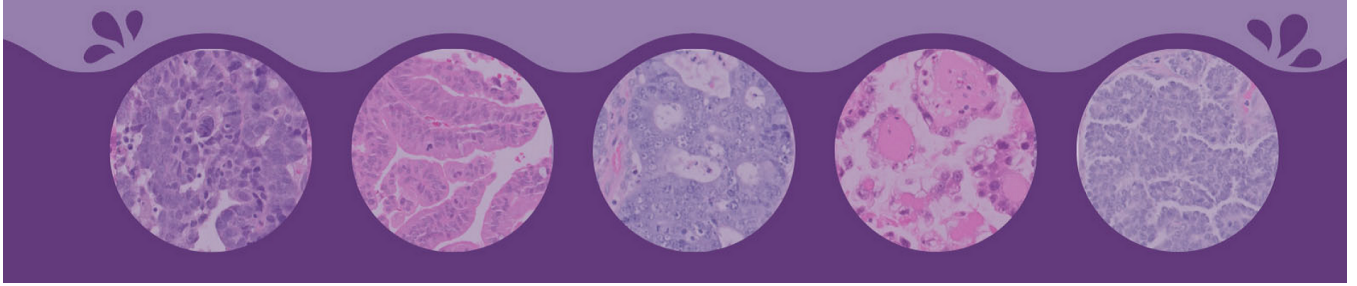# Ovarian Cancer Subtype Classification and Outlier Detection (UBC-OCEAN)

WU Fei
HKUST(GZ)
Guangzhou, China
fwu750@connect.hkust-gz.edu.cn

SUNG Ming Shin Cindy
HKUST(GZ)
Guangzhou, China
mscsung830@connect.hkust-gz.edu.cn

XING Daiyan
HKUST(GZ)
Guangzhou, China
dxing004@connect.hkust-gz.edu.cn

Figure 1. Complexity of cancer subtypes

Abstract

Ovarian cancer stands as the most lethal disease affecting the female reproductive system. It comprises five prevalent subtypes, and each represents a distinct entity characterized by differing morphology, molecular biology, and clinical outcomes, necessitating tailored treatment approaches. Effective classification is imperative in this context. Using a diverse dataset from 20 global centers, the study employs image processing techniques, such as patch extraction, stain normalization, data augmentation, and noise filtering. Experimental results demonstrate the effectiveness of the proposed models and ensemble strategies in accurately classifying OC subtypes, with stain normalization and patch extraction playing crucial roles. The findings underscore the potential of AI in transforming OC diagnosis, providing a robust method for subtype classification. Future work may explore the integration of these models into clinical workflows for broader applications in medical imaging tasks.

Keywords: Ovarian carcinoma,histology, image classification, outlier detection, convolution neural network

1 Introduction

Ovarian Carcinoma (OC) ranks as the eighth most prevalent cancer among women worldwide [7]. The criteria for classification can be into five major subtypes, including High-grade serous carcinoma (HGSC), Clear-cell ovarian carcinoma (CCOC), Endometrioid (ENOC), Low-grade serous carcinoma (LGSC), and Mucinous carcinoma (MC), shown in Figure 1. Also, there are rare cases that differs from the generalised cases, thus are considered outliers. Current research

vision shows that each major subtype is linked from unique former lesions and molecular pathways, to morphological and clinical characteristics [14]. Accurate subtype classification done by specialists, i.e. gynecologic oncologists, is crucial in precise diagnosis for delivering subtype-specific treatment, which can significantly improve patient outcomes. Case studies showed that specialised care for OC patients achieved better efficacy compared to general gynecologist [15], but are limited by human and physical resources for decision making in many hospitals [11]. One recent heightened awareness and maturation in precision medicine and prognosis using artificial intelligence (AI) to extract relevant clinical information using computer-aided detection (CAD) . This is exemplified by biomarker extraction from histology cancer images through deep learning (DL) techniques [6]. This has led to a critical demand for employing novel machine learning methods for biomarker discovery and identification, emerging as a key driving force in numerous clinical applications, such as workflow automation, guiding decision making. This report explores the application of AI to enhance the accuracy of diagnosing OC subtypes and managing the scarcity of specialized resources.

