



UvA-DARE (Digital Academic Repository)

PriCE: Privacy-Preserving and Cost-Effective Scheduling for Parallelizing the Large Medical Image Processing Workflow over Hybrid Clouds

Wang, Y.; Kanwal, N.; Engan, K.; Rong, C.; Grosso, P.; Zhao, Z.

DOI

[10.1007/978-3-031-69577-3_15](https://doi.org/10.1007/978-3-031-69577-3_15)

Publication date

2024

Document Version

Final published version

Published in

Euro-Par 2024: Parallel Processing

License

Article 25fa Dutch Copyright Act (<https://www.openaccess.nl/en/policies/open-access-in-dutch-copyright-law-taverne-amendment>)

[Link to publication](#)

Citation for published version (APA):

Wang, Y., Kanwal, N., Engan, K., Rong, C., Grosso, P., & Zhao, Z. (2024). PriCE: Privacy-Preserving and Cost-Effective Scheduling for Parallelizing the Large Medical Image Processing Workflow over Hybrid Clouds. In J. Carretero, S. Shende, J. Garcia-Blas, I. Brandic, K. Olcoz, & M. Schreiber (Eds.), *Euro-Par 2024: Parallel Processing: 30th European Conference on Parallel and Distributed Processing, Madrid, Spain, August 26–30, 2024 : proceedings* (Vol. I, pp. 210-224). (Lecture Notes in Computer Science; Vol. 14801), (Advanced Research in Computing and Software Science). Springer.
https://doi.org/10.1007/978-3-031-69577-3_15

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (<https://dare.uva.nl>)



PriCE: Privacy-Preserving and Cost-Effective Scheduling for Parallelizing the Large Medical Image Processing Workflow over Hybrid Clouds

Yuandou Wang¹(✉) , Neel Kanwal² , Kjersti Engan² , Chunming Rong² ,
Paola Grosso¹ , and Zhiming Zhao¹(✉) 

¹ Multiscale Networked Systems, University of Amsterdam,
Amsterdam, The Netherlands

{y.wang,p.grosso,z.zhao}@uva.nl

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

{neel.kanwal,kjersti.engan,chunming.rong}@uis.no

Abstract. Running deep neural networks for large medical images is a resource-hungry and time-consuming task with centralized computing. Outsourcing such medical image processing tasks to hybrid clouds has benefits, such as a significant reduction of execution time and monetary cost. However, due to privacy concerns, it is still challenging to process sensitive medical images over clouds, which would hinder their deployment in many real-world applications. To overcome this, we first formulate the overall optimization objectives of the privacy-preserving distributed system model, i.e., minimizing the amount of information about the private data learned by the adversaries throughout the process, reducing the maximum execution time and cost under the user budget constraint. We propose a novel privacy-preserving and cost-effective method called PriCE to solve this multi-objective optimization problem. We performed extensive simulation experiments for artifact detection tasks on medical images using an ensemble of five deep convolutional neural network inferences as the workflow task. Experimental results show that PriCE successfully splits a wide range of input gigapixel medical images with graph-coloring-based strategies, yielding desired output utility and lowering the privacy risk, makespan, and monetary cost under user's budget.

Keywords: Privacy · Cost-effectiveness · Hybrid Clouds · Medical Image splitting · Multi-Objective Optimization · Scheduling

1 Introduction

Modern medical image processing techniques utilize deep neural networks to extract hidden patterns and make predictions; however, running such machine

learning-based inferences for large medical images is resource-hungry and time-consuming when computing resources are limited. Cloud computing can provide ample and highly scalable storage, computational resources, and ubiquitous access for distributed processing tasks. Hybrid Clouds (HCs) combine the economies and efficiencies of public cloud with the security and control of private cloud [1]. However, privacy concerns significantly complicate the development of an optimal cloud resource allocation plan for outsourcing computations on sensitive data processing tasks: (1) medical images often contain privacy-sensitive information in their metadata, which cannot be directly outsourced to the public cloud due to the risk of data leakage, (2) assigning distributed processing workloads to available cloud resources to meet multiple user requirements such as the reduction of time and monetary cost, known as Multi-Objective Optimal (MOO) workflow scheduling, is a typical NP-hard problem [2].

The problem of privacy-preserving and cost-effective scheduling in large (e.g., gigapixel) medical image processing over HCs has not yet been studied in detail, and we observed that this problem exhibits unique characteristics. For instance, although patching the original large image in a grid and dividing the patch-level dataset into multiple sub-datasets are common practices, the strategy employed for these practices is critical for both privacy preservation and resource provisioning for deployment in the cloud. Furthermore, although several studies have addressed workflow privacy in the context of cloud technology [2,3], there is a need for precise measurements of privacy metrics that align with the specific privacy-preserving approaches used for workflow scheduling.

This work aims to overcome those limits and solve the problem of privacy-preserving and cost-effective distributed inference tasks over clouds. To address this, we first formulate the research problem that minimizes the vulnerability, reduces the monetary cost, and minimizes the maximum execution time of the privacy-preserving distributed system under constraints. We propose a novel privacy-preserving and cost-effective algorithm called PriCE. Our experimental evaluation reveals the benefits of PriCE in privacy-preserving and cost-effective workflow scheduling when answering the following three sub-research questions: (I) “What are the trade-offs among privacy, cost, and execution time as split strategies related to the distributed processing of large images?” (II) “Can privacy-preserving split strategies improve resource planning and lead to Pareto optimality?” and (III) “How different are the Pareto optimal solutions from different split strategies?”

