Final Project Report - Deep Learning MVA: Classifying Ovarian Cancer Subtypes

Raphaël Romand-Ferroni
CentraleSupélec
raphael.romandferroni@student-cs.fr

Arthur Gallois
CentraleSupélec
arthur.gallois@student-cs.fr

Abstract

This proposal outlines a project aimed at classifying ovarian cancer subtypes using deep learning techniques on histopathology images. The project leverages a dataset from the UBC-OCEAN Kaggle competition [1], aiming to enhance diagnostic accuracy and contribute to personalized treatment strategies.

1. Motivation

Ovarian cancer is a profoundly impactful disease affecting millions of women worldwide. Despite advancements in medical treatment, it continues to be the deadliest form of cancer within the female reproductive system, boasting a meager five-year survival rate of only 47%. One of the primary factors contributing to this bleak prognosis lies in the intricate challenge of precisely diagnosing ovarian cancer subtypes. These subtypes exhibit distinct cellular morphologies, etiologies, molecular and genetic profiles, as well as clinical characteristics. The emergence of subtype-specific treatment strategies offers hope, yet their successful implementation hinges on the accurate identification of subtypes—a task that can be substantially enhanced through the application of data science. Presently, the diagnostic process relies heavily on pathologists to discern subtypes, a practice fraught with difficulties, including interobserver variability and diagnostic reproducibility. Moreover, underserved communities often lack access to specialized pathologists, while even well-developed regions grapple with a shortage of pathologists with expertise in gynecological malignancies. Although deep learning models have showcased remarkable proficiency in scrutinizing histopathology images for ovarian cancer subtype identification, they are not without their challenges. These include the imperative need for expansive and diverse datasets for training and validation, as well as navigating technical, ethical, and financial constraints. Our endeavor is dedicated to surmounting these challenges and contributing to the advancement of more precise and dependable diagnostic tools for ovarian cancer. By harnessing the largest and most diverse repository of ovarian cancer histopathology images from over 20 centers across four continents, we aspire to elevate the accuracy and reproducibility of ovarian cancer subtype identification. Ultimately, this pursuit aims to translate into improved treatment outcomes for patients.

2. Problem Definition

2.1. An overview of the dataset

The dataset for this competition comprises intricate microscopy scans of biopsy samples procured from ovarian cancer patients. These scans are initially categorized into two distinct classes: whole slide images (WSI) and tissue microarrays (TMA). The WSIs are captured at a magnification of 20x and can be quite expansive in dimensions, encompassing intricate details. In contrast, the TMAs are more compact, typically measuring around 4,000x4,000 pixels, but are acquired at a higher magnification of 40x, enabling a closer examination of cellular structures. The training dataset is a compilation of images collected from diverse hospitals, each thoughtfully labeled to denote the specific subtype of ovarian cancer present within the image. In stark contrast, the test dataset comprises images sourced from different medical facilities, featuring notably large images, some exceeding dimensions of nearly 100,000 x 50,000 pixels. This test set encompasses approximately 2,000 images, with the majority being TMAs, and it boasts a substantial size of 550 GB. It's noteworthy that the training set also includes thumbnail representations of the WSIs, accompanied by supplemental masks that delineate regions within the respective WSIs as cancerous, healthy, or necrotic. The central challenge posed by this competition is to proficiently categorize the type of ovarian cancer depicted in the images, while concurrently identifying any outliers that defy classification within the provided subtypes.

We encountered several significant challenges due to

the inherent nature of the dataset, which are summarized below:

- One of the foremost challenges we confronted was the imbalance within the dataset. This imbalance manifested in two distinct aspects. Firstly, there was an uneven distribution between whole slide images (WSI) and tissue microarray (TMA) images. The former, captured at 20x magnification, were notably more prevalent than the latter, which were smaller and captured at 40x magnification. Secondly, there existed a notable disparity in the representation of different ovarian cancer subtypes within the dataset.
- Another formidable obstacle we had to contend with was the sheer size of the images. Processing and analyzing such high-dimensional images demanded substantial computational resources and posed a considerable computational burden.
- The presence of outliers in the test set posed a unique challenge during our training process. These outliers were instances that did not conform to any of the predefined ovarian cancer subtype categories. Effectively handling these outliers and ensuring that our models could account for them was a crucial aspect of our approach.

