Towards Equitable Kidney Tumor Segmentation: Bias Evaluation and Mitigation

Muhammad Muneeb Afzal MUNEEB.AFZAL@NYU.EDU
Muhammad Osama Khan OSAMA.KHAN@NYU.EDU
Shujaat Mirza SHUJAAT.MIRZA@NYU.EDU
New York University, New York, USA

Abstract

Kidney tumors, affecting over 400,000 individuals annually, require accurate segmentation for effective treatment and surgical planning. Yet, manual segmentation is time-consuming, steering the medical community towards automated methods. While computer-aided diagnostic tools promise improvements, their transition into the real world mandates an understanding of their performance across diverse population subgroups. Our study is the first to investigate fairness concerning kidney and tumor segmentation, particularly focusing on sensitive attributes like sex and age. Our findings show an existence of bias in performance across both attributes. In particular, despite a male-dominated training dataset, females showed superior segmentation performance. Age groups 60-70 and above 70 also deviated significantly from the average performance for all ages. To address these biases, we comprehensively explore bias mitigation strategies - encompassing pre-processing techniques (Resampling Algorithm and Stratified Batch Sampling) and in-processing methods (Fair Meta-learning and architectural adjustments). Specifically, Attention U-Net was identified as the optimal model for balancing fairness across both attributes while maintaining high segmentation performance. We present a crucial insight that the architecture itself could be a source of inherent biases, and careful selection of the network design can inherently reduce these biases. Our assessment of UNet variants challenges the prevailing paradigm of model selection predicated solely on segmentation performance, especially considering the profound implications biases can have in clinical outcomes.

Keywords: Fair AI, Segmentation, Kidney Tumors, Bias Mitigation

1. Introduction

Kidney tumors constitute a significant health concern with an annual incidence exceeding 400,000 cases (Sung et al. (2021)). For formulating treatment strategy and surgery planning (Taha et al. (2018); Kutikov and Uzzo (2009)) for the patient, accurate segmentation of kidney and tumor using medical images is essential. Since manual delineation remains a daunting task that requires radiologists to annotate hundreds of slices, the medical imaging community has focused on developing automatic segmentation methods that improve segmentation quality.

While the advantages of AI-aided diagnostic tools seem evident, transitioning these methods from research to real-world application demands careful consideration. Typically, performance evaluation for deep learning models is based on the general population without considering diverse subgroups. However, these models may exhibit inconsistent performance across certain sensitive subgroups (e.g. sex or race), leading to differential treatments among these subgroups. Prior to deploying models in clinical settings, it is crucial to minimize inherent bias. The medical imaging community has increasingly focused on ensuring fairness in models across various modalities (e.g., MRI Ribeiro et al. (2022), X-Ray Seyyed-Kalantari et al. (2021)), anatomical regions (e.g., brain Ioannou et al. (2022), chest Cherepanova et al. (2021), heart Puyol-Antón et al. (2021)), and considers sensitive attributes (e.g., sex Petersen et al. (2022), age Brown et al. (2022), race Zhang et al. (2018)).

Previous research has indicated a higher prevalence of kidney cancer in males (Rampersaud et al. (2014)), and this gender disparity in renal cell carcinoma (RCC) incidence decreases with increasing age (Korn and Shariat (2017)). Given the observed influence of sex and age on kidney cancer, a significant

question arises regarding the fairness of segmentation tasks related to these sensitive attributes.

Surprisingly, despite kidney and tumor segmentation being a well-recognized challenge (Heller et al. (2019)) in the medical imaging community, no previous study has explored the fairness aspect of kidney and tumor segmentation. To bridge this gap, we investigate whether the segmentation methods, trained on the publicly available kidney tumor dataset, exhibit fairness across different subgroups defined by sensitive attributes: sex and age. In our study, we employ the nnU-Net network, recognized for its success in winning the Kidney and Kidney Tumor Segmentation 2019 (KiTS19) challenge, and train it using the KiTS19 dataset (Heller et al. (2019)). Our approach is one of the initial endeavors in the relatively unexplored area of fairness in medical segmentation (Ioannou et al. (2022); Puyol-Antón et al. (2022); Salahuddin et al. (2023)).

Our results reveal a pronounced bias in performance based on sex and age. Notably, despite the training data being predominantly male, the female subgroup exhibits significantly better performance. In terms of age, the model significantly deviates from the average score for groups between 60 to 70 and those above 70, performing worst for the former and best for the latter.

To mitigate these biases, we comprehensively experiment with four mitigation approaches: two *pre-processing* methods (Resampling Algorithm and Stratified Batch Sampling) and two *in-processing* techniques (Fair Meta-learning and altering architectural design). While all four methods reduced bias to varying degrees, choosing the appropriate network architecture was the most effective way to debias. Specifically, in terms of fairness, Attention U-Net performs the best in the sex attribute whereas U-Net performs the best in the age attribute. To balance out fairness across both attributes while maintaining segmentation performance comparable to nnUNet, we identify Attention U-Net as the most suitable model.

To summarize, our key contributions in this work are:

- We are the first to investigate fairness in kidney and tumor segmentation. Our analysis reveals notable biases in performance across sensitive attributes, namely sex and age.
- Through evaluating four bias mitigation approaches, we find that pre-processing techniques, such as Resampling Algorithm and Stratified Batch Sampling, outperform explicit fairness training methods like Fair Meta-learning.
- Unlike other fairness studies in medical imaging that center on mitigation strategies within a single architecture, our research explores the notion that the architecture itself could be the root of inherent biases. Our findings suggest that judicious architecture selection could serve as an intrinsic de-biasing mechanism.
- Our analysis reveals a trade-off between fairness and segmentation performance, highlighting the risk of prioritizing performance without addressing algorithmic bias in clinical contexts.