Our main contributions can be summarized as follows:

- We design and implement PriCE, which consists of multiple image-splitting strategies, image label perturbation, and multi-objective optimization procedures that can seek the Pareto front of resource provisioning for the privacy-preserving and cost-effective system model.
- We demonstrate how to analyze, quantify, compare, and understand different split strategies within PriCE and obtain the final assignment, estimated objective values of the cloud instances, by conducting experiments based on a use case for artifact detection tasks on gigapixel medical images.

The remainder of this paper is structured as follows. Section 2 presents related work to privacy-aware workflow scheduling in HCs. In Sect. 3, we propose our system model, provide critical metrics used for the methodology and evaluation, and formulate the research problem. Section 4 illustrates our proposed approach for problem-solving and Sect. 5 details the experimentation and evaluation. Finally, we conclude our work in Sect. 6.

2 Related Work

The scheduling of privacy-aware workflows has garnered increased attention in recent years, especially aiming to minimize costs and processing time while ensuring compliance with privacy requirements. Sharif *et al.* [2] target a resource allocation map based on privacy privileges over HCs that combine private, community, and public clouds, while using a healthcare workflow that consists of private, semi-private, and public tasks as a case for problem modeling. Their objective is to minimize the overall execution cost of workflows while satisfying concerns about their privacy and deadline. Zhou *et al.* [4] study the European (EU) data protection regulation—GDPR¹ in the geo-distributed cloud and formulate the geo-distributed process mapping problem that minimizes the cost of workflow applications while meeting data privacy constraints regarding the restrictions of the data movement between different cloud data centers. Lei *et al.* [3] define a deadline-constrained cost optimization problem in a HC under the deadline and privacy constraints. Similarly, Wen *et al.* [5] propose a multi-objective privacy-aware scheduling algorithm to obtain a set of Pareto trade-off solutions between execution makespan and cost reductions while meeting the set of privacy protection constraints. As presented in Table 1, these works are close to our research problem in terms of data and task privacy constraints, time performance improvement, and cost reduction for resources planning over HCs; nonetheless, our work significantly differs from their works.

Table 1. Comparisons of the problem formulation.

Problem Model	Optimization Objective			Constraint(s)
	Time	Cost	Privacy	
Sharif <i>et al.</i> [2]	✗	✓	✗	task/data privacy, deadline
Lei <i>et al.</i> [3]	✗	✓	✗	deadline, privacy
Zhou <i>et al.</i> [4]	✗	✓	✗	data privacy, GDPR, deadline
Wen <i>et al.</i> [5]	✓	✓	✗	privacy protection
Ours	✓	✓	✓	user's budget

¹ General Data Protection Regulation (GDPR). <https://gdpr.eu/>.

We study the scheduling problem that minimizes the quantified amount of information about private data learned by the adversaries, lowers financial cost, and reduces the maximum execution time with different data-split strategies. To the best of our knowledge, this is a unique problem statement since the majority of the related works leave privacy as a constraint, instead of a quantified optimization objective.

3 Problem Formulation

3.1 System Model

We consider a system as depicted in Fig. 1. We remove the privacy-sensitive metadata from the original image \mathcal{D} before splitting the large image into many image tiles to introduce data parallelism. We apply a privacy-preserving data splitting procedure in step ①, in conjunction with image label encryption and image object serializing to preserve the data privacy. Data splitting aims to protect data privacy by fragmenting sensitive data and storing the fragments in different locations so that individual parts do not disclose identities or confidential information [6]. In step ②, we transform the Convolutional Neural Network (CNN) inference model into several reusable fine-grained computational tasks; configuring the available resources such as private and public cloud resources in step ③ can facilitate making plans for resource provision.

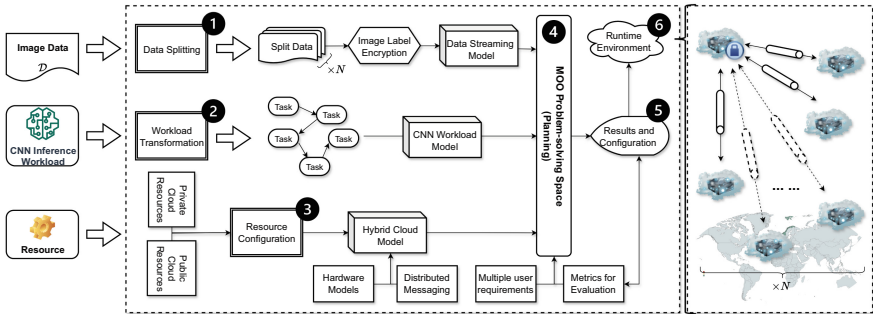


Fig. 1. Workflow of the system model and its related application scenario.

Based on the prepared data, workload, and HC model, step ④ maps available resources to various workloads in a manner that optimizes their utilization and satisfies user requirements. The privacy-preserving and cost-effective problem is a classic MOO scheduling problem, where N workloads and split privacy-preserved datasets have to be scheduled on M identical cloud instances over HCs, with the multi-objective functions that minimize the information about sensitive data learned by adversaries, the financial expense of the total instances consumed, and the maximum execution time of the instance that completes

the last workload, i.e., makespan. After obtaining the estimated results and configuration in step ⑤, the application will be deployed and executed onto the runtime environment. The system should ensure that data storage and workload 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 ⑥ [7]. Note that when constructing this model, we refer to the prior research experience and user requirements from several EU projects such as CLARIFY², BlueCloud2026³, ENVRI-Hub NEXT⁴, and LifeWatch ERIC⁵.

