# Investigating a Feature Unlearning Bias Mitigation Technique for Cancer-type Bias in AutoPet Dataset

Duc Thang HOANG[1] Quentin FERRE[1] Elsa SCHALCK[1] Olivier HUMBERT[2] Rosana EL JURDI[1]

[1]Euranova

[2]Université Côte d'Azur, France

**Résumé –** La mise en commun de jeux de données contenant différents types de cancer est cruciale pour améliorer la précision du diagnostic, en particulier dans les situations où la quantité de données est limitée, car les modèles d'apprentissage profond s'améliorent généralement lorsque la quantité de données d'entraînement augmente. Cependant, cet avantage inhérent est souvent entravé par l'introduction de variance due à des biais, tels que la sous-représentation ou la sur-représentation d'une maladie. Dans cet article, nous proposons un modèle invariant au type de cancer, capable de segmenter efficacement les tumeurs du lymphome et du cancer du poumon indépendamment de leur fréquence ou de leur représentation. Pour ce faire, nous formulons le problème comme une tâche d'apprentissage par transfert, en utilisant un réseau discriminateur et une fonction de perte de confusion pour préserver les caractéristiques génériques tout en désapprenant celles spécifiques au domaine. Nous démontrons que la méthode atteint des performances état de l'art tout en améliorant l'équité entre les sous-groupes sensibles et en étant efficace dans les scénarios présentant un fort déséquilibre entre les sous-groupes.

**Abstract –** The integration of datasets from different cancer type sub-groups is crucial for enhancing diagnostic accuracy, particularly in situations where data is limited, as deep learning models generally improve with more data. However, this inherent benefit is often hindered by the introduction of variance due to biases, such as the under-representation of one disease or the over-representation of another. In this paper, we propose a cancer-type-invariant model capable of effectively segmenting tumors from both lymphoma and lung cancer, irrespective of their frequency or representation. To this end, we frame the problem as a transfer learning task, utilizing a discriminator and a confusion loss to preserve generic features while unlearning domain-specific ones. We demonstrate that the method achieves state-of-the-art performance, improves fairness across sensitive subgroups, and is effective in scenarios with high subgroup imbalance.

## 1 Introduction

Medical image segmentation, which involves generating per-pixel predictions, plays a crucial role in early disease detection, diagnosis, and follow-up. While convolutional neural networks have achieved significant success in this area, a critical challenge remains : ensuring fair performance across patient subgroups, beyond simply achieving state-of-the-art results. Unfortunately, many models exhibit biases, leading to performance disparities among these subgroups. The often-overlooked unfairness aspects of CNNs can significantly compromise the reliability and equity of medical diagnoses.

Bias in medical datasets, including gender, racial, or scanner bias, can be categorized by their source or stage within the learning pipeline, as outlined in [3]. Training data bias occurs when the data doesn't accurately represent the target population, causing a mismatch in subgroup frequency or representation. This issue is exemplified by public datasets like AutoPET [4], a large-scale resource for lesion segmentation in Fluorodeoxyglucose Positron emission tomography / Computed Tomography (FDG-PET/CT) scans, containing over 1,000 scans from lymphoma, melanoma, and lung cancer patients. A model trained primarily on lung cancer cases in AutoPET performs well in detecting pulmonary nodules but struggles with lymphoma, which involves different anatomical regions like lymph nodes. This discrepancy, caused by the over-representation of certain diseases, leads to biased segmentation performance. Thus, addressing training data bias,