2. Related Work

A limited number of studies have explored fairness in medical image segmentation. Ioannou et al. (2022) addressed demographic bias in CNN-based brain agnetic Resonance(MR) segmentation, shedding light on the influence of demographic variables on segmentation outcomes. Puyol-Antón et al. (2021, 2022) examined potential biases in cardiac magnetic resonance imaging, particularly focusing on sex and racial discrepancies influenced by data imbalances. In a more expansive scope, Salahuddin et al. (2023) presented an end-to-end framework for head and neck tumor Positron Emission Tomography(PET)/Computed Tomography(CT) imaging, incorporating fairness alongside uncertainty and multimodal radiomics considerations. Previous works on kidney and tumor segmentation have solely focused on segmentation task or integrating clinical characteristics Lund and van der Velden (2021) to improve segmentation performance. However, fairness in kidney and tumor segmentation remains unexplored in existing literature, an oversight we address in this study.

Regarding mitigation strategies, previous research has identified interventions at three phases: pre-processing, in-processing, and post-processing techniques. Pre-processing methods adjust data using techniques like data resampling (Puyol-Antón et al. (2021); Brown et al. (2022)), GAN-based sample synthesis (Pakzad et al. (2022); Joshi and Burlina (2021)), and data aggregation from various sources (Seyyed-Kalantari et al. (2020); Zhou et al. (2021)). However, these methods can face challenges due to limited data or potential data skewing (Maluleke et al. (2022)). In-processing methods

focus on altering the model's architecture. Strategies such as adversarial learning reduce the impact of sensitive data on feature vectors (Adeli et al. (2021)), while disentanglement learning divides feature vectors (Deng et al. (2023)). Other methods, like the one proposed by Du et al. (Du et al. (2022)), adjust feature vector distances. Their effectiveness can vary, especially when sensitive attributes are closely linked to target tasks. Post-processing methods, though less prevalent, refine the outputs of models. They employ calibration for specific subgroup thresholds (Pleiss et al. (2017)) and pruning to eliminate certain neurons (Marcinkevics et al. (2022); Wu et al. (2022)), making the most of pre-trained models with minimal alterations. Beyond examining these mitigations, our study is the first to explore how network architecture itself might influence biases in medical imaging.

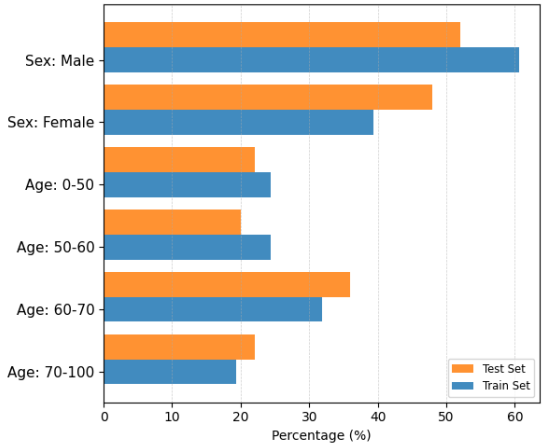

Figure 1: Distribution of gender and age groups within the KiTS19 dataset's training and test sets.

3. Methods

In the segmentation of kidneys and tumors, the model is required to output segmentations for both the kidney and the tumor using the input CT image $X \in \mathbb{R}^{H \times W \times C}$. We consider sex and age groups as sensitive attributes, s, aiming to achieve optimal segmentation performance that is unaffected by s.

3.1. Dataset

We utilized the KiTS 2019 dataset (Heller et al. (2019)) from the Kidney Tumor Segmentation Challenge. This dataset comprises volumetric CT scans of 210 patients who underwent either partial or radical nephrectomy at the University of Minnesota Medical Center between 2010 and 2018. These preoperative abdominal CT images, captured during the late-arterial phase, provide a distinct representation of kidney tumor voxels in the ground truth. The dataset, presented in the anonymized Neuroimaging Informatics Technology Initiative (NIFTI) format, includes imaging data alongside corresponding ground truth labels. Accompanying each scan is metadata detailing patient age, sex, and other pertinent clinical details. For our study, following Wang et al. (2020), the data was randomly divided into training and test sets of 160 and 50 samples. Figure 1 provides an overview of the distribution of gender and age groups within the KiTS19 dataset's training and test sets. Notably, similar patterns are observed across both data splits with slight variations, ensuring a consistent foundation for our subsequent analyses.

3.2. Preprocessing and Data Augmentations

3.2.1. Preprocessing

The KiTS dataset, like most large CT datasets, exhibits non-uniform voxel spacings, particularly in the voxel dimensions. Such variability can hinder the efficacy of 3D convolutions, often leading to performance akin to 2D models. Since CNN based architectures like nnUNet inherently struggle with inconsistent voxel spacings, preprocessing becomes crucial.

Following the recommended practices from nnUnet Isensee et al. (2021), we resampled all samples to a consistent voxel spacing. It is worth noting that the choice of voxel spacing plays a pivotal role in determining the amount of contextual information a 3D CNN can capture, as well as the overall voxel count of the image. However, a larger voxel spacing can compromise image detail. To strike a balance, we standardized all cases to a voxel spacing of $3.22 \times 1.62 \times 1.62$ mm for training samples.