3.2 Privacy-Preserving and Cost-Effective Metrics

To satisfy user demands, privacy-preserving and cost-effective data processing should consider goals such as privacy preservation, makespan, and monetary cost. Additionally, we use the semi-honest threat model to evaluate the robustness of the system [6]. HC service providers honestly fulfill their role in the storage and processing tasks, but may inspect the information that users store or process. The attacker typically tries to collude with a number of processing nodes storing the datasets to infer and reason sensitive information of the other nodes to gain insights without directly altering the data or system. We assume that once the attacker obtains the image datasets stored on the nodes, he will try to reconstruct the original image data by using sensitive information from image labels, e.g., the coordinates $(x, y)_{\text{coord}}$. The system should evaluate the results and configuration of the MOO problem-solving before the sub-datasets and CNN inference models are assigned to planned infrastructures.

To quantify the privacy-preserving goals, we introduce information-theoretic metrics since they assume a stronger adversary and are more efficient concerning both communication and computational demands [8,9]. Let \mathcal{N} be the number of datasets $\{d_{p,1}, d_{p,2}, \dots, d_{p,N}\}$ split from the entire image \mathcal{D} ; the set $\{d_{e,1}, d_{e,2}, \dots, d_{e,N}\}$ denotes the encrypted datasets. Let Z_i and S_i be the encrypted and private information of $d_{e,i}$ and $d_{p,i}$ ($i \in \mathcal{N}$), respectively. We denote the entropy that measures the randomness of S_i as $H(S_i) = -\sum_{j=1}^n p_j \log_2 p_j$, where p_j denotes the probability of the unique coordinate information in S_i . The mutual information $I(S; Z) = H(S) - H(S|Z)$ between two random variables S and Z measures the dependence between S and Z , quantifying the average reduction in uncertainty about S that results from learning the value of Z .

Output Utility. The output utility is to measure how close the estimated value \hat{Y}_i of a privacy-preserving distributed processing algorithm is to its desired output Y_i , for each node $i \in \mathcal{N} \subset M$,

$$u_i = I(Y_i; \hat{Y}_i), \quad \forall i \in \mathcal{N} \quad (1)$$

where $0 \leq u_i \leq I(Y_i; Y_i)$ and $u_i = I(Y_i; Y_i)$ indicates perfect output utility [9].

² CLARIFY project. <http://www.clarify-project.eu/>.

³ Blue-Cloud2026 project. <https://blue-cloud.org/about-blue-cloud-2026>.

⁴ ENVRI-Hub NEXT. <https://envri-hub.envri.eu/>.

⁵ LifeWatch ERIC. <https://www.lifewatch.eu/>.

Privacy Risk. Let \mathcal{V} denote the set of random variables containing all information collected by the adversaries throughout the whole process. The individual privacy of honest node $i \in \mathcal{N}_h$ quantifies the amount of information about the private data S_i learned by the adversaries, which is given by,

$$\rho_i = I(S_i; \mathcal{V}), \quad \forall i \in \mathcal{N}_h \subset \mathcal{N} \quad (2)$$

The smaller ρ_i , the more private the data is, hence, the lower the privacy risk the data is. Based on the definitions of the adversary model, the lower bound on individual privacy risk is formally stated by,

$$\rho_{i,min} = I(S_i; \{S_j, \hat{Y}_j\}_{j \in \mathcal{N}_c}), \quad \mathcal{N}_c = \mathcal{N} - 1 \quad (3)$$

where adversaries always have knowledge of the private data $\{s_j\}_j = S_j$ and estimated outputs over corrupted nodes $\{\hat{y}_j\}_j = \hat{Y}_j$, in which the maximum number of corrupted nodes \mathcal{N}_c denotes $\mathcal{N} - 1$ out of \mathcal{N} [9].

Total Cost and Makespan. To measure the total cost and makespan of employing cloud instances \mathcal{N} over the HC model, we consider the pay-as-you-go pricing model. Let p_k be the unit price of the k^{th} cloud instance over the HC that consists of commercial and private cloud resources M . The total monetary cost is given by,

$$\text{Cost} = \sum_{k=1}^{\mathcal{N}} (T_k^{(\text{compt.})} + T_k^{(\text{comm.})}) \times p_k \times x_k, \quad k \in M \quad (4)$$

where $x_k \in \mathbf{x}$ is a boolean value that if sub-dataset d_k from \mathcal{D} and the inference model are mapping to the cloud instance k , then x_k equals 1; otherwise, 0. If k belongs to commercial cloud instances, then $p_k \in R^+$; otherwise, $p_k = 0$. $T_k^{(\text{comm.})}$ is the communication time when transferring the inference model and encrypted sub-dataset $d_{e,k}$ to the cloud instance k and $T_k^{(\text{compt.})}$ is the computation time when running the workload on k , respectively. Similarly, the makespan over distributed processing nodes is given by,

$$\text{Makespan} = \max((T_k^{(\text{compt.})} + T_k^{(\text{comm.})}) \times x_k), \quad k \in M \quad (5)$$

where the compute capacity, geographical location, and network bandwidth of the instance k impact the time of executing the privacy-preserving distributed application that typically consists of the time cost of on-site computation and communication overhead.

3.3 Multi-objective Optimization

The overall optimization objective of the system model intends to minimize the average lower bound on privacy risk f_1 , the total monetary cost f_2 , and the maximum completion time f_3 over a HC:

$$\min f_1 = \text{Average minimal privacy risk} = \bar{\rho}_{min}(s), \quad \forall s \in \mathcal{S} \quad (6)$$

$$\min f_2 = \text{Cost} = \sum_{k=1}^{\mathcal{N}} (T_k^{(\text{compt.})} + T_k^{(\text{comm.})}) \times p_k \times x_k, \quad \forall k \in M \quad (7)$$

$$\min f_3 = \text{Makespan} = \max((T_k^{(\text{compt.})} + T_k^{(\text{comm.})}) \times x_k), \quad \forall k \in M \quad (8)$$

with regards to the following constraints:

- The average privacy risk $\bar{\rho}_{min}$ is sensitive to privacy-preserving data-splitting strategy s in the set \mathcal{S} , since each strategy s generates unique sub-datasets of different sizes and image labels.
- The total number of split datasets \mathcal{N} is expected to be the minimal number of the employed distributed processing nodes, which is limited to the maximum available resources of the HC model M .
- Each split dataset and its corresponding inferences should be mapped to only one instance at a time. Meanwhile, one instance only runs one processing task per time. The total monetary cost f_2 is limited by the user’s *budget*.