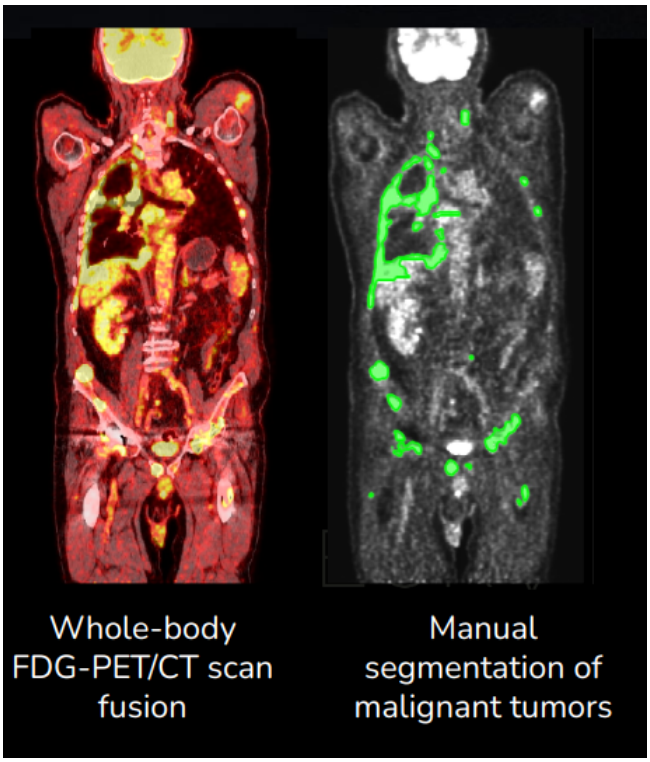


Figure 1 – Example of a whole-body FDG-PET/CT scan in AutoPET, where (left) shows a fused PET/CT scan and (right) illustrates the manual segmentation of malignant tumors [4].

especially concerning subgroup frequency, is crucial for ensuring accurate and unbiased diagnostic segmentation across diverse diseases and populations.

One approach to mitigating bias in medical imaging is framing the problem as a transfer learning task, treating each subgroup as a separate domain. This involves developing a domain adaptation framework that preserves common features while discarding domain-specific ones as proposed in [2, 6]. For instance, [2] harmonized data from different scanners, and [6] adapted a skull-stripping model from adults to newborns. In this paper, we apply similar techniques to bridge the gap between lung cancer and lymphoma datasets, addressing representation and prevalence biases. These cancers are chosen due to their similar image intensity characteristics but anatomical differences, with lung cancer appearing as pulmonary nodules and lymphoma as lymph nodes. This divergence and underlying biases allow us to explore developing models robust to both similarities and differences across diseases.

The paper is organized as follows: Section 2 presents the AutoPET dataset, Section 3 describes the proposed debiasing method, Section 4 provides experimental results on the AutoPET dataset with varying data imbalance frequencies, and Section 5 concludes with future work and perspectives.

## 2 Autopet Dataset

**Dataset Description.** The AutoPET dataset contains 1,014 FDG-PET/CT scans from 900 patients, including 489 scans with malignant melanoma, lymphoma, lung cancer, and negative controls. Acquired at University Hospital Tübingen and LMU Munich, it was part of the 2022 MICCAI challenge [4]. While the dataset includes 489 annotated scans, we focused our experiments on 154 lung cancer and 132 lymphoma scans, as they present a well-defined bias scenario. The melanoma and negative control cases were excluded to limit domain complexity and focus our study on subgroup fairness.

**Dataset Preprocessing.** PET scans were normalized to [0, 1], and CT scans to [-1000, 1000] using Hounsfield units [1] and SUV [9]. The dataset was resampled to a uniform resolution of $[2.62mm, 2.62mm, 2.62mm]$ with B-spline interpolation for CT and PET scans, and nearest neighbor for segmentation maps. These preprocessing steps were applied to 154 lung cancer and 132 lymphoma scans, which were used in this study. Nuclear medicine physician feedback indicates that these cancers share similar intensity values but differ anatomically, with lung cancer presenting as pulmonary nodules and lymphoma in lymph nodes (neck, mediastinum, abdomen). This study focuses on these two diseases to address biases from data imbalance.

## 3 Bias Mitigation pipeline

In this section, we present the explored debiasing model for medical image segmentation, detailing its components: the U-Net segmentation network [8], the discriminator [2], and the batch resampling algorithm [7]. We also discuss its training mechanism and optimization strategy. The adopted architecture is shown in Figure 2.

