

Towards a privacy-preserving distributed cloud service for preprocessing very large medical images

Yuandou Wang¹, Neel Kanwal², Kjersti Engan², Chunming Rong², Zhiming Zhao^{1,3}

¹Multiscale Networked Systems, University of Amsterdam, the Netherlands

²Department of Electrical Engineering and Computer Science, University of Stavanger, Norway

³LifeWatch ERIC Virtual Lab and Innovation Center (VLIC), Amsterdam, the Netherlands

Email: {y.wang8, z.zhao}@uva.nl; {neel.kanwal, kjersti.engan, chunming.rong}@uis.no

Abstract—Digitized histopathology glass slides, known as Whole Slide Images (WSIs), are often several gigapixels large and contain sensitive metadata information, which makes distributed processing unfeasible. Moreover, artifacts in WSIs may result in unreliable predictions when directly applied by Deep Learning (DL) algorithms. Therefore, preprocessing WSIs is beneficial, e.g., eliminating privacy-sensitive information, splitting a gigapixel medical image into tiles, and removing the diagnostically irrelevant areas. This work proposes a cloud service to parallelize the preprocessing pipeline for large medical images. The data and model parallelization will not only boost the end-to-end processing efficiency for histological tasks but also secure the reconstruction of WSI by randomly distributing tiles across processing nodes. Furthermore, the initial steps of the pipeline will be integrated into the Jupyter-based Virtual Research Environment (VRE) to enable image owners to configure and automate the execution process based on resource allocation.

Index Terms—Computational Pathology, Cloud Computing, Privacy-preserving, Image Preprocessing, Virtual Research Environment, Infrastructure Planning

I. INTRODUCTION

Deep Learning (DL) approaches have advanced and innovated automatic diagnostics, such as quantifying the presence of cancerous cells in digitized histopathology glass slides, called Whole Slide Images (WSIs). However, running these diagnostic services over a large scale requires a significant infrastructure capacity for storing and processing images using complex DL models, e.g., cloud computing and High-Performance computing (HPC), are often needed. The owners of the medical images, e.g., hospitals, often do not have such an infrastructure and have to rely on collaborators with remote infrastructure resources.

Establishing a DL-based pipeline for medical images on a remote infrastructure is challenging; for instance, 1) WSIs often contain privacy-sensitive information in their metadata and cannot be directly sent out to the public cloud from the hospitals; 2) WSIs are often very large and require high network bandwidth to upload; and 3) WSIs are split into tiles to process [1], [2] and require specialized hardware, e.g., GPUs, to run complex DL models. Furthermore, it is often complicated to deploy an end-to-end pipeline and create an efficient re-configurable workflow [3] on remote infrastructure.

In this paper, we will tackle these challenges by proposing a cloud-based service that will be integrated into a collaborative

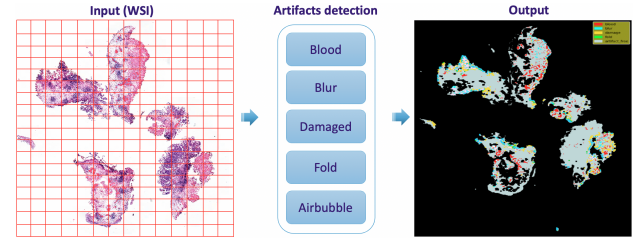


Fig. 1. An overview of the preprocessing pipeline for detecting artifacts in whole slide images.

virtual research environment based on the works [4], [5], and we will present a use case of a medical image preprocessing application from digital pathology domain to testify our methodology.

II. CASE STUDY

Digital pathology overcomes the hurdles of traditional histopathology by facilitating the diagnostic process using a WSI [1]. The preparation of histological glass slides may result in the appearance of artifacts on the obtained WSI due to improper handling of the tissue specimen during the tissue processing stages. These histological artifacts are diagnostically irrelevant and are usually ignored by pathologists in the diagnosis process [2], [6]. Therefore, it is vital to detect and remove them before applying diagnostic or prognostic algorithms. Some frequently appearing artifacts are damaged tissue, folded tissue, blur, air bubbles, and diagnostically irrelevant blood [2], [6]. Computational pathology (CPATH) researchers may run DL-based artifact preprocessing algorithms over thousands of WSIs before applying diagnostic algorithms, requiring powerful computational resources to process WSIs efficiently. Fig. 1 presents an overview of such artifact preprocessing pipeline, which is an ensemble of five DL models for blood, blur, damaged tissue, folded tissue, and air bubbles detection tasks from WSI in a binary fashion.

