Exploring the Limits of Foundation Models in Medical Image Segmentation: A Case Study With SAM and Genetic Algorithms

Juan D. Gutiérrez[1]*, Nuria Lozano-García[2], Emilio Delgado[3], Álvaro Rubio-Largo[4], Roberto Rodriguez-Echeverria[3]

[1] Department of Electronics and Computer Science, Universidade de Santiago de Compostela, Rúa Benigno Ledo, 27002, Lugo (Spain)
[2] Department of Technologies of Computers and Communications, Universidad de Extremadura, Escuela Politécnica, 10003, Cáceres (Spain)
[3] Instituto de Investigación en Tecnologías Informáticas Aplicadas (INTIA), Universidad de Extremadura, Av. Universidad s/n, 10003, Cáceres (Spain)
[4] Department of Computers and Telematics Systems Engineering, Universidad de Extremadura, Escuela Politécnica, 10003, Cáceres (Spain)

* Corresponding author: juandiego.gutierrez@usc.es

Received 8 November 2024 | Accepted 19 December 2025 | Early Access 24 February 2026

Abstract

This paper investigates the limits of foundation models in medical image segmentation, mainly focusing on SAM by Meta. While previous research demonstrated SAM's potential for cost-efficient segmentation, this study explores its performance enhancement through integration with prompt enhancement optimization and genetic algorithms, aiming to minimize user input further. As a proof of concept, we apply this novel approach to lung segmentation tasks using public axial lung CT scans, frontal chest X-ray datasets, and spleen MRIs. Our findings reveal that the genetic algorithm optimization significantly improves SAM's segmentation accuracy, bringing it closer to the state-of-the-art performance achieved by specifically trained models. In particular, when compared with our previous approach, this technique reaches a 94.85 % Jaccard Index (+3.77 delta) and a 97.17 % Dice Score (+2.50 delta) for lung CT scans, a 93.39 % Jaccard Index (+5.95 delta) and a 96.57 %Dice Score (+3.38 delta) for chest X-rays, and a 91.00 % Jaccard Index (+6.51 delta) and a 95.07 % Dice Score (+4.12 delta) for spleen MRIs. Notably, this improvement is achieved without retraining or modifying SAM's architecture. However, our analysis also identifies an inherent limitation in this optimization approach, revealing a performance ceiling that cannot be surpassed despite further genetic algorithm iterations. The implications of these findings emphasize the potential of combining foundation models with non-intrusive optimization techniques for cost-effective and accessible medical image segmentation. While dataset-related limitations may affect generalizability, validating the approach across broader clinical scenarios remains essential. Future work should explore applications to additional organs, diverse datasets, and the integration of expert-in-the-loop strategies to enhance clinical utility.

Keywords

Deep Learning, Foundation Models, Genetic Algorithms, Image Segmentation, Medical Imaging, Zero-Shot Learning.

DOI: 10.9781/ijimai.2026.2223

I. Introduction

Image segmentation, the process of partitioning an image into meaningful regions [1], is a cornerstone of computer vision, especially in medical imaging, where it focuses on identifying and delineating regions of interest such as anatomical structures or pathological areas. This task is crucial for accurate diagnosis, treatment planning, and disease monitoring. While Deep Learning (DL) has revolutionized medical image segmentation [2], [3], developing and deploying these sophisticated models requires substantial computational resources, specialized expertise, and extensive datasets, potentially limiting their accessibility and practicality in real-world healthcare settings.

The emergence of foundation models, such as Segment Anything Model (SAM) by Meta, introduced in 2023, signified a paradigm shift

Please cite this article as: J. D. Gutiérrez, N. Lozano-García, E. Delgado, Á. Rubio-Largo, R. Rodriguez-Echeverria. Exploring the Limits of Foundation Models in Medical Image Segmentation: A Case Study With SAM and Genetic Algorithms, International Journal of Interactive Multimedia and Artificial Intelligence, (2026), http://doi.org/10.9781/ijimai.2026.2223

in image segmentation [4]. Trained on an unprecedented scale of data, SAM exhibited remarkable zero-shot transfer capabilities, enabling it to segment images from domains on which it had not been explicitly trained. This breakthrough has sparked interest in leveraging SAM's capabilities for efficient and user-friendly medical image segmentation [5]. As we showed in a previous work [6], SAM without modifications can deliver high-quality medical image segmentation, producing results comparable to state-of-the-art models designed for this task. SAM's performance was significantly improved in that work by providing additional input as prompts, such as points or bounding boxes that are not arbitrarily chosen. These prompts guide the model's attention to specific regions of interest, improving its accuracy, especially in cases where the target objects have unclear boundaries or the model struggles to distinguish between foreground and background.

A key strength of the proposed framework lies in its real-world applicability, particularly in clinical workflows where manual image segmentation is both time-intensive and expertise-dependent. By integrating SAM with Genetic Algorithms (GAs), the system enables specialists to rapidly generate preliminary segmentations, which can then be iteratively refined with minimal additional input. This approach not only reduces annotation time by over 90 %compared to traditional manual methods but also alleviates the workload of highly trained professionals, allowing them to focus on tasks that require expert judgment.

Improving the performance of Artificial Intelligence (AI) models often necessitates going beyond the initial training phase. Several approaches have emerged to achieve this goal. Fine-tuning involves adapting a pre-trained model to a specific task or dataset by further training it on new data. Model alteration delves into the model's architecture, potentially adding layers, modifying connections, adjusting hyperparameters, or including new components. Input preprocessing aims to optimize the data fed into the model, which might involve cleaning, formatting, or augmenting the input to aid the model achieve its intended goal. Lastly, prompt engineering focuses on crafting effective input prompts to elicit desired outputs. Each of these approaches offers distinct advantages and may be employed independently or in conjunction to enhance the efficacy of AI models in various applications.

Although medical image segmentation is a critical process in healthcare diagnostics and treatment planning, no single algorithm can effectively segment all types of medical images across different modalities. Hybrid segmentation approaches (e.g., combining DL models with evolutionary methods) address this challenge by integrating two or more techniques to leverage their complementary strengths while minimizing their individual limitations [7].

Further enhancing SAM's performance, particularly in achieving accurate segmentation with fewer user prompts, remains a compelling research direction. However, it raises the question of how much improvement can be made using non-modifying, cost-effective techniques. Despite various approaches to enhance SAM, there may be an ultimate limit to how much its segmentation results can be improved. This paper explores the innovative combination of SAM with GAs to address this question.

GAs, inspired by the principles of natural evolution [8], offer a powerful optimization technique for exploring complex search spaces. Combining GAs and AI is common and is ideal for refining the prompts used in SAM's segmentation process. GAs and AI have been used together to perform Human Activity Recognition (HAR) [9], or predict wine quality [10], just to mention a couple of them.

Building on prior research into SAM's capabilities for cost-efficient medical image segmentation [6], this study aims to further minimize the need for user input. Our goal is to potentially achieve segmentation of an entire medical image volume with a single, optimally engineered prompt. By integrating GAs, we aim to optimize the prompt engineering process, enabling SAM to reach its highest possible segmentation accuracy with minimal user intervention. GAs will search through the space of possible prompts to find the optimal one, while also investigating whether a performance ceiling exists.

In summary, the key contributions of this paper are as follows:

- Introduces a novel approach combining SAM with GAs to enhance segmentation accuracy with minimal user input.
- Demonstrates the potential for optimizing prompt engineering using GAs to achieve more precise and efficient segmentation results.
- Explores the theoretical limits of improving SAM's performance using non-modifying, cost-effective techniques.
- Advances previous research in cost-efficient medical image segmentation by further reducing the need for user intervention.

The remainder of the paper is organized as follows to address the questions above and to stimulate the discussion on whether the results obtained by a foundation model as SAM can be improved using proven algorithm techniques that do not involve model modification or retraining. Section II explores the techniques applied in other works to improve SAM's performance. Section III describes the process we propose to improve SAM's performance, and the environment where the experiments were performed. The results obtained in the experimental tests are compared in Section IV with those achieved by our previous work. These results are further commented on in Section V, where some interesting particularities are analyzed. Threats to validity and limitations of this study are presented in Section VI. Finally, in Section VII, we analyze these results and propose future lines of research.

II. Related Works

Since its publication on April 2023, SAM has been the subject of numerous studies. SAM's Magnetic Resonance Imaging (MRI) scans of brain tumor segmentation capabilities are tested in [11]. The authors conclude that while SAM demonstrates the potential for automating specific tasks within Medical Image Segmentation (MIS), its application for diagnostic purposes, particularly in complex cases such as the analysis of brain tumor MRI scans, requires further training and implementation refinements. Instead of focusing on a particular modality and segmentation task, in [5], the authors present a comprehensive comparison of SAM's performance within MIS across multiple modalities and segmentation tasks. Their goal was to guide researchers on appropriately using and developing SAM. Among the paper's conclusions, its authors argue that SAM performed remarkably in some specific object segmentation. However, it was unstable, imperfect, or even failed in other situations. Nevertheless, fine-tuning SAM on specific medical tasks could improve its average performance.

Fine-tuning is a common technique in machine learning, taking a pre-trained model and further training it on a smaller, domain-specific dataset. This technique allows the model to adapt its learned representations to the specific task and perform better. Multiple efforts fine-tune SAM, further training it for MIS tasks [12]–[20]. This process helps the model adapt to the specific characteristics of the new data and improve its performance on the target task. However, fine-tuning can be computationally expensive, requiring large annotated datasets, significant computational resources, and teams of experts to manage the process.

When adapting large models like SAM, fine-tuning the entire

model can be computationally expensive. Parameter-efficient fine-tuning techniques address this by only updating a small subset of the model's parameters, making the process more efficient while maintaining performance [13], [15], [16], [18], [21]. These techniques are particularly useful when adapting SAM for medical imaging, where large annotated datasets are often scarce.

The original SAM's mask decoder might not be optimal for specific medical image segmentation tasks. Redesigning the decoder, for instance, by incorporating a U-shaped structure, has shown potential in enhancing SAM's performance [16]. Morton Colbert et al. [22] aims to develop and compare methods for refining traditional U-Net segmentations by repurposing them for automated SAM prompting.

A similar approach to the one we followed in our previous work (*i. e.*, looking for improvements that do not require altering the model) is taken by [23]. SAM's performance is improved by enhancing the images to segment before feeding them to the model instead of altering it.