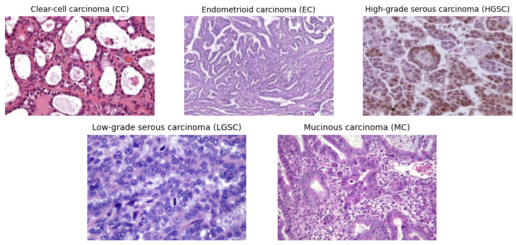

Figure 1. Ovarian cancer subtype images with their labels.

The competition revolves around the classification of ovarian cancer subtypes, with each subtype representing a distinct category. These subtypes are comprehensively described based on the valuable work conducted by Koshiyama M. and colleagues [5]. Here, we provide detailed insights into each of these subtypes:

- Clear-cell carcinoma (CC) is characterized by the presence of clear cells and is classified as a rare form of cancer. It is notable for the distinct appearance of its cells and is one of the subtypes encountered in ovarian cancer.
- Endometrioid carcinoma (EC) Derived from endometriosis, endometrioid carcinoma accounts for approximately 10% of all ovarian epithelial cancers. It represents a specific histological subtype with unique characteristics.
- High-grade serous carcinoma (HGSC) stands out as the most lethal histotype among ovarian cancers. It is also the second most common gynecologic malignancy. This subtype is associated with particularly aggressive features.
- Low-grade serous carcinoma (LGSC) is a rare histological subtype of ovarian carcinoma. In contrast to HGSC, LGSC tends to occur at a younger age and typically follows a less aggressive or indolent course. It exhibits distinct clinical characteristics.
- Mucinous carcinoma (MC) is characterized by its histological features, including a complex papillary growth pattern with microscopic cystic glands. It is considered a rare and malignant epithelial tumor of the ovary.

3. Related Work

In recent years, there has been a burgeoning interest in the classification of ovarian cancer subtypes and the detection of outliers. This heightened interest has led to extensive research into employing various deep learning techniques to classify ovarian carcinoma based on cell morphology and histopathological features, as outlined in the work by Ziyambe in 2023 [10]. This area of research has garnered substantial attention due to the immense potential of deep learning algorithms in enhancing the accuracy and efficiency of ovarian cancer diagnosis. The utilization of cutting-edge technology holds promise in revolutionizing our approach to diagnosing and understanding ovarian cancer subtypes.

An approach that has been extensively explored involves the use of Deep Convolutional Neural Networks (DCNNs). For instance, in a study conducted by Wu et al. in 2018 [9], a DCNN based on the AlexNet architecture was employed to classify different types of ovarian cancers using cytological images. The findings from this study revealed that the DCNN, constructed based on the AlexNet model, demonstrated the capability to recognize a majority of ovarian cancer cells. However, it's worth noting that certain images were misclassified, primarily attributed to issues such as poor image clarity and overlapping cells. This underscores the significance of image quality in the accuracy of DCNN-based classification methods, highlighting the need for high-quality data in achieving more precise results in the field of ovarian cancer diagnosis.

Another alternative method in the classification of ovarian tumors involves the utilization of Convolutional

Neural Networks (CNNs) combined with a Convolutional Autoencoder (CAE). In a study conducted by Jung et al. in 2022 [4], they developed a model known as CNN-CAE for the purpose of classifying ovarian tumors, using ultrasound images as the dataset. Remarkably, this model achieved high accuracy in the classification of ovarian tumors. Furthermore, the study conducted a comparison with other machine learning methods, and the results demonstrated that the CNN-CAE model outperformed traditional machine learning algorithms both in terms of accuracy and efficiency. This highlights the potential of deep learning techniques, specifically the CNN-CAE architecture, in enhancing the accuracy and effectiveness of ovarian tumor classification, particularly in the context of medical imaging.