**Segmentation model.** The model uses a patch-based 3D U-Net for tumor segmentation (Figure 2), with an encoder of

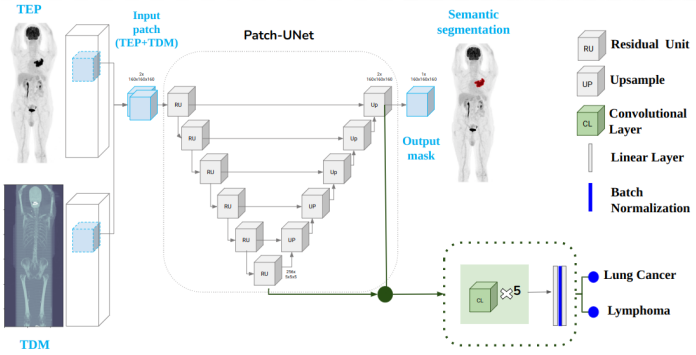


Figure 2 – Unlearning Architecture: Patch-UNet performs the segmentation task. The green block, a discriminator, concatenates (green dot) features from the U-Net's bottleneck and output to classify the bias domain. During unlearning, the discriminator is frozen, and a confusion loss is applied to enforce uniform distribution at the input level, thus unlearning cancer-type features in the U-Net layer.

convolutional layers (stride 2) and a decoder with transpose convolutions for upsampling. It processes parallel 3D patches [160, 160, 160] from CT and PET modalities, producing a binary tumor segmentation map, with training focused on positive (tumor) and negative (healthy tissue) patches.

**The discriminator** The Discriminator is a CNN for 3D data, with five convolutional layers and fully connected layers for classification [2, 6]. It uses instance normalization and progressively increases the number of channels. The model starts with 32 channels and expands to 512 before concatenating an auxiliary input y, output of the U-Net's bottleneck. After further processing, it outputs a binary classification of the sub-groups : Lung cancer and Lymphoma.

**Batch Resampling (**)** This strategy modifies the training sampling process to eliminate discrimination before training [7]. During each training batch, data is stratified by sensitive subgroups, and samples are selected to ensure equal representation of each protected group. Specifically, the batch resampling algorithm adjusts the sampling probabilities for each subgroup to prevent bias caused by imbalanced data distributions, ensuring fairness in the training process.

**Loss Functions** To train the network, we use three main loss functions. The **downstream task loss** $L_p$ optimizes the segmentation task and is defined as: $\mathcal{L} = (1-\lambda)\mathcal{L}Dice + \lambda\mathcal{L}ce$ This loss combines Dice and cross-entropy losses, with $\lambda$ dynamically adjusted during training, increasing with each epoch. The **domain loss** $L_d$ trains the descriptor to distinguish between sub-groups, guiding the unlearning process, and is based on the classification cross-entropy loss from [2]. Finally, the **confusion loss** encourages the model to "confuse" the subgroup identities, unlearning group-specific features. Defined as $L_{\text{conf}} = -\frac{1}{N}\sum_{n=1}^{N}\log(p_n)$, it uses the output probability $p_n$ from the discriminator, where $N$ is the number of sub-groups.

**Experimental Setting**. To train the unlearning pipeline, we randomly split the data into train, test (15%), and validation sets (See Table 1). We establish a baseline by training on one subgroup (**Subgroup-only**) and evaluating on the other. The unlearning pipeline is evaluated in three training schemes: **Balanced Scheme** trains on balanced data, while **Moderate Exclusion** and **Limited Access** involve training with 50% and

| | Training | | Validation | | Testing | |
|---|---|---|---|---|---|---|
| | Lung Cancer | Lymphoma | Lung Cancer | Lymphoma | Lung Cancer | Lymphoma |
| LungCancer-Only | 108 | - | 23 | - | 23 | 20 |
| Lymphoma-Only | - | 92 | - | 20 | 23 | 20 |
| Balanced Scheme | 108 | 92 | 23 | 20 | 23 | 20 |
| Moderate Exclusion | 108 | 50 | 23 | 10 | 23 | 20 |
| Limited Access | 108 | 23 | 23 | 5 | 23 | 20 |