Recent years have witnessed a surge of interest in hybrid segmentation approaches that combine DL models with evolutionary or metaheuristic optimization techniques. JiMing et al. [24] proposed SAOBL-IA, a hybrid method integrating simulated annealing, opposition-based learning, and an island algorithm with 2D OTSU fusion segmentation, demonstrating strong performance in medical image tasks. Hosny et al. [25] introduced COVID-HHOA, which employs hybrid metaheuristics for multilevel thresholding in both 2D and 3D medical image segmentation, leveraging Otsu's and Kapur's entropy-based fitness functions.

III. Methods

In our previous work, we proposed an approach that involved building an optimal prompt based on the characteristics of the target organ. This prompt, along with the corresponding image, was input into SAM for segmentation, and the results were compared with the ground truth. The segmentation output was close to state-of-the-art performance.

In this paper, we refine the previous pipeline even more by including a post-processing step. Fig. 1 shows our new approach based on genetic algorithms. The prompt provided by our previous approach contains a bounding box and some points. The positive points mark the organ to segment, meaning that whatever is in that location is significant for the segmentation process. The other one is negative, meaning the opposite. In this new approach, said prompt will be used to get an initial segmentation: the first generation. SAM's output is evaluated using the metrics described in Section D. That way, the fitness of this first generation is evaluated. Using this fitness, a new generation (i.e., a new prompt) is obtained. This process repeats until no significant improvement is achieved.

Using GAs to improve SAM's medical image segmentation results is a novel avenue of research. As far as we know, no work has explored this optimization technique at the time of this writing. Since there is no state-of-the-art to compare, the results must be analyzed systematically and precisely. Therefore, to ensure that this is a self-contained work, in this section, we will briefly discuss GAs and which of the different existing modalities we have used in our experiments. We will also describe the environment used for these experiments, including the data, the software, and the hardware.

A. Genetic Algorithms

GAs are a branch of artificial intelligence that uses evolutionary techniques to search for solutions in a broad search space. These algorithms mimic the process of natural selection, reproduction and mutation that occurs in biological evolution, but on a much faster and controlled scale [8].

Evolution starts from a population of individuals generated randomly or by specific heuristics and is an iterative process. These individuals represent possible solutions to the problem at hand, and are encoded in what has been called a chromosome. In each generation (iteration), the fitness of each individual in the population is evaluated, using the objective function of the optimization problem in question. The fittest individuals of the current population are selected, and their chromosome is modified (by crossover and/or mutation) to form a new generation, which is used in the next iteration of the algorithm. The algorithm terminates when a predetermined criterion is reached (e.g., maximum number of generations, no improvement, or satisfactory fitness level for the population, just to mention a few).

GAs use the principles of selection, reproduction and mutation to explore the search space and find optimal or near-optimal solutions to complex problems. This ability to search for solutions in parallel and explore the search space efficiently is what makes GAs a powerful tool in a variety of fields. Specifically in the field of medicine, they have been applied to very diverse objectives, such as prediction of enzymatic function [26], development of auxiliary systems for the diagnosis of Parkinson's disease [27], identification of carcinogenic genes from microarray data [28], and population pharmacokinetic/pharmacodynamic model selection [29].

B. Proposal

To optimize SAM prediction, a basic single-objective GA is applied with the classical genetic operators selection, crossover, and mutation. The objective function is the Jaccard index or Dice score, calculated by comparison with the provided masks of the lungs. Optimization is applied to the xy coordinates of the prompt's input points in the image. Two of them are positive, meaning that whatever is in that location is significant for the segmentation process. The other one is negative, meaning the opposite. Thus, the chromosome is a vector of length 6 whose components are the x and y coordinates of each entry point.

1. Formal Definition of the Problem

The problem addressed here can be formulated as a single-objective optimization problem, as described in (1):

$$\begin{aligned} &\text{minimize } F(x) = \{f_1(x) \vee f_2(x)\} \\ &\text{subject to } x \in \Omega \end{aligned} \tag{1}$$

where x is a vector with the horizontal and vertical coordinates of each input point, $f_1(x)$ is the opposite of Jaccard Index and $f_2(x)$ the opposite of Dice Score. Ω contains the lower and upper limits to which each component of x is subject, defining its decision space.

The objective functions $f_1(x)$ and $f_2(x)$ are defined as shown in (2):

$$\begin{aligned} f_1(x) &= -\mathrm{JCI}(A, B) \\ f_2(x) &= -\mathrm{DSC}(A, B) \end{aligned} \tag{2}$$

Here, JCI denotes the Jaccard Index (3), and DSC represents the Dice Score (4):

$$\mathrm{JCI}(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{3}$$

$$\mathrm{DSC}(A, B) = \frac{2|A \cap B|}{|A| + |B|} \tag{4}$$

where A is the segmentation predicted by SAM using the input points defined in x, and B is the ground truth mask. A more detailed explanation of these functions can be found in Section D.

2. Chromosome Definition

The individual's chromosome (X) has been defined as a vector of six

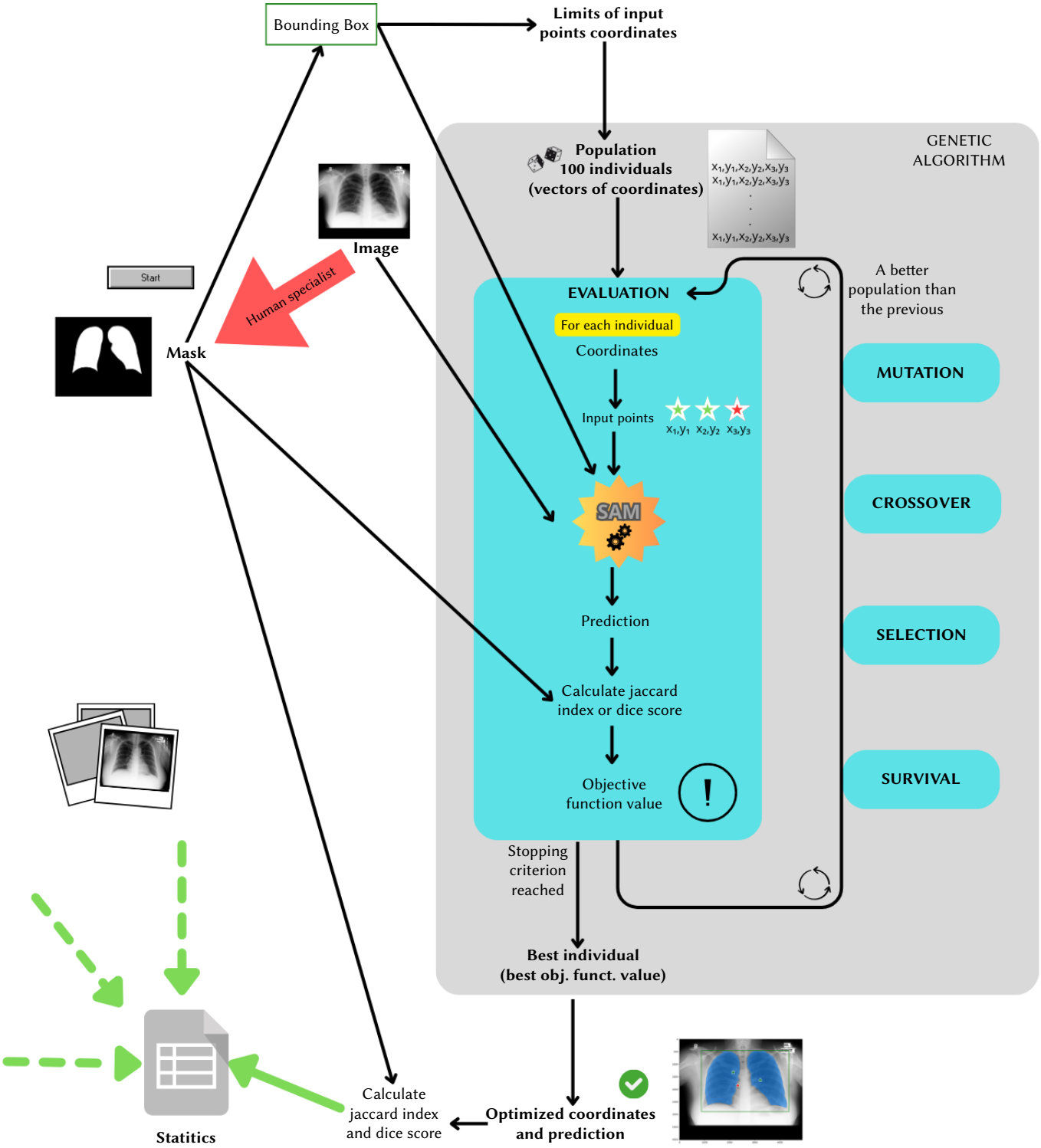


Fig. 1. Genetic algorithm optimization of medical image segmentation with SAM.

elements: $X = (x_1, x_2, ..., x_6)$, where each pair of consecutive elements (1-2, 3-4, 5-6) represents the horizontal and vertical coordinates in the input image of an input point. The first two (x_1, x_2) and (x_3, x_4) are positive points, and the last (x_5, x_6) is a negative point.

3. Procedure

For each image, first the bounding box that completely surrounds the mask area provided in the datasets (the lungs) is obtained, which will be the input bounding box for SAM. Furthermore, throughout the optimization process individuals may only be located within that same area. Next, the optimization process (GA) begins:

- The population is randomly initialized, with coordinate values constrained so that points (individuals) lie within the bounding box when working with lung slices.
- The evaluation function of the individuals is the opposite of the Jaccard index or Dice score, as it will be minimized. To calculate it, SAM is launched individually with all individuals in the population as input points (together with the bounding box and the image to

be segmented). From the prediction generated by SAM and the ground truth mask, the Jaccard index or Dice score is calculated, depending on which objective function is being used.

- The individuals are ordered by the evaluation function (fitness), and survival of the fittest is applied.
- For selection, a Tournament is applied. The crossover operator is set as Simulated Binary Crossover (SBX) and the mutation as Polynomial Mutation (PM).
- If after the evaluation step the stopping criterion is reached, the best individual (the one with the best value in the objective function) is returned, together with its corresponding SAM segmentation.

