	0.79	0.71	**0.75**	0.76	**0.89**	**0.85**	**0.86**	0.86	0.67	0.61	0.54	0.54
50	0.75	0.74	0.72	0.73	0.88	**0.85**	0.80	0.80	0.67	0.61	0.54	**0.68**
100	0.71	0.65	0.66	0.66	0.86	0.81	0.77	0.87	0.66	0.61	**0.61**	0.61

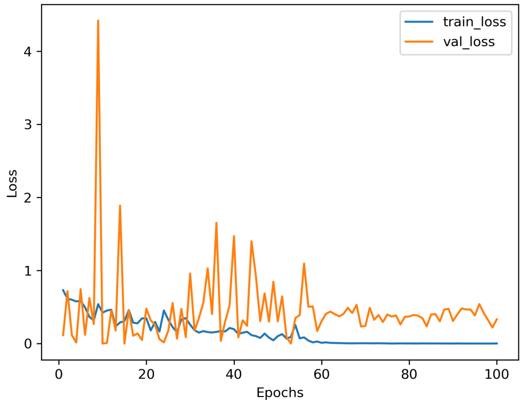

Fig. 5. Training loss and validation loss.

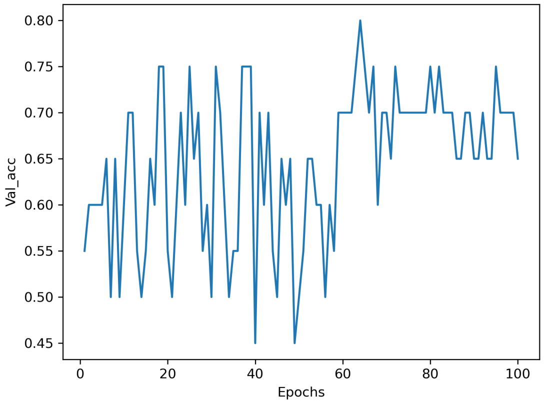

Fig. 6. Validation accuracy in ResNet50 re-training process.

and SVM, which was 0.86. In every dimensionality reduction method, the resulting AUC in SVM, KNN, RF, and NN were very close.

In this case, PCA was not able to improve the performance, the AUCs got worse after applying PCA in classification models. ANOVA showed an excellent improvement in classification again. In this case, even the lowest AUC with ANOVA can reach 0.80.

D. *Pre-trained SwinUNETR (10 Features)*

Using pre-trained SwinUNETR, 768 features were extracted from every sample. An extra comparison in pre-trained SwinUNETR was implemented to select better features. The extra comparison was between features extracted from the combination and those from only original CT images. Results in table II indicated that features extracted from only original CT images resulted in better AUCs. Therefore, in later processes, this study only used features extracted from original 3D CT images to compare the performance of different ML classifiers.

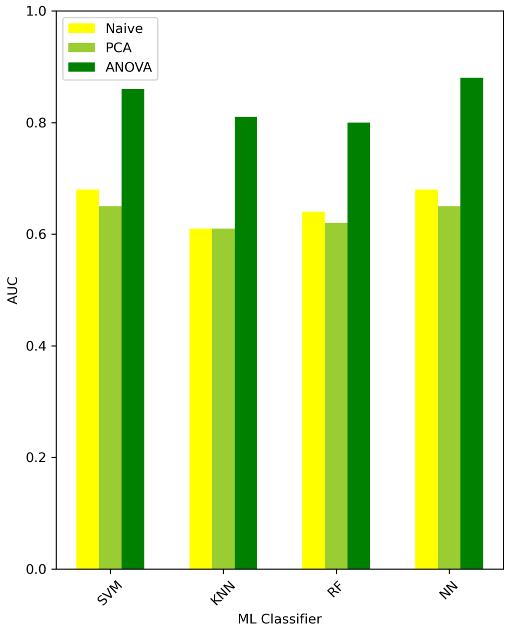

Fig. 7. AUCs for various ML classifiers in Resnet50.

TABLE II
$k = 10$, AUCs FOR ML CLASSIFIERS USING VARIOUS FEATURES.

Feature source	SVM	KNN	RF	NN
Combination	0.60	0.52	0.61	0.51
Original CT image	0.68	0.74	0.57	0.58

In the naive method, 768 features were fed into the classification learner as predictor variables. Other operations were similar to the previous method.