Radiomics has emerged as a promising field of research for classifying ovarian cancer subtypes and detecting outliers. In a study conducted by Tardieu et al. in 2022 [6], the researchers aimed to investigate the associations between Magnetic Resonance Microscopy (MRM)-derived radiomic features and histopathology in ovarian cancer. The findings of this study revealed a strong correlation between MRM images and whole-slide histology images, with a Dice index of 0.77. This significant correlation suggests that radiomics has the potential to offer valuable insights for the diagnosis and prognosis of ovarian cancer. It highlights the potential of leveraging radiomic features to improve our understanding of ovarian cancer and enhance diagnostic and prognostic capabilities.

A groundbreaking deep learning framework was introduced by Ziyambe in 2023 [10], aimed at predicting and diagnosing ovarian cancer. This innovative algorithm exhibited an impressive F1-score of 0.94 and demonstrated the ability to predict and diagnose images in less than 5 seconds. These remarkable results underscore the tremendous potential of deep learning algorithms to enhance both the speed and accuracy of ovarian cancer diagnosis. Such advancements hold promise for improving patient outcomes and streamlining the diagnostic process in the field of ovarian cancer research and healthcare.

Furthermore, the exploration of Extracellular Vesicle-Associated Protein Biomarkers has offered promising insights into early detection methods for high-grade serous ovarian cancer. Trinidad et al. (2023) [7] conducted a comprehensive study that yielded impressive results, with Area Under the Curve (AUC) values ranging from 0.85 to 0.98 for these biomarkers. Such high AUC values signify a substantial distinction between individuals with the disease and healthy controls. These findings underscore the potential significance of extracellular vesicle-associated protein biomarkers as a valuable tool for the early detection of ovarian cancer. This research opens up new avenues for early intervention and improved outcomes in the diagnosis and treatment of this malignancy.

In conclusion, the studies highlighted in this discussion underscore the considerable potential of deep learning techniques in the classification of ovarian cancer subtypes and the detection of outliers, utilizing microscopic scans of biopsy specimens. These advancements represent significant progress in the field of ovarian cancer diagnosis. However, it is crucial to acknowledge that further research is needed to refine and optimize these techniques for clinical applications. Future studies should aim to address several key areas to enhance the accuracy and reliability of these methods. Firstly, efforts should be made to overcome the challenge of requiring high-quality images, as image clarity plays a vital role in the effectiveness of deep learning models. Additionally, the acquisition of larger and more diverse datasets is essential to improve the robustness of these models and ensure their applicability to a broader range of cases. Moreover, it is imperative to conduct comprehensive clinical evaluations to determine the real-world usefulness of these methods. Assessing their impact on patient outcomes and the clinical workflow is essential to gauge their practicality in healthcare settings. As research in this field continues to evolve, it holds the promise of revolutionizing the early diagnosis and treatment of ovarian cancer, ultimately improving patient care and outcomes.

Deep learning models have demonstrated a remarkable ability to analyse histopathology images. However, certain challenges remain, including the need for a significant amount of training data, preferably from a single source. Technical, ethical and financial limitations, coupled with confidentiality concerns, are significant barriers to the training process. In the competition, participants will have access to the largest and most diverse ovarian cancer dataset of histopathology images from over 20 centres across four continents. This dataset aims to address the above challenges and facilitate the development of more robust and accurate deep learning models for histopathology image analysis.

4. Methodology

4.1. Problem simplification

The decision to use resized thumbnails of images instead of their full resolution is a practical approach to simplify the implementation of the model and reduce training time and resource requirements. It's a common practice in deep learning to work with lower-resolution images, especially when computational resources are limited. The assumption

that relatively good results can still be obtained from lower-quality images is reasonable, and it allows for more efficient model training.

Regarding the handling of outliers, it's understandable that dealing with outliers can be challenging, and focusing on classifying the five basic classes is a substantial task in itself. It's essential to make strategic decisions based on the available resources and project constraints. Ignoring outliers in this context is a valid choice, even though it may result in a lower score on the competition leaderboard due to false predictions for outliers.

