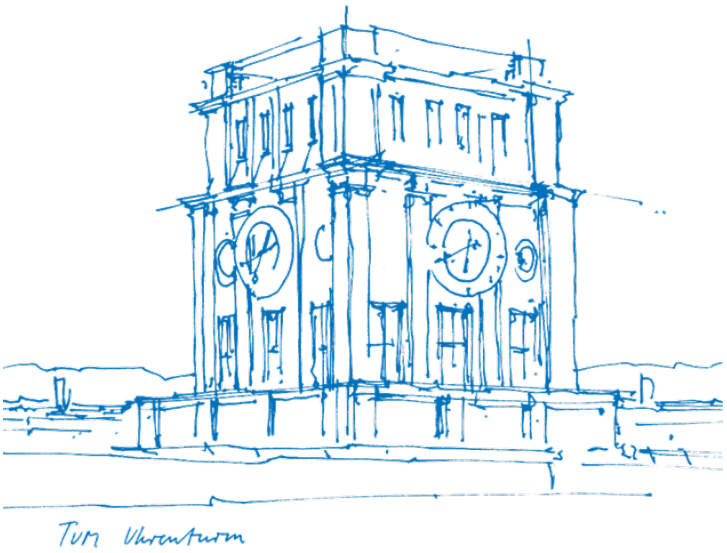
Chair for Computer Aided Medical Procedures & Augmented Reality
TUM School of Computation, Information and Technology
Technical University of Munich

Evaluating Fairness for Semi-Supervised Cardiac Magnetic Resonance Image Segmentation

Bewertung der Fairness für semi-supervidierte Segmentierung von kardialen Magnetresonanztomographie-Bildern

Muhammad Khattab



Evaluating Fairness for Semi-Supervised Cardiac Magnetic Resonance Image Segmentation

Bewertung der Fairness für semi-supervidierte Segmentierung von kardialen Magnetresonanztomographie-Bildern

Muhammad Khattab

Chair for Computer Aided Medical Procedures & Augmented Reality
TUM School of Computation, Information and Technology
Technical University of Munich

Evaluating Fairness for Semi-Supervised Cardiac Magnetic Resonance Image Segmentation

Bewertung der Fairness für semi-supervidierte Segmentierung von kardialen Magnetresonanztomographie-Bildern

Muhammad Khattab

Thesis for the attainment of the academic degree

Biomedical Computing Master of Science (M.Sc.)

at the TUM School of Computation, Information and Technology of the Technical University of Munich.

Examiner:
Prof. Dr. Daniel Rückert

Supervisor:
PhD Candidate Haifa Beji

Submitted:
Munich, 13.06.2024

I hereby declare that this thesis is entirely the result of my own work except where otherwise indicated. I have only used the resources given in the list of references.

Munich, 13.06.2024 Muhammad Khattab

Abstract

Leveraging AI models in healthcare offers significant benefits to healthcare providers, as these models provide various tools to save time and resources across different domains and tasks. However, AI models can have harmful effects if not properly audited for bias and fairness. Failure to mitigate bias could result in disparate performance among protected groups, subgroups, or intersectional groups. In this study, we train a semi-supervised learning model to utilize unlabeled data, particularly in low-data environments. We assess the effectiveness of semi-supervised learning and representation learning as potential strategies for bias mitigation. Additionally, we design various experiments to examine the impact of multiple factors on fairness in the UK Biobank dataset. Notably, we do not explicitly include protected attributes in the training process. Furthermore, our results indicate that semi-supervised models alone are insufficient to mitigate bias, despite their impressive performance in image segmentation tasks. We also find that balancing the training set does not necessarily eliminate disparities between groups. Moreover, we demonstrate that overall disparities may be obscured by aggregating results across different groups. The model may favor one group in some instances and the opposite group in others, resulting in an overall balanced outcome. Lastly, we emphasize the significance of test sets in bias analysis.

Contents

1 Introduction

Artificial intelligence (AI) models have provided unprecedented advancements in various computer vision tasks across many domains, especially in tasks like semantic segmentation, image classification, and object detection [16, 70, 26, 47, 19, 54]. Furthermore, recent advancements in AI models extend to cover more complex challenges in multi-modal tasks like image captioning [83, 46, 76] and text-to-image generation [82, 17]. In healthcare, the workload is always increasing, and the workforce is limited, which poses a huge challenge. An example showcasing this idea is Rwanda, where in 2015, there were only 11 radiologists available for a population of 12 million [62]. The challenges posed to healthcare providers provide enough motivation for transferring these breakthroughs to healthcare to help extend the tools available for healthcare providers so that they can automate certain tasks and save time and resources. There are plenty of use cases that can benefit from AI models, such as disease detection, classification, and image segmentation [88, 40, 65, 42]. Moreover, image captioning models are extended and used in radiology report generation, where the model is given an input X-ray image to produce a report of findings evaluating the presence of different pathologies [18, 86, 38, 78]. Furthermore, text-to-image generative models could be used to augment the training set to overcome the scarcity of images for certain conditions [42]. With these advancements, AI models can not only surpass human performance but also save a lot of time and resources.

1.1 Motivation

In this work, we are concerned with providing tools that assist in the cardiac magnetic resonance image segmentation task, which can help healthcare save time and resources. Cardiac magnetic resonance is considered the gold standard for analyzing cardiac function by assessing ejection fractions (EF) for the left ventricle (LV) and right ventricle (RV), stroke volumes (SV), left ventricle mass, and myocardium thickness. Measuring these values requires accurate segmentation or delineation of the short-axis slices for both end-diastolic (ED) and end-systolic (ES) phases. However, this analysis and segmentation task is time-consuming and can only be done by experts, which is why providing helpful tools for this task is of utmost importance. Davies et al. [24] showed that it takes an expert around 13 minutes to provide left ventricle (LV) analysis, whereas their pipeline took only around 20 seconds for the same task.

Another important factor affecting AI models in medicine is that, in various scenarios, the models operate in low-data regimes due to the scarcity of data or certain classes, expensive annotations, or privacy constraints. The recent huge leap in semi-supervised learning makes it possible to address this issue using recent data-efficient state-of-the-art models.

Bias is a major concern in AI models, especially in healthcare, where the models are used to make decisions that affect human lives. There are many definitions for bias in the literature; however, in this work, we will refer to bias as the disparate performance between protected groups, subgroups, or intersectional groups. Numerous papers have shown disparate performance between protected groups or subgroups across various imaging modalities [58, 57, 65]. An interesting study by [6] demonstrated that the model could infer race from chest X-ray images without being trained for that task and without even being given the labels.

Not all discrimination is morally unacceptable, as Hellman [37] discussed, such as refusing to continue to employ pilots over the age of 62 or issuing driver's licenses to people under 16. However, in other cases, discrimination is harmful, such as refusing to hire people from a certain ethnic group. In medicine, certain diseases have different thresholds for different groups. For example, the threshold for HbA_{1c}, which is used to diagnose diabetes, is higher for the Black ethnic group than for the White ethnic group [29].

Moreover, some diseases are rare in younger populations and more common in older populations, like age-related macular degeneration (AMD) [28]. Due to the complexity of each medical task, we emphasize the importance of bias analysis and considering the nature and conditions surrounding each task.

We consider three aspects of approaching the bias problem: identifying bias, understanding the root causes of bias, and mitigating bias. Identifying bias is the first step of the process and is crucial for medical applications. Bias and fairness analysis for AI models not only ensure fair results for all subgroups but also help avoid causing harm to under-represented groups. For example, if a model systematically under-diagnosed a certain group, it would cause significant harm by resulting in those patients receiving less or no care. In rare cases, the analysis could also provide an answer or insight into the causes of the bias; however, this is a complex challenge. The complexity arises because, in practice, there are many sources of bias and confounding variables for given tasks, making it difficult to separate each of them. Moreover, bias can sometimes be mitigated without understanding its causes. Mitigating bias is also not straightforward, as algorithmic approaches to bias mitigation do not always work. The impact of bias and fairness in medical AI models highlights the importance of bias analysis and mitigation to avoid or at least minimize harm inflicted on under-represented groups.