4 PriCE: Privacy-Preserving and Cost-Effective Solution

This section presents the proposed PriCE solution and its algorithm pseudocode, as detailed in Algorithm 1. The solution primarily involves two key components: (1) privacy-preserving image splitting using graph-coloring, and (2) 3D Pareto trade-off solutions for resource planning.

4.1 Privacy-Preserving Image Splitting with Graph-Coloring

To cope with the diverse image samples of the privacy-preserving data-splitting procedure, we abstract the entire image as a grid graph where different patches with pixel size $p \times p$ are cropped from the original image \mathcal{D} , containing sensitive image labels and objects. The image label contains sensitive coordinate information to reconstruct the image and guide the outcome. One essential hypothesis is that the more adjacent image patches an attacker obtains, the higher the probability he could succeed in restoring the entire original image.

Let $G = (V, E)$ be a graph extracted from the entire patch dataset D cropped from the original image \mathcal{D} . Each patch is represented as a vertex $v \in V$. Two vertices v and μ of V such that $(v, \mu) \in E$ are called to be adjacent. Let $v = (x_i, y_i)$ and $\mu = (x_{i+1}, y_{i+1})$, we denote all possible adjacent relationships between v and μ as: (1) horizontal: $|x_{i+1} - x_i| = p$; (2) vertical: $|y_{i+1} - y_i| = p$; and (3) diagonal: $\sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2} = \sqrt{2} \times p$. With these characteristics, the positions of the patches can be identified in the original image. The graph-coloring-based splitting procedure is written in pseudo-code from step 1 to step 4.

Based on the assumption, we study different split strategies to scramble these identifications and reduce the risk of restoring the original dataset from the image fragments by the adversary. On the one hand, we adopt the graph-coloring-based

Algorithm 1: PriCE Method

Input: The original image \mathcal{D} , patch size p , split strategies \mathcal{S} , inference workloads Θ , and cloud instances M , budget B

Output: A new label system A_e , Pareto optimal solutions \mathbf{x}^* , and encrypted split datasets $d_{e,1}, d_{e,2}, \dots, d_{e,N}$.

- 1 Patch set $D \leftarrow \text{create_patches}(\mathcal{D}, p)$
- 2 $G \leftarrow \text{abstract_a_graph}(D)$
- 3 $\mathcal{N}, C \leftarrow \text{graph_coloring}(G, s)_{s \in \mathcal{S}}$
- 4 $d_{p,1}, \dots, d_{p,N} \leftarrow \text{divide } D_p \text{ into } \mathcal{N} \text{ sub data sets}$
- 5 **for** $d_{p,k} \in d_{p,1}, \dots, d_{p,N}$ **do**
- 6 Data matrix A_p of size $(a \times b) \leftarrow S_k \leftarrow d_{p,k}$
- 7 $\bar{x} \leftarrow \text{mean}(A_p)$
- 8 $\sigma \leftarrow \text{std}(A_p)$
- 9 $A_{p,c} \leftarrow (A_p - \bar{x})/\sigma$ // Normalize the data
- 10 $\text{Cov}(A_{p,c}) \leftarrow \frac{1}{a-1} A_{p,c}^T A_{p,c}$ // Compute the covariance matrix of $A_{p,c}$
- 11 $\lambda, \mathbf{V} \leftarrow \text{eig}(\text{Cov}(A_{p,c}))$ // Compute eigenvalues and eigenvectors
- 12 $\lambda_{\text{sort}}, \mathbf{V}_{\text{sort}} \leftarrow \text{sort_eig}(\lambda, \mathbf{V})$
- 13 $\mathbf{V}_k \leftarrow \text{first } k \text{ columns of } \mathbf{V}_{\text{sort}}$
- 14 $A_e \leftarrow A_{p,c} \mathbf{V}_k$ // Transform the original coordinate data
- 15 $d_{e,k} \leftarrow \text{rename}(d_{p,k}, A_e, r)$ // Transform labels with random values r and perturbed coordinates A_e
- 16 $\bar{\rho}_{\min}(s)_{s \in \mathcal{S}} \leftarrow \rho_{i, \min}(s)_{s \in \mathcal{S}}, i \in \mathcal{N} \leftarrow \text{Eq. 3}$
- 17 $\text{totalC}, \text{maxTime}, \mathbf{x} \leftarrow \text{prob.solve}(\mathcal{N}, \Theta, M, B)$, derived from **Eqs. 6, 7, and 8**
- 18 $\mathbf{x}^* \leftarrow \text{plot_Pareto_3D}(\mathbf{x})$ // plot the 3D Pareto trade-off solutions
- 19 **return** $\mathbf{x}^*, A_e, \{d_{e,1}, d_{e,2}, \dots, d_{e,N}\}$

split strategies [10–12], including ‘largest_first’, ‘random_sequential’, ‘smallest_last’, ‘independent_set’, ‘connected_sequential’, ‘saturation_largest_first’, to split the entire dataset D into different sub-datasets $d_{p,1}, \dots, d_{p,N}$, such that no two adjacent vertices share the same color or dataset. On the other hand, we introduce a random data perturbation to preserve the sensitive coordinates on split datasets’ labels by inserting random noise.