CT images inherently offer quantitative consistency, meaning an organ should exhibit uniform intensity values across scans, even from varied scanners. Leveraging this property, we set intensity levels within an organ-specific range. In line with Isensee and Maier-Hein (2019), we constrained each case's intensity to the range $[-79, 304]$. These values were then normalized by subtracting 101 and dividing by 76.9, preparing them for processing within the nnUNet architecture.

3.2.2. Data Augmentations

To enhance our model's robustness and adaptability, we incorporated a myriad of data augmentation techniques during training using the MONAI framework (Cardoso et al. (2022)). We adjusted the spatial dimensions of both images and labels to match a specified patch size through spatial padding. We applied random cropping to regions based on positive and negative labels, ensuring a balance between the two. The images underwent random zooming between 0.9 to 1.2 times their original size with a 15% likelihood. Additionally, Gaussian noise, with a standard deviation of 0.01, and Gaussian smoothing–with varying sigma values across the x, y, and z dimensions–were introduced at a 15% chance. The intensity of the images was randomly scaled by a factor of 0.3 with a 15% probability. We also incorporated random flipping of images and labels across each of the three spatial axes, each with a 50% probability.

3.3. Model and Training

We adopted the nnU-Net architecture Isensee et al. (2021), renowned for its achievements in several medical segmentation challenges Ma (2021), including the Kidney Tumor Segmentation Challenge 2019 (KiTS19). In our study, this model served as the baseline for segmentation comparisons, trained without referencing protected attributes like race and gender.

Given GPU memory limitations, our approach aligned with conventional practices for 3D segmentation in CT data, training the model with patches of size $160 \times 160 \times 80$ voxels. Utilizing the stochastic gradient descent (SGD) optimizer, the model was trained for 2000 epochs to ensure convergence, with a learning rate set at $1e^{-3}$ and momentum at 0.9, with a batch size of 4. The training process incorporated both multi-class Dice loss and cross-entropy loss. In the specific instance of the Fair Meta-learning bias mitigation approach, we employed a hybrid of segmentation and classification loss as described in Equation 1. Our experiments demonstrate the impact of varying the parameters α and β on segmentation and fairness performance. Notably, deep supervision was employed, computing losses at every decoder stage, which inherently facilitates gradients to flow deeper into the network. All our methods were implemented using Pytorch (Paszke et al. (2019)) and MONAI framework (Cardoso et al. (2022)) on a single NVIDIA Tesla V100 GPU.

3.4. Metrics

We employed the Dice Similarity Score (DSC) as our segmentation metric, gauging the overlap between predicted and actual segmentations. We report DSC on kidney, kidney overlap, and their aggregated average.

In alignment with established fairness research (Puyol-Antón et al. (2021); Wang and Deng (2020)), we adopted the Standard Deviation (SD) and Skewed Error Rates (SER) as our fairness metrics. The SD quantifies the dispersion in mean DSC values across different sensitive groups. The SER is determined by the ratio of the maximum to the minimum error rate among these groups. It is mathematically represented as:

$$\text{SER} = \frac{\max_g(1 - \text{DSC}_g)}{\min_g(1 - \text{DSC}_g)}$$

where g denotes the protected groups.

The fairness metrics SD and SER were initially formulated for classification tasks as outlined in Wang and Deng (2020). Their applicability, however, extends beyond classification, having been effectively utilized in fairness evaluations for medical imaging segmentation, as evidenced in Puyol-Antón et al. (2021).

3.5. Fairness Evaluation

Our objective was to assess whether the baseline model performed consistently, without favoring one sex or age group over the other. To this end, we began by training the network on the entire training set without accounting for any attribute labels. Following this initial training, we delved into the model's predictions on protected group subsets within the test set, aiming to identify any performance disparities.

For sex-based fairness, we scrutinized the model's outcomes for both male and female subsets in the test set. For age, we segmented the test set into distinct age brackets, as delineated by Salahuddin et al. (2023): $[0, 50)$, $[50, 60)$, $[60, 70)$ and > 70. This granular approach facilitated an in-depth analysis of the model's consistency across various age groups.

3.6. Bias Mitigation Techniques

We implement bias mitigation approaches for our segmentation task and evaluate fairness by examining the performance across various subgroups, defined by

sensitive attributes. Note that we compare the results of these mitigation methods with our baseline nnU-Net model which is blinded to the sensitive attributes (sex and age). In particular, we conduct a comprehensive comparison of the baseline framework (nnU-Net) against four mitigation strategies: two pre-processing methods: Resampling and Stratified Batch Sampling, and two in-processing techniques: Fair Meta-learning and changes in architectural design.

3.6.1. Fair Meta-learning

This mitigation strategy is designed to address inherent biases in model predictions by making the network aware of the sensitive attributes like sex. This is achieved by integrating an additional classification branch dedicated to identifying the sensitive attribute alongside the primary segmentation network. Drawing from insights in prior research (Xu et al. (2020); Puyol-Antón et al. (2021)), the core intuition is to reduce spurious correlations between sensitive attributes and the representations learned for the segmentation task.

For this attribute classification, we employ a DenseNet network (Huang et al. (2017)) that processes the original CT image. The setup is treated as a multi-task learning problem, jointly optimizing both segmentation and classification networks. The combined loss function is defined as:

$$L_{\text{total}} = \alpha L_{\text{segmentation}} + \beta L_{\text{classification}} \quad (1)$$

where α and β are used to balance the contributions of the segmentation and classification losses, respectively. In this context, $L_{\text{segmentation}}$ is a combination of dice and cross-entropy Loss, while $L_{\text{classification}}$ is the standard cross-entropy loss computed from classification labels.