C. Datasets

In our previous work [6], SAM's performance when segmenting medical images was tested using two public lung datasets: axial lung scans and frontal chest X-rays. These datasets were selected based on their accessibility, prior publication, and established use in scientific research. The proposed optimization technique founded on GAs should employ the same datasets to be adequately compared with our prior method.

The first dataset, comprising 20 axial lung scans in NIfTI format [30], is divided into two subsets: Coronacases and Radiopaedia. The Coronacases subset exhibits superior Computed Tomography (CT) slice quality compared to the Radiopaedia subset, which consists of lower-resolution cone beam CTs. Furthermore, the Coronacases subset contains only 512 px × 512 px slices. In contrast, the Radiopaedia subset is more heterogeneous, featuring nine 630 px × 630 px slices and one 630 px × 401 px slice. With an average of 176 slices per CT scan, this dataset provides approximately 3520 images for evaluating the performance of SAM. This dataset was also used by [31] to evaluate the performance of their DMDF-Net.

The second dataset, Montgomery, originates from Montgomery County, Maryland, USA, and consists of 138 frontal chest X-rays in Portable Network Graphics (PNG) format [32]. The dataset includes both normal cases and cases with pulmonary abnormalities, primarily related to tuberculosis. Corresponding left and right lung masks are also provided in PNG format. Chen et al. [33] used this dataset to evaluate the performance of their TransAttUnet model.

Our previous work's homonymous section contains more information about these datasets.

To complement the main evaluation conducted on CT and X-ray lung images, we further assess model performance on a different imaging modality and anatomical structure. Specifically, we incorporate a subset of the Amos dataset [34], a clinical abdominal organ segmentation benchmark comprising 500 CT and 100 MRI volumes. Each volume includes multiple segmentation annotations for various abdominal organs, including the spleen. We selected the spleen as the target structure to evaluate model performance on an organ distinct from the lungs, and we used MRI data to assess performance on a modality different from the CT images used in the main experiments. To specifically evaluate SAM on MRI data, we selected 14 spleen-labeled MRI volumes. To enhance representativeness and mitigate potential biases associated with scanner variability, we further refined the selection to 8 volumes acquired using a Siemens Ingenia scanner. Slice dimensions range from 260 px × 80 px to 320 px × 260 px, with variations in between. These are the smallest dimensions across all the datasets used.

D. Metrics

The segmentation performance of SAM will be assessed using two widely employed similarity indices: the Jaccard Index (*JCI*) and the Dice Score (*DSC*), as defined in (3) and (4), respectively. These metrics quantify the spatial overlap between two binary images, A and B, typically a segmentation result and the corresponding ground truth.

As noted by [35], both the Jaccard Index and Dice Score are considered robust and reliable metrics for evaluating medical image segmentation. While the Dice Score balances sensitivity and accuracy, the Jaccard Index penalizes under-segmentation and over-segmentation more heavily, making it particularly suitable when high segmentation precision is paramount.

Fig. 2 illustrates the difference in severity between the two metrics. Using hypothetical square areas representing ground truth (side length: 100 px) and a prediction (side length: 70 px), the Jaccard Index (49.00 %) is lower than the Dice Score (65.77 %) due to its greater sensitivity to the mismatch in size.

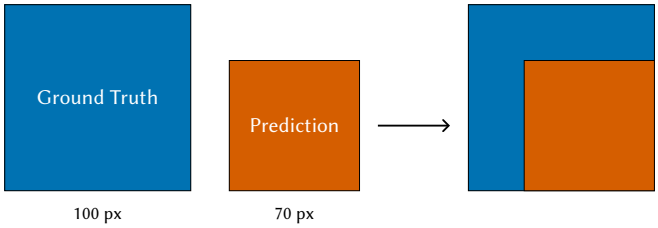


Fig. 2. Example comparing ground truth and predicted segmentation.

Given the lack of established threshold values for determining segmentation effectiveness [36], SAM results will be compared to those reported in relevant literature to assess its performance.

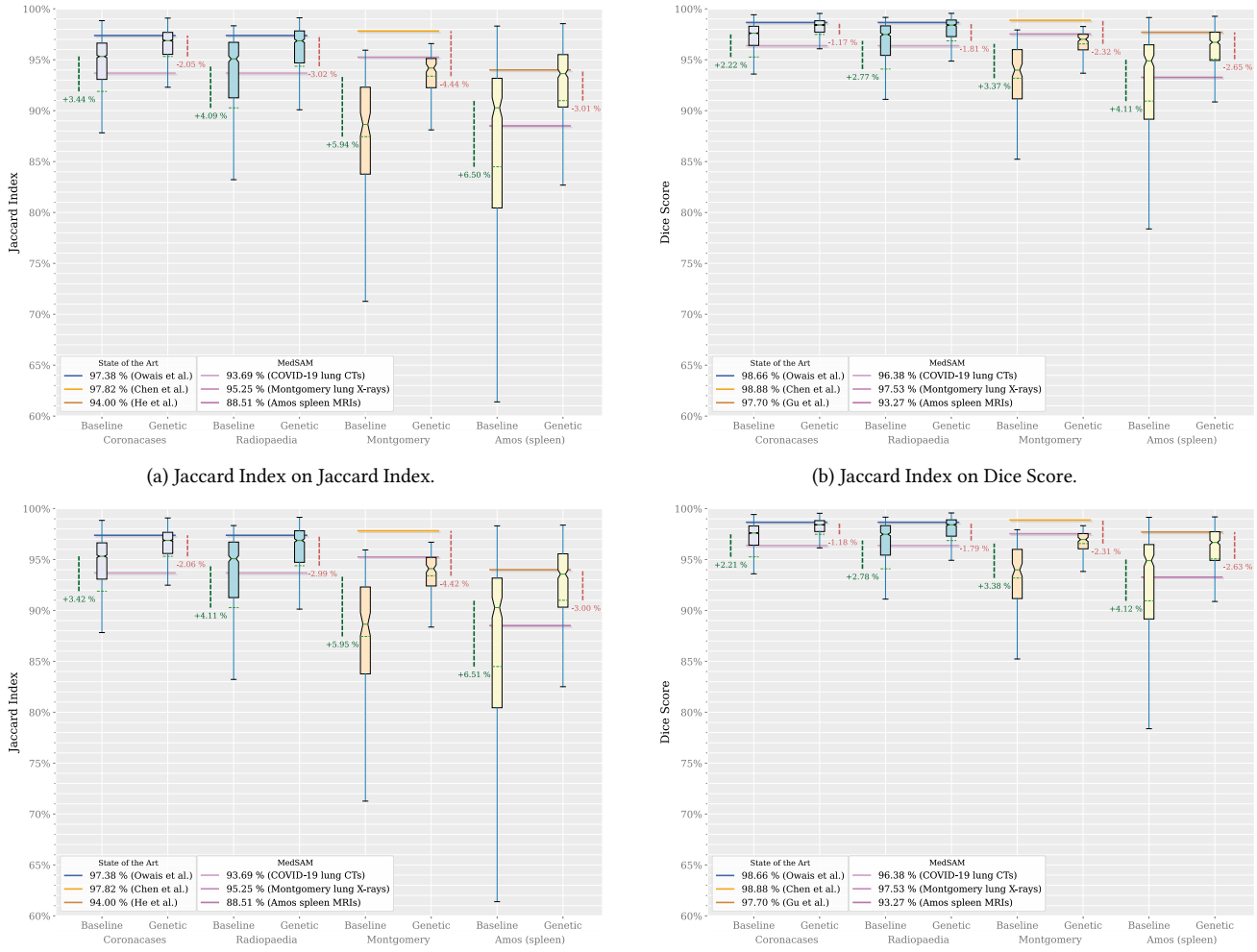
E. Experimental Setup Environment

The experiments were performed on a 2x Intel(R) Xeon(R) CPU E5-2630 v3 @ 2.40 GHz and 78 GB of RAM with an NVIDIA GeForce GTX TITAN Graphics Processing Unit (GPU), CUDA version 10.1 running under Ubuntu 14.04.6 and Python 3.9.0. For the implementation of the GA, the version 0.6.1.1 of the Python framework pymoo[1] for multi-objective optimization was used.

IV. Results

The genetic algorithm approach selected for this paper uses the segmentation evaluation metrics (Jaccard Index and Dice Score) as its objective. At the same time, we evaluate the segmentation performance using these metrics. How does using a given metric as an objective impact the results? For example, does using the Jaccard Index as an objective benefit the Jaccard Index results while hurting the Dice Score results? Fig. 3 shows the influence that using a given metric as the objective in the genetic algorithm optimization has on the other metrics. Fig. 3b, *e.g.*, shows the Dice Score obtained when using the Jaccard Index as the objective metric, *i.e.*, the influence of the Jaccard Index on the Dice Score. The results obtained in [6] are used as baseline. For each dataset, the baseline and the genetic optimization results are shown side by side. These box plots include the median of the distribution as a continuous green line, and its average as a dotted green line. The notch represents the confidence interval of the median. A darker green dotted line is shown between each pair of box plots, representing the difference between the average of the genetic optimization and the baseline. The value of said difference is shown at the top of each dotted line. The current state-of-the-art is shown above each pair of plots. CT results (Coronacases and Radiopaedia) are compared with [31], X-rays are compared with [33], while spleen MRI results are compared with [21] for the Jaccard Index and [16] for the Dice Score. The percentage still needed for SAM + the genetic algorithm optimization to reach the state-of-the-art is shown at the

[1] https://pymoo.org/



(a) Jaccard Index on Jaccard Index.

(b) Jaccard Index on Dice Score.

(c) Dice Score on Jaccard Index.

(d) Dice Score on Dice Score.

Fig. 3. Objective metric influence on other metrics.

TABLE I. Objective Metric Influence on Other Metrics (Averages)

	Baseline	Genetic	Delta
Coronacases	**91.89 %**	**95.33 %**	+3.44 %
Radiopaedia	90.27 %	94.36 %	+4.09 %
Montgomery	87.45 %	93.38 %	+5.94 %
Amos (spleen)	84.49 %	90.99 %	**+6.50 %**

(a) Jaccard on Jaccard.

	Baseline	Genetic	Delta
Coronacases	**95.27 %**	**97.49 %**	+2.22 %
Radiopaedia	94.08 %	96.85 %	+2.77 %
Montgomery	93.19 %	96.56 %	+3.37 %
Amos (spleen)	90.94 %	95.05 %	**+4.11 %**

(b) Jaccard on Dice.