2 Related work

2.1 Image Processing

This study builds upon existing research in the field of OC classification and detection. Previous works have predominantly focused on microscopic examination of hematoxylin and eosin (H&E) stained images for cancer diagnosis. The conventional approach, while effective, faces challenges in terms of accuracy and efficiency when dealing with a large

volume of images. Recent studies, such as those by Jeffrey Boschman et al.[5], have explored the utility of color normalization in AI-based diagnosis of pathology images. This report further extends these efforts by incorporating advanced computational methods to enhance classification accuracy and reduce pathologists' workload.

2.2 Predictive Modeling in Cancer Classification

In the realm of predictive modeling for cancer classification, significant progress has been made using various DL models. Recent reviews have shown that among convolutional neural networks (CNNs), generative adversarial models, autoencoders and recurrent nerural networks, deep CNN has demonstrated best performance and has high potential on demonstrating the effectiveness of deep in cancer detection [12]. Therefore, our study focuses on deep CNN, due to their ability to capture intricate features in medical images and ultimately contributes to an improved accuracy in classification. A variety of models were employed for transfer learning in regard to the problem scenario; each selected for its specific strengths in image classification and feature extraction, as well as their histroical success in cancer image modelling. This includes ResNet, VGG-16, EfficientNet, MobileNet, and InceptionNet.

Beyond experimenting with different designs of CNN kernels, we proposed an Ensembled learning-based approach to build a voting classifier, which involves aggregating predictions from multiple CNN models. This strategic approach was driven by the need to harness the strengths of diverse CNN architectures by complementing their limitations to another, thus providing a more robust and accurate prediction compared to singular CNN models[8]. This is evident in recent development of lung cancer, breast cancer and colorectal cancer detection to be desribed later. The study will test is efficiency in mitigating individual model biases and enhancing overall diagnostic accuracy for deployment of diverse data across medical centres, shown in Figure 2.
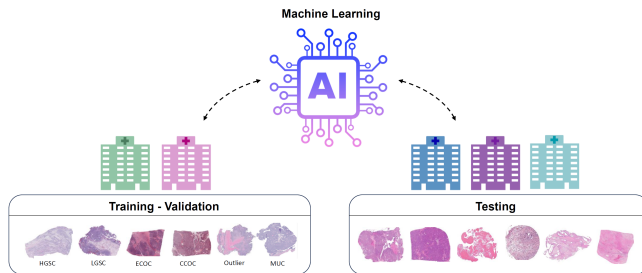
Figure 2. Schematic diagram of general deployment for training and testing. Note that there are six classes in test set instead of five.

3 Problem Setups

The primary challenge of this study is the classification of OC subtypes and the detection of outliers. The focus lies on subtypes such as HGSC, CCOC, ENOC, LGSC, and MC, along with several rare subtypes and normal cases.

Given the limitations of conventional microscopic examination of H&E-stained images, the project aims to leverage computational models to enhance accuracy and efficiency. The objective is to classify the five main subtypes and detect outliers, thereby addressing the lack of specialist decision-making resources in many hospitals and improving patient outcomes, as illustrated in figure 2 .

4 Workflow

Our workflow illustrated in Figure 3 can be divided into 2 main parts: data pre-processing and CNN modelling.
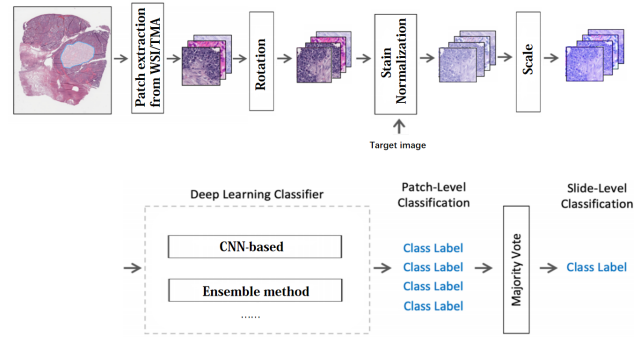
Figure 3. Workflow showing data processing pipeline.