3.6.2. Resampling Algorithm (RESM)

The Resampling Algorithm (RESM) (Du et al. (2022); Kamiran and Calders (2012)) is a pre-processing strategy that balances the dataset by adjusting sample counts based on sensitive attribute groups. Specifically, it oversamples from underrepresented groups and undersamples from overrepresented ones to achieve a balanced dataset. This approach encourages the model to treat all groups equitably. In our experiments, we employed equal sampling weights, ensuring each group is represented equally in training.

3.6.3. Stratified Batch Sampling

Stratified Batch Sampling, a pre-processing technique, aims to eradicate biases at the batch sampling phase of training. By categorizing data according to sensitive attributes within each training batch, this approach ensures that every sensitive group is equally represented. By doing so, the model is consistently exposed to a diverse set of data, reducing the risk of bias towards any particular subgroup. Such stratification has been previously employed to bolster fairness in both classification and segmentation (Kamiran and Calders (2012); Puyol-Antón et al. (2022)).

3.6.4. Altering Architectural Design

While traditional methods for de-biasing in medical imaging rely on a consistent neural network architecture, we probe deeper to question if inherent model biases might originate from the architecture itself. To this end, we delved into the exploration of various U-Net variants, a prevalent architecture widely used in medical imaging tasks.

Owing to its exceptional performance, as our baseline, we employed nnU-Net, a network that was employed to win the Kidney and Kidney Tumor Segmentation Challenge 2019. This baseline was evaluated against other prominent architectures: the classic U-Net, V-Net, and the Attention U-Net.

4. Results

In this section, we examine the results concerning the prevalence of bias in relation to the sensitive attributes of sex and age, while also evaluating the effectiveness of various mitigation strategies deployed to address these biases. Specifically, Section 4.1 provides an evaluation of model fairness for sex and age attributes. Section 4.2 investigates a variety of bias mitigation techniques designed to enhance model fairness. We also examine the trade-off between achieving optimal segmentation performance and upholding fairness criteria, synthesizing the insights gained to identify the most effective approach across all attributes and mitigation strategies.

4.1. Fairness Evaluation

Table 1 provides an overview of our assessment of sex and age bias for the state-of-the-art approach for kidney and kidney tumor segmentation. Across both protected attributes, we observe that the baseline

Table 1: Performance and Fairness Evaluation of Kidney Tumor Segmentation Across Sensitive Groups on our baseline method. The table shows Dice Similarity Coefficient (DSC) values for Kidney and Tumor segmentations and their mean, across the entire dataset and further divided by gender and age groups. For Fairness Evaluation, we use Standard Deviation - SD (lower is better) and Skewed Error Rate - SER (1 is optimal) metrics. The high values of SD and SER (boldfaced) signify high bias. The average and standard deviation scores with three random seeds are reported.

Attributes	**Group**	**DSC**			**Fairness**	
		Kidney ↑ (%)	**Tumor ↑ (%)**	**Mean ↑ (%)**	**SD ↓**	**SER ↓**
All	-	94.9 ± 0.05	78.0 ± 0.90	86.5 ± 0.65	-	-
Gender	male	95.1 ± 0.05	73.4 ± 0.40	84.2 ± 0.20	2.32 ± 0.38	1.42 ± 0.09
	female	94.7 ± 0.10	83.0 ± 1.45	88.9 ± 1.25		
Age	0 - 50	95.1 ± 0.05	79.8 ± 0.20	87.4 ± 0.05	3.22 ± 0.49	2.08 ± 0.13
	50 - 60	95.0 ± 0.01	77.0 ± 0.30	86.0 ± 0.25		
	60 - 70	95.1 ± 0.25	70.5 ± 2.25	82.8 ± 1.15		
	> 70	94.4 ± 0.30	89.6 ± 0.40	91.8 ± 0.10		

nnUnet-based model exhibits biases, with the fine-grained analysis presented next.

4.1.1. Fairness Assessment for Sex

We observe a notable disparity in performance (mean DSC) between females and males, with females exhibiting significantly higher performance. Furthermore, a high standard deviation (SD) and Skewed Error Rate (SER) clearly indicates the existence of bias among the sensitive group (Table 1). This result is particularly surprising considering the composition of the training set, which was predominantly male (61%) as opposed to female (39%). Lifestyle disparities, particularly in smoking and alcohol usage, as noted in our dataset, might correlate with various health conditions and complicate medical diagnosis. Higher incidence of smoking and alcohol usage among males could partially explain why a model trained on this dataset might underperform on the male subgroup despite their majority presence. The male subgroup exhibits lifestyle habits that correlate with health risks, potentially leading to a broader range of medical presentations and outcomes that a model would need to generalize across.

4.1.2. Fairness Assessment for Age

There exists a significant variation in segmentation performance across different age groups. Specifically, the mean DSC scores for the age groups 60-70 and above 70 exhibits a noticeable deviation from the average DSC score computed across all age demographics (Table 1). This variation is supported by a high SD and SER, confirming the presence of bias in the age attribute. These results suggest that our baseline method exhibits biases across different age groups, with a tendency to yield better segmentation results for patients who are either below 50 or above 70. This finding is particularly important as it highlights the necessity to address age-related biases in the model to ensure equitable performance across all age groups.

To reduce sex and age bias in the baseline segmentation model, we experiment with various bias mitigation techniques next (Section 4.2).

4.2. Bias Mitigation Approaches

Tables 2 and 3 provide overviews of the comparisons between the baseline approach and four bias mitigation methods, focusing on the attributes of sex and age, respectively. We will discuss the specifics in the following sections.