	Baseline	Genetic	Delta
Coronacases	**91.89 %**	**95.32 %**	+3.42 %
Radiopaedia	90.27 %	94.39 %	+4.11 %
Montgomery	87.45 %	93.40 %	+5.95 %
Amos (spleen)	84.49 %	91.00 %	**+6.51 %**

(c) Dice on Jaccard.

	Baseline	Genetic	Delta
Coronacases	**95.27 %**	**97.48 %**	+2.21 %
Radiopaedia	94.08 %	96.87 %	+2.78 %
Montgomery	93.19 %	96.57 %	+3.38 %
Amos (spleen)	90.94 %	95.07 %	**+4.12 %**

(d) Dice on Dice.

right of each pair of plots, this time using a red dotted line. The actual values for each test are shown in Table I, with the best values in bold.

Additionally, we have included the mean values obtained by segmenting the different datasets using MedSAM [37] as an additional reference baseline. As this fine-tuned model was trained specifically on medical images and uses only bounding boxes as prompts, we adopted the same prompt configuration to generate results for our datasets, in alignment with the methodology described in MedSAM.

The general trend is that the genetic algorithm optimization improves the results obtained by SAM in all datasets, regardless of the metric selected as an objective for the genetic algorithm optimization. Besides, SAM's results are now closer to the state-of-the-art than before.

Furthermore, the results of SAM with genetic optimization are now closer to the state of the art and surpass those achieved by MedSAM across most datasets. The only exception is the Montgomery dataset, where the results obtained through SAM with genetic optimization closely approach—but do not surpass—those of MedSAM and other

state-of-the-art methods. In this case, it is reasonable that CT and MRI modalities yield higher Jaccard Index and Dice Score values, as they provide more granular spatial information and finer anatomical detail than X-ray. These modalities are designed to better visualize and delineate soft tissues. However, a more in-depth study is required to confirm this explanation.

When working with the Coronacases dataset, the genetic algorithm is able to reduce the gap with the state-of-the-art by 3.44 % using the Jaccard Index as objective, and by 2.22 % using the Dice Score. That way, the distance from the state-of-the-art has narrowed down to 2.05 % and 1.19 %, respectively.

On the other hand, when working with the Radiopaedia dataset, the genetic algorithm is able to reduce the gap with the state-of-the-art by 4.09 % using the Jaccard Index and by 2.78 % using the Dice Score. Now, the distance from the state-of-the-art is 3.02 % and 1.79 %, respectively.

In the Montgomery dataset, the genetic algorithm reduces the gap with the state-of-the-art by 5.94 % in terms of the Jaccard Index and by 3.38 % in terms of the Dice Coefficient. As a result, the average distance between SAM and the state-of-the-art is 4.44 % (Jaccard Index) and 2.31 % (Dice Score).

A more substantial improvement is observed in the spleen subset of the Amos dataset, where the genetic algorithm narrows the gap by 6.51 % for the Jaccard Index and by 4.12 % for the Dice Coefficient. In this case, SAM's average distance from the state-of-the-art is reduced to 3.01 % (Jaccard Index) and 2.63 % (Dice Score).

Not only the average values are closer to the state-of-the-art, but the distribution of the results is considerably better. In every instance of the results obtained by the genetic algorithm optimization, the gap between the median and the average is shorter than before, indicating that the genetic algorithm can improve the segmentation obtained by SAM. The Interquartile Range (IQR) and the confidence interval have narrowed, indicating that the results are more consistent than before.

A. Ablation Study

To study the impact of the different parameters of the GA, an ablation study was conducted in which specific characteristics of the algorithm were systematically modified. The original configuration employs a mutation probability (p_m) of 0.9, SBX as the crossover operator, and Tournament as the selection method. For this study, seven additional experiments were performed using a subset of Radiopaedia images, with the Jaccard Index as the objective function. In each experiment, one parameter from the standard configuration was altered. The parameters evaluated include mutation probabilities of $p_m = 0.5$, 0.7, 1.0, Uniform, One-Point, and Exponential crossover methods, and Random selection. Table II presents the results of these configurations in comparison with the standard setup, using the same image subset. As shown, the standard configuration yields the best performance.

TABLE II. Ablation Study: Average Jaccard Index for Each Parameter Configuration

	Objective (Jaccard Index)
Standard	**93.01 %**
$p_m = 0.5$	92.84 %
$p_m = 0.7$	92.66 %
$p_m = 1.0$	92.72 %
Random selection	92.95 %
Uniform crossover	92.90 %
Single Point crossover	92.81 %
Exponential crossover	92.97 %

However, since the differences are small, we can conclude that the effect of modifying these parameters is not particularly significant.

V. Discussion

This section provides a detailed interpretation of the experimental results, emphasizing the impact of GA optimization on segmentation quality. The evolution of objective metrics over time is analyzed in Section A to assess convergence and consistency across datasets. The dynamics of prompt displacement and movement trajectories are examined in Sections B and C to understand their contribution to performance gains. Section D evaluates SAM's internal confidence score as a potential surrogate metric for optimization and discusses its relevance to the real-world application of our proposal, when no ground truth is available. Practical implications for clinical workflows, including execution time and usability considerations, are addressed in Sections E to G. Finally, Section H contrasts the efficiency of the proposed approach with the resource demands of training conventional segmentation models, and Section I reflects on the limitations imposed by the observed performance ceiling.

A. Objective Metric Evolution

As illustrated in Fig. 3, the choice of objective metric (Jaccard or Dice) has little impact on the final results, confirming the robustness of the optimization process. Also, the improvement reached its limit, as the results are similar across all datasets and metrics. Fig. 4 shows that all metrics rapidly improve during early generations and then stabilize, indicating convergence. This pattern is consistent across datasets.

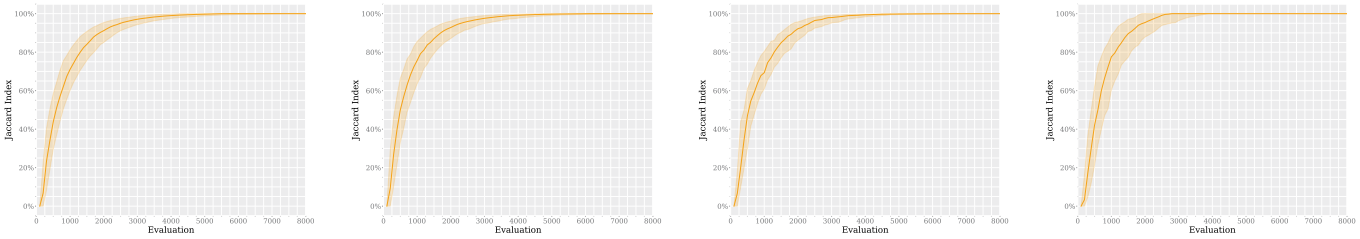
The evolution of the objective metric in every test is similar. All start with a low value, increase rapidly, and stabilize. All datasets show a rapid initial improvement followed by stabilization, with consistent patterns in the IQR indicating convergence. Nevertheless, that is the point of the genetic algorithm: to improve the results obtained by SAM until it is no longer possible. Interestingly, the limit reached in each case, regardless of the objective metric used by the genetic algorithm optimization, is quite similar.

B. Distance Distribution

Fig. 5 highlights that, in the lung datasets, positive points tend to move more than negative ones, with greater displacements correlating with higher optimization gains—especially in the Montgomery dataset. In contrast, the opposite pattern emerges in the spleen datasets, where the displacement of negative points from their original location exceeds that of positive points.

Suppose we focus on the positive points in each lung plot, where there are two positive points. In the spleen dataset, by contrast, there is only one positive point. In that case, the positive points travel similar distances across datasets, regardless of the objective metric used for the genetic algorithm optimization. The negative point, however, travels significantly less than the positive prompts. A notable distinction arises with the Montgomery dataset, where the overall prompt displacements are longer compared to the other datasets. Interestingly, larger prompt displacements are associated with greater optimization gains in the Montgomery dataset. This behavior may be attributed to the nature of the images in that dataset, which are lower-resolution cone beam CTs compared to Coronacases and Radiopaedia. A possible explanation for the shorter distance traveled by the positive prompt point in the spleen dataset is the smaller size of the target region, especially when compared to the lungs. Similarly, the larger surrounding area outside the spleen may account for the greater variability in the movement of the negative prompt point.

C. Travel

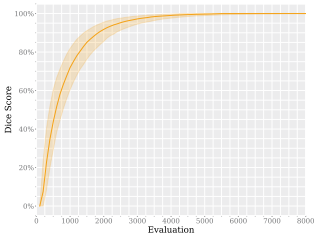


(a) Coronacases Jaccard Index evolution.

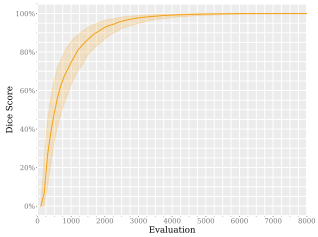
(b) Radiopaedi Jaccard Index evolution.

(c) Montgomery Jaccard Index evolution.

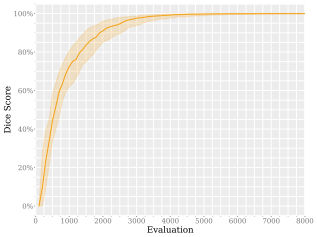
(d) Amos (spleen) Jaccard Index evolution.



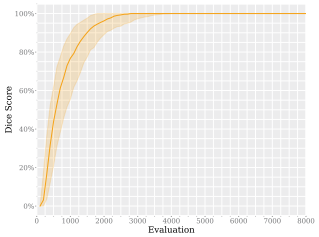
(e) Coronacases Dice Score evolution.



(f) Radiopaedi Dice Score evolution.

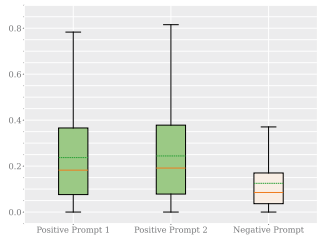


(g) Montgomery Dice Score evolution.

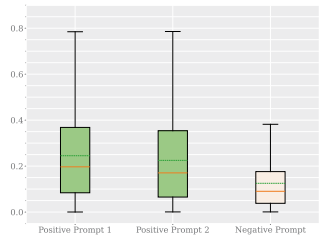


(h) Amos (spleen) Dice Score evolution.