1.2 Objective

In order to provide a fair model for the cardiac magnetic resonance image segmentation task, we propose the use of a semi-supervised learning model with the intuition that representation learning will help with generalization and capturing more features of the under-represented groups, especially in low data regimes. Namely, we would like to test how representation learning models perform for protected groups, subgroups, and intersectional groups. Moreover, we want to touch upon the modifications they require to improve the overall performance, avoid bias, and harm. We try to investigate other factors that can affect the model other than under-representation. To the best of our knowledge, there are not any works investigating the fairness of semi-supervised models for cardiac magnetic resonance image segmentation. The closest work to ours is [44], which investigated supervised models based on vision transformers [26] and convolutional neural networks (CNNs) [55]. In our results, we find disparities between the sex groups, ethnic groups, and intersectional groups in some cases and no disparities in other cases. Moreover, we find interesting results where it would seem that there are no disparities between the groups, but with further investigation, we see that the disparities between the intersectional groups alternate, which causes the overall performance to look balanced. Moreover, we provide an analysis of the effects of training set balancing and test set sampling. Our contribution could be summarized as follows:

- Train the semi-supervised model UniMatch [80] on the UK Biobank dataset [56] and evaluate performance for a baseline model which employs fairness through unawareness.
- Devise different experiments to investigate the effect of different factors on the model's performance.
- Perform bias analysis on protected groups and intersectional groups and investigate the factors that affect the model and analysis itself.

2 Related Work

2.1 Bias Analysis in Medical Imaging

There are numerous papers exploring fairness and bias in different medical applications covering various imaging modalities. These works focus on analyzing performance between different protected groups and subgroups and attempt to explain the reasons behind the discrepancies. In the cardiac magnetic resonance image segmentation task, there have been a few papers which perform bias analysis.

Puyol et al.[57] focused only on healthy patients to minimize the confounding factor related to anatomical changes from diseases. Their results still suggested that the model has racial bias but no sex bias. They tried explaining the disparity in performance using other factors like risk factors and patient characteristics using multivariate linear regression. Their results showed that even after adjusting for possible confounders such as cardiovascular risk factors, the bias still persists when the data contains unhealthy patients with cardiomyopathy. This in turn points to the effect of data balance on the model performance which they confirmed in their previous paper[58].

Lee et al. [45] varied the proportions of protected groups during training to analyze performance and used two test sets balanced for sex and ethnicity. The results for sex imbalance showed no significant disparity except for intersectional groups like black vs. white females. They also state that accuracy parity does not necessarily occur at a 50%-50% split for the protected groups which they interpret as the segmentation task could be harder for certain groups.

In another paper from Lee et al. [44], they investigated the role of the model architecture on bias where they performed a number of experiments while varying the proportions of protected groups for different supervised segmentation models. They reported the results using various metrics and showed that data imbalance affects performance for under-represented sex and ethnic groups and that the disparities change according to the model.

For chest X-ray imaging modality, a number of papers have reported disparities between protected groups. The work from Seyyed-Kalantari et al.[65] investigated the effects of age, sex, ethnicity, and socioeconomic status where they used health insurance type as a proxy to obtain this label. They defined under-diagnosis using the false-positive rate (FPR) and showed that the model had a higher under-diagnosis rate for certain protected and intersectional groups, especially patients who belong to two under-represented groups. They also explained that the data collection might affect the model performance; for example, the NIH dataset[77] uses selective admission which makes certain labels, such as "no finding," scarce, which in turn affects the FPR.

Banerjee et al. [6] conducted various experiments to assess the capability of AI models in detecting race. Additionally, they explored anatomical and phenotypic confounders such as BMI, breast density, bone density, the location of specific anatomical features, and environmental factors. They trained models directly on different chest X-ray datasets for the task of race detection, as well as on other imaging modalities for different body parts. Furthermore, they employed models trained for other tasks and evaluated their performance in race identification. The results indicated that no reliable explanation could be derived from the proposed confounders, aligning with the medical consensus that there are no reliable medical biomarkers correlated with identifying race. This underscores the understanding that race is primarily a social and political construct. Moreover, to evaluate the internal decision-making strategy of the model, they attempted to examine images in the frequency domain after applying various filters. Additionally, they tested the impact of image quality on performance and investigated whether racial information was localized in specific parts of the image using the grad-cam methodology [64]. Furthermore, they examined the effect of removing a patch from the image or utilizing only one patch during training. Ultimately, they

concluded that models could learn to classify race irrespective of modality, dataset, task, or objective, and thus, the bias could not be explained using the proposed confounders.

Glocker et al. [31] endeavored to decode the information utilized by AI models to make decisions, aiming to comprehend the rationale behind their behavior. Additionally, they highlighted the limitations of previous studies, which relied on random sampling of train and test sets. Consequently, they employed strategic resampling with replacement to obtain a balanced test set. Moreover, they utilized a combination of test-set resampling, multi-task learning, and unsupervised exploration of feature representations. These approaches provided a framework to detect and evaluate the relationship between disease detection and the encodings of protected patient attributes. The authors reported that certain disparities disappeared when the test sets were resampled. They also confirmed other disparities between groups, but they lacked clear evidence that race was being used by the model. Identifying the cause of disparate performance remains challenging.

2.2 Bias Mitigation in Medical Imaging

Several papers have investigated strategies to improve the fairness of models by incorporating certain techniques. For example, Obermeyer et al. [53] shed light on certain aspects of prediction algorithms that rely on expected income per patient, which disproportionately affects poor populations. Hence, they suggested combining health prediction variables with cost prediction, which improved performance.

Puyol et al. [58] conducted bias analysis and found a significant disparity in performance between ethnic groups but not between sex groups. Hence, they proposed three bias mitigation strategies and reported their results. First, stratified sampling ensures each batch during training has an equal number of protected group members. Second, fair meta-learning for segmentation involves adding a classifier for protected attributes similar to multi-task learning. Third, protected group models entail training a model for each protected group. Their results show that all three approaches improve the fairness of the model, with the last approach yielding the best results; however, it requires the protected attribute during inference time.

Zhang et al. [84] expanded on the work of Seyyed-Kalantari et al. [65] and showed that empirical Risk Minimization (ERM) models yield statistically significant performance gaps across many metrics and protected attributes. They also demonstrated that no method outperforms simple data balancing. They benchmarked three baseline methods based on ERM, four methods that do not explicitly seek to improve protected group performance, and lastly three methods that aim to improve the performance of the worst-case group. The results indicate that the best method is simple balanced-ERM, suggesting that balancing classes is the most straightforward factor to adjust in order to achieve fairness and that no other method outperforms it on the worst-case group for any evaluation metric. Moreover, they showed that group fairness, when applied to chest X-rays, worsens the performance for all groups. They also stated that the CheXpert labeler provides poor label predictions and was a source of bias. Furthermore, they made an important remark that improving worst-case group performance is not good enough and that equalized odds methods worsen the performance of all groups. Additionally, they noted that de-biasing the model using equalized odds results in a change of the decision threshold for groups. The threshold could also be adjusted to improve the effectiveness of the model, posing a risk of not aligning with clinical practice guidelines. Another point they mentioned is that some disparities in performance are justified when the task is harder for certain groups, such as older people due to comorbidity.

Zong et al. [90] proposed a comprehensive benchmark for fairness, employing various algorithms, datasets, protected attributes, and model selection strategies. They evaluated the models in both in-distribution and out-of-distribution settings using various metrics. Their findings revealed performance gaps between protected groups across different Empirical Risk Minimization (ERM) models for various modalities and demonstrated that model selection strategies affect the performance of the worst-case group. They concluded that no method outperforms ERM in either in-distribution or out-of-distribution settings, which aligns with the findings of Zhang et al. [84].

Ktena et al. [42] utilized a generative AI model to augment the training set with generated synthetic data in three medical tasks, conditioning the model on either diagnostic labels or protected attributes. During

sampling, they achieved an equal distribution for protected groups while preserving disease prevalence. Their results demonstrated that the mentioned pipeline improves robustness both in-distribution and out-of-distribution, as well as the fairness of the model. However, the generative model could exhibit biases from the training data, resulting in outputs that are biased towards a specific range of the distribution, incorrect for the given conditions, or unable to generate an image at all.