We extract $(x, y)_{\text{coord}}$ as a data matrix A_p of size $(a \times b)$, $a < b$, from $d_{p,k} \subset D$. After normalization, we compute the covariance matrix of the normalized matrix $A_{p,c}$, and then computed the eigenvalues λ and eigenvectors \mathbf{V} so that we can get the top- k eigenvectors \mathbf{V}_k to calculate A_e . Moreover, we transform the data into a new coordinate system and encrypt it into datasets $\{d_{e,1}, d_{e,2}, \dots, d_{e,N}\}$. The pseudo-code is illustrated in step 5 to step 15. From the perturbed data, since we know the noise variance, we obtain the estimate coordinates \hat{Y} from decryption by inversely transforming the eigenvector matrix \mathbf{V}_k and A_e , i.e., $\hat{Y} = A_e \cdot \mathbf{V}_k^T$. Note that the size of the transformed data matrix A_e might differ from that of the original data matrix A_p . To address this discrepancy, we introduce random values r to compensate for the size difference. Consequently, we can obtain a set of split image datasets with encrypted labels $\{d_{e,1}, \dots, d_{e,k}, \dots, d_{e,N}\}$. Besides, since we know the mappings of original labels and their corresponding encrypted labels,

it is easy to measure the output utility shown in Eq. 1. Furthermore, we calculate the average minimal privacy risk over the distributed datasets by Eq. 3. Empirical evidence indicates that the computation of eigenvalues and eigenvectors remains lightweight, even with up to 10,000 patches.

4.2 Pareto Trade-Off Solution Among Privacy, Cost, and Time

The number of split datasets \mathcal{N} is directly related to the number of cloud instances that need to be rented. Given split data $\{d_{e,1}, d_{e,2}, d_{e,N}\}$, CNN inference workload Θ , hybrid cloud instances M , and user’s budget B , our PriCE establishes a decision-making process based on MOO for resource planning.

The MOO problem is a classical integer programming problem since the variables \mathbf{x} are restricted to be integers. We first find available optimal solutions of the bi-objective optimization problem that minimizes f_2 (Eq. 7) and f_3 (Eq. 8) under the *budget*, and together with all available split strategies that generate f_1 (Eq. 6), obtaining all feasible solutions as \mathbf{X} . Due to the trade-offs among minimizing lower bound on privacy risk f_1 , monetary cost reduction f_2 , and makespan minimization f_3 , we then seek Pareto trade-off solutions $\mathbf{x}^* \in \mathbf{X}$, where no solution is superior to another in all objectives (See Sect. 3.3). Pareto front is the set of all such non-dominated Pareto optimal solutions. From the mathematical point of view, the definition of the dominance between two candidate solutions can be expressed as \mathbf{x}_1 dominates \mathbf{x}_2 if $f_i(\mathbf{x}_1) \leq f_i(\mathbf{x}_2)$, $\forall i = 1, 2, \dots, n$. Therefore, a solution $\mathbf{x}^* \in \mathbf{X}$ is called to be nondominated or Pareto optimal if and only if there does not exist any other point $\mathbf{x} \in \mathbf{X}$, such that $\mathbf{F}(\mathbf{x}) \leq \mathbf{F}(\mathbf{x}^*)$ and $f_i(\mathbf{x}) < f_i(\mathbf{x}^*)$ for at least one function [13], in which $\mathbf{F}(\mathbf{x}) = [f_1(\mathbf{x}), f_2(\mathbf{x}), f_3(\mathbf{x})]$. The MOO problem-solving procedure is written in the steps 16 to 18, as shown in Algorithm 1.

5 Experiments and Evaluation

This section details the experimental setup, demonstrates the visualized results, and evaluates the outcomes for validation. We implemented our algorithm PriCE in Python and evaluated the capability and quality of response of our algorithm via simulation when used to answer the questions in Sect. 1. It is important to note that the results presented here are, in whole or in part, based on data generated by the TCGA Research Network: <https://www.cancer.gov/tcga>. The original data and experimental results are available online⁶.

5.1 Experimental Setup

We conducted extensive experiments on a dedicated remote server equipped with 6 cores/12 threads@3.6GHz, 64GB DDR4 RAM, and 2×512 GB NVMe SSD and a private GPU server equipped with a Tesla T4 16 GB device. To collect

⁶ The source code is available online. <https://github.com/yuandou168/PriCE>.

benchmark data, we used a real-world CNN ensemble application for artifact detection developed by Kanwal *et al.* [14] as our CNN workloads. Specifically, it is the ensemble of five CNNs. The complexity of the inference workloads has been measured by the total parameters, FLOPs, batch size, and memory usage.

For the evaluation of solutions, we have investigated 25 commercial GPU servers offered by Fluidstack⁷ and 2 private GPU servers offered by universities, located in different cities across the Netherlands, Norway, USA, Iceland, and India. Since these cloud instances have different configurations, we investigated their relative performance based on TPU review data about GPUs⁸ and collected the workload performance over the private T4 GPU server offered by the university of Amsterdam. Besides, since the bandwidths are various with different geo-locations, we refer to fixed upload Internet speeds provided by SpeedTest⁹ to simplify the network environments. More technical details have been presented in the source code.

5.2 Visualization

In this study, we utilized a WSI named ‘TCGA-E9-A1N3-01Z-00-DX1’ for demonstration, which is available on the TCGA repository¹⁰. In Fig. 2, we demonstrate that PriCE can effectively split a large medical image into different sub-datasets using a graph-coloring-based strategy, ensuring that no two adjacent patches are placed in the same sub-dataset.