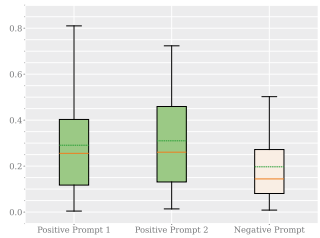
Fig. 4. Objective metric evolution over time for each dataset.



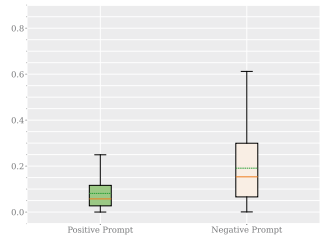
(a) Coronacases Jaccar Index.



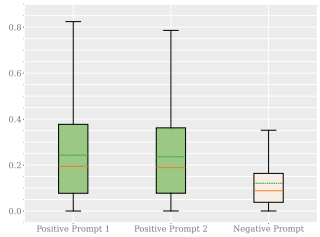
(b) Radiopaedia Jaccard Index.



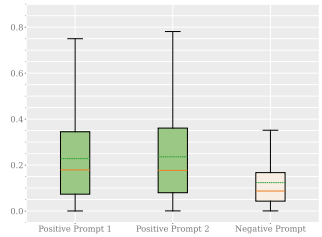
(c) Montgomery Jaccard Index.



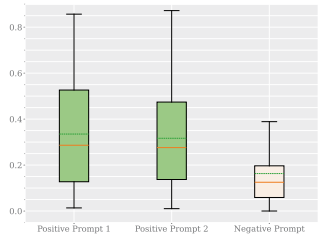
(d) Amos (spleen) Jaccard Index.



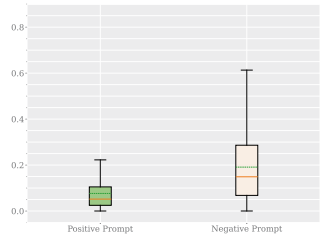
(e) Coronacases Dice Score.



(f) Radiopaedia Dice Score.

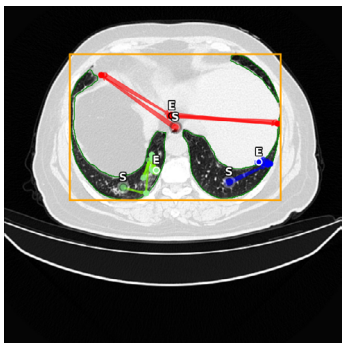


(g) Montgomery Dice Score.

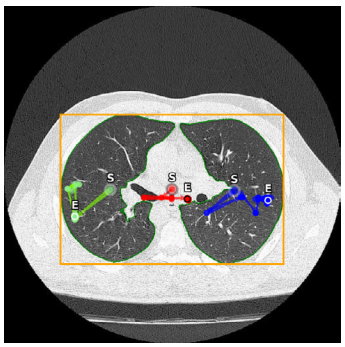


(h) Amos (spleen) Dice Index.

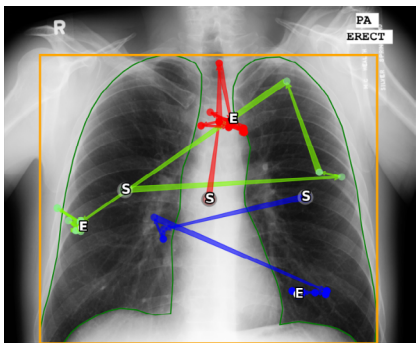
Fig. 5. Normalized pixel distance distribution for each dataset.



(a) Coronacases.



(b) Radiopaedia.



(c) Montgomery.

Fig. 6. Travel described by the prompts during the genetic optimization.

Prompt movement trajectories, shown in Fig. 6, reveal more substantial shifts in the Montgomery dataset, which aligns with its higher improvement gains. This suggests greater optimization potential where initial prompts are suboptimal.

Until an optimal result is achieved, the points used to mark each lung in Fig. 6a are moved within the lung. The prompt for the right lung ends in a location very close to its edge. The negative prompt moves to extreme locations but ends in a position very close to the initial one. In Fig. 6b, prompt movements are gradual, reflecting steady convergence toward optimal values. However, the positive and negative prompts in Fig. 6c describe trajectories that take them through very different positions.

One possible interpretation of the travel made by the positive and negative prompts in these images is that it is more difficult to improve the results in both Coronacases and Radiopaedia, which are already close to the maximum. However, Montgomery has more room for improvement, and is the one that benefits the most from the genetic algorithm optimization. The reason behind this is the nature of the images in the dataset. This interpretation aligns with the results shown in Fig. 3 and Table I, where the highest deltas for the lung images optimized using GAs are observed in the Montgomery dataset.

D. SAM Score as Estimated Metric

The results presented in Section IV show a significant improvement over those obtained by SAM in our previous work when a genetic algorithm is applied to it, without the need to retrain the model or make any adjustments. Using this method, SAM remains even closer to the state-of-the-art, competing with models specifically trained for this task. However, it is necessary to consider a fundamental aspect: for the genetic algorithm to perform its task, it is necessary to have the masks of the organ to be segmented and to use the corresponding metric as an objective. This information is not available in an actual environment, so applying this system would not be possible.

Fortunately, when performing the prediction, SAM returns a confidence value called SAM Score. This value predicts the Jaccard Index of the estimated segmentation [4]. The question is whether this confidence value improves with the genetic algorithm. The answer is yes, as can be seen in Fig. 7.

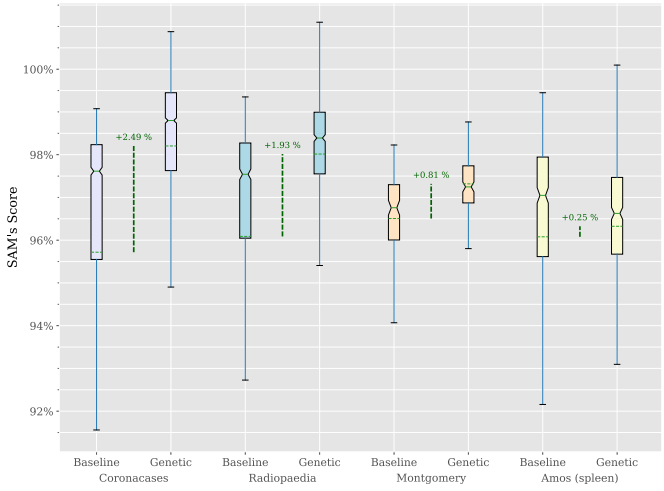


Fig. 7. SAM Score as objective metric.

As shown in Fig. 7, SAM's internal confidence score improves after optimization, with tighter value distributions and higher consistency across datasets. Not only that, but the evolution of the SAM Score over time follows the same pattern as the Jaccard Index and the Dice Score exhibited in Fig. 4, as Fig. 8 shows. In summary, the SAM Score could be used as a reliable estimator for genetic algorithm optimization, as its behavior seems similar to that of the Jaccard Index and the Dice Score.

E. Real-World Application

A potential application of this framework involves a specialist interacting with the system to segment a series of images iteratively. Initially, the specialist would obtain an image requiring segmentation and provide input to the SAM model by marking a bounding box encompassing the region of interest and specifying positive and negative prompts. The SAM model would then generate a preliminary segmentation. Importantly, even without access to ground truth masks, the GA can optimize the segmentation using the SAM Score as an estimated metric, as its behavior closely mirrors that of established overlap measures, as discussed in Section D. If deemed satisfactory, the specialist could leave this segmentation as background input for the genetic algorithm to refine over time. This process would be repeated for subsequent images, allowing the specialist to efficiently provide initial segmentations for a large dataset. Upon completion of this preliminary phase, the specialist could then review and evaluate the refinements made by the genetic algorithm to the initial segmentations.

F. Execution Time

The execution time required to complete the GA varies depending on the type of image. For the Montgomery dataset, a maximum processing time of 10 min per image was imposed to ensure that the entire set of experiments could be completed within a manageable timeframe, as individual runs exceeded this duration without such a limit. For the other datasets, it was not necessary to enforce a time constraint, as the execution times were consistently shorter. The average execution times per image and for the complete processing of each dataset (including additional required processing tasks) are reported in Table III.

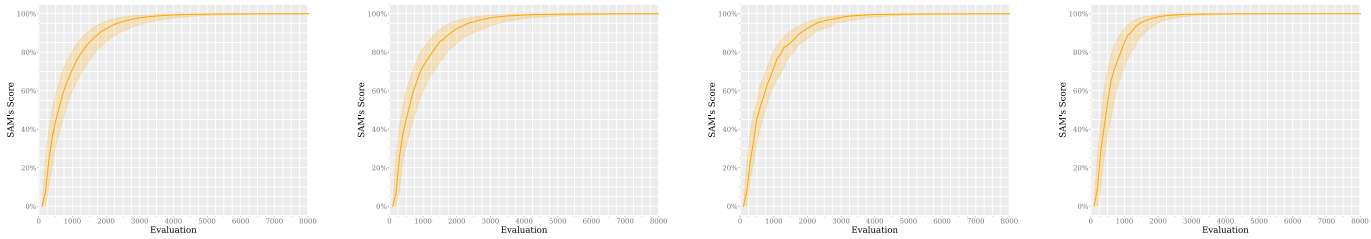
TABLE III. Average Execution Times Per Image and Per Dataset (Standard Deviation in Parentheses), in Minutes

	Per image	Total
Coronacases	1.69 (0.55)	3655.79 (511.51)
Radiopaedia	1.86 (0.58)	1530.30 (96.42)
Amos (spleen)	1.08 (0.40)	444.56 (89.57)

In view of the runtimes reported in Table III, GA-based refinement appears appropriate in scenarios where, after an initial segmentation using real-time applications with SAM, a radiologist requests the enhancement of specific images. While the runtime of the GA (approximately 1.50 min per image on average) is not suitable for real-time and highly responsive user interfaces, it represents a significant improvement over manual segmentation, which takes an average of 4.27 min per image [5]. This performance makes the approach well suited for batch processing and potentially viable for semi-interactive systems, particularly if further runtime optimizations are explored. Moreover, the proposed system could display progressively improved segmentations to the physician while the GA optimization runs in the background, allowing them to stop the process once a satisfactory result is reached. Alternatively, a target score could be defined so that the optimization process terminates automatically upon achieving it. Another option could be to use GA optimization as an on-demand enhancement method for SAM segmentations, which the physician could invoke when the initial result is not satisfactory.