4.2.1. Fair Meta-learning

For the sex attribute, making the network cognizant of this attribute by concurrently performing classification of both sexes improves fairness, as indicated by the reduced SD and SER compared to the baseline (Table 2). Our findings corroborate previous studies (Xu et al. (2020); Puyol-Antón et al. (2021))

Table 2: Comparison of Bias Mitigation Techniques for Sex: Performance and Fairness Metrics Evaluation

Mitigation	**DSC**			**Fairness**	
	Kidney ↑ (%)	**Tumor ↑(%)**	**Mean ↑(%)**	**SD ↓**	**SER ↓**
Baseline	94.9	78.0	86.5	2.32	1.42
Fair Meta-learning	94.4	78.3	86.3	1.55	1.26
Stratified Batch Sampling	94.7	76.6	85.6	1.20	1.18
RESM Algorithm	94.3	76.3	85.3	0.75	1.11
Architecture: Attention U-Net	94.8	75.6	85.2	**0.40**	**1.06**

Table 3: Comparison of Bias Mitigation Techniques for Age: Performance and Fairness Metrics Evaluation

Mitigation	**DSC**			**Fairness**	
	Kidney ↑ (%)	**Tumor ↑(%)**	**Mean ↑(%)**	**SD ↓**	**SER ↓**
Baseline	94.9	78.0	86.5	3.22	2.08
Fair Meta-learning	94.6	79.4	87.0	3.24	2.02
Stratified Batch Sampling	94.2	75.1	84.6	3.33	1.80
RESM Algorithm	94.5	76.6	85.6	2.52	1.69
Architecture: U-Net Network	94.6	73.0	83.8	**0.80**	**1.10**

that have demonstrated that explicitly encoding sensitive attribute information with a classification head enhances network fairness.

Conversely, for the age attribute, the *Fair Meta-learning* did not yield the expected improvement in fairness. As detailed in Table 3, the increased SD value for age groups indicates that explicitly encoding sensitive attributes does not universally guarantee improved network fairness. This observation highlights the complexity of the relationship between sensitive attributes and network fairness. To explore various parameters for the loss function (Equation 1), we conducted an ablation study, detailed in Appendix C, to identify the optimal settings. Additionally, Appendix A contains Table 5, where the detailed results are presented.

4.2.2. Stratified Batch Sampling

The *Stratified Batch Sampling* method ensures equal selection of each sensitive attribute in every learning batch. For the sex attribute, the method is successful at reducing bias by providing a more balanced sample of males and females, making the network less likely to be skewed towards one sensitive attribute (Table 2).

As for the age attribute, the *Stratified Batch Sampling* method provided marginal improvements in fairness by providing balanced samples of each age group in every learning batch. Similar to *Fair Meta-learning*, this method was not effective in mitigating biases amongst age groups (cf. SD value in Table 3). This suggests that simply offering a balanced batch for learning during the pre-processing phase is insufficient. To effectively mitigate bias in age groups, there's a need for advanced methods that alter the learning algorithm. Guided by this insight, we delved into in-processing mitigation methods, as detailed in Sections 4.2.1 and 4.2.4, which involve modifications to the network architecture.

4.2.3. RESM Algorithm

The RESM approach notably improves fairness for both sensitive attributes, as shown in Tables 2 and 3. For a detailed breakdown, see Table 7 located in Appendix A. Unlike Stratified Batch Sampling, which ensures each batch has an equal number of samples from each subgroup, the RESM Algorithm samples the training dataset to maintain an equal number of samples for each sensitive subgroup in the entire training set.

Compared to the baseline, we see a significant reduction in bias for the sex attribute (Table 3) and a noticeable reduction for the age attribute (Table 4). In particular, achieving balanced representation

Table 4: Fairness Evaluation for Bias Mitigation using Different Segmentation Architectures

Architecture	DSC			Sex		Age Group	
	Kidney (%)	Tumor (%)	Mean(%)	SD	SER	SD	SER
U-Net	94.6	73.0	83.8	0.80	1.10	**0.88**	**1.16**
V-Net	94.6	73.6	84.1	1.25	1.17	2.73	1.61
Attention U-Net	94.8	75.6	85.2	**0.40**	**1.06**	1.66	1.31
nnUNet	94.9	78.8	**86.9**	2.45	1.46	2.94	2.00

in the training dataset resulted in improved performance and fairer outcomes for males. This contrasts with the baseline scenario where, despite their over-representation, they faced under-diagnoses. The improved fairness emphasizes the importance of assembling datasets with comparable proportions of sensitive attributes, a practice often overlooked in many datasets. However, this improvement in fairness might have come at the cost of relatively decreased performance for females, a phenomenon highlighted by Suriyakumar et al. (2022). In light of this apparent trade-off, it is crucial to develop methods that enhance overall fairness without significantly reducing the performance of any particular group. Refer to Table 7 in Appendix A for detailed results.

4.2.4. Altering Architectural Design

To assess the impact of architectural design on fairness, we conducted experiments with several U-Net variations. Our findings underscore that architectural modifications markedly influence the model's fairness across both sensitive attributes, as evidenced by Tables 2 and 3 (note that an SER value of 1 denotes optimal fairness).

Comprehensive findings related to architectural adjustments are detailed in Table 4 (alongside Table 8 in Appendix B). Among the tested architectures, the Attention U-Net emerges as a favorable choice for fairness concerning the sex attribute, while the classic U-Net is better suited for age-related fairness. However, this age-related fairness in U-Net comes at the expense of some segmentation performance.