Figure 2a represents the thumbnail picture of the original WSI and Fig. 2b is the binary mask pictures after removing the background of the large image. The patch nodes from the original image represent a set of colored images with the size of 224×224 pixels; each of them has a unique label that contains the coordinate information of the image to identify its position in the original image. First, we can see that PriCE can perform graph-coloring-based splitting with the coloring assignment to distribute image samples for the WSI. For example, we obtain eight sub-datasets with the ‘random sequential’ graph-coloring strategy. Then, we adopt the data perturbation method to the split patch labels to hide the sensitive coordinate information; meanwhile, we examine the output utility that measures how close the estimated label of the privacy-preserving algorithm is to its desired output. In Fig. 2c, we plot the reconstructed graph with identified coordinates after the decryption procedure.

For more complex and diverse WSIs, we can apply our method to extract various graphs by adding patch nodes and edges, choose, and run different strategies to generate different split datasets. Practically, we tested our PriCE for a number of TCGA WSIs. The outcomes always match the original coordinates, achieving the perfect output utility.

⁷ Cloud GPU servers. <https://console2.fluidstack.io/virtual-machines>.

⁸ GPU Database–TechPowerUp. <https://www.techpowerup.com/gpu-specs/>.

⁹ Network Speed Test. <https://www.speedtest.net/performance>.

¹⁰ <https://portal.gdc.cancer.gov/image-viewer/MultipleImageViewerPage?caseId=03c143e0-d8a1-4d60-a4a3-df0501fc6b6e>.

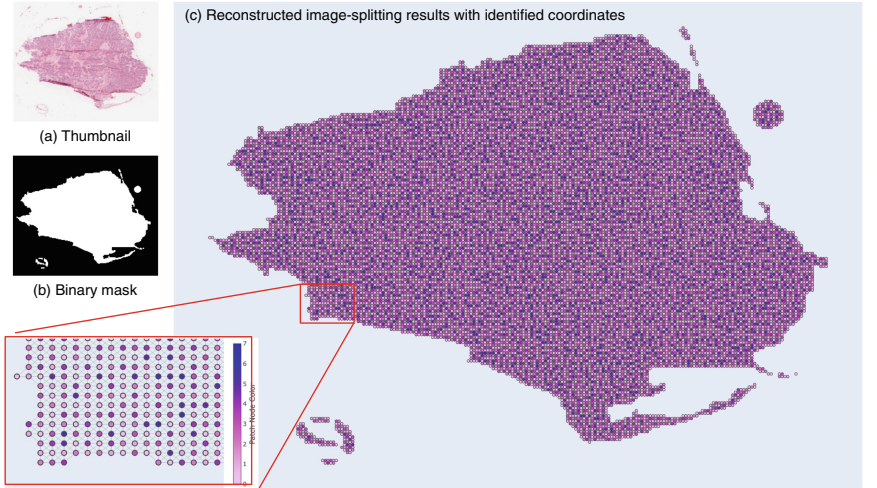


Fig. 2. Visualization of the image-splitting: (a) the thumbnail picture of the original medical image, (b) the binary mask picture, and (c) the reconstructed graph from estimated coordinates after decryption.

5.3 Reduction of the Average Lower Bound on Privacy Risk

We bring a unique perspective by introducing our novel PriCE algorithm with ‘graph-coloring-based split’ strategies (i.e., \mathcal{S}_{graph}), which we compare with the commonly used ‘average split with shuffle or without shuffle’ ones (i.e., \mathcal{S}_{avg}). Table 2 presents the overall comparisons of the number of split dataset and the average minimal privacy risk scores of the image-splitting strategies. Using graph-based split strategies, we achieved the minimum number \mathcal{N} of split datasets of varying sizes. However, the average split strategy is not able to obtain that number except for customizing it by users. To make the comparisons fair, we custom the same \mathcal{N} when splitting the datasets in the average cases.

The results show that (1) in the average split cases for the same number \mathcal{N} , the minimal privacy risk scores with shuffle (‘is_shuffled_True’) and without shuffle (‘is_shuffle_False’) are very close to each other; (2) when the number of split datasets is four, the privacy risk scores of all cases are almost the same, though the one with ‘saturation_largest_first’ is higher than the others; (3) as the number of datasets increases, the corresponding average minimal privacy risk score decreases; and (4) for the same number of split datasets, the average minimal privacy risk obtained by most graph-coloring-based split methods are generally lower but accompanied by a larger standard deviation. When $\mathcal{N} = 7$, the difference between graph and average-based strategies is over 0.04, which is higher than when $\mathcal{N} = 5, 6$, or 8, where the differences are below 0.02, 0.04, and 0.03, respectively. As a result, the number of datasets \mathcal{N} and corresponding split strategies significantly affect the privacy risk.

Table 2. Comparisons of split strategies, the number of split datasets, and the average minimal privacy risk scores.

	Strategy	\mathcal{N}	$\bar{\rho}_{min}$		
			x_{coord}	y_{coord}	$\sum(x, y)$
Graph-based Split	saturation_largest_first	4	0.1835±0.001	0.1389±0.0007	0.3224
	smallest_last	5	0.1523±0.0564	0.1169±0.0409	0.2692
	connected_sequential	6	0.1281±0.0738	0.0987±0.055	0.2268
	independent_set	7	0.1108±0.0798	0.0859±0.0597	0.1967
	largest_first	8	0.1101±0.058	0.0879±0.0445	0.198
	random_sequential	8	0.1106±0.0571	0.0881±0.0437	0.1987
Average Split w/wo Shuffle	is_shuffled_True	4	0.1832±0.0004	0.1386±0.0005	0.3218
	is_shuffled_True	5	0.1613±0.0003	0.1255±0.0002	0.2868
	is_shuffled_True	6	0.1448±0.0005	0.1145±0.0005	0.2593
	is_shuffled_True	7	0.132±0.0003	0.1061±0.0005	0.2381
	is_shuffled_True	8	0.1213±0.0007	0.0989±0.0007	0.2202
	is_shuffled_False	4	0.1831±0.0009	0.1386±0.0002	0.3217
	is_shuffled_False	5	0.1615±0.0011	0.1251±0.0008	0.2866
	is_shuffled_False	6	0.1448±0.001	0.1147±0.0006	0.2595
	is_shuffled_False	7	0.1318±0.0007	0.106±0.0006	0.2378
	is_shuffled_False	8	0.1213±0.0005	0.0985±0.0007	0.2198