4.1 Datasets

The study employs an extensive dataset of medical images from over 20 centers across four continents, capturing variations in data quality and subtypes. The whole dataset includes a mix of whole slide images (WSIs) at 20X and tissue microarray (TMA) images at 40X magnification.

The whole dataset can be categorised into training and test set. The training set has 538 images, adding up to a total of 750GB. As for the test set, it contains around 550GB, totaling roughly 2000 images, mostly TMAs. The large data volume presents unique challenges in terms of memory management and computational efficiency. Importantly, the test set has been intentionally designed with inclusion of non-cancerous biopsies and rare subtypes. This is to assess the generalization capabilities of models, emphasizing the need for robust capturing of features and adaptable algorithms to handle the diverse conditions presented by the data. For a visual outline, refer to Figure 2.

4.2　Modelling

Building on the current The study employs various deep learning models for analysis, such as ResNet, VGG-16, EfficientNet, MobileNet, and InceptionNet. These models were chosen for their ability to capture intricate features in medical images. The ensemble approach combines multiple models, enhancing robustness and accuracy. Extensive preprocessing, including patch extraction, stain normalization, and data augmentation, was applied to optimize the models for high performance.

Our proposed CNN models are described below:

ResNet: Known for its deep architecture and residual connections using skip connections, ResNet, developed in 2016, is adept at learning from complex image data without suffering from the vanishing gradient problem. It has shown to be reproducible with reliable outcomes in classifying malignant and benign tumors in colorectal cancer with 80% accuracy, 87% sensitivity and 83% specificity [13].

VGG-16: This model, with its uniform architecture of 3x3 convolutional layers, excels in capturing detailed features from medical images, making it highly effective for intricate pattern recognition.

EfficientNet: It scales the network depth, width, and resolution in a balanced manner for achieving high efficiency and accuracy, and has demonstrated good results in multiclass classification of lung cancer. [1].

MobileNet: Designed for mobile and resource-limited environments, MobileNet uses depthwise separable convolutions technique, ensuring a lightweight yet effective model performance with efficient feature exraction. Adapted in breast cancer [10].

InceptionNet: Features a unique architecture with inception modules. It contains multiple kernel sizes on the same layer level to capture broader information at various scales and complexities, ideal for diverse medical image analyses. It was shown when incorporated with ResNet, it has promising improvement in performance (2-3%) against other benchmark datasets for breast cancer detection [2].

5　Title Image Processing Techniques

5.1　Patch Extraction

Addressing the substantial variation in widths and heights within the training dataset is crucial for ensuring a uniform size. However, given the distinct magnifications between Tissue Microarray (TMA) at 40x and Whole Slide Images (WSI) at 20x, a separate processing approach was implemented [4].

For TMAs, a region of 1028x1028 pixels was extracted, and utilizing the Lanczos filter, it was resized to 512x512 pixels, matching the 20x magnification. Subsequently, all images, including both TMAs and WSIs, were uniformly resized to 256x256 pixels, maintaining the 20x magnification. This standardized sizing procedure ensures consistency across the

dataset despite the initial differences in magnification and dimensions.