After the models were trained, the AUCs of every model were compared in figure 8. The overview results in SwinUNETR were obviously lower than the other two feature extraction methods. The highest AUC was shown in the combination of ANOVA and KNN, which was 0.74. The second one was shown in the combination of ANOVA and SVM, which was 0.68. However, there were some AUCs lower than 0.5, which indicated the classification performance of these models was even worse than random binary classification. In this case, PCA only helped in the SVM classifier model, while others did not. ANOVA kept its great improvement in binary differentiation.

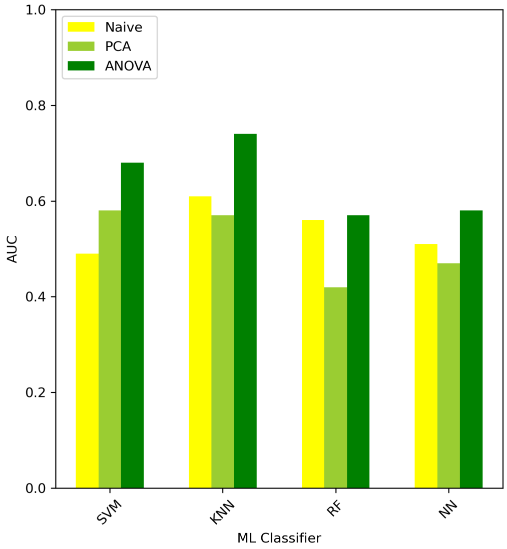

Fig. 8. AUCs for various ML classifiers in SwinUNETR.

IV. Discussion

A. Feature Extraction Limitations

The Pyradiomics and SwinUNETR extracted features much more slowly than ResNet50 because they used 3D images. Additionally, pre-trained SwinUNETR was only trained on 30 3D CT images, which is a relatively small dataset compared to ImageNet-1K. The ResNet50 model was re-trained on the ovarian dataset and got new weights, while SwinUNETR did not.

B. Dimensionality Reduction Limitations

PCA is an unsupervised process on the high-dimension feature data, and the target category is not required in this case. However, the feature selection method ANOVA is supervised learning using labeled data. Hence the information of the target was also used in selecting significant features. In this case, the selected k features might change if the ovarian tumor dataset is updated. Therefore, the result difference among dimensionality reduction methods is understandable. Furthermore, this study did not test the k value automatically in the ANOVA method, which implied the global optimization might not be reached. There might be a better k that can represent more information than the current selection.

V. Conclusion

In this study, the classification process of tumors is divided into three main parts: feature extraction, dimensionality reduction, and machine learning classification. These steps were explored to search for ways that could improve ovarian tumor classification.

The 2D-CNN ResNet50 features were more effective for tumor classification than the other two feature extraction methods, in which Pyradiomics performed better than pre-trained SwinUNETR. It is also found that features from only original images were better in pre-trained SwinUNETR. Therefore, the 2D-CNN ResNet50 was the best choice in the feature extraction step, and Pyradiomics was the second option.

In the dimensionality reduction part, k values in ANOVA were firstly analyzed, and results illustrated that $K = 10$ should be chosen. Furthermore, the performance of supervised learning feature selection ANOVA was much more excellent than the naive and PCA method. PCA method was not stable in changing the model's AUC. Both increase and decrease of AUC compared to the naive method occurred. Therefore, ANOVA was the best choice to select essential features from all extracted features.

In the machine learning classification section, algorithms only had a little difference. SVM and KNN were stably better than RF and NN, and NN showed great classification results in some cases. The Random Forest showed no superiority in classification.

In summary, 2D-CNN ResNet50 and ANOVA feature selection can help with the differentiation between benign and malignant tumors, and effective machine learning classifiers can be chosen from SVM, KNN, and NN. These effective feature extraction, selection, and classification help categorize ovarian tumors, which can further help ovarian cancer diagnosis earlier in medical practice and even reduce the death rate depending on the early-stage diagnosis.

In future research, the 2D neural network can be trained on a larger dataset to improve the model. The 3D-CNN could be developed more to enhance the classification performance. For example, the SwinUNETR can be re-trained on the ovarian dataset, and weights can be updated according to this dataset. Additionally, a better ANOVA implementation could be developed in future work to improve the final classification.

References

[1] J. T. Chambers, M. J. Merino, E. I. Kohorn, and P. E. Schwartz, "Borderline ovarian tumors," *American journal of obstetrics and gynecology*, vol. 159, no. 5, pp. 1088–1094, 1988.