Table 1 – Dataset distribution for each training scheme: **Balanced Scheme:** Training with nearly equal samples from Lung Cancer and Lymphoma groups. **Moderate Exclusion:** Training with a moderate reduction in Lymphoma samples (50% of Lymphoma samples). **Limited Access:** Training with significantly fewer Lymphoma samples (25% of Lymphoma samples)

| **Subgroup test set (number of patients)** | | | **Lung Cancer (23)** | **Lymphoma (20)** | **Lung Cancer+Lymphoma (43)** | | | |
|---|---|---|---|---|---|---|---|---|
| **Method / metrics** | Batch Res. | Unlearn | Mean dice ± std | Mean dice ± std | Av. | SER | STD | ESSP |
| **Subgroup-only** | | | | | | | | |
| **(1)** LungCancer-Only | | | 0.72 ± 0.18 | 0.45 ± 0.28 | 0.59 | 0.13 | 1.92 | 0.47 |
| **(2)** Lymphoma-Only | | | 0.57 ± 0.19 | 0.64 ± 0.25 | 0.59 | 0.03 | 1.19 | 0.55 |
| **Balanced Scheme** | | | | | | | | |
| **(3)** Mixed Train-RandomRes. | | | 0.74 ± 0.15 | 0.64 ± 0.23 | 0.69 | 0.04 | 1.37 | 0.63 |
| **(4)** Mixed Train-BatchRes | ✓ | | 0.73 ± 0.15 | 0.63 ± 0.23 | 0.68 | 0.04 | 1.37 | 0.62 |
| **(5)** Mixed Train-Unlearn | | ✓ | 0.76 ± 0.15 | 0.67 ± 0.22 | 0.71 | 0.04 | 1.35 | 0.66 |
| **(6)** Mixed Train-Unlearn-BatchRes | ✓ | ✓ | **0.74 ± 0.13** | **0.68 ± 0.24** | **0.71** | **0.03** | **1.25** | **0.67** |
| **Moderate Exclusion** | | | | | | | | |
| **(7)** Mixed Train-RandomRes. | | | 0.76 ± 0.14 | 0.64 ± 0.24 | 0.70 | 0.05 | 1.48 | 0.62 |
| **(8)** Mixed Train-BatchRes | ✓ | | 0.72 ± 0.15 | 0.64 ± 0.25 | 0.68 | 0.03 | 1.25 | 0.63 |
| **(9)** Mixed Train-Unlearn | | ✓ | 0.71 ± 0.17 | 0.64 ± 0.25 | 0.67 | 0.03 | 1.23 | 0.63 |
| **(10)** Mixed Train-Unlearn-BatchRes | ✓ | ✓ | **0.73 ± 0.14** | **0.67 ± 0.21** | **0.70** | **0.02** | **1.19** | **0.67** |
| **Limited Access** | | | | | | | | |
| **(11)** Mixed Train-RandomRes. | | | 0.75 ± 0.15 | 0.54 ± 0.27 | 0.65 | 0.10 | 1.85 | 0.54 |
| **(12)** Mixed Train-BatchRes | ✓ | | 0.74 ± 0.14 | 0.56 ± 0.26 | 0.66 | 0.09 | 1.73 | 0.55 |
| **(13)** Mixed Train-Unlearn | | ✓ | 0.72 ± 0.18 | 0.58 ± 0.26 | 0.65 | 0.07 | 1.52 | 0.57 |
| **(14)** Mixed Train-Unlearn-BatchRes | ✓ | ✓ | **0.73 ± 0.15** | **0.60 ± 0.25** | **0.67** | **0.06** | **1.48** | **0.59** |