5.2　Stain Normalization

The normalization process strives to enhance the consistency and comparability of histopathology images across various time frames and centers. Previous studies have compared outcomes with and without stain normalization, revealing significantly improved performance in images subjected to stain normalization. All stain normalization methods have been systematically compared based on their performance[4]. In our approach, we opted for the Macenko method, a well-established color deconvolution technique, to standardize image color and facilitate machine analysis. This method is recognized for its maturity and effectiveness. The utilization of the torchstain software provides a convenient and efficient means to implement this technique[3]. It is shown in Figure 4.
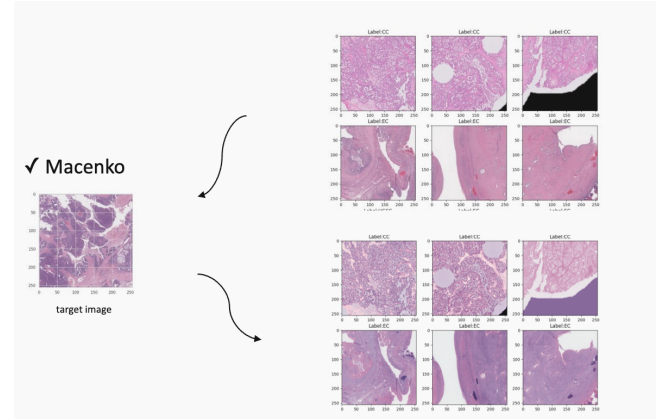

Figure 4. Stain Normalization

5.3　Data Augmentation

The image labels exhibit an evident imbalance, necessitating efforts to achieve a more balanced distribution. This adjustment is crucial to prevent overfitting of HGSC and ensure the machine can effectively predict the rare types. Striking a balance is essential, and thus, we opted for a partial balancing approach[4].

Specifically, we randomly selected EC and all LGSC, CC, and MC images and applied a 90-degree rotation. This not only mitigated the imbalance but also expanded the dataset. Given the well-known data-hungry nature of deep neural networks, this augmentation strategy enhances the model's ability to generalize effectively.

5.4　Noises Filtering and Scaling

Segmenting images into pieces introduces challenges, particularly when some patches contain minimal tissue. To address this, a proven strategy employed by leading teams in

the Prostate cANcer graDe Assessment (PANDA) challenge was adopted[9]. Patches were sorted based on the sum of all pixel values, and a subset with the lowest sum was selected. This approach ensures the retention of patches with a higher tissue percentage, as white pixels typically represent the background. Figure 5 shows the noises we filtered out.
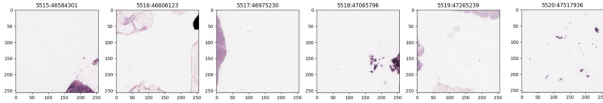
Figure 5. Noises filtered out

Additionally, scaling was performed to normalize the images using the average and variance from IMAGENET, optimizing them for computational efficiency. This normalization enhances the compatibility of the images for subsequent computational processes.

6 Model Architecture, training and hyperparameter tuning

6.1 Baseline classifiers

We performed a multi-class classification of OC subtypes that aims to detect image features with minimal bias. Therefore, we chose to perform a 6-class classification because the training dataset was imbalanced without the normal tissue or rare subtype samples. An additional sixth class was added with initialised minimum weight, in cases where the extracted feature negatively correlates with that of the five classes. To commence with, the sliced non-overlapping patches from data preprocessing undergo random shuffling before splitting into 85% for training set and 15% as validation set. The last layer of the proposed models, described in table 1, were removed and set with 6 outputs, and convolution layers set as trainable. Traing epochs were evaluated by comparing with one-hot encoded true labels with sparse categorical cross entropy as training loss. Models were trained with 10 epochs plus a diminishing learning rate to prevent over-fitting.

6.2 Ensemble and Voting Strategies

To enhance the accuracy and reliability of cancer subtype classification, the study employs an ensemble approach with innovative voting strategies.

Patch-Level Voting: Each model analyzes individual image patches, and their predictions are aggregated at the patch level. The final classification of an image is determined by the majority vote across all patches, ensuring a comprehensive analysis of the entire image.

Model-Level Voting: Predictions from each model (ResNet, VGG-16, etc.) are combined at the model level. The final decision for an image's classification is made based on the

consensus among these models, leveraging the strengths of each to enhance overall prediction accuracy.