Azizi et al. [2] utilized multi-view images in a self-supervised learning setting to create natural augmentation pairs, aiming to narrow the gap between in-distribution and out-of-distribution performance. They combined supervised pre-training with self-supervised learning to avoid domain-specific customizations. Their strategy is data-efficient and requires 6-33% less data for retraining compared to other supervised models. Furthermore, they evaluated the model performance in three settings: in-distribution with fine-tuning, zero-shot out-of-distribution, and data-efficient out-of-distribution with fine-tuning. Their results demonstrated that self-supervision improves generalization, especially in the data-efficient setting, and requires fewer labels to reach a level where it could be used in clinical applications. Moreover, it provides better in-distribution performance that exceeds supervised baselines on five of six medical tasks.

2.3 Cardiac Magnetic Resonance Image Segmentation

Various cardiac magnetic resonance (CMR) imaging applications utilize AI models. There are works that automate heart localization and image acquisition [30, 36], as well as optimize frequency adjustments for CMR [32]. Moreover, numerous models address myocardial tissue characterization and texture analysis [74, 72]. Additionally, AI models can utilize CMR images for meta-analysis, used in prognosis and characterizing differences between healthy and unhealthy participants [1, 87]. However, in this work, we focus solely on the task of cardiac magnetic resonance image segmentation, which is well-defined in the Automatic Cardiac Diagnosis Challenge (ACDC) [9]. This challenge served as a primary driver for many works to provide state-of-the-art models for CMR image segmentation. Furthermore, it provided a dataset with short-axis cine-MRI images for 150 patients evenly balanced into 5 classes spanning normal, systolic heart failure with infarction, cardiomyopathy, hypertrophic cardiomyopathy, and abnormal right ventricle. Most of the methods addressing this challenge were fully-supervised; however, some recent works have attempted to incorporate semi-supervised learning to address the challenge, as we will discuss in the following sections.

2.3.1 Supervised Learning

The first class of papers that addresses CMR image segmentation employs fully supervised learning approaches. A recent work by Tragakis et al. [73] introduced the Fully Convolutional Transformer (FCT), which combines the convolutional power of UNet [61] with the long-term dependencies of transformers [75]. Additionally, they attempted to alleviate the linear nature of attention projection and feature processing. The model achieves new state-of-the-art results on ACDC [9] while being three to five times smaller than other state-of-the-art models.

nnFormer [88] is a 3D transformer that combines convolution and self-attention in an interleaved manner. It utilizes skip attention, local and global volume-based multi-head self-attention. This approach is similar to TransUNet [15], which combines convolution and transformers to leverage detailed high-resolution spatial features and global context encoded by the latter.

Swin-Unet [13] exclusively employs transformers in a U-shaped encoder-decoder architecture similar to U-net [61]. The basic unit of the model consists of a Swin Transformer block [48]. Furthermore, Agile-Former [59] builds upon Swin-Unet and proposes spatially dynamic self-attention [75], as well as novel multi-scale deformable positional encoding.

nn-Unet [40] automatically configures preprocessing steps, architecture choices, and post-processing steps for new segmentation tasks in the medical domain. Its success is attributed to its systemic automatic configuration setup, which avoids relying on empirical approaches.

2.3.2 Semi-supervised Learning

The other class of papers addressing CMR image segmentation employs semi-supervised learning approaches. Semi-supervised learning models, in general, aim to leverage unlabeled data to improve the performance of the model. In the context of CMR image segmentation, there are few models that use semi-supervised learning and report their results on ACDC [9].

Recently, UniMatch [80] introduced two techniques to leverage unlabeled images on top of FixMatch [66], which we will discuss in more detail in the background chapter since we use them in this work. Bai et al. [5] proposed using a Mean Teacher architecture [71] to learn common semantics from labeled and unlabeled images by employing bidirectional copy-pasting [83, 85]. Basak and Yin [7] introduced a patch-based semi-supervised contrastive learning approach without using a pretext task. Additionally, they proposed a novel contrastive loss. CauSSL [52] proposed the use of a causal diagram which also contains intermediate representations.

Apart from the models that report their results on CMR image segmentation, there are plenty of semi-supervised models in natural image segmentation. Here, we mention some of them that are related to UniMatch [80]. MixMatch [11] introduced a pipeline where it generates K augmentations and averages them to predict the label for each unlabeled image. Later, ReMixMatch [10] introduced two improvements over MixMatch: Distribution Alignment and Augmentation Anchoring, which we will discuss in more detail in Section 3.4.1. Unsupervised Data Augmentation (UDA) [79] is similar to ReMixMatch [10] in that neither of them use pseudo-labels, and they enforce consistency between the artificial labels from weak perturbations and strong perturbations. On the other hand, FixMatch [66] combines consistency regularization and pseudo-labeling while maintaining a simple pipeline. Furthermore, Cross teaching between CNN and Transformers [49] combines architecture-level and output-level perturbation. It uses cross-teaching loss, which is a bidirectional loss function between CNN and Transformer streams. Additionally, the transformer only complements the training rather than the final prediction.

3 Background

The generalization of AI models is a crucial aspect in the development of AI systems. In various works, models are evaluated in an out-of-distribution setting to assess their generalization. In-distribution evaluation for images involves training and evaluating the model on the same combination of datasets, population distributions, sensitive attributes, and protocols used for image acquisition. On the other hand, out-of-distribution evaluation would have one of these parameters different during evaluation compared to training, making the input new to the model.

The idea of in-distribution and out-of-distribution evaluation provides insight into the model's generalization capabilities, which could in some way indicate how biased or unbiased the model is. However, this evaluation alone is not the only way of defining fairness. There are various fairness definitions and types in the literature, which we will discuss next. Furthermore, there are numerous sources of bias in medical imaging, and understanding these sources is crucial for the development of fair models.

Finally, we explore self-supervised models, which are a form of representation learning and can have a significant impact on moving towards fairer models. This is why we discuss key concepts related to the topic.

3.1 Fairness Definitions

Following [51], we present some of the fairness definitions in the literature. Furthermore, the choice of which fairness definition to apply could depend on the task and which protected attributes are important to monitor. It is also important to note that it is impossible to satisfy all fairness definitions simultaneously [41].

Equalized Odds

"A predictor $\hat{Y}$ satisfies equalized odds with respect to protected attribute A and outcome Y, if $\hat{Y}$ and A are independent conditional on Y. $P(\hat{Y} = 1|A = 0, Y = y) = P(\hat{Y} = 1|A = 1, Y = y), y \in \{0, 1\}$" [35]. This means that the probability of a person being assigned correctly a positive outcome and the probability of a person being assigned incorrectly a positive outcome should be the same regardless of the person's protected group membership.

Equal Opportunity

"A binary predictor $\hat{Y}$ satisfies equal opportunity with respect to A and Y if $P(\hat{Y} = 1|A = 0, Y = 1) = P(\hat{Y} = 1|A = 1, Y = 1)$" [35]. This means that the probability of a person being assigned correctly to a positive outcome should be the same regardless of the person's protected group membership.

Demographic Parity

"A predictor $\hat{Y}$ satisfies demographic parity if $P(\hat{Y}|A = 0) = P(\hat{Y}|A = 1)$" [43]. This means that the likelihood of the outcome should be the same regardless of the person's protected group membership.

Fairness Through Awareness

"An algorithm is fair if it gives similar predictions to similar individuals" [43]. This means that individuals with similar values of the task-dependent similarity metric score used should get a similar outcome.

Fairness Through Unawareness

"An algorithm is fair as long as any protected attributes A are not explicitly used in the decision-making process" [43]. This means that there is no particular strategy to be followed as long as the protected attributes are not explicitly used.

Treatment Equality

"Treatment equality is achieved when the ratio of false negatives and false positives is the same for both protected group categories" [8]. This means that the probability of a person being incorrectly assigned a negative outcome and the probability of a person being incorrectly assigned a positive outcome should be the same regardless of the person's protected group membership.

Test Fairness

"A score $S = S(x)$ is test fair (well-calibrated) if it reflects the same likelihood of recidivism irrespective of the individual's group membership, R. That is, if for all values of $s, P(Y = 1|S = s, R = b) = P(Y = 1|S = s, R = w)$" [20].

Counterfactual Fairness

"Predictor $\hat{Y}$ is counterfactually fair if under any context $X = x, A = a, P(\hat{Y}_{A\leftarrow a}(U) = y|X = x, A = a) = P(\hat{Y}_{A\leftarrow a'}(U) = y|X = x, A = a)$, for all y and for any value a' attainable by A" [43]. This means a decision is fair if it is the same for an individual in both the actual world and the counterfactual world, where the individual belongs to a different protected group.