Table 2 – Comparison of Dice coefficient values across methods with different training schemes: **LungCancer-Only**: Training exclusively on lung cancer samples. **Lymphoma-Only**: Training exclusively on lymphoma samples. **Mixed Train-RandomRes.**: Training on a mixture of lymphoma and melanoma samples. **Mixed Train-batchRes**: Training on a mixture of lymphoma and lung cancer samples with batch resampling. **Mixed Train-unlearn**: Training on a mixture of lung cancer and lymphoma samples with random resampling using the unlearning model. **Mixed Train-unlearn-BatchRes**: Training on a mixture of lung cancer and lymphoma samples using the unlearning model with batch resampling. Unlearn is using the model and optimization scheme mentioned in (*) whereas batch resampling is adopting the batch resampling method in (**)

25% lymphoma samples in the training set, respectively. Importantly, the test set remains consistent across all experiments to ensure fair evaluation of the unlearning pipeline.

**Optimization Strategy (*).** An iterative optimization strategy is used to address the conflicting objectives of domain discrimination and confusion losses. For each data batch, three forward and backward passes are performed: First, the main task loss $L_p$ is optimized to improve downstream task performance. Next, the domain discrimination loss $L_d$ is optimized to enhance the discriminator's ability to distinguish domains. Finally, the confusion loss $L_{\text{conf}}$ is optimized, but instead of updating the discriminator, it updates the feature extractor to hinder the discriminator's performance. This adversarial approach makes the feature representation domain-invariant, as the feature extractor learns to confuse the discriminator. Alternating between these losses ensures a balance between task performance and domain invariance, resulting in a robust and generalizable model. The unlearning model is hence defined as the combination of the U-Net and discriminator trained via the above optimization technique.

**Evaluation Metrics**. To validate segmentation performance, we use Dice accuracy [5] and three additional fairness metrics. These include the standard deviation between the average accuracy of each sub-group and the Skewed Error Rate (SER) [7], which represents the ratio of maximum to minimum error across subgroups; The Equity-Scaled Segmentation Performance (ESSP) metric from [10] is defined as

ESSP $= \frac{I((z',y))}{1+\Delta}$ where $I((z',y))$ represents the overall accuracy metric adopted for the segmentation task and $\Delta$ denotes the disparity in performance between the sub-groups. ESSP adjusts the Dice value by factoring in performance disparities, penalizing models that show significant accuracy differences between groups. ESSP encourages models that provide more equitable outcomes across all sub-groups. The ESSP value is always equal to or smaller than the average Dice value.

# 4 Results and Analysis

A summary of the results are reported in Table 4. Fairness metrics are reported to valorize the rhobustness of the method in different imbalanced schemes.

**Subgroup-only**: Models trained exclusively on either lymphoma or lung cancer (lines 1 and 2 in Table 4 ) show significantly worse segmentation accuracy on the other group, emphasizing that training on only one subgroup fail to generalize when applied to other sub-groups.

**Balanced training**: The average Dice coefficient across all mixed training methods appears similar. This suggests that the balanced scheme effectively bridges the performance gap seen in the baselines. Nevertheless, a deeper dive into the fairness metrics reveals inherent difference in these performance values. Specifically, the ESSP, SER, and STD metrics highlight significant disparities in fairness across the methods. Thus, the proposed method (line 6) demonstrates a 9 $\%$ reduction in the standard deviation between subgroups compared to the mixed training with random sampling (line 3). Furthermore, the penalized Dice score, as reflected by the ESSP, shows a $3\%$ increase, reinforcing the method's effectiveness in mitigating subgroup disparities, which were very clear in the baseline results, while maintaining strong segmentation performance.

**Moderate Exclusion**: The average Dice coefficient converges to the perfectly balanced scenario even with a moderate access to lymphoma class (line 6 vs line 10). A deeper examination of the fairness metrics, particularly the ESSP, SER, and STD, reveals that combined application of unlearning and batch resampling (line 10) yields the highest ESSP, demonstrating its capacity to enhance fairness even with the moderate data exclusion. Comparing Method 10 to Method 7 (the random resampling baseline of the Moderate Exclusion scheme), the ESSP shows an 8.06 $\%$ increase (from 0.62 to 0.67) and exhibits a significant reduction in the standard deviation between subgroups ( decrease STD from 1.48 to 1.19  19.59 $\%$ decrease) and therefore a significant increase in fairness.