This approach, integrating multiple models and sophisticated voting mechanisms, significantly improves the robustness and reliability of the classification system

7 Experiment Results

The experimental phase evaluated the performance of individual deep learning models and their ensemble using accuracy as metrics The results showed significant improvement in the accuracy of OC subtype classification.

Each model's performance, quantified in terms of accuracy, precision, recall, and F1 score, varied.

The ensemble strategy, particularly the patch-level and model-level voting methods, remarkably improved classification accuracy. The patch-level voting effectively handled variations within individual images, while model-level voting utilized the collective strengths of different models for a more accurate final classification.

Image preprocessing techniques played a crucial role in enhancing model performance. The stain normalization and patch extraction processes significantly improved image quality, leading to more accurate model predictions.

In summary, the experiments confirmed the effectiveness of the selected deep learning models and ensemble strategies in accurately classifying OC subtypes.

8 Conclusion

This study successfully demonstrated the effectiveness of employing advanced deep learning models and ensemble strategies in the classification of OC subtypes. The tailored use of ensemble models combined with voting significantly improved the accuracy of subtype classification. The findings underscore the potential of AI in medical diagnostics, offering a path toward more accurate and efficient cancer subtype identification.

Future work can explore further optimisation and integration of these models into clinical workflows, potentially revolutionizing OC diagnosis and treatment planning. This research not only provides a robust method for cancer subtype classification but also sets a precedent for the application of AI in other complex medical imaging tasks.

Contribution

WU Fei played a primary role in the data preprocessing phase and contributed to the execution of the report and the submission job.

SUNG Ming Shin Cindy played a primary role in introduction and model architecture , as well as execution of the report and submission.

XING Daiyan played a primary role in the voting strategies and the submission job, as well as model construction and the execution of the report.

Table 1. models xxxxxx

Strategy	Fine-tuned Model	Top 5 accuracy	Parameters	Time per inference step (GPU)
Interpretability-based	Baseline CNN-7	-	-	-
	VGG-16	90.1%	138.4M	4.2ms
Resource Efficiency-based	EfficientNetB0	93.3%	5.5M	4.9ms
	MobileNet	89.5%	4.3M	3.4ms
Deeper CNNs	ResNet50v2	92.1%	25.6M	4.6ms
	InceptionV3	93.7%	23.9M	6.9ms

Table 2. Model Evaluation

Strategy	Fine-tuned Model	Accuracy
Interpretability-based	7-layer CNN baseline	0.3308
	VGG-16	0.4941
Resource Efficiency-based	EfficientNet	0.5867
	MobileNet	0.5746
Deeper CNNs	ResNet_v2_	0.5413
	Inception-Net	0.6089
Ensemble-based	EfficientNet + MobileNet + Inception-Net	0.6275

Table 3. voting

Model	Accuracy	Voting Accuracy	Balanced Accuracy
7-layer CNN baseline	0.3308	0.3723	0.3697
VGG-16	0.4941	0.5372	0.4547
EfficientNet	0.5867	0.5747	0.6114
MobileNet	0.5642	0.5612	0.5411
ResNet_v2_	0.5413	0.5463	0.5686
InceptionNet	0.6089	0.6337	0.6277
Ensemble-based	0.6275	0.6494	0.6525

References

[1] Karar Ali, Zaffar Ahmed Shaikh, Abdullah Ayub Khan, and Asif Ali Laghari. 2022. Multiclass skin cancer classification using efficientnets – a first step towards preventing skin cancer. *Neuroscience Informatics* 2, 4 (2022), 100034. https://doi.org/10.1016/j.neuri.2021.100034

[2] Md Zahangir Alom, Chris Yakopcic, Mst. Shamima Nasrin, Tarek M. Taha, and Vijayan K. Asari. 2019. Breast cancer classification from histopathological images with inception recurrent residual convolutional neural network. *Journal of Digital Imaging* 32, 4 (2019), 605–617. https://doi.org/10.1007/s10278-019-00182-7