Fairness in Relational Domains

Farnadi et al. [27] used first-order logic to define fairness. They aimed to capture the complexity of discrimination and to define and capture relational structures like social and organizational structures.

Conditional Statistical Parity

"For a set of legitimate factors L, predictor $\hat{Y}$ satisfies conditional statistical parity if $P(\hat{Y}|L = 1, A = 0) = P(\hat{Y}|L = 1, A = 1)$" [21]. In this definition, they proposed unconstrained fairness with a single threshold instead of race-specific thresholds.

Max-Min Fairness

The strategy of max-min fairness [60] seeks to maximize the performance of the (max) of the worst-case group (min) and treats the model achieving this as the fairer model. This strategy is appropriate for medical diagnosis models. However, when using the minimax fairness definition it results in worsening the performance for other groups other than the worst-case group [89].

3.2 Fairness Types

Individual Fairness

Individual fairness is concerned with providing similar outcome to similar individuals [43]. Fairness definitions that fall into this type include fairness through unawareness, fairness through awareness, and counterfactual fairness [51].

Group Fairness

Group fairness aims to treat different groups equally [51] [43]. Additionally, it tries to achieve parity of predictive performance between protected groups which is useful for some medical applications with limited resources which are considered zero-sum game problems [90]. However, this strategy often worsens the performance of the best-case group or even worsens the performance of the worst-case group as well [89]. Fairness definitions that fall into this type include demographic parity, conditional statistical parity, equalized odds, equal opportunity, treatment equality, and test fairness.

Subgroup Fairness

Subgroup fairness combines group and individual fairness [51] and tries to treat subgroups equally.

3.3 Bias Sources in Medical Imaging

There are a numerous sources of bias in general, however, we only discuss here the ones that could potentially lead to bias in AI models in medical imaging.

Measurement Bias

Measurement bias emerges from the way certain features are measured [69]. Automatic labelers could output poor labels for certain groups and become even worse for subgroups. This results in poor results when these labels are used further for training classification models like in the case of CheXpert Labeler [39] as reported in [84].

Representation Bias

This is one of the most common bias sources in medical imaging where the population in the dataset have few samples for certain groups [69]. For example, over 90% of the the subjects in the UK Biobank have European ancestries [68].

Aggregation Bias

Aggregation bias happens when conclusions for individuals are based on the observations from the entire population especially for cases when different behaviors emerge according to the underlying subgroups in the population [69]. This is a potential source of bias in medicine, as seen in cases such as diabetes imaging [29, 69]. However, we are not aware of any studies reporting similar results for cardiac magnetic resonance imaging.

Sampling Bias

Sampling bias is similar to representation bias and it is a consequence of sampling from a group in a non-random fashion [51]. For example, NIH dataset [77] has such bias because the institute uses selective admission for the diseases it is interested in researching which in turn causes certain labels to be scarce especially the "No Finding" label.

Algorithmic Bias

Algorithmic bias emerges when the data has no bias, however, the outcome of the model is biased towards certain groups which could go back to any internal component of the algorithm or model like objective functions [4]. For example, The results reported by [58] showed that the accuracy parity depends on the model or the architecture used when the data is balanced for protected groups.

3.4 Semi-supervised Learning

In this work we will be using UniMatch [80] which builds on top of FixMatch [66]. We also try to incorporate a fairness component based on ReMixMatch [10] that is why we provide a brief background about the three of them next. Furthermore, we provide a brief overview of some key concepts that are used by these models, namely Consistency Regularization, Entropy Minimization, and Contrastive Learning.

Consistency Regularization

Consistency Regularization [3] is a important strategy in recent state-of-the-art semi-supervised learning which leverages unlabeled data to improve the model's performance. It aims at generating models robust to perturbation by assuming that the model should output similar predictions for perturbations of the same image. A loss function is used to measure consistency between the predictions of the original images and their corresponding perturbations.

Entropy Minimization

Entropy Minimization leverages unlabeled images and encourages the model to have low entropy for the output distribution of unlabeled data [33]. One approach that is related to entropy minimization is Pseudo-Label [50, 63] which predicts hard labels for unlabeled images above a predefined threshold and uses them as targets to train the model.

Contrastive Learning

Contrastive Learning is a strategy that learns representations through maximizing the agreement between positive pairs and minimizing the agreement between negative pairs [34].

3.4.1 ReMixMatch

ReMixMatch [10] introduced two improvements over MixMatch [11]. The first improvement is distribution alignment, which uses the labeled data distribution as a reference for the unlabeled data. This enforcement of the aggregate predictions on the unlabeled data was an improvement and an extension of an idea introduced a long time ago [12]. This approach introduces a notion of fairness into the model called distribution alignment, which encourages the model to use the class distribution rather than making equal predictions per class. In essence, distribution alignment maintains a running average of the unlabeled data prediction during training, which is then used to scale and normalize the prediction to construct a probability distribution. This probability distribution is used as a label guess for the unlabeled image.

Given that X represents the labeled data, U represents the unlabeled data, q_b represents the guess for the unlabeled image, $p(y|X)$ represents the class distribution of the labeled data, and $p_m(y|U)$ represents the model prediction for the unlabeled image, the distribution alignment is as follows:

$$Normalize(x)_i = \frac{x_i}{\sum_j x_j} \tag{3.1}$$

$$\tilde{q}_b = Normalize(q_b \times \frac{p(y|X)}{p_m(y|U)}) \tag{3.2}$$

They proposed augmentation anchoring as a second improvement, which uses the weakly augmented unlabeled images' predictions as the label for strong augmentations of the same image. The unlabeled image receives a guessed label after applying distribution alignment and sharpening. Then, K strong augmentations of the image are generated. The label for these strong augmentations is the one predicted from the weak augmentation (anchor). Moreover, they introduced Control Theory Augment (CTAugment) to produce high-performance augmentation strategies. The advantage of CTAugment is that it does not

require optimization on a supervised proxy task, unlike AutoAugment [22], which learns a policy through reinforcement learning.

Labeled augmentations are represented by X'. Similarly, U' represents unlabeled augmentations, which are obtained using MixUp [85]. $\hat{U}_1$ represents a single strong augmentation of an unlabeled image and its label guesses without MixUp. $Rotate$ is used to rotate images by $r \in \{0, 90, 180, 270\}$, and then the model predicts the rotation. λ represents the weight of each loss term. The labels and guesses are evaluated against the model predictions using cross-entropy loss H. The total loss is as follows:

$$\sum_{x,p \in X'} H(p, p_{model}(y|x;\theta)) + \lambda_u \sum_{u,q \in U'} H(q, p_{model}(y|u;\theta)) \\ +\lambda_{\hat{U}_1} \sum_{u,q \in \hat{U}_1} H(p, p_{model}(y|u;\theta)) + \lambda_r \sum_{u \in U'} H(r, p_{model}(r|Rotate(u,r);\theta)) \tag{3.3}$$

3.4.2 FixMatch

FixMatch [66] combines consistency regularization and pseudo-labeling and utilizes weak-to-strong perturbations. Furthermore, the loss consists of two terms. The first part of the loss function is the supervised loss term, which is the cross-entropy loss on weakly augmented labeled images. The second part is the unsupervised loss term, which is the cross-entropy against the model output for the strong augmentation of the unlabeled image. To calculate the unsupervised loss term, the class distribution is predicted using weak augmentation of the unlabeled image. Then the distribution is used to get an artificial label. This approach enforces consistency regularization through using weak perturbations for getting the pseudo-label then using the loss against the strong perturbation. FixMatch [66] uses weak augmentations which leverage the flip-and-shift strategy. Additionally, it uses strong perturbations using RandAugment [23] and CTAugment [10], followed by CutOut [25]. However, strong perturbations require domain-specific design and hyper-parameter fine-tuning [80], especially for the medical domain.