**Limited access**: the average Dice coefficient converges, indicating that even under constrained data availability, the mixed unlearn training approaches effectively bridge the performance gap observed in the baselines. However, the performance is lower compared to the balanced scheme, highlighting the challenges of drastic limited data. A deeper examination of the fairness metrics, particularly the ESSP, SER, and STD, reveals notable variations : the combined application of unlearning and batch resampling (line 14) yields the highest ESSP, demonstrating its capacity to enhance fairness even with limited data. Comparing mixed training with unlearning and bias mitigation (line 14) vs mixed training on the segmentation model with random resampling (line 11), the ESSP shows a $9.26\%$ increase (from 0.54 to 0.59) and significant reduction in the standard deviation between subgroups. Thus, the STD decreases from 1.85 to 1.48, representing a $20\%$ decrease in STD, and therefore a significant increase in fairness.

As a result, one can say that the combined application of unlearning and batch resampling yields the highest ESSP value while maintaining state of the art performance, indicating an enhanced ability to balance performance and fairness.

# 5 Conclusion and Future Work

In this study, we investigate a feature unlearning bias mitigation transfer learning technique to address Cancer-type Bias. Our experiments demonstrated that the approach improves segmentation accuracy for underrepresented sub-groups, even with limited data. Future work will focus on extending our approach to other datasets and bias types. In addition, we aim to explore normalizing flows methods for common feature embeddings.

# 6 Acknowledgement

The authors acknowledge the ANR – FRANCE (French National Research Agency) for its financial support under reference ANR-23-CE23-0019 (FAMOUS)

# References

[1] Tami D. DenOtter and Johanna Schubert. *Hounsfield Unit*. StatPearls Publishing, Treasure Island (FL), 2023.

[2] Nicola K. Dinsdale, Mark Jenkinson, and Ana I. L. Namburete. Unlearning Scanner Bias for MRI Harmonisation in Medical Image Segmentation. In *Medical Image Understanding and Analysis*, pages 15–25, Cham, 2020. Springer International Publishing.

[3] Karen Drukker, Weijie Chen, Judy Gichoya, Nicholas Gruszauskas, Jayashree Kalpathy-Cramer, Sanmi Koyejo, Kyle Myers, Rui C. Sá, Berkman Sahiner, Heather Whitney, Zi Zhang, and Maryellen Giger. Toward fairness in artificial intelligence for medical image analysis. *Journal of Medical Imaging*, 10(6):061104, November 2023.

[4] Sergios Gatidis and Thomas Kuestner. A whole-body FDG-PET/CT dataset with manually annotated tumor lesions (FDG-PET-CT-Lesions), 2022.

[5] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571, 2016.

[6] Abbas Omidi, Aida Mohammadshahi, Neha Gianchandani, Regan King, Lara Leijser, and Roberto Souza. Unsupervised Domain Adaptation of MRI Skull-stripping Trained on Adult Data to Newborns. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7703–7712, Waikoloa, HI, USA, January 2024. IEEE.

[7] Esther Puyol-Antón, Bram Ruijsink, Stefan K. Piechnik, Stefan Neubauer, Steffen E. Petersen, Reza Razavi, and Andrew P. King. Fairness in Cardiac MR Image Analysis: An Investigation of Bias Due to Data Imbalance in Deep Learning Based Segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 413–423, Cham, 2021. Springer International Publishing.

[8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015.

[9] Joseph A. Thie. Understanding the standardized uptake value, its methods, and implications for usage. *Journal of Nuclear Medicine*, 45(9):1431–1434, 2004.

[10] Yu Tian, Min Shi, Yan Luo, Ava Kouhana, Tobias Elze, and Mengyu Wang. Fairseg: A large-scale medical image segmentation dataset for fairness learning using segment anything model with fair error-bound scaling. In *The Twelfth International Conference on Learning Representations*, 2024.