Traditionally, the artifact preprocessing pipeline runs over a single machine, bringing the disadvantages of a single security breach or system failure. Besides, handling gigapixel WSIs is time-consuming and resource-intensive, which raises the demand for parallel distributed computing. Nevertheless, WSIs processed on private clouds in research environments are de-identified or pseudonymized under various regulations. This

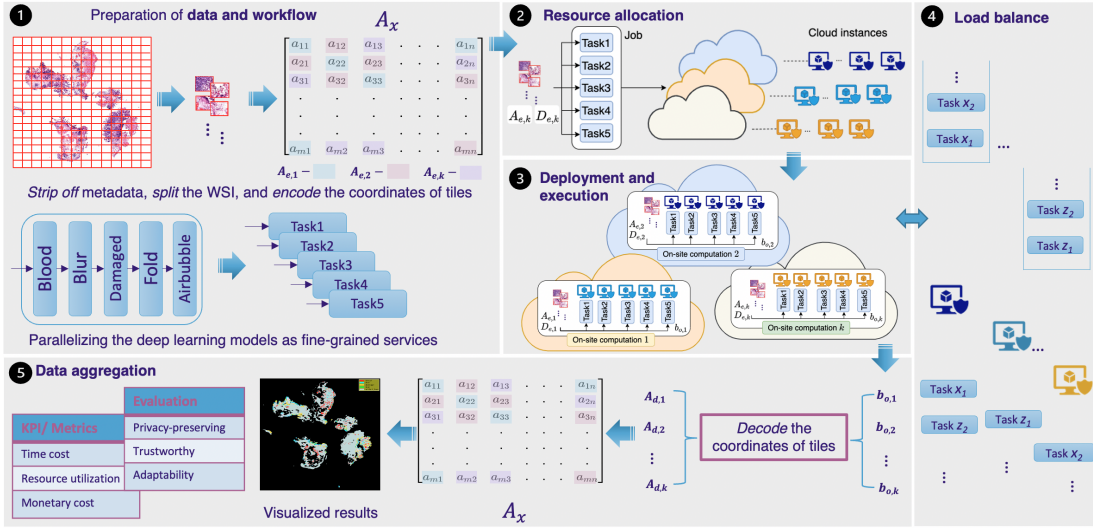


Fig. 2. The methodology proposed in this work. Step ❶ - preparation of data and workflow; step ❷ - resource allocation; step ❸ - deployment and execution; step ❹ - load balance; and step ❺ - aggregating distributed data in a summary view to visualize the final results.

raises concerns about the embedded privacy-sensitive metadata while making distributed processing over public clouds.

III. METHODOLOGY

We introduce a methodology to cope with the highlighted challenges, as shown in Fig. 2. It consists of five main steps: preparation of data and workflow, resource allocation, deployment and execution, load balance, and data aggregation.

A. Data and workflow preparation

Step ❶ aims to introduce parallelism and encryption available for the next steps. To guarantee privacy-preserving requirements, we remove the *metadata* from WSI before splitting the gigapixel image into many image tiles to introduce data parallelism. Meanwhile, containerizing the computational tasks as several reusable fine-grained services can improve scalability and security since they are isolated from each other and from the host system. Besides, we apply a matrix A_x to record the distribution of the tiles over a grid which can be encoded to hide its coordinates and divided into sub-matrices $A_{e,1}, A_{e,2}, \dots, A_{e,K}$. Each sub-matrix is considered as an index of distributed dataset $D_{e,k}$ for each service-based task.

B. Resource allocation

Based on the prepared data and service-based tasks, step ❷ is to map available resources (e.g., clusters at universities and commercial clouds) to various tasks in a manner that optimizes their utilization and satisfies requirements [7]. Related methods, such as IC-PCP [8] and machine learning-based approach [9], can be improved for workflow scheduling. For bi-objective optimization, such as reducing execution time and cost, there are trade-offs with time performance and monetary cost over the cloud. On this basis, this paper looks into workflow scheduling problems under the influence of privacy requirements and the split data sets, so the research problem is more challenging.