Given that $\hat{q}_b$ is the pseudo-label, τ is the threshold, μB represents the relative unlabeled batch size, $\hat{q}_b = \text{argmax}(q_b)$, A represents strong perturbations, and α represents weak perturbations, then the total loss is as follows:

$$\frac{1}{B} \sum_{b=1}^{B} H(p_b, p_m(y|\alpha(x_b))) + \frac{1}{\mu B} \sum_{b=1}^{\mu B} 1(max(q_b) \geq \tau) H(\hat{q}_b, p_m(y|A(u_b))) \tag{3.4}$$

Total loss after applying Distribution Alignment from ReMixMatch [10] as mentioned in Equation (3.2):

$$\frac{1}{B} \sum_{b=1}^{B} H(p_b, p_m(y|\alpha(x_b))) + \frac{1}{\mu B} \sum_{b=1}^{\mu B} 1(max(\tilde{q}_b) \geq \tau) H(\hat{\tilde{q}}_b, p_m(y|A(u_b))) \tag{3.5}$$

3.4.3 UniMatch

UniMatch [80] proposed to increase the perturbation space, noting that FixMatch [66] only employs image-level perturbations, thus constraining the model's effectiveness due to the limited perturbation space. Another limitation they mentioned is the reliance on manually designing strong perturbations by researchers according to the domain [80]. This design process requires domain knowledge and time-consuming fine-tuning. UniMatch [80] proposed a pipeline that utilizes feature-level perturbations, which complements the image-level perturbations. The weakly perturbed image x_w is used to acquire a feature-level perturbed image, and each level of perturbation is separated into an independent feed-forward stream. Furthermore, they propose dual-stream perturbation to leverage unlabeled data using the idea of multi-view learning similar to SwAV [14] and ReMixMatch [10]. The dual stream perturbation is yielded from the weakly perturbed image by the strong perturbation pool A_s, which produces two strong image-level perturbations x_1 and x_{s2} from x_w.

(a) FixMatch **(b)** UniMatch

Figure 3.1 Comparison of FixMatch and UniMatch architectures [80].

The model consists of an encoder g and a decoder h which obtains three signals: p_w for the image-level weakly perturbed image, p_{fp} for the feature-level perturbation of the weakly perturbed image, and p_s for the strong image-level perturbed image. They used separate streams for different perturbations to reduce the complexity of the learning process and allow the model to learn the signals directly. Additionally, the separation enforces consistency of the model at the image and feature level and helps the student achieve target consistency for each stream in a direct manner. The feature perturbation used in UniMatch is dropout [67] since they did not aim to introduce new strategies in their work. The unsupervised loss term is as follows:

$$\frac{1}{\mu B}\sum 1(max(p^w) \geq \tau)(\lambda H(p^w, p^{fp}) + \frac{\mu}{2}(H(p^w, p^{s1}) + H(p^w, p^{s2}))) \tag{3.6}$$

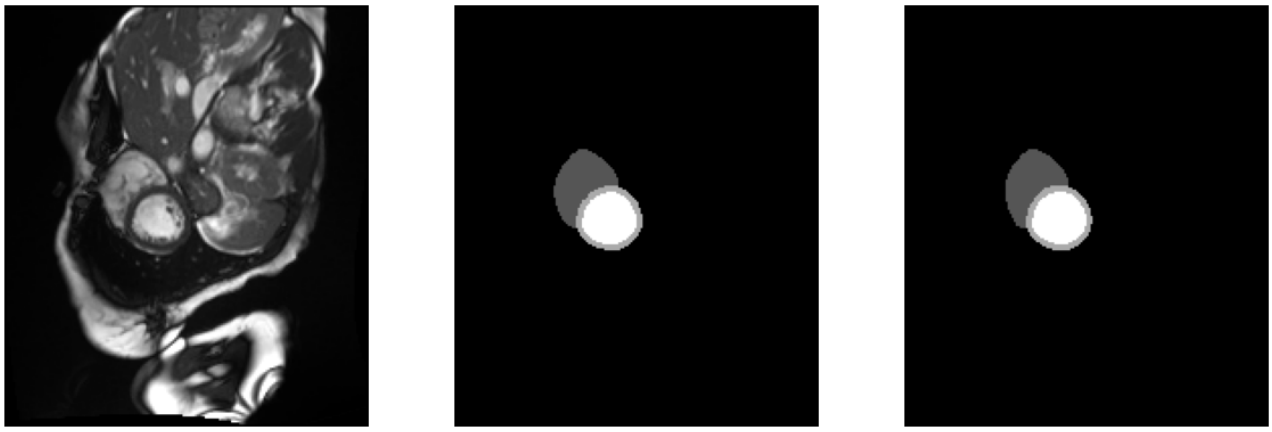
4 Methodology

In this work, we are interested in the task of cardiac magnetic resonance image segmentation where the model is given an input image and is required to output a mask representing the segmentation as shown in Figure 4.1. Furthermore, we would like to investigate bias between protected groups and uncover the presence of disparities. We aim to check the effect of balancing data, the effect of the test set, and the effect of increasing the data on the performance of the model and the disparities between groups.

To examine the effects of these conditions between different protected groups, we devise four experiments representing four models. Models 2, 3, and 4 represent clusters of models where we sample the training set for each model 10 times by changing the seed. We use each new set to repeat the training and obtain a new model for the same experiment to increase confidence in the results and eliminate the factor of the model receiving easier samples. This results in 10 models for each experiment and a total of 30 models, as well as a baseline model. For each experiment, the training set is balanced for zero or more protected attributes and then used for training, resulting in a different model for each experiment. Therefore, in the discussion, we can use the words "model" and "experiment" interchangeably.

We employ the semi-supervised model UniMatch [80] to assess the extent to which representation learning or semi-supervised learning can aid in bias mitigation. In the original work on UniMatch, the model was tested on ACDC [9]. Therefore, we adapt the code accordingly to evaluate it on the UK Biobank [56] dataset.

To streamline the experimentation process, we automate the repetition of experiments by adjusting the configuration and running the training and testing scripts. This automation helps to avoid the tedious manual process. We maintain the same configuration for all experiments and only alter the seed responsible for the random sampling of the training set. Additionally, we automate the analysis of statistical significance between each pair of groups or models, which we discuss in detail in the following sections.



(a) Input **(b)** Ground Truth **(c)** Prediction

Figure 4.1 Baseline Model - Input vs. Ground Truth vs. Prediction on Short Axis End Diastolic Slice

We employ the Dice Similarity Coefficient (DSC) to evaluate the segmentation performance. The DSC is defined as the intersection of the prediction and the ground truth divided by the sum of the prediction and the ground truth. To ensure numerical stability, we add $\epsilon = 1 \times 10^{-9}$ to both the numerator and the denominator. The DSC is calculated using the following formula:

$$DSC = \frac{2 \times |A \cap B|}{|A| + |B|} \tag{4.1}$$

4.1 UK Biobank Dataset

The UK Biobank (UKBB) [56] is a cohort study comprising half a million participants aged between 40 and 69, with a mean age of 56.5. Females constitute 54.4% of the data, and the distribution of ethnicities is as follows: 94.4% White, 1.9% Asian, and 1.6% Black. The dataset encompasses various tabular data and imaging modalities, with our focus specifically on CMR data. CMR imaging was conducted using a 1.5 Tesla scanner (MAGNETOM Aera, Syngo Platform VD13A, Siemens Healthcare, Erlangen, Germany). We selected the UKBB for this study due to its comprehensive subject information, facilitating bias analysis. Initially, we explored the ACDC [9] dataset to investigate the impact of BMI and age. However, we transitioned to the UKBB dataset for our primary analysis.

Each subject possesses a set of CMR images including long axis, short axis, and 4-chamber view images, yet our focus lies solely on the short axis images for this study. For each subject, there are two short axis images available: one for end-diastole (ED) and the other for end-systole (ES), stored in NIfTI format. Typically, each image contains approximately 10 slices, though the exact number may vary. These images are annotated by experts to delineate the right ventricle (RV), left ventricle (LV), and myocardium (MYO), with the annotations stored separately. The data has not undergone preprocessing or normalization.

Ethnicity in the UKBB is coded according to the codes presented in Table 4.1. For our investigation, we focus on the white, Asian, and Black ethnic groups. Individuals with codes 1, 1001, 1002, and 1003 are classified as belonging to the white ethnic group. Similarly, for Asians, we group codes 3, 3001, 3002, and 3003, excluding Chinese individuals to align with their coding scheme. The Black group includes subjects with codes 4, 4001, 4002, and 4003. We do not explore other ethnic groups to maintain study focus within the given time constraints. The full list of codes is provided in the appendix. Additionally, sex representation is binary, with code 0 for females and 1 for males.