G. Impact of SAM on Clinical Workflows

Manual segmentation of medical images is a laborious task that demands the expertise of highly trained professionals. Multiple recent investigations have analyzed the time investment necessary for



(a) Coronacases SAM Score evolution.

(b) Radiopaedi SAM Score evolution.

(c) Montgomery SAM Score evolution.

(d) Amos (spleen) SAM Score evolution.

Fig. 8. SAM Score evolution over time for each dataset.

segmenting anatomical regions across different imaging modalities. For example, in a study involving 100 images spanning nine distinct modalities, annotations were performed by three physicians with more than a decade of experience, requiring an average of 4.27 min per image [5]. In a similar vein, another work using COVID-19 CT images reported that segmenting a single scan composed of 250 slices took approximately 400 ± 45 min, equating to an average of 1.6 min per slice [38]. This segmentation effort was divided into three stages: initial labeling by junior specialists, refinement by senior radiologists, and a final verification by a highly experienced radiologist.

The adoption of SAM for medical image segmentation has considerably shortened annotation times while preserving, and in some cases enhancing, segmentation accuracy. In a multi-modality annotation task, SAM cut the average annotation duration from 4.27 to 2.96 min per image and simultaneously improved segmentation precision, reflected by a decrease in the Human Correction Effort (HCE) index from 5.07 to 4.80 [5]. Similarly, in the aforementioned work where manual annotation averaged 1.6 min per slice [38], SAM lowered the required time to approximately ~1 s, as detailed in Section F, achieving a time reduction exceeding 96 %. These results underscore the value of integrating SAM in clinical workflows: not only does it streamline segmentation tasks, but it also mitigates the burden on medical professionals.

SAM performs segmentation in roughly ~1 s. Given that a specialist-led segmentation process can span several minutes, the computational time required by SAM is practically negligible. The GA optimization batch processing described in Section E takes, on average, 1.69 min for the Coroncases dataset, 1.86 min for Radiopaedia, and 1.08 min for the Amos (spleen) dataset, as summarized in Table 3. However, it is crucial to note that following SAM's automated segmentation, clinicians are expected to revise and adjust the results based on their domain knowledge and the specific clinical scenario. Thus, SAM is not intended to supplant expert judgment in annotation or diagnosis; instead, it functions as a supportive tool that accelerates the segmentation process, reduces workload, and contributes to higher-quality outcomes through its advanced capabilities.

H. Requirements for Training Segmentation Models

As discussed in Section I, developing segmentation models for medical imaging is inherently resource-intensive, requiring considerable computational power, domain-specific knowledge, and financial support. These demands have been increasingly emphasized in recent research.

For example, DMDF-Net [31], a dual multiscale dilated fusion architecture, was developed to segment COVID-19 lesions from lung CT images. Its training relied on the MosMed [39] and COVID-19-CT-Seg [30] datasets, comprising a total of 70 CT scans and 5569 annotated slices. The model was trained for 15 epochs using a batch size of eight and a learning rate of 0.001, with computations performed on an NVIDIA GTX 1070 GPU. The overall project cost, including data preparation, network design, model training and validation, as well as salaries for a team of three engineers, was estimated at approximately 61 500 e over a six-month period. This figure is based on the average annual salary for a Machine Learning Engineer in Spain, which is around 41 000 €[2].

Another example is TransAttUnet [33], a U-Net variant enhanced with multi-level attention and transformer-based modules for medical image segmentation. The model was trained on 4255 images compiled from multiple public datasets, including ISIC-2018 [40], JSRT [41], Montgomery [32], NIH [42], Clean-CC-CCII [43], Bowl [44], and GLaS [45]. Training was performed over 100 epochs using a batch size of four and a learning rate of 0.0001, on an NVIDIA Titan XP GPU. With a development team comprising five engineers, the total project expenditure was estimated at 102 500 € for a six-month duration, based on the same average salary benchmark.

A further case is a lung CT segmentation method built upon the Mask R-CNN framework [46], which was trained on a dataset comprising 1265 annotated images. Model training was carried out on an NVIDIA GTX 1050 Ti GPU.

The overall project cost was estimated at 101 600 € over a six-month period, reflecting the high expenses often linked to extended training cycles and the tuning of hyperparameters. Assuming a team of four engineers, personnel costs alone would account for approximately 82 000 €, based on the average salary reference cited earlier. DS-TransUNet [47], a residual U-Net architecture tailored for medical image segmentation tasks, was trained using a combination of the LUNA, VESSEL12 [48], and HUG-ILD [49] datasets, totaling 11 325 images. The model underwent training for 50 epochs with a batch size of eight and a learning rate of 0.0001, using an NVIDIA GTX 1060 GPU. The development effort engaged a team of six engineers, with estimated personnel expenses reaching 123 000 € over a six-month period, according to the previously noted salary benchmark.

Moreover, the time required to achieve functional results is considerable. Training and refining foundational models, such as SAM, demands a large number of GPUs and extensive datasets to reach optimal performance. For instance, MedSAM, the SAM fine-tuned model for medical applications, utilized 20 NVIDIA A100 GPUs for distributed training [37]. MedSAM was trained on a large-scale dataset comprising over 1.57 million image-mask pairs spanning multiple imaging modalities and anatomical structures. This dataset was curated by aggregating images from publicly available medical image segmentation resources. The time required to assemble the source datasets and to construct the dataset used for MedSAM represents a substantial investment in data curation and preprocessing, including tasks such as harmonizing formats, annotating structures, and standardizing intensity ranges. This setup underscores the challenges of resource allocation and highlights the importance of optimizing efficiency in segmentation model development.

[2] https://www.glassdoor.es/Sueldos/spain-machine-learning-engineer-sueldo-SRCH_IL.0,5_IN219_KO6,31.htm, accessed on 23 April 2025.

These prior efforts illustrate the considerable computational load, extensive data requirements, and high financial investment typically involved in training medical image segmentation models. In contrast, our proposed methodology capitalizes on pretrained segmentation models and advanced prompting techniques, substantially lowering both computational and development overhead. Instead of building complex architectures from the ground up, we focus on efficient adaptation strategies that reduce GPU usage and eliminate the need for large-scale dataset assembly. Moreover, updating to newer versions of SAM or switching to a different foundation model is more straightforward than fine-tuning the model itself. This approach presents a cost-effective, scalable, and accessible solution for advancing medical image segmentation.

I. Ceiling

As demonstrated, using GAs to explore the space of possible prompts in search of optimal values achieves its goal. This combination of techniques leaves SAM even closer to the state of the art than our previous work. These results justify the use of combinations between foundation models and improvements that do not alter those models, making results close to the state of the art available to those who cannot afford to develop their models or alter them with expensive techniques.

However, the results confirm a performance ceiling: as seen in Fig. 4 and Fig. 8, the optimization quickly reaches an asymptote beyond which no further improvement is observed. From iteration 6000 onwards, there is no significant improvement in any of the experiments.

VI. Threats and Limitations

While introducing a novel approach for optimizing the segmentation of medical images with SAM using GAs, this study has several limitations. We distinguish here between *internal validity threats*, which concern the soundness of the conclusions drawn within the context of the study, and *external validity threats*, which relate to the generalizability of the findings beyond the study setting. These limitations highlight future research directions to strengthen and broaden the applicability of the presented findings.

A. Internal Validity Threats

Dataset selection bias. The use of datasets containing COVID-19 lesions introduces potential biases. Although the presence of these lesions does not appear to meaningfully affect the results, the reliance on such specific datasets may limit the robustness of the conclusions within this context. Future work should validate the approach on a wider range of lung conditions to confirm its reliability.

Prompt configuration bias. We employed a single, albeit effective, prompt configuration (a bounding box with two positive and one negative point), determined by our algorithm following predefined criteria. While this configuration proved successful and was verified visually, the study does not explore alternative prompt configurations. This may introduce biases related to prompt selection and limit the internal validity of conclusions about the method's adaptability.

B. External Validity Threats

Organ-specific focus. The study's focus on lung and spleen segmentation may restrict the generalizability of its findings to other anatomical structures. Although this focus allows a fair comparison with prior work, the applicability of the proposed approach to additional organs remains untested and should be addressed in future research.

Clinical setting generalizability. The results were obtained under controlled experimental conditions and may not directly translate to diverse clinical settings or imaging devices. Future studies should examine the method's performance across different institutions, populations, and imaging protocols to strengthen external validity.

VII. Conclusion

This study explores the potential of enhancing SAM by Meta for medical image segmentation using GAs. Our results demonstrate that integrating GAs significantly improves SAM's segmentation accuracy, achieving results closer to state-of-the-art performance on both axial lung CT scans and frontal chest X-ray datasets. This improvement is achieved without requiring retraining or modifying SAM's architecture, highlighting its potential for cost-effective and accessible medical image analysis.

However, our analysis also reveals a performance ceiling for this optimization technique, beyond which further improvement through genetic algorithm iterations becomes negligible. This result emphasizes the importance of exploring alternative optimization strategies and hybrid approaches to further bridge the gap between foundation and specialized segmentation models.

Future research should investigate the generalizability of our proposed approach to other organ segmentation tasks and diverse clinical datasets. Additionally, exploring different prompt configurations and incorporating expert-in-the-loop refinement strategies could further enhance the accuracy and robustness of this framework for real-world clinical applications. In addition to GAs, other optimization techniques could be explored and their suitability evaluated for this task. Potential alternatives include other evolutionary algorithms such as differential evolution; machine learning approaches, such as reinforcement learning; stochastic optimization methods like simulated annealing; swarm intelligence techniques, including particle swarm and ant colony optimization; or even custom metaheuristics designed *ad hoc* for this specific problem. While our method demonstrates strong performance in optimizing SAM segmentations through prompt evolution, future work could explore comparisons with fine-tuned segmentation models. Developing compatible evaluation strategies could offer valuable insights into the relative benefits of prompt tuning versus model fine-tuning in medical imaging applications, as well as how effectively both approaches can be combined. Despite its limitations, this study underscores the promising potential of combining foundation models with intelligent optimization techniques to democratize access to cutting-edge medical image analysis tools. Although this is a promising step, further research is needed.

Author Contributions

All authors contributed equally to the conception, design, implementation, and writing of this paper.