Ultimately, the project's success should be evaluated based on its ability to accurately classify the main classes of ovarian cancer subtypes, and this decision allows for a more focused approach to achieving that goal within the project's time frame. It's important to prioritize the primary objective while considering the available resources and constraints.

4.2. Data preprocessing

Handling the imbalance in your training dataset is a crucial step in ensuring that your deep learning model performs well across all classes. The decision to use a stratified split of 80% and 20% for training and validation sets is a sound approach to address this issue. By doing so, you ensure that the same distribution of labels is maintained in both sets, which helps in training a model that can generalize well and make accurate predictions for all classes, including the less represented ones like Mucinous carcinoma (MC) and Low-grade serous carcinoma (LGSC). This approach helps mitigate the impact of class imbalance on the model's performance and ensures fair evaluation during validation.

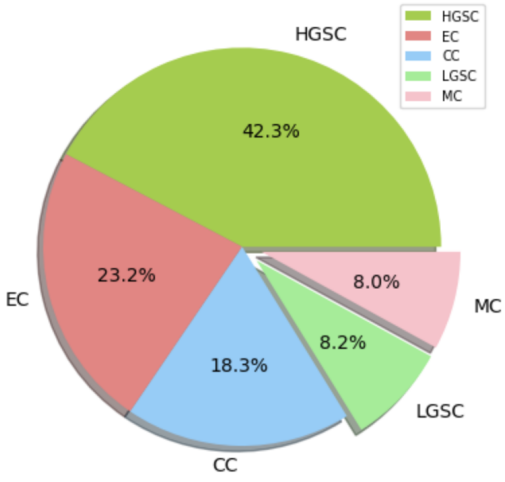

Figure 2. Imbalanced training dataset visualization.

Addressing the significant over-representation of whole slide images (WSI) over tissue microarray (TMA) images is indeed a critical consideration, especially when the test set is expected to contain a majority of TMA images. Developing a pipeline to classify between these two different image types is a smart strategy to ensure that your model can handle both types effectively. By doing so, you are preparing your model to perform well on the test set, which is representative of the real-world scenario.This approach acknowledges the importance of adapting to the data distribution in the test set, which is a key aspect of building a robust and reliable deep learning model for this competition. It also demonstrates your team's awareness of the challenges posed by dataset characteristics and your proactive efforts to address them.

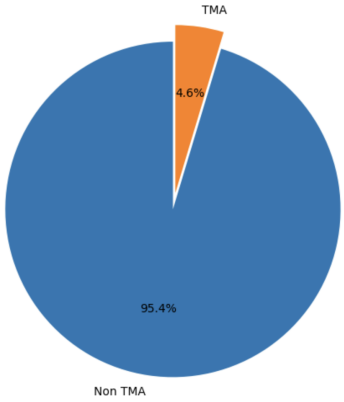

Figure 3. Repartition of WSI and TMA images in the training set.

Using the presence of black pixels as a criterion to distinguish between WSI and TMA images is a practical approach, especially given the difference in magnification. This heuristic provides a simple yet effective way to make this classification. By setting a threshold of 5% black pixels, you are creating a clear rule for determining the image type, which can be automated and integrated into your pipeline.

This approach takes advantage of visual characteristics in the data to make an informed decision, and it aligns with the practical constraints of the competition. It's a valuable step in preprocessing the data and ensuring that your model is properly trained on both WSI and TMA images.

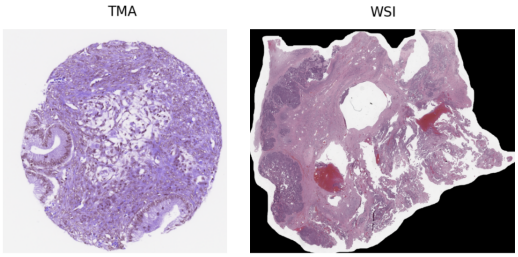

Figure 4. TMA and WSI samples from the training set.