Code	Meaning
1001	British
1002	Irish
1003	Any other white background
3001	Indian
3002	Pakistani
3003	Bangladeshi
3004	Any other Asian background
4001	Caribbean
4002	African
4003	Any other Black background

Table 4.1 Ethnicity Codes

4.2 Experiment 1 - Baseline

The goal of this experiment is to evaluate the overall performance of the model for the segmentation task and gain an understanding of its capabilities. This experiment represents the baseline model and is not balanced for any attribute. We sample 4000 subjects from the white ethnic group and concatenate them with available Asian and Black subjects, resulting in a total of 4819 subjects in the dataset used throughout this work.

From this dataset, we sample a train-validation set of 3794 subjects, with 3301 participants forming the training set and the remaining 493 forming the validation set. The complete test set comprises 1025 subjects, sampled from two test sets: one controlled for sex groups containing 962 subjects, and the other controlled for ethnicity containing 108 subjects. The overall split used is 68.6% for training, 10.3% for

validation, and 21.1% for testing, as shown in Table 4.2. The same test sets are used for all experiments to ensure fair comparison.

To improve generalization, we shuffle the inputs for all sets. For the training set, we form a pool of slices from both end-diastole (ED) and end-systole (ES) images by loading each image and listing the available slices. Since each image contains a different number of slices, we use 10% of the slices for labeled images, and the rest are used for unlabeled images. The distribution of age per ethnic group and sex group counts is available in Table 4.3 and Figure 4.2.

Model	Train	Validation	Test	Labeled Slices	Unlabeled Slices
1	3301	493	1025	7073	63655
2	805	122	1025	1748	15736
3	264	39	1025	569	5125
4	182	27	1025	390	3506

Table 4.2 Train Validation Test Splits, and Labeled and Unlabeled Slices

Ethnicity	Mean Age	STD	Males	Females	Total
White	56.14	7.95	1504	1674	2178
Asian	52.67	9.11	258	127	385
Black	49.77	7.76	98	133	231
Total			1860	1934	3794

Table 4.3 Baseline Distribution

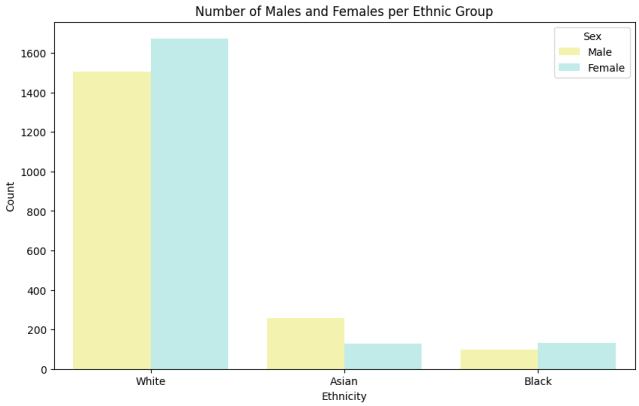


Figure 4.2 Baseline - Subjects Count per Sex Group per Ethnic Group

4.3 Experiment 2 - Age Sex Controlled Train Set

The goal of this experiment is to evaluate the impact of balancing the sex groups in the training set. We hypothesize that the model will perform similarly across different sex groups. In this experiment, the train set is balanced for age and sex, sampling from the same distribution curated for Experiment 1. The train set comprises an equal number of male and female subjects, totaling 796. Additionally, it contains an equal number of males and females within each ethnic group. We use 10% of the slices as labeled images. The distribution of age per ethnic group and subject count per group are available in Table 4.4 and Figure 4.3.

Ethnicity	Mean Age	STD	Males	Females	Total
White	51.52	3.74	398	398	796
Asian	50.13	3.47	35	35	70
Black	50.28	3.50	36	36	72
Total			469	469	938

Table 4.4 Experiment 2 Distribution

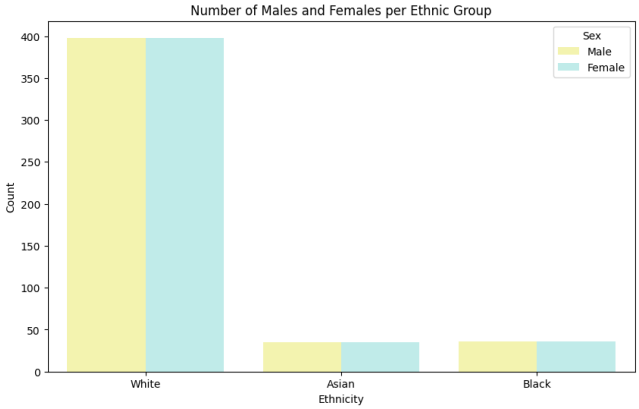


Figure 4.3 Experiment 2 - Subjects Count per Sex Group per Ethnic Group for 1 Traing Set out of 10

4.4 Experiment 3 - Age Ethnicity Controlled Train Set

The goal of this experiment is to evaluate the impact of balancing the ethnic groups in the training set. We hypothesize that the model will perform similarly across different ethnic groups. In this experiment, the train set is balanced for age and ethnicity, sampling from the same distribution curated for Experiment 1. We use 10% of the slices as labeled images. The number of subjects per ethnic group is the same across all the 10 repeated runs since all of them are controlled for age and ethnicity. However, the number of subjects per sex group may vary, and we show one example in the following tables. The distribution of age per ethnic group and subject count per ethnic group are available in Table 4.5 and Figure 4.4.

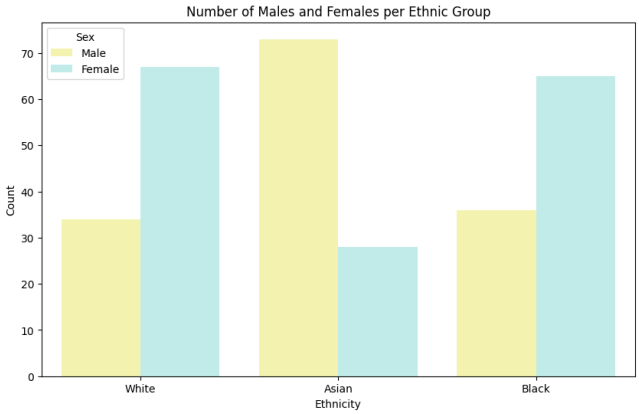


Figure 4.4 Experiment 3 - Subjects Count per Sex Group per Ethnic Group for 1 Traing Set out of 10

Ethnicity	Mean Age	STD	Males	Females	Total
White	50.86	3.77	38	63	101
Asian	50.67	3.67	78	23	101
Black	50.28	3.55	36	65	101
Total			152	151	303

Table 4.5 Experiment 3 Distribution

4.5 Experiment 4 - Age Sex Ethnicity Controlled Train Set

The goal of this experiment is to check the effect of balancing both sex and ethnicity on the performance of the model. Hence, the train set is balanced for age, sex, and ethnicity, sampling from the same distribution curated for Experiment 1. We use 10% of the slices as labeled images. The distribution of age per ethnic group and subject count per group are available in Table 4.6 and Figure 4.5. The number of subjects per sex and ethnic groups is the same across all the 10 repeated runs, as all of them are controlled for age, sex, and ethnicity.

Ethnicity	Mean Age	STD	Males	Females	Total
White	51.91	3.69	35	35	70
Asian	50.13	3.47	35	35	70
Black	50.31	3.47	35	35	70
Total			105	105	210

Table 4.6 Experiment 4 Distribution

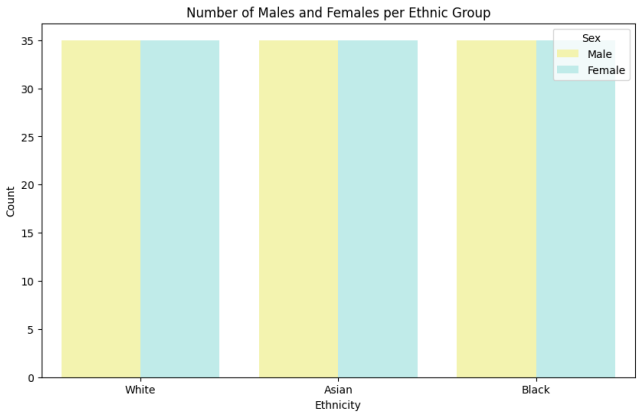


Figure 4.5 Experiment 4 - Subjects Count per Sex Group per Ethnic Group

4.6 Test set 1 - Age Sex Controlled

The goal of this test set is to evaluate the model's performance on sex groups. Hence, this test set is balanced for age and sex. We choose from each ethnic group a sample within $age = 51.0 \pm 7.0$. The distribution of age per ethnic group and subject count per ethnic group are available in Table 4.7 and Figure 4.6.