C. Deployment and execution

After a deployment plan is created at step ❷, the datasets and service-based tasks will be assigned to planned infrastructures equipped with computation, communication, and storage resources. The system should ensure that data storage and task execution remain in place and continue to be effective even among changes (such as downtime, errors, or attackers) to the system or emerging threats, according to step ❸. Due to distributed processing, it reduces the burden of a single machine and avoids a single security breach.

D. Load Balance

Considering that computing nodes may unpredictably slow down or fail during their execution, step ❹ aims to improve the performance, reliability, and load balance of task-based applications [10]. This approach asymptotically achieves near-ideal load balancing and computation cost in the presence of slow nodes (stragglers), which could also be complementary to workflow scheduling.

E. Data aggregation

Step ❺ takes the predicted distributed output, $b_{o,k}$ for each encoded tile, from step ❸ for every service task, to reconstruct the encoded distributions as $A_{d,k}$. Privacy preservation can be guaranteed by the random-value perturbation technique [11]. This approach has solid theoretical foundations and is easier to apply for the reconstruction of the encrypted data than others (e.g., differential privacy and secure multiparty computation [12]) especially considering a data matrix manipulation. Then by tracing back coordinates, we can create a segmentation mask for detected artifacts. It presents the results of the DL-based artifacts detection in a summary view, incl. visualization, evaluation, and metrics.

IV. SYSTEM MODEL

We sketch out a privacy-preserving distributed processing pipeline for the medical application, shown in Fig. 3. It is composed of three main steps - viz *Splitting*, *Computation*, and

Aggregation. Both Splitting (See in step ❶) and Aggregation (See in step ❷) will be executed on a *Trusted Server*. Such distributed data processing application can be defined as a tuple:

$$\mathcal{A} = (\mathcal{M}, \varepsilon, \mathcal{D}, \mathcal{R}, \mathcal{I}, req) \quad (1)$$

where \mathcal{M} denotes a set of lightweight interconnected *microservices*. A *source microservice* m_{src} processing the data stream produced by the source dataset \mathcal{D} . A *sink microservice* m_{snk} representing its final results \mathcal{R} . ε indicates a set of *data streams* $d_{u,i}$ flowing from an upstream microservice m_u to a downstream microservice $m_i \in \mathcal{M}$. \mathcal{I} denotes a set of *cloud infrastructure* and req is a set of user requirements.

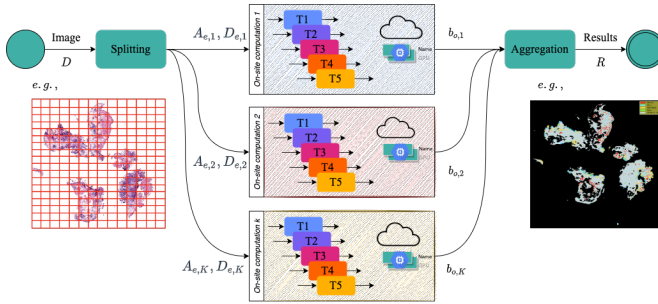


Fig. 3. An overview of the privacy-preserving preprocessing pipeline for whole slide image in a distributed processing manner.

Along with these lines, the research problem has turned the emphasis of studying on privacy-preserving service orchestration – Or more precisely, *how to customize a virtual infrastructure and schedule the workflow execution under privacy-preserving constraints while reducing its time and monetary cost?*

Privacy requirements: Reconstruction of a WSI from distributed resource nodes can lead to finding similar medical images using content-based image retrieval and extrapolating possible patient information from other sources. Therefore, using this distributed scheme, their privacy will be preserved during the process.

Bi-objective optimization: Reducing the execution time of the application over the cloud can be crucial for many stakeholders as it can lead to significant cost savings and improve overall processing time for WSI. We aim to reduce the monetary cost f_1 and minimize the application's maximum completion time (i.e., makespan) f_2 . Let us denote $ET(m_{src})$ and $ET(m_{snk})$ as the execution time of splitting and aggregation services over the trusted server. m_{det} for the artifact detection microservices, which will be deployed to the cloud where $ET(m_{det}(A_{e,k}, \mathcal{D}_{e,k}), \mathcal{I}_k)$ denotes its total execution time. Then, the bi-objective optimization problem can be formulated as follows:

$$\min f_1 = \sum_{k=1}^K ET(m_{det}(A_{e,k}, \mathcal{D}_{e,k}), \mathcal{I}_k) \times p_k \times x_k \quad (2)$$

$$\min f_2 = ET(m_{src}(\mathcal{D})) + makespan(m_{det}) + ET(m_{snk}(\mathcal{R})) \quad (3)$$