4.2.5. Bias Mitigation: Outcomes and Recommendations

Upon evaluation of various bias mitigation techniques, clear patterns emerge in their effectiveness. For the sex attribute, we observe that every mitigation strategy improves fairness as reflected by SD and SER values compared to the baseline model (Table 2). For the age attribute, all strategies effectively mitigate the bias if we consider SER as the sole fairness metric. However, if we take SD into account, *Fair Meta-learning* and *Stratified Batch Sampling* fall short in improving fairness.

Interestingly, a consistent pattern emerges regarding the efficacy of mitigation strategies across both attributes. Specifically, *Fair Meta-learning* demonstrates the most modest improvement in fairness. This is followed by balanced representation approaches, namely *Stratified Batch Sampling* and *RESM Algorithm*. Modifying the architectural design stands out as the most effective technique.

The comparative success of techniques like Stratified Batch Sampling and RESM Algorithm as opposed to Fair Meta-learning hints at an important insight: sometimes, fairness might be more effectively achieved at the data level rather than trying to force the model to learn it. Pre-processing techniques, which aim to balance the data before it even reaches the model, may offer a more foundational approach to fairness.

Our findings suggest that the prevailing trend of selecting architectures based purely on segmentation performance can adversely impact fairness. Our data indicates that the selected architecture plays a pivotal role in shaping the biases. Indeed, an appropriate selection of architecture could serve as an intrinsic de-biasing mechanism.

We show that although nnU-Net has achieved significant recognition in medical segmentation challenges, it might not always be the optimal selection when prioritizing fairness. To strike a balance, we recommend Attention U-Net as the preferred choice, as it outperforms nnUNet in fairness for both sex and age attributes while maintaining comparable segmentation performance (see Table 4). We hypothesize that attention gates in Attention U-Net (Oktay et al. (2018)) contribute to its notable fairness, as they in-

herently learn to suppress irrelevant image regions while emphasizing the salient features vital for kidney and tumor identification and localization.

We conclude that selecting models based solely on segmentation performance may compromise fairness. Our exploration with variants of UNet based architectures highlights the need for evaluation criteria that balance performance and fairness. Leveraging Neural Architecture Search (NAS) specifically tailored for fairness could be pivotal in this endeavor. As medical imaging advances, prioritizing architectures that guarantee both performance and equity is essential, especially considering the grave consequences of bias in clinical decisions.

5. Conclusion

In this study, we are the first to investigate fairness in the widely recognized Kidney and Kidney Tumor Segmentation task focusing on the sensitive attributes of sex and age. Our findings showed that while the current models, such as nnU-Net, offer promising high segmentation performance, they exhibit significant biases across both sensitive attributes. In particular, although the data is dominated by male subgroup, female subgroups exhibited superior performance. Furthermore, age-based discrepancies in segmentation performance were evident, particularly among the 60-70 and above 70 age groups. To counter these biases, we rigorously evaluated four mitigation techniques, concluding that an informed choice of network architecture emerges as the most potent bias mitigator. Notably, Attention U-Net excelled in balancing fairness and segmentation performance. As we usher these tools into clinical practice, our study emphasizes the critical need for awareness and mitigation of potential biases.

References

Ehsan Adeli, Qingyu Zhao, Adolf Pfefferbaum, Edith V Sullivan, Li Fei-Fei, Juan Carlos Niebles, and Kilian M Pohl. Representation learning with statistical independence to mitigate bias. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2513–2523, 2021.

Alexander Brown, Nenad Tomasev, Jan Freyberg, Yuan Liu, Alan Karthikesalingam, and Jessica Schrouff. Detecting and preventing shortcut learning for fair medical ai using shortcut testing (short). *arXiv preprint arXiv:2207.10384*, 2022.

M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.

Valeriia Cherepanova, Vedant Nanda, Micah Goldblum, John P Dickerson, and Tom Goldstein. Technical challenges for training fair neural networks. *arXiv preprint arXiv:2102.06764*, 2021.

Wenlong Deng, Yuan Zhong, Qi Dou, and Xiaoxiao Li. On fairness of medical image classification with multiple sensitive attributes via learning orthogonal representations. In *International Conference on Information Processing in Medical Imaging*, pages 158–169. Springer, 2023.

Siyi Du, Ben Hers, Nourhan Bayasi, Ghassan Hamarneh, and Rafeef Garbi. Fairdisco: Fairer ai in dermatology via disentanglement contrastive learning. In *European Conference on Computer Vision*, pages 185–202. Springer, 2022.

Nicholas Heller, Niranjan Sathianathen, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, et al. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445*, 2019.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Stefanos Ioannou, Hana Chockler, Alexander Hammers, Andrew P King, and Alzheimer's Disease Neuroimaging Initiative. A study of demographic bias in cnn-based brain mr segmentation. In *International Workshop on Machine Learning in Clinical Neuroimaging*, pages 13–22. Springer, 2022.

Fabian Isensee and Klaus H Maier-Hein. An attempt at beating the 3d u-net. *arXiv preprint arXiv:1908.02182*, 2019.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Neil Joshi and Phil Burlina. Ai fairness via domain adaptation. *arXiv preprint arXiv:2104.01109*, 2021.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33 (1):1–33, 2012.

Stephan M Korn and Shahrokh F Shariat. Gender differences in bladder and kidney cancers. In *Principles of Gender-Specific Medicine*, pages 603–610. Elsevier, 2017.

Alexander Kutikov and Robert G Uzzo. The renal nephrometry score: a comprehensive standardized system for quantitating renal tumor size, location and depth. *The Journal of urology*, 182(3):844–853, 2009.