Ethnicity	Mean Age	STD	Males	Females	Total
White	51.40	3.59	400	400	800
Asian	50.89	3.73	50	50	100
Black	50.44	3.71	31	31	62
Total			481	481	962

Table 4.7 Test Set 1 Distribution

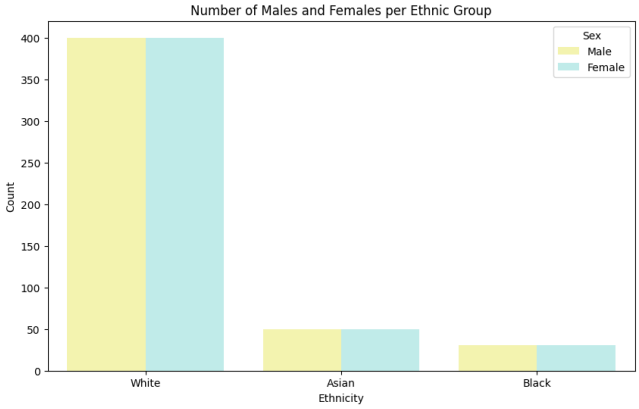


Figure 4.6 Test Set 1 - Subjects Count per Sex Group per Ethnic Group

4.7 Test set 2 - Age Sex Ethnicity Controlled

The goal of this test set is to evaluate the model's performance on ethnic groups. Hence, this set is balanced for age, sex, and ethnicity. The distribution of age per ethnic group and subject count per ethnic group are available in Table 4.8 and Figure 4.7.

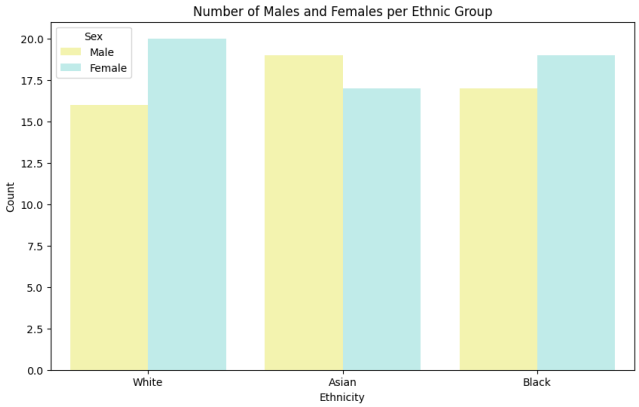


Figure 4.7 Test Set 2 - Subjects Count per Sex Group per Ethnic Group

Ethnicity	Mean Age	STD	Males	Females	Total
White	50.42	3.75	16	20	36
Asian	50.61	3.67	19	17	36
Black	50.89	3.42	17	19	36
Total			52	56	108

Table 4.8 Test Set 2 Distribution

4.8 Training

For this work, we utilize the UniMatch [80] as the overall semi-supervised framework with U-Net [61] as the backbone segmentation model. The strong perturbations in the UniMatch pipeline are based on ST++ [81] and CutMix [83]. Weak perturbations are obtained through simple crop and flip operations, where the raw image is resized between 0.5 and 2.0, cropped, and flipped. Feature perturbations employ a channel dropout of 50% using a built-in PyTorch function.

We use 10% labeled data for all experiments, following the setup in UniMatch [80], and adopt their choice of hyperparameters from ACDC [9]. Specifically, we set the batch size to 12, learning rate to 0.01, confidence threshold to 0.95, crop size to 256, and the number of classes to 4. The models are trained for 55 epochs in Experiment 1, 50 epochs in Experiment 2, 100 epochs in Experiment 3, and 200 epochs in Experiment 4.

For optimization, we use stochastic gradient descent with a momentum of 0.9 and weight decay of 0.0001. Additionally, we preprocess the images by rotating them 90 degrees counterclockwise, normalizing to the range [0, 1], and swapping the RV and LV mask labels to match the preprocessing of the ACDC dataset [9].

Each model's training time is approximately 1 hour, except for the baseline model, which took around 17 hours. Therefore, we do not repeat training for the baseline model. The total training time for all models is approximately 45 hours. Inference time for each model is in the order of seconds for each test set. The code for these experiments can be found in the following repository: `https://github.com/msskzx/UniMatch`.

5 Results

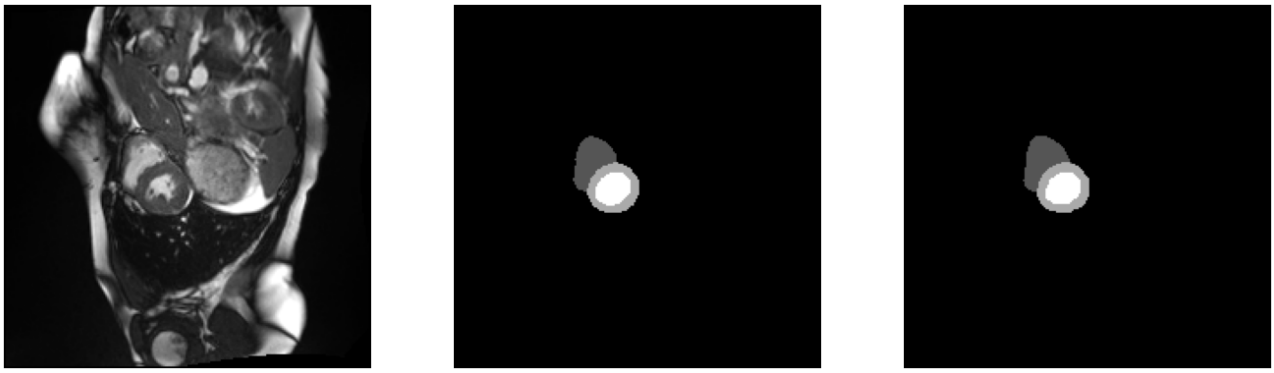
In order to evaluate the fairness of the models, we compare their performance across different groups, intersectional groups, and segmentation classes (RV, LV, MYO). We employ independent Student's t-tests to determine if the difference in performance is statistically significant, using a significance level of $\alpha = 0.05$. When comparing three groups, we conduct t-test on each pair of them (e.g., white vs. asian, white vs. black, and asian vs. black). We report the most interesting results in this chapter, while the complete results can be found in the appendix. Additionally, we compare the performance across different test sets to evaluate the effect of bias analysis.

5.1 Experiment 1

This model serves as a baseline since we do not control for any protected attributes. It provides insight into the model's general performance and its potential for the task on a new dataset. The model achieves a mean DSC score of 92.19 and 91.11 on the two sampled test sets, respectively. Additionally, it demonstrates strong performance visually, as shown in Figure 5.1. Table 5.1 includes the most important DSC scores, and the complete list is available in the appendix.