5.4 Evaluation by Simulations

For Pareto optimal resource planning, we successfully obtained the Pareto trade-off solutions out of all feasible solutions from various split strategies under budget constraints through simulation experiments. Figures 3 and 4 depict the Pareto trade-off solutions selected using the 3D Pareto front, comparing them across graph-based and average-based split strategies and two budget constraints.

When the budget is limited to 120, we have identified four Pareto optimal solutions out of ten feasible solutions (4/10) in Fig. 3a, within the scenario using only average split strategies ($s \in \mathcal{S}_{avg}$). In contrast, we have found four out of six Pareto trade-off solutions (4/6) within the scenarios using only graph-based split strategies ($s \in \mathcal{S}_{graph}$), as shown in Fig. 3b. When decreasing the

budget to 100, there are eight feasible solutions in the only average split case ($s \in \mathcal{S}_{avg}$), compared to ten solutions under the *budget*=120, as seen in Fig. 3c. This reduction is attributed to the absence of bi-objective optimal solutions for f_2 and f_3 identified by the problem solver when the number of average split datasets is eight. Remarkably, for the graph-coloring-based split, our PriCE method could still identify all Pareto trade-off resource planning solutions under the *budget*=100, as depicted in Fig. 3d.

When combining both average and graph-based strategies ($s \in \mathcal{S}_{all}$), there are four Pareto trade-off solutions out of sixteen (4/16) when the budget is constrained by 120, as illustrated in Fig. 4a. Notably, when compared to the separate

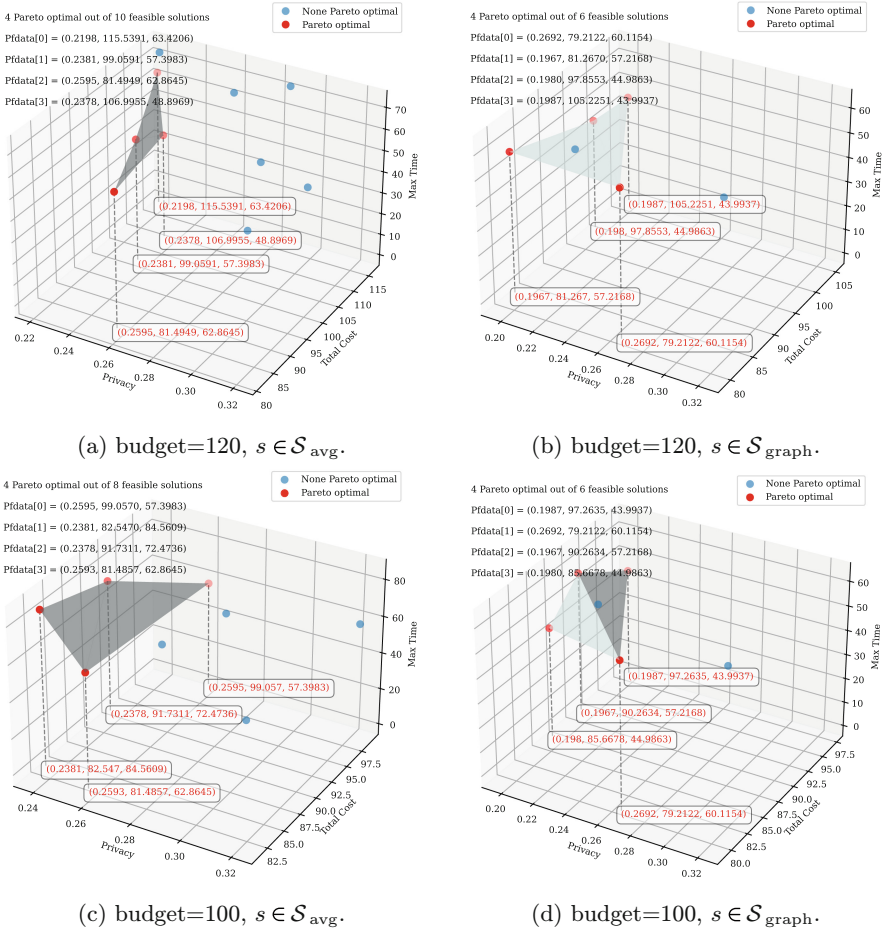


Fig. 3. The 3D Pareto trade-off solutions are compared as follows: (a) and (c) solutions fall within the only average split strategies \mathcal{S}_{avg} , while (b) and (d) solutions are derived from the only graph-coloring-based split strategies \mathcal{S}_{graph} .