Funding

This work was supported in part by Grant CPP2021-008491 funded by MICIU/AEI/10.13039/50100011033 and by the European Union NextGeneration EU/PRTR, in part by the AEI (State Research Agency, Spain), the MCIN (Ministry of Science and Innovation, Spain), and the ERDF (European Regional Development Fund, EU), as part of the project PID2022-137275NA-I00 (X-BIO project) funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe".

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. Kirillov, K. He, R. Girshick, C. Rother, P. Dollar, "Panoptic Segmentation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, June 2019, pp. 9396–9405, IEEE.

[2] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, July 2017, doi: https://doi.org/10.1016/j.media.2017.07.005.

[3] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, D. Terzopoulos, "Image Segmentation Using Deep Learning: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, pp. 3523–3542, Feb. 2021, doi: https://doi.org/10.1109/TPAMI.2021.3059968.

[4] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, R. Girshick, "Segment Anything." Web Page, Apr. 2023. doi: https://doi.org/10.48550/arXiv.2304.02643.

[5] Y. Huang, X. Yang, L. Liu, H. Zhou, A. Chang, X. Zhou, R. Chen, J. Yu, J. Chen, C. Chen, S. Liu, H. Chi, X. Hu, K. Yue, L. Li, V. Grau, D.-P. Fan, F. Dong, D. Ni, "Segment anything model for medical images?," *Medical Image Analysis*, vol. 92, p. 103061, Dec. 2023, doi: https://doi.org/10.1016/j.media.2023.103061.

[6] J. D. Gutiérrez, R. Rodriguez-Echeverria, E. Delgado, M. Á. S. Rodrigo, F. Sanchez-Figueroa, "No More Training: SAM's Zero-Shot Transfer Capabilities for Cost-Efficient Medical Image Segmentation," *IEEE Access*, vol. 12, pp. 24205–24216, Jan. 2024, doi: https://doi.org/10.1109/ACCESS.2024.3353142.

[7] S. N. Kumar, A. L. Fred, P. S. Varghese, "An Overview of Segmentation Algorithms for the Analysis of Anomalies on Medical Images," *Journal of Intelligent Systems*, vol. 29, pp. 612–625, June 2018, doi: https://doi.org/10.1515/jisys-2017-0629.

[8] J. H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. MIT Press, Apr. 1992.

[9] K. K. Verma, B. M. Singh, "Deep Multi-Model Fusion for Human Activity Recognition Using Evolutionary Algorithms," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, p. 44, Dec. 2021, doi: https://doi.org/10.9781/ijimai.2021.08.008.

[10] T. Hui-Ye Chiu, C. Wu, C.-H. Chen, "A Generalized Wine Quality Prediction Framework by Evolutionary Algorithms," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, p. 60, Sept. 2021, doi: https://doi.org/10.9781/ijimai.2021.04.006.

[11] L. Ali, F. Alnajjar, M. Swavaf, O. Elharrouss, A. Abd-alrazaq, R. Damseh, "Evaluating segment anything model (SAM) on MRI scans of brain tumors," *Scientific Reports*, vol. 14, p. 21659, Sept. 2024, doi: https://doi.org/10.1038/s41598-024-72342-x.

[12] T. Cai, H. Yan, K. Ding, Y. Zhang, Y. Zhou, "WSPolyp-SAM: Weakly Supervised and Self-Guided Fine-Tuning of SAM for Colonoscopy Polyp Segmentation," *Applied Sciences*, vol. 14, p. 5007, June 2024, doi: https://doi.org/10.3390/app14125007.

[13] C. Chen, J. Miao, D. Wu, A. Zhong, Z. Yan, S. Kim, J. Hu, Z. Liu, L. Sun, X. Li, T. Liu, P.- A. Heng, Q. Li, "MA-SAM: Modality-agnostic SAM adaptation for 3D medical image segmentation," *Medical Image Analysis*, vol. 98, p. 103310, Aug. 2024, doi: https://doi.org/10.1016/j.media.2024.103310.

[14] G. Dong, Z. Wang, Y. Chen, Y. Sun, H. Song, L. Liu, H. Cui, "An efficient segment anything model for the segmentation of medical images," *Scientific Reports*, vol. 14, p. 19425, Aug. 2024, doi: https://doi.org/10.1038/s41598-024-70288-8.

[15] S. Gong, Y. Zhong, W. Ma, J. Li, Z. Wang, J. Zhang, P.-A. Heng, Q. Dou, "3DSAM-adapter: Holistic adaptation of SAM from 2D to 3D for promptable tumor segmentation," *Medical Image Analysis*, vol. 98, p. 103324, Aug. 2024, doi: https://doi.org/10.1016/j.media.2024.103324.

[16] Y. Gu, Q. Wu, H. Tang, X. Mai, H. Shu, B. Li, Y. Chen, "LeSAM: Adapt Segment Anything Model for medical lesion segmentation," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–11, May 2024, doi: https://doi.org/10.1109/JBHI.2024.3406871.

[17] X. Liu, Y. Zhao, S. Wang, J. Wei, "G-SAM: GMM-based segment anything model for medical image classification and segmentation," *Cluster Computing*, July 2024, doi: https://doi.org/10.1007/s10586-024-04679-x.

[18] Z. Ren, Y. Zhang, S. Wang, "Large Foundation Model for Cancer Segmentation," *Technology in Cancer Research & Treatment*, vol. 23, p. 15330338241266205, July 2024, doi: https://doi.org/10.1177/15330338241266205.

[19] P. Shi, J. Qiu, S. M. D. Abaxi, H. Wei, F. P.- W. Lo, W. Yuan, "Generalist Vision Foundation Models for Medical Imaging: A Case Study of Segment Anything Model on Zero-Shot Medical Segmentation," *Diagnostics*, vol. 13, p. 1947, June 2023, doi: https://doi.org/10.3390/diagnostics13111947.

[20] N. Ndipenoch, A. Miron, Y. Li, "Performance Evaluation of Retinal OCT Fluid Segmentation, Detection, and Generalization Over Variations of Data Sources," *IEEE Access*, vol. 12, pp. 31719–31735, Feb. 2024, doi: https://doi.org/10.1109/ACCESS.2024.3369913.

[21] D. He, Z. Ma, C. Li, Y. Li, "Dual-Branch Fully Convolutional Segment Anything Model for Lesion Segmentation in Endoscopic Images," *IEEE Access*, vol. 12, pp. 125654–125667, Aug. 2024, doi: https://doi.org/10.1109/ACCESS.2024.3449428.

[22] Z. Morton Colbert, D. Arrington, M. Foote, J. Gårding, D. Fay, M. Huo, M. Pinkham, P. Ramachandran, "Repurposing traditional U-Net predictions for sparse SAM prompting in medical image segmentation," *Biomedical Physics & Engineering Express*, vol. 10, p. 025004, Jan. 2024, doi: https://doi.org/10.1088/2057-1976/ad17a7.

[23] C. Wang, H. Chen, X. Zhou, M. Wang, Q. Zhang, "SAM-IE: SAM-based image enhancement for facilitating medical image diagnosis with segmentation foundation model," *Expert Systems with Applications*, vol. 249, p. 123795, Sept. 2024, doi: https://doi.org/10.1016/j.eswa.2024.123795.

[24] M. A. JiMing, D. HongYu, W. YuFan, W. LiNa, "Medical image segmentation based on simulated annealing and opposition-based learning island algorithm," *PLOS ONE*, vol. 19, p. e0307278, July 2024, doi: https://doi.org/10.1371/journal.pone.0307278.

[25] K. M. Hosny, A. M. Khalid, H. M. Hamza, Mirjalili, "Multilevel segmentation of 2D and volumetric medical images using hybrid Coronavirus Optimization Algorithm," *Computers in Biology and Medicine*, vol. 150, p. 106003, Nov. 2022, doi: https://doi.org/10.1016/j.compbiomed.2022.106003.

[26] D. R. Reis, B. C. Santos, L. Bleicher, L. E. Zárate, C. N. Nobre, "Prediction of enzymatic function with high efficiency and a reduced number of features using genetic algorithm," *Computers in Biology and Medicine*, vol. 158, p. 106799, Mar. 2023, doi: https://doi.org/10.1016/j.compbiomed.2023.106799.

[27] M. Chen, Z. Sun, F. Su, Y. Chen, D. Bu, Y. Lyu, "An Auxiliary Diagnostic System for Parkinson's Disease Based on Wearable Sensors and Genetic Algorithm Optimized Random Forest," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 2254–2263, Aug. 2022, doi: https://doi.org/10.1109/TNSRE.2022.3197807.

[28] M. Ghosh, S. Adhikary, K. K. Ghosh, A. Sardar, Begum, R. Sarkar, "Genetic algorithm based cancerous gene identification from microarray data using ensemble of filter methods," *Medical & Biological Engineering & Computing*, vol. 57, pp. 159–176, Aug. 2019, doi: https://doi.org/10.1007/s11517-018-1874-4.

[29] M. Sale, E. A. Sherer, "A genetic algorithm based global search strategy for population pharmacokinetic/pharmacodynamic model selection," *British Journal of Clinical Pharmacology*, vol. 79, pp. 28–39, June 2013, doi: https://doi.org/10.1111/bcp.12179.

[30] M. Jun, G. Cheng, W. Yixin, A. Xingle, G. Jiantao, Y. Ziqi, Z. Minqing, L. Xin, D. Xueyuan, C. Shucheng, W. Hao, M. Sen, Y. Xiaoyu, N. Ziwei, L. Chen, T. Lu, Z. Yuntao, Z. Qiongjie, D. Guoqiang, H. Jian, "COVID-19 CT lung and infection segmentation dataset." Web Page, Apr. 2020. doi: https://doi.org/10.5281/zenodo.3757476.

[31] M. Owais, N. R. Baek, K. R. Park, "DMDF-Net: Dual multiscale dilated fusion network for accurate segmentation of lesions related to COVID-19 in lung radiographic scans," *Expert Systems with Applications*, vol. 202, p. 117360, May 2022, doi: https://doi.org/10.1016/j.eswa.2022.117360.

[32] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, P.-X. Lu, G. Thoma, "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases," *Quantitative Imaging in Medicine and Surgery*, vol. 4, p. 475, Dec. 2014, doi: https://doi.org/10.3978/j.issn.2223-4292.2014.11.20.