Here K and p_k represent the number of the split data sets and the unit price of cloud infrastructure \mathcal{I}_k , subject to,

$$makespan(m_{det}) = \max\{ET(m_{det}(A_{e,k}, \mathcal{D}_{e,k}), \mathcal{I}_k) \times x_k\}$$

$$ET(m_{det}(A_{e,k}, \mathcal{D}_{e,k}), \mathcal{I}_k) > 0$$

$$\text{where } x_k = \begin{cases} 1, & \text{if } m_{det} \text{ is mapping to } \mathcal{I}_k, \\ 0, & \text{otherwise.} \end{cases}$$

V. DISCUSSION AND FUTURE WORK

This work-in-progress paper presents the methodology for privacy-preserving task-based parallel applications for distributed cloud environments. Our method enables domain-specific users to handle gigapixel medical images efficiently, maintaining privacy among distributed nodes. In future work, we will develop prototypes and demonstrate the benefits of our pipeline using datasets from different hospitals and integrating the method with a Jupyter-based virtual research environment.

ACKNOWLEDGMENT

This work has been funded by the European Union project CLARIFY (860627), ENVRI^{FAIR} (824068), BlueCloud-2026 (101094227) and LifeWatch ERIC.

REFERENCES

- [1] N. Kanwal, F. Pérez-Bueno, A. Schmidt, R. Molina, and K. Engan, "The devil is in the details: Whole slide image acquisition and processing for artifacts detection, color variation, and data augmentation. a review." *IEEE Access*, 2022.
- [2] N. Kanwal, T. Eftestøl, F. Khoraminia, T. C. M. Zuiverloon, and K. Engan, "Vision transformers for small histological datasets learned through knowledge distillation," in *Advances in Knowledge Discovery and Data Mining*. Springer Nature Switzerland, 2023, pp. 167–179.
- [3] Z. Zhao, A. Belloum, and M. Bubak, "Special section on workflow systems and applications in e-Science," *Future Generation Computer Systems*, vol. 25, no. 5, pp. 525–527, May 2009. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0167739X08001726>
- [4] Y. Wang, S. Koulouzis, R. Bianchi, N. Li, Y. Shi, J. Timmermans, W. D. Kissling, and Z. Zhao, "Scaling notebooks as re-configurable cloud workflows," *Data Intelligence*, vol. 4, no. 2, pp. 409–425, 2022.
- [5] Z. Zhao, S. Koulouzis, R. Bianchi, S. Farshidi, Z. Shi, R. Xin, Y. Wang, N. Li, Y. Shi, J. Timmermans *et al.*, "Notebook-as-a-vre (naavre): From private notebooks to a collaborative cloud virtual research environment," *Software: Practice and Experience*, 2022.
- [6] N. Kanwal, S. Fuster, F. Khoraminia, T. C. Zuiverloon, C. Rong, and K. Engan, "Quantifying the effect of color processing on blood and damaged tissue detection in whole slide images," in *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*. IEEE, 2022, pp. 1–5.
- [7] L. Hurwicz, "The design of mechanisms for resource allocation," *The American Economic Review*, vol. 63, no. 2, pp. 1–30, 1973.
- [8] A. Taal, J. Wang, C. de Laat, and Z. Zhao, "Profiling the scheduling decisions for handling critical paths in deadline-constrained cloud workflows," *Future Generation Computer Systems*, vol. 100, pp. 237–249, 2019.
- [9] Y. Wang, H. Liu, W. Zheng, Y. Xia, Y. Li, P. Chen, K. Guo, and H. Xie, "Multi-objective workflow scheduling with deep-q-network-based multi-agent reinforcement learning," *IEEE access*, vol. 7, pp. 39 974–39 982, 2019.
- [10] E. Soljanin, "Technical perspective: Balancing at all loads," *Communications of the ACM*, vol. 65, no. 5, pp. 110–110, 2022.
- [11] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "Random-data perturbation techniques and privacy-preserving data mining," *Knowledge and Information Systems*, vol. 7, pp. 387–414, 2005.
- [12] Q. Li, J. S. Gundersen, R. Heusdens, and M. G. Christensen, "Privacy-preserving distributed processing: metrics, bounds and algorithms," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 2090–2103, 2021.