Christina B Lund and Bas HM van der Velden. Leveraging clinical characteristics for improved deep learning-based kidney tumor segmentation on ct. In *International Challenge on Kidney and Kidney Tumor Segmentation*, pages 129–136. Springer, 2021.

Jun Ma. Cutting-edge 3d medical image segmentation methods in 2020: Are happy families all alike? *arXiv preprint arXiv:2101.00232*, 2021.

Vongani H Maluleke, Neerja Thakkar, Tim Brooks, Ethan Weber, Trevor Darrell, Alexei A Efros, Angjoo Kanazawa, and Devin Guillory. Studying bias in gans through the lens of race. In *European Conference on Computer Vision*, pages 344–360. Springer, 2022.

Ricards Marcinkevics, Ece Ozkan, and Julia E Vogt. Debiasing deep chest x-ray classifiers using intra- and post-processing methods. In *Machine Learning for Healthcare Conference*, pages 504–536. PMLR, 2022.

Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

Arezou Pakzad, Kumar Abhishek, and Ghassan Hamarneh. Circle: Color invariant representation learning for unbiased classification of skin lesions. In *European Conference on Computer Vision*, pages 203–219. Springer, 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Eike Petersen, Aasa Feragen, Maria Luise da Costa Zemsch, Anders Henriksen, Oskar Eiler Wiese Christensen, Melanie Ganz, and Alzheimer's Disease Neuroimaging Initiative. Feature robustness and sex differences in medical imaging: A case study in mri-based alzheimer's disease detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 88–98. Springer, 2022.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *Advances in neural information processing systems*, 30, 2017.

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: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 413–423. Springer, 2021.

Esther Puyol-Antón, Bram Ruijsink, Jorge Mariscal Harana, Stefan K Piechnik, Stefan Neubauer, Steffen E Petersen, Reza Razavi, Phil Chowienczyk, and Andrew P King. Fairness in cardiac magnetic resonance imaging: assessing sex and racial bias in deep learning-based segmentation. *Frontiers in cardiovascular medicine*, 9: 859310, 2022.

Edward N. Rampersaud, Tobias Klatte, Geoffrey Bass, Jean-Jacques Patard, Karim Bensaleh, Malte

Böhm, Ernst P. Allhoff, Luca Cindolo, Alexandre De La Taille, Arnaud Mejean, Michel Soulie, Laurent Bellec, Jean Christophe Bernhard, Christian Pfister, Marc Colombel, Arie S. Belldegrun, Allan J. Pantuck, and Daniel George. The effect of gender and age on kidney cancer survival: Younger age is an independent prognostic factor in women with renal cell carcinoma. *Urologic Oncology: Seminars and Original Investigations*, 32(1):30.e9–30.e13, 2014. ISSN 1078-1439. doi: https://doi.org/10.1016/j.urolonc.2012.10.012.

Fernanda Ribeiro, Valentina Shumovskaia, Thomas Davies, and Ira Ktena. How fair is your graph? exploring fairness concerns in neuroimaging studies. In *Machine Learning for Healthcare Conference*, pages 459–478. PMLR, 2022.

Zohaib Salahuddin, Yi Chen, Xian Zhong, Henry C Woodruff, Nastaran Mohammadian Rad, Shruti Atul Mali, and Philippe Lambin. From head and neck tumour and lymph node segmentation to survival prediction on pet/ct: An end-to-end framework featuring uncertainty, fairness, and multi-region multi-modal radiomics. *Cancers*, 15(7):1932, 2023.

Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. In *BIOCOMPUTING 2021: proceedings of the Pacific symposium*, pages 232–243. World Scientific, 2020.

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew BA McDermott, Irene Y Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in underserved patient populations. *Nature medicine*, 27 (12):2176–2182, 2021.

Wen-Wei Sung, Po-Yun Ko, Wen-Jung Chen, Shao-Chuan Wang, and Sung-Lang Chen. Trends in the kidney cancer mortality-to-incidence ratios according to health care expenditures of 56 countries. *Scientific Reports*, 11(1):1479, 2021.

Vinith M Suriyakumar, Marzyeh Ghassemi, and Berk Ustun. When personalization harms: Reconsidering the use of group attributes in prediction. *arXiv preprint arXiv:2206.02058*, 2022.

Ahmed Taha, Pechin Lo, Junning Li, and Tao Zhao. Kid-net: convolution networks for kidney vessels segmentation from ct-volumes. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11*, pages 463–471. Springer, 2018.

Mei Wang and Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9322–9331, 2020.

Yixin Wang, Yao Zhang, Jiang Tian, Cheng Zhong, Zhongchao Shi, Yang Zhang, and Zhiqiang He. Double-uncertainty weighted method for semi-supervised learning. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pages 542–551. Springer, 2020.

Yawen Wu, Dewen Zeng, Xiaowei Xu, Yiyu Shi, and Jingtong Hu. Fairprune: Achieving fairness through pruning for dermatological disease diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 743–753. Springer, 2022.

Tian Xu, Jennifer White, Sinan Kalkan, and Hatice Gunes. Investigating bias and fairness in facial expression recognition. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 506–523. Springer, 2020.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

Yuyin Zhou, Shih-Cheng Huang, Jason Alan Fries, Alaa Youssef, Timothy J Amrhein, Marcello Chang, Imon Banerjee, Daniel Rubin, Lei Xing, Nigam Shah, et al. Radfusion: Benchmarking performance and fairness for multimodal pulmonary embolism detection from ct and ehr. *arXiv preprint arXiv:2111.11665*, 2021.