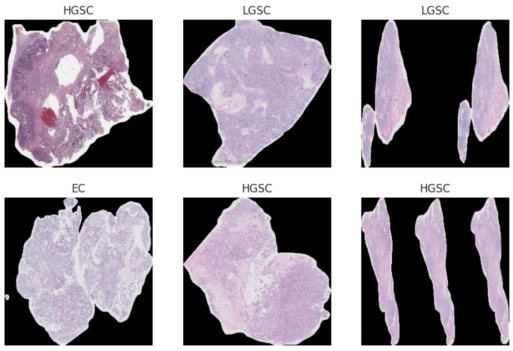

Figure 5. WSI image samples.

4.3. Model choice

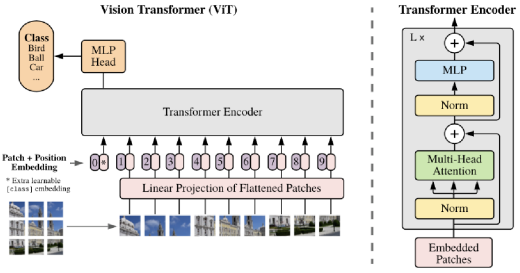

Figure 6. ViT Architecture.

The model selected for our final submission was a pretrained version of **ViT** : Vision Transformers (2021, Google Brain) [3], that make use of the advantages brought by the transformer architecture introduced in 2017 by Vaswani and al. [8] into Computer Vision. Following the illustration of the model's architecture in Figure 6, ViT split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder made of L stacked transformer block that make use of the Multi-Head attention representations combined with Layer Normalization. In order to perform classification, we use the standard approach of adding an extra learnable classification token to the sequence. By splitting an image in small patches as we would have done with a sentence by splitting it in tokens and by using multiple self-attention layers, ViT is able to embed the information globally across the overall image and learn the relative location of the image patches to reconstruct the structure of the original image.

We used a pretrained version of ViT that was trained on ImageNet, a dataset of 14 million images of resolution 224x224 pixels and 21 different classes.

Our training dataset is composed of high-dimensional images of width 3000 and a height ranging from 530 to 7800 pixels as illustrated in the violin plot of the thumbnails heights in 7. We thus needed to handle this high dimensionality to feed our ViT model that was pretrained on 224x224 pixel images.

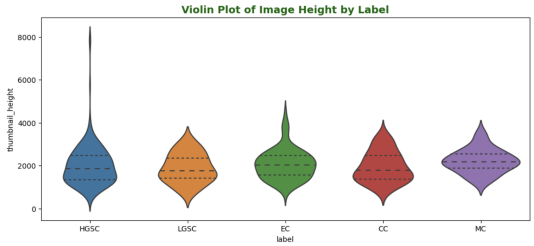

Figure 7. Violin plot of image height by label.

The original image of high dimension is thus first split into smaller patches and we applied a padding if necessary. As input, we give our function the training images with tile size set to 256. Padding is added to make the image dimensions divisible by the tile size and we eventually end up with images of shape $(height + pad_h, w + pad_h, channels)$. We can visualize the different patches of the image in the grid just below on Figure 8. We ordered the patches based on their pixel intensity, with the assumption that patches with higher intensity contain more information. We concatenate these patches to form a final single image of patches.

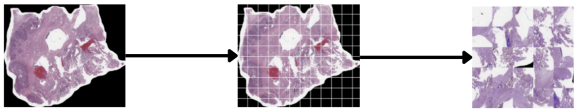

Figure 8. Tiling process.

Using PyTorch's torchvision library, we applied the following transformations to our training images:

- Resizing the image to 224×224 to fit into our ViT model.
- Randomly rotating the image by 15° degrees
- Flipping the image upside-down and left-right with a probability of 0.5
- Randomly changes the brightness, contrast and hue of the image
- Convert the image into a PyTorch tensor which automatically scales the image between 0 and 1.
- We eventually normalize the pixel values using the parameters used for ImageNet dataset [2]

4.3.1 Training

The training process last for a maximum of 30 epochs. The validation set was used for early stopping (using the validation balanced accuracy as stopping criterion) with a patience parameter of 4 epochs. In addition to this, we used the AdamW optimizer with a constant learning rate of 0.0001 and weight decay of 0.001. The loss used was the Cross-Entropy (CE) loss adjusted with the class weights of the training set to help the model with the low-represented labels. In addition to that, we created a Weighted Random Sampler to ensure that each batch during the training process had a balanced number from each class.