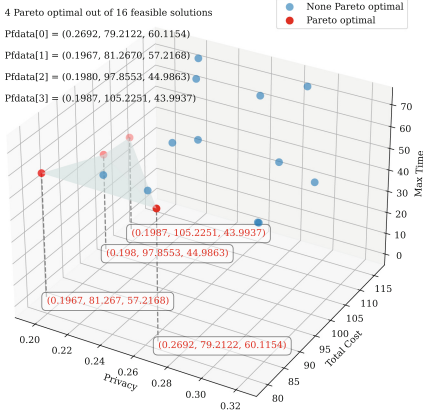
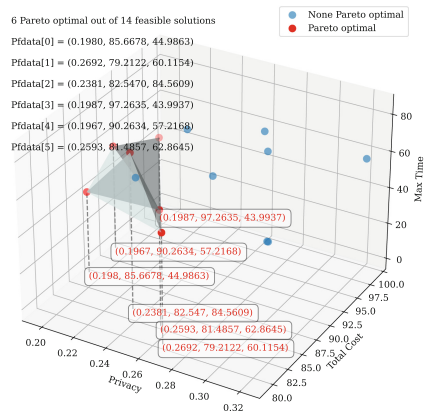
(a) budget=120, $s \in \mathcal{S}_{\text{all}}$.(b) budget=100, $s \in \mathcal{S}_{\text{all}}$.

Fig. 4. Comparisons of the 3D Pareto trade-off solutions with all combined strategies \mathcal{S}_{all} under two budget constraints: (a) budget = 120, and (b) budget = 100.

cases, all four Pareto trade-off solutions (4/4) originate from the graph-coloring-based split strategies. Furthermore, when combining all feasible solutions from all split strategies and budget = 100, as depicted in Fig. 4b, the majority of Pareto trade-off solutions (4/6) for the privacy-preserving and cost-effective problem are derived from the graph-coloring-based split strategies, with the remaining 2/6 solutions originating from the average split strategies. It can be inferred that when the budget becomes more constrained, privacy will be sacrificed. We have observed when the split datasets are unbalanced, the resource allocation plan exhibits greater resilience in matching with heterogeneous cloud instances. This increased resilience arises from the enhanced diversity in execution times and costs, which maximizes the opportunities for efficient task assignments.

6 Conclusion and Future Work

This paper investigates the workflow scheduling problem of privacy-preserving and cost-effective distributed inference using multiple GPU servers over hybrid clouds. We propose a novel solution, PriCE, which employs various image splitting strategies to enhance privacy and cost-efficiency. To the best of our knowledge, this is the first approach to address a privacy-aware scheduling problem that minimizes privacy risk while reducing makespan and cost within a budget in a privacy-preserving distributed system. We conducted a comprehensive experimental evaluation using a real-world application for medical image artifact detection. The results demonstrate that PriCE successfully takes the large number of patches from a gigapixel image and splits them using multiple graph-coloring-based strategies, yielding the desired output utility while lowering privacy risk, makespan, and monetary cost under the user's budget. Further improvements in

implementation might include making better use of secure distributed messaging or secrets handling across multiple clouds with high-level automated operations. Additionally, exploring the trade-offs between privacy overhead versus time and monetary cost in more complex scenarios could provide valuable insights.

Acknowledgment. We thank Mr. Zongxiong Chen for discussing the methodology and Mr. Aditya Shankar for reviewing the manuscript. This work has been partially funded by the European Union’s Horizon research and innovation program by CLARIFY (860627), BlueCloud-2026 (101094227), ENVRI-Hub Next (101131141), OSCARS (101129751), EVERSE (101129744), BioDT (101057437, via LifeWatch ERIC), by the LifeWatch ERIC and by the Dutch NWO LTER-LIFE project.

References

1. Mazhelis, O., Tyrväinen, P.: Economic aspects of hybrid cloud infrastructure: User organization perspective. *Inf. Syst. Front.* **14**, 845–869 (2012)
2. Sharif, S., et al.: Privacy-aware scheduling SaaS in high performance computing environments. *IEEE Trans. Parallel Distrib. Syst.* **28**(4), 1176–1188 (2016)
3. Lei, J., Wu, Q., Xu, J.: Privacy and security-aware workflow scheduling in a hybrid cloud. *Future Gener. Comput. Syst.* **131**, 269–278 (2022)
4. Zhou, A.C., et al.: Privacy regulation aware process mapping in geo-distributed cloud data centers. *IEEE Trans. Parallel Distrib. Syst.* **30**(8), 1872–1888 (2019)
5. Wen, Y., et al.: Scheduling workflows with privacy protection constraints for big data applications on cloud. *Future Gener. Comput. Syst.* **108**, 1084–1091 (2020)
6. Domingo-Ferrer, J., et al.: Privacy-preserving cloud computing on sensitive data: a survey of methods, products and challenges. *Comput. Commun.* **140**, 38–60 (2019)
7. Wang, Y., et al.: Towards a privacy-preserving distributed cloud service for pre-processing very large medical images. In: 2023 IEEE International Conference on Digital Health (ICDH), pp. 325–327. IEEE (2023)
8. Kraskov, A., Stögbauer, H., Grassberger, P.: Estimating mutual information. *Phys. Rev. E* **69**(6), 066138 (2004)
9. Li, Q., et al.: Privacy-preserving distributed processing: metrics, bounds and algorithms. *IEEE Trans. Inf. Forensics Secur.* **16**, 2090–2103 (2021)
10. Matula, D.W., Beck, L.L.: Smallest-last ordering and clustering and graph coloring algorithms. *J. ACM (JACM)* **30**(3), 417–427 (1983)
11. Kubale, M.: *Graph Colorings*, vol. 352. American Mathematical Society (2004)
12. Deo, N., Kowalik, J.S., et al.: *Discrete optimization algorithms: with Pascal programs*. Courier Corporation (2006)
13. Marler, R.T., Arora, J.S.: Survey of multi-objective optimization methods for engineering. *Struct. Multidiscip. Optim.* **26**, 369–395 (2004)
14. Kanwal, N., et al.: Equipping computational pathology systems with artifact processing pipelines: a showcase for computation and performance trade-offs. *arXiv preprint arXiv:2403.07743* (2024)