[3] Carlo Alberto Barbano and André Pedersen. 2022. *EIDOSLAB/torchstain: v1.2.0-stable.* https://doi.org/10.5281/zenodo.6979540

[4] Jeffrey Boschman. 2022. Generalizable deep learning models for epithelial ovarian carcinoma classification. 49-100 (2022), 138. https://dx.doi.org/10.14288/1.0416550

[5] Jeffrey Boschman, Hossein Farahani, Amirali Darbandsari, Pouya Ahmadvand, Ashley Van Spankeren, David Farnell, Adrian B. Levine, Julia R. Naso, Andrew Churg, Steven J.M. Jones, and et al. 2021. The utility of color normalization for AI-based diagnosis of hematoxylin and eosin-stained pathology images. *The Journal of Pathology* 256, 1 (Jan. 2021), 15–24. https://doi.org/10.1002/path.5797

[6] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L. Siegel, Lindsey A. Torre, and Ahmedin Jemal. 2007. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Journal of the American College of Medicine* 54, 2, Article 5 (April 2007), 50 pages. https://doi.org/10.1145/1219092.1219093

[7] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L. Siegel, Lindsey A. Torre, and Ahmedin Jemal. 2018. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians* 68, 6 (2018), 394–424. https://doi.org/10.3322/caac.21492

[8] Joseph L. Breeden and Eugenia Leonova. 2021. Creating unbiased machine learning models by design. *Journal of Risk and Financial Management* 14, 11 (2021), 565. https://doi.org/10.3390/jrfm14110565

[9] Wouter Bulten, Kimmo Kartasalo, Peter H.C. Chen, Patrik Ström, Hans Pinckaers, Kumar Nagpal, and et al. 2022. Artificial Intelligence for Diagnosis and Gleason Grading of Prostate Cancer: The PANDA Challenge. *Nature Medicine* 28, 1 (13 Jan 2022), 154–163. https://doi.org/10.1038/s41591-021-01620-2

[10] Huang et al. 2022. BM-net: CNN-based MobileNet-V3 and bilinear structure for breast cancer detection in whole slide images. *Bioengineering* 9, 6 (2022), 261. https://doi.org/10.3390/bioengineering9060261

[11] M. Fung-Kee-Fung, E.B. Kennedy, J. Biagi, T. Colgan, D. D'Souza, L.M. Elit, A. Hunter, J. Irish, R. McLeod, and B. Rosen. 2015. The Optimal Organization of Gynecologic Oncology Services: A systematic review. *Current Oncology* 22, 4 (2015), 282–293. https://doi.org/10.3747/co.22.2482

[12] Khushboo Munir, Hassan Elahi, Afsheen Ayub, Fabrizio Frezza, and Antonello Rizzi. 2019. Cancer diagnosis using Deep Learning: A bibliographic review. *Cancers* 11, 9 (2019), 1235. https://doi.org/10.3390/cancers11091235

[13] Devvi Sarwinda, Radifa Hilya Paradisa, Alhadi Bustamam, and Pinkie Anggia. 2021. Deep learning in image classification using residual network (ResNet) variants for detection of colorectal cancer. *Procedia Computer Science* 179 (2021), 423–431. https://doi.org/10.1016/j.procs.2021.01.025

[14] Ankita Singh, Khairun Nisaa, Sudipta Bhattacharyya, and Amirul Islam Mallick. 2019. Immunogenicity and protective efficacy of mucosal delivery of recombinant HCP of *Campylobacter jejuni* type VI secretion system (T6SS) in chickens. *Molecular Immunology* 111 (2019), 182–197. https://doi.org/10.1016/j.molimm.2019.04.016

[15] Flora Vernooij, Peter Heintz, Els Witteveen, and Yolanda van der Graaf. 2007. The outcomes of ovarian cancer treatment are better when provided by gynecologic oncologists and in Specialized Hospitals: A systematic review. *Gynecologic Oncology* 105, 3, Article 5 (April 2007), 12 pages. https://doi.org/10.1016/j.ygyno.2007.02.030