5. Results

5.1. Training results

The model's performance is evaluated using balanced accuracy, as per the competition's guidelines. We achieve a promising **44.5%** on our final model. The confusion matrix can be found on Figure 9. We see that our model still struggle to detect the low-represented labels MC and LGSC and that EC is often wrongly detected as being HGSC.

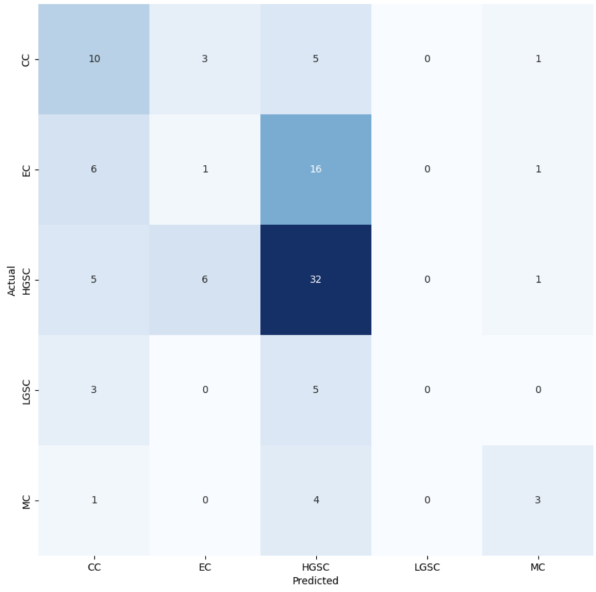

Figure 9. ViT Confusion Matrix.

In addition to the Vision Transformer (ViT) model, I also experimented with ResNet-50 and VGG-16 architectures, both trained under the same parameters and transformations. While these two models are renowned for their effectiveness in computer vision tasks, their performance on the specific task at hand yielded balanced accuracy scores of **0.427%** and **0.33%**, respectively. ResNet-50, short for Residual Network with 50 layers, is a deep convolutional neural network architecture. It is characterized by its residual blocks, which allow for training very deep networks while mitigating the vanishing gradient problem. VGG-16, on the other hand, is a classic convolutional neural network architecture known for its straightforward structure, comprising 16 weight layers. Although it has been widely used in various computer vision applications, it may struggle with capturing complex patterns and hierarchical features compared to more modern architectures like ResNet and ViT. The different balanced accuracy yielded by these models have been summarized on Figure 10

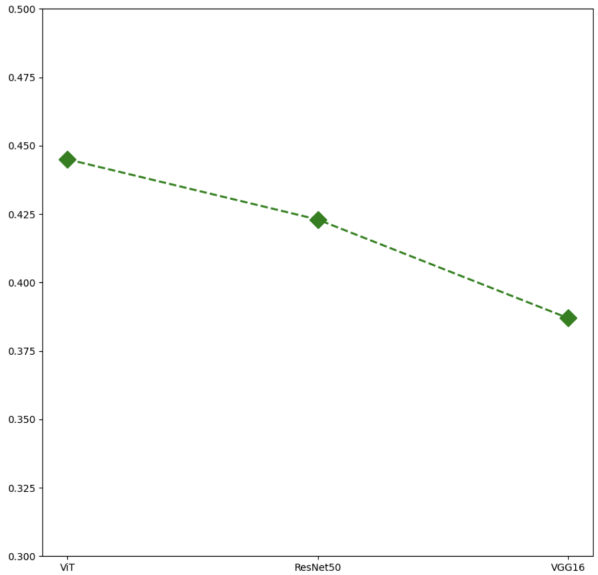

Figure 10. Balanced Accuracy of the models used.

5.2. Competition results

The achieved balanced accuracy on the final dataset is 0.35, positioning our team at 455th place out of 1327 participants. We find this result quite satisfying, especially considering that we joined the competition after it had already begun on October 6, 2023. Our team officially entered the competition around November 20.