[2] A. K. Höhn, C. E. Brambs, G. G. R. Hiller, D. May, E. Schmoeckel, and L.-C. Horn, "2020 who classification of female genital tumors," *Geburtshilfe und Frauenheilkunde*, vol. 81, no. 10, pp. 1145–1153, 2021.

[3] A. Patel, "Benign vs malignant tumors," *JAMA oncology*, vol. 6, no. 9, pp. 1488–1488, 2020.

[4] Z. Su, W. S. Graybill, and Y. Zhu, "Detection and monitoring of ovarian cancer," *Clinica chimica acta*, vol. 415, pp. 341–345, 2013.

[5] C. Stewart, C. Ralyea, and S. Lockwood, "Ovarian cancer: an integrated review," in *Seminars in oncology nursing*, vol. 35, no. 2. Elsevier, 2019, pp. 151–156.

[6] M. Wu, C. Yan, H. Liu, and Q. Liu, "Automatic classification of ovarian cancer types from cytological images using deep convolutional neural networks," *Bioscience reports*, vol. 38, no. 3, 2018.

[7] M. Aditya, I. Amrita, A. Kodipalli, and R. J. Martis, "Ovarian cancer detection and classification using machine leaning," in *2021 5th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT)*. IEEE, 2021, pp. 279–282.

[8] M. Coccia, "Deep learning technology for improving cancer care in society: New directions in cancer imaging driven by artificial intelligence," *Technology in Society*, vol. 60, p. 101198, 2020.

[9] J. J. Van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. Aerts, "Computational radiomics system to decode the radiographic phenotype," *Cancer research*, vol. 77, no. 21, pp. e104–e107, 2017.

[10] V. Spruyt, "The curse of dimensionality in classification," *Computer vision for dummies*, vol. 21, no. 3, pp. 35–40, 2014.

[11] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Transactions on systems, man, and cybernetics*, no. 6, pp. 610–621, 1973.

[12] G. Thibault, B. Fertil, C. Navarro, S. Pereira, P. Cau, N. Levy, J. Sequeira, and J. Mari, "Texture indexes and gray level size zone matrix," *Application to Cell Nuclei Classification. PRIP*, pp. 140–145, 2009.

[13] M. M. Galloway, "Texture analysis using grey level run lengths," *NASA STI/Recon Technical Report N*, vol. 75, p. 18555, 1974.

[14] A. Chu, C. M. Sehgal, and J. F. Greenleaf, "Use of gray value distribution of run lengths for texture analysis," *Pattern recognition letters*, vol. 11, no. 6, pp. 415–419, 1990.

[15] L. Zhang, J. Huang, and L. Liu, "Improved deep learning network based in combination with cost-sensitive learning for early detection of ovarian cancer in color ultrasound detecting system," *Journal of medical systems*, vol. 43, no. 8, pp. 1–9, 2019.

[16] K. Gupta and N. Chawla, "Analysis of histopathological images for prediction of breast cancer using traditional classifiers with pre-trained cnn," *Procedia Computer Science*, vol. 167, pp. 878–889, 2020.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] Y. Tang, D. Yang, W. Li, H. R. Roth, B. Landman, D. Xu, V. Nath, and A. Hatamizadeh, "Self-supervised pre-training of swin transformers for 3d medical image analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 730–20 740.

[19] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *International MICCAI Brainlesion Workshop*. Springer, 2022, pp. 272–284.

[20] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[21] D. Ballabio, "A matlab toolbox for principal component analysis and unsupervised exploration of data structure," *Chemometrics and intelligent laboratory systems*, vol. 149, pp. 1–9, 2015.

[22] M. Kuhn, K. Johnson *et al.*, *Applied predictive modeling*. Springer, 2013, vol. 26.

[23] T. Y. Chen, F.-C. Kuo, and R. Merkel, "On the statistical properties of the f-measure," in *Fourth International Conference onQuality Software, 2004. QSIC 2004. Proceedings.* IEEE, 2004, pp. 146–153.

[24] R. Kumari and S. K. Srivastava, "Machine learning: A review on binary classification," *International Journal of Computer Applications*, vol. 160, no. 7, 2017.

[25] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," *arXiv preprint arXiv:2010.16061*, 2020.

[26] J. A. Hanley, B. J. McNeil *et al.*, "A method of comparing the areas under receiver operating characteristic curves derived from the same cases," *Radiology*, vol. 148, no. 3, pp. 839–843, 1983.

[27] S. Wang and X. Yao, "Using class imbalance learning for software defect prediction," *IEEE Transactions on Reliability*, vol. 62, no. 2, pp. 434–443, 2013.