Appendix A. Bias Mitigation: Detailed Results

Tables 5, 6 and 7 report results for Fair Meta-learning, Stratified Batch Sampling and RESM Algorithm approaches respectively across protected attributes of Sex and Age.

Table 5: Results for Fair Meta-learning with classification branches for sex and age

Attributes	Group	DSC Kidney (%)	DSC Tumor (%)	Mean DSC (%)	SD	SER
Gender	all	94.4	78.3	86.3	-	-
	male	95.0	74.7	84.8	1.55	1.26
	female	93.9	81.9	87.9		
Age Group	all	94.6	79.4	87.0	-	-
	0 - 50	95.1	80.9	88.0	3.24	2.02
	50 - 60	95.0	78.0	86.5		
	60 - 70	94.0	70.4	82.2		
	> 70	94.2	88.2	91.2		

Table 6: Fairness on Stratified Batching (equal number of samples in each batch)

Attributes	Group	DSC Kidney (%)	DSC Tumor (%)	Mean DSC (%)	SD	SER
Gender	all	94.7	76.6	85.6	-	-
	male	94.9	74.1	84.5	1.2	1.18
	female	94.4	79.4	86.9		
Age Group	all	94.2	75.1	84.6	-	-
	0 - 50	94.8	80.1	87.4	3.33	1.8
	50 - 60	94.6	72.8	83.7		
	60 - 70	94.0	67.3	80.6		
	> 70	93.3	85.2	89.2		

Table 7: Results for RESM Algorithm Across Sex and Age

Attributes	Group	DSC Kidney (%)	DSC Tumor (%)	Mean DSC (%)	SD	SER
Gender (63 samples)	all	94.3	76.3	85.3	-	-
	male	94.9	74.3	84.6	0.75	1.11
	female	93.6	78.6	86.1		
Age Group (31 samples)	all	94.5	76.6	85.6	-	-
	0 - 50	94.4	78.3	86.3	2.52	1.69
	50 - 60	94.8	76.4	85.6		
	60 - 70	94.9	70.2	82.6		
	> 70	93.8	85.6	89.7		

Appendix B. Effect of Modifications to Architectural Design

Table 8 report results for various variants of U-Net architecture across protected attributes of Sex and Age.

Table 8: Detailed Fairness Evaluation for Sex and Age across Different Network Architectures

Architecture	Characteristics	Group	DSC Kidney	DSC Tumor	Mean DSC	SD	SER
UNet	Total	-	94.6	73.0	83.8	-	-
		all	94.6	73.0	83.8		
	Gender	male	94.7	71.3	83.0	0.80	1.10
		female	94.3	74.9	84.6		
		all	94.6	73.0	83.8		
		0 - 50	94.7	73.9	84.3		
	Age	50 - 60	94.6	72.7	83.7	0.88	1.16
		60 - 70	95.5	70.0	82.7		
		> 70	92.7	77.5	85.1		
VNet	Total	-	94.6	73.6	84.1	-	-
		all	-	-	-		
	Gender	male	94.5	71.2	82.9	1.25	1.17
		female	94.7	76.1	85.4		
		all	-	-	-		
		0 - 50	94.4	72.5	83.4		
	Age	50 - 60	94.5	70.7	82.6	2.73	1.61
		60 - 70	95.3	69.3	82.3		
		> 70	93.7	84.2	89.0		
Attention Unet	Total	-	94.8	75.6	85.2	-	-
		all	-	-	85.2		
	Gender	male	95.0	76.2	85.6	0.40	1.06
		female	94.5	75.1	84.8		
		all	-	-	85.6		
		0 - 50	94.8	75.0	84.9		
	Age	50 - 60	95.1	79.1	87.1	1.66	1.31
		60 - 70	95.5	70.9	83.2		
		> 70	93.4	80.9	87.2		

Appendix C. Effect of Different Loss Parameters on Fair Meta-learning for Bias Mitigation

Tables 9 and 10 show results for different loss parameters in Equation 1 for Fair Meta-learning mitigation approach across protected attributes of sex and age. To achieve high-quality segmentation along with effective bias mitigation, we selected the parameters $\alpha = 1.0$ and $\beta = 2.0$ for the sex attribute, and $\alpha = 1.0$ and $\beta = 1.5$ for the age attribute.

Table 9: Comparison of Loss Parameters from Equation 1: Fair Meta-learning Approach for Sex Attribute

Loss Parameters		**DSC**			**Fairness**	
α	β	**Kidney ↑ (%)**	**Tumor ↑(%)**	**Mean ↑(%)**	**SD ↓**	**SER ↓**
1.0	2.0	94.4	78.3	**86.3**	**1.55**	**1.26**
1.5	1.0	94.3	77.2	85.8	1.15	1.18
1.0	1.0	94.0	76.6	85.3	1.40	1.21
1.0	1.5	94.2	77.4	85.8	1.60	1.25
2.0	1.0	94.1	78.4	86.2	1.65	1.27

Table 10: Comparison of Loss Parameters from Equation 1: Fair Meta-learning Approach for Age Attribute

Loss Parameters		**DSC**			**Fairness**	
α	β	**Kidney ↑ (%)**	**Tumor ↑(%)**	**Mean ↑(%)**	**SD ↓**	**SER ↓**
1.0	2.0	94.6	78.6	86.6	3.52	2.07
1.5	1.0	94.6	79.2	86.9	3.70	2.20
1.0	1.0	94.4	77.4	85.9	3.35	2.00
1.0	1.5	94.6	79.4	**87.0**	**3.24**	**2.02**
2.0	1.0	94.4	78.5	86.5	4.12	2.47