As anticipated, our balanced accuracy decreased when the model was submitted for the competition compared to the performance on the test set during training. This reduction in accuracy can be attributed to several factors. Firstly, we did not address the challenge of outliers in our model, making it likely to produce incorrect predictions when faced with outlier classifications. Additionally, a significant source of performance loss is the fact that the test images are sourced from different hospitals than the training images. This introduces variability in the data distribution and can impact the model's ability to generalize effectively. Another important point to highlight is that the competition explicitly states that the test dataset is designed to evaluate the generalization capabilities of models. This means that

the test set is intentionally crafted to challenge the models' ability to adapt to diverse data patterns and variations. Furthermore, it's crucial to emphasize that the test phase remains entirely concealed from participants. The evaluation of solutions occurs in a "black box" manner, where participants have no access to the test images or any output generated during the evaluation process. This adds an element of unpredictability to the competition, making it imperative for participants to build models that can generalize well to new and unseen data patterns.

6. Conclusion

We participated in this competition for several reasons. Firstly, it enabled us to apply a machine learning model we had studied during the course and compare its performance with that of other data scientists, both professional and amateur. Moreover, although our contribution may not have revolutionised scientific research, the competition addresses a crucial issue: the classification of breast cancers. We are pleased to have participated and helped raise the visibility of this competition, along with other participants, in the hope of attracting more experienced data scientists to these important issues. Additionally, working on these challenges is more engaging than working on 'toy problems'. This competition has presented many challenges, and we are glad to have been able to face them.

References

[1] OTTA Consortium Anthony Karnezis Ardalan Akbari Sirim Kim Ashley Chow Sohier Dane Allen Zhang Maryam Asadi Ali Bashashati, Hossein Farahani. Ubc ovarian cancer subtype classification and outlier detection (ubc-ocean), 2023. 1

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 5

[4] Yuyeon Jung, Taewan Kim, Mi-Ryung Han, Sejin Kim, Geunyoung Kim, Seungchul Lee, and Youn Jin Choi. Ovarian tumor diagnosis using deep convolutional neural networks and a denoising convolutional autoencoder. *Scientific Reports*, 12(1):17024, 2022. 3

[5] Matsumura N. Konishi I. Koshiyama, M. Subtypes of ovarian cancer and ovarian cancer screening. *Diagnostics (Basel, Switzerland)*, 2017. 2

[6] Marion Tardieu, Yulia Lakhman, Lakhdar Khellaf, Maida Cardoso, Olivia Sgarbura, Pierre-Emmanuel Colombo, Mireia Crispin-Ortuzar, Evis Sala, Christophe Goze-Bac, and Stephanie Nougaret. Assessing histology structures by ex vivo mr microscopy and exploring the link between mrm-derived radiomic features and histopathology in ovarian cancer. *Frontiers in Oncology*, 11:771848, 2022. 3

[7] Camille V Trinidad, Harsh B Pathak, Shibo Cheng, Shin-Cheng Tzeng, Rashna Madan, Mihaela E Sardiu, Leonidas E Bantis, Clayton Deighan, Andrea Jewell, Sagar Rayamajhi, et al. Lineage specific extracellular vesicle-associated protein biomarkers for the early detection of high grade serous ovarian cancer. *Scientific Reports*, 13(1):18341, 2023. 3

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. 5

[9] Miao Wu, Chuanbo Yan, Huiqiang Liu, and Qian Liu. Automatic classification of ovarian cancer types from cytological images using deep convolutional neural networks. *Bioscience reports*, 38(3):BSR20180289, 2018. 2

[10] Blessed Ziyambe, Abid Yahya, Tawanda Mushiri, Muhammad Usman Tariq, Qaisar Abbas, Muhammad Babar, Mubarak Albathan, Muhammad Asim, Ayyaz Hussain, and Sohail Jabbar. A deep learning framework for the prediction and diagnosis of ovarian cancer in pre-and post-menopausal women. *Diagnostics*, 13(10):1703, 2023. 2, 3