[33] B. Chen, Y. Liu, Z. Zhang, G. Lu, A. W. K. Kong, "TransAttUnet: Multi-Level Attention-Guided U-Net With Transformer for Medical Image Segmentation," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 8, pp. 55–68, Sept. 2023, doi: https://doi.org/10.1109/TETCI.2023.3309626.

[34] Y. Ji, H. Bai, C. GE, J. Yang, Y. Zhu, R. Zhang, Z. Li, L. Zhanng, W. Ma, X. Wan, P. Luo, "AMOS: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation," Nov. 2022. doi: https://doi.org/10.5281/zenodo.7262581.

[35] D. Müller, I. Soto-Rey, F. Kramer, "Towards a guideline for evaluation metrics in medical image segmentation," *BMC Research Notes*, vol. 15, p. 210, June 2022, doi: https://doi.org/10.1186/s13104-022-06096-y.

[36] F. Kofler, I. Ezhov, F. Isensee, F. Balsiger, C. Berger, M. Koerner, B. Demiray, J. Rackerseder, J. Paetzold, H. Li, S. Shit, R. McKinley, M. Piraud, S. Bakas, C. Zimmer, N. Navab, J. Kirschke, B. Wiestler, B. Menze, "Are we using appropriate segmentation metrics? Identifying correlates of human expert perception for CNN training beyond rolling the DICE coefficient," *Machine Learning for Biomedical Imaging*, vol. 2, pp. 27–71, May 2023, doi: https://doi.org/10.59275/j.melba.2023-dg1f.

[37] J. Ma, Y. He, F. Li, L. Han, C. You, B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, p. 654, Jan. 2024, doi: https://doi.org/10.1038/s41467-024-44824-z.

[38] J. Ma, Y. Wang, X. An, C. Ge, Z. Yu, J. Chen, Q. Zhu, G. Dong, J. He, Z. He, T. Cao, Y. Zhu, Z. Nie, X. Yang, "Toward data-efficient learning: A benchmark for COVID-19 CT lung and infection segmentation," *Medical Physics*, vol. 48, pp. 1197–1210, Dec. 2020, doi: https://doi.org/10.1002/mp.14676.

[39] S. P. Morozov, A. E. Andreychenko, I. A. Blokhin, P. B. Gelezhe, A. P. Gonchar, A. E. Nikolaev, N. A. Pavlov, V. Y. Chernina, V. A. Gombolevskiy, "MosMedData: Data set of 1110 chest CT scans performed during the COVID-19 epidemic," *Digital Diagnostics*, vol. 1, pp. 49–59, Dec. 2020, doi: https://doi.org/10.17816/DD46826.

[40] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, H. Kittler, A. Halpern, "Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)," Mar. 2019. doi: https://doi.org/10.48550/arXiv.1902.03368.

[41] J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K.-i. Komatsu, M. Matsui, H. Fujita, Y. Kodera, K. Doi, "Development of a Digital Image Database for Chest Radiographs With and Without a Lung Nodule: Receiver Operating Characteristic Analysis of Radiologists' Detection of Pulmonary Nodules," *American Journal of Roentgenology*, vol. 174, pp. 71–74, Nov. 2012, doi: https://doi.org/10.2214/ajr.174.1.1740071.

[42] Y.-B. Tang, Y.-X. Tang, J. Xiao, R. M. Summers, "XLSor: A Robust and Accurate Lung Segmentor on Chest X-Rays Using Criss-Cross Attention and Customized Radiorealistic Abnormalities Generation," in *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*, May 2019, pp. 457–467, PMLR.

[43] X. He, S. Wang, X. Chu, S. Shi, J. Tang, X. Liu, C. Yan, J. Zhang, G. Ding, "Automated Model Design and Benchmarking of Deep Learning Models for COVID-19 Detection with Chest CT Scans," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 4821–4829, May 2021, doi: https://doi.org/10.1609/aaai.v35i6.16614.

[44] J. C. Caicedo, A. Goodman, K. W. Karhohs, B. A. Cimini, J. Ackerman, M. Haghighi, C. Heng, T. Becker, M. Doan, C. McQuin, M. Rohban, S. Singh, A. E. Carpenter, "Nucleus segmentation across imaging experiments: The 2018 Data Science Bowl," *Nature Methods*, vol. 16, pp. 1247–1253, Oct. 2019, doi: https://doi.org/10.1038/s41592-019-0612-7.

[45] P. Malík, K. Knapová, Š. Krištofík, "Instance Segmentation Model Created from Three Semantic Segmentations of Mask, Boundary and Centroid Pixels Verified on GlaS Dataset," in *Proceedings of the 2020 Federated Conference on Computer Science and Information Systems*, Sept. 2020, pp. 569–576.

[46] Q. Hu, L. F. De F. Souza, G. B. Holanda, S. S. Alves, F. H. Dos S. Silva, T. Han, P. P. Rebouças Filho, "An effective approach for CT lung segmentation using mask region-based convolutional neural networks," *Artificial Intelligence in Medicine*, vol. 103, p. 101792, Jan. 2020, doi: https://doi.org/10.1016/j.artmed.2020.101792.

[47] A. Khanna, N. D. Londhe, S. Gupta, A. Semwal, "A deep Residual U-Net convolutional neural network for automated lung segmentation in computed tomography images," *Biocybernetics and Biomedical Engineering*, vol. 40, pp. 1314–1327, July 2020, doi: https://doi.org/10.1016/j.bbe.2020.07.007.

[48] R. D. Rudyanto, S. Kerkstra, E. M. van Rikxoort, C. Fetita, P.-Y. Brillet, C. Lefevre, W. Xue, X. Zhu, J. Liang, İ. Öksüz, D. Ünay, K. Kadipaşaog˘lu, R. S. J. Estépar, J. C. Ross, G. R. Washko, J.- C. Prieto, M. H. Hoyos, M. Orkisz, H. Meine, M. Hüllebrand, C. Stöcker, F. L. Mir, V. Naranjo, E. Villanueva, M. Staring, C. Xiao, B. C. Stoel, A. Fabijanska, E. Smistad, A. C. Elster, F. Lindseth, A. H. Foruzan, R. Kiros, K. Popuri, D. Cobzas, D. Jimenez-Carretero, A. Santos, M. J. Ledesma-Carbayo, M. Helmberger, M. Urschler, M. Pienn, D. G. H. Bosboom, A. Campo, M. Prokop, P. A. de Jong, C. Ortiz-de-Solorzano, A. Muñoz-Barrutia, B. van Ginneken, "Comparing algorithms for automated vessel segmentation in computed tomography scans of the lung: The VESSEL12 study," *Medical Image Analysis*, vol. 18, pp. 1217–1232, July 2014, doi: https://doi.org/10.1016/j.media.2014.07.003.

[49] A. Depeursinge, A. Vargas, A. Platon, A. Geissbuhler, P.-A. Poletti, H. Müller, "Building a reference multimedia database for interstitial lung diseases," *Computerized Medical Imaging and Graphics*, vol. 36, pp. 227–238, July 2011, doi: https://doi.org/10.1016/j.compmedimag.2011.07.003.

Juan D. Gutiérrez

Assistant professor at the Universidad de Santiago de Compostela (USC). With more than twenty years of experience in the computer world, his research focuses on the application of AI to various fields of knowledge, with a particular interest in exploring low-cost optimizations for adapting foundation models to specialized tasks. His current work aims to systematize the identification of the operational limits of these models. His training includes programming in different languages, system administration, application design, and databases and the Internet. He has written more than twenty computer science books and translated another ten from English into Spanish. What began as a fun experience in the mid-nineties has ended up being a real passion for him. Juan Diego enjoys computing but, above all, learning new things.

Nuria Lozano-Garcia

She received the BSc + MSc degree in Biology from the University of Valencia in 2003, the Technical Engineering of Computer Systems university degree from the National Distance Education University (UNED) in 2012, and the MSc degree in Bioinformatics and Computational Biology from the Complutense University of Madrid in 2012, all of them in Spain. She is currently a member of the Department of Computer and Communications Technologies, and previously of the Department of Computers and Telematics Systems Engineering, University of Extremadura, in a Scientific and Research Staff position as a bioinformatician, and has held similar positions in Universities and Research Centers since 2012. She has co-authored or authored 11 Journal Citation Report (JCR) papers. Her research interests cover a wide range in the field of bioinformatics, as metagenomics, 16S rDNA, bacterial genomes and SNPs detection, Chip-Seq, and more recently evolutionary computation and multiobjective optimization applied to bioinformatics and other real-world problems.

Emilio Delgado

Researcher at the Universidad de Extremadura, whose primary focus is Machine Learning, particularly the study of DL. Currently, his research is at the intersection of artificial intelligence and healthcare, where he is applying DL techniques to solve medical problems. His work aims to use these algorithms to process and analyze large amounts of clinical and medical imaging data to improve people's standard of living. He is constantly looking for ways to improve and optimize DL algorithms for application in medicine, striving to ensure that they are accurate, efficient, and useful for healthcare professionals. He is exploring how machine learning can be used to improve medical diagnoses and treatments and investigating how these systems can be designed and trained to respect patient privacy and data security.

Álvaro Rubio-Largo

He received his Ph.D. in Computer Engineering from the University of Extremadura, Spain, in 2013. He is currently an Associate Professor in the Department of Computers and Telematics Systems Engineering at the University of Extremadura. With a strong academic and research background, Dr. Rubio-Largo has authored or coauthored over 70 publications, including more than 35 articles in journals indexed in the Journal Citation Reports (JCR). His research interests span big data, machine learning, and evolutionary computation, with a particular focus on multiobjective optimization for real-world applications. Dr. Rubio-Largo is active in the academic community, having co-organized several international workshops and served on the technical program committees of numerous international conferences. Additionally, he has co-edited several special issues for JCR-indexed journals and serves as a reviewer for various high-impact international journals.

Roberto Rodriguez-Echeverria

Professor of software architecture at the Computer Languages and Systems Department of the Universidad de Extremadura (UEx), Spain. His research interests include software engineering, model-driven engineering, data-driven software development, machine learning, web engineering, and legacy software modernization. He is currently head of the Applied Informatic Technology Institute. Moreover, he truly believes in local socioeconomic value generation through entrepreneurship, so he has recently created a new UEx spin-off company, named MetrikaMedia, which defines itself as a SaaS solution for multimedia content measurement.