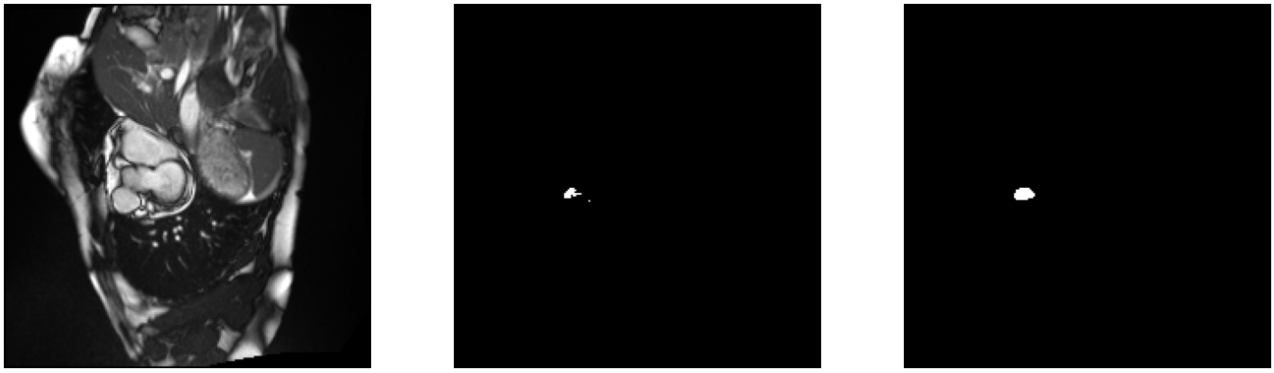
Although bias analysis is conducted for this model, it is not as detailed as in other experiments, since we do not repeat this experiment. In the bias analysis for this model, we find no significant difference in DSC scores between males and females (*p-value* $= 0.97$) for test set 1 and (*p-value* $= 0.99$) for test set 2. Similarly, we find no significant difference in DSC scores between Asian and Black groups for either of the test sets. However, we observe a significant difference in DSC scores between White and Asian groups (*p-value* $= 0.01$) only for test set 1. For White and Black groups, a significant difference in DSC scores is observed (*p-value* $= 0.0004$) only for test set 1.

Additionally, analysis of low-scoring slices reveals that the model's performance drops for slices near the apex or base of the heart (slices near 0 or 12) where the heart is not visible or the images contain irrelevant segmentation, as depicted in Figures 5.2 and 5.3.



(a) Input **(b)** Ground Truth **(c)** Prediction

Figure 5.1 Baseline Model - Input vs. Ground Truth vs. Prediction on Short Axis End Systolic Slice



(a) Input **(b)** Ground Truth **(c)** Prediction

Figure 5.2 Baseline Model - Input vs. Ground Truth vs. Prediction on Short Axis End Systolic Slice Near the Apex

(a) Input **(b)** Ground Truth **(c)** Prediction

Figure 5.3 Baseline Model - Input vs. Ground Truth vs. Prediction on Short Axis End Systolic Slice Near the Base

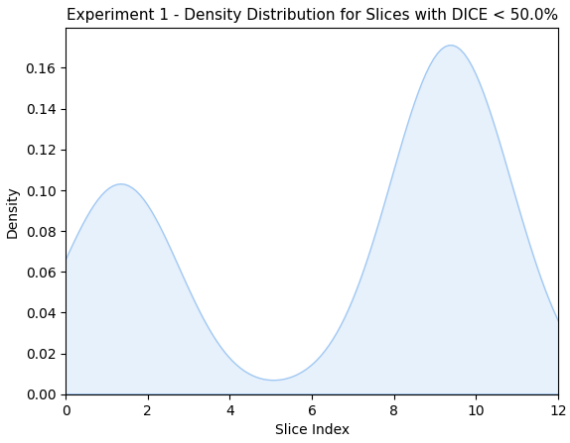


Figure 5.4 Input

Figure 5.5 Baseline Model - Slices with Low Performance

Table 5.1 Baseline Model Dice Scores

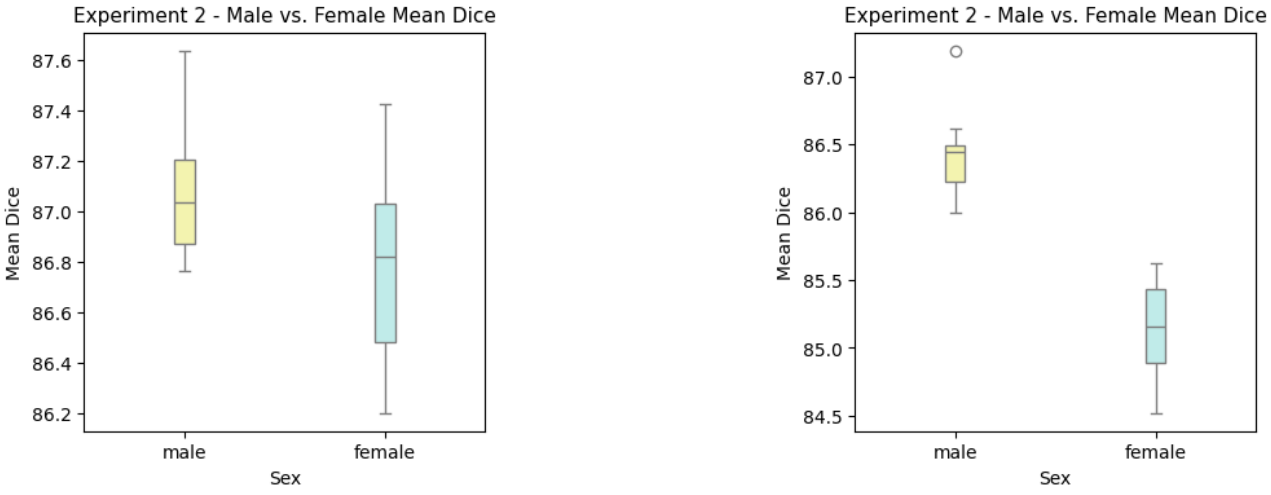
Category	Test Set 1	Test Set 2
dice mean	92.194169	91.106028
male	92.190938	91.555495
female	92.197605	90.639029
white	92.470077	91.986081
asian	91.719302	92.155741
black	90.982253	91.406135
white male	92.397797	90.431073
asian male	91.350297	92.060546
black male	90.751102	92.080667
white female	92.376524	91.315370
asian female	91.212682	89.507563
black female	91.398476	90.879934

5.2 Experiment 2

5.2.1 Test Set 1 Results

Sex Groups

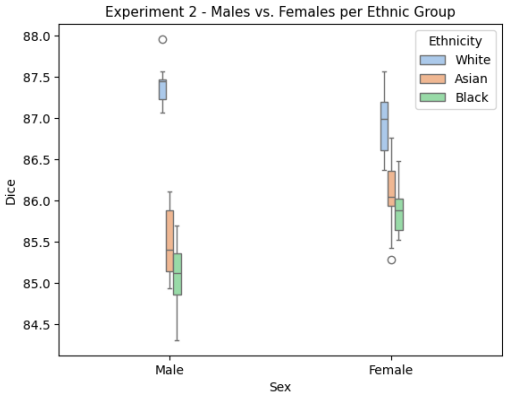
For experiment 2 on test set 1, we find no sufficient evidence to conclude a statistically significant difference in the overall mean DSC between male and female sex groups. However, we find significant disparities between males and females within the same ethnic group for all the tested groups, where white males score higher than white females ($p\text{-value} = 0.007$), asian females score higher than asian males ($p\text{-value} = 0.014$), and black females score higher than black males ($p\text{-value} = 0.0002$). The performance is higher for males sometimes and for females in others, which explains why the overall mean is balanced out and shows no significant difference, as shown in figures 5.6 and 5.7. Additionally, we find a statistically significant difference for RV, where the model performance for females is higher ($p\text{-value} = 0.007$). The same goes for the mean dice for LV ($p\text{-value} = 0.003$) and MYO ($p\text{-value} = 0.0003$), where the model performance for males is higher.



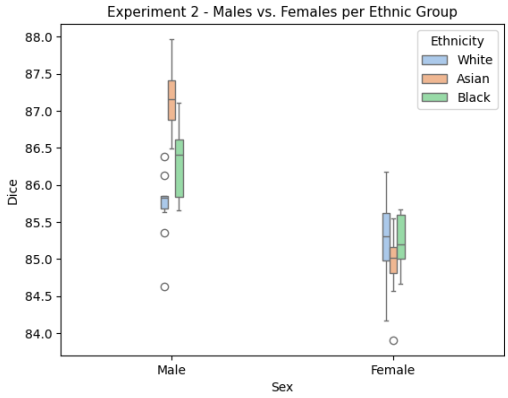
(a) Test Set 1

(b) Test Set 2

Figure 5.6 Experiment 2 - Males vs. Females



(a) Test Set 1



(b) Test Set 2

Figure 5.7 Experiment 2 - Males vs. Females

Ethnic Groups

In order to check how the test set also affects the result, we share the results of the model on ethnic groups, although this test set does not contain a balanced representation of ethnic groups. We find that there is a statistically significant difference in the overall mean Dice similarity scores between the white and Asian ethnic groups, where the white group has a higher performance (*p-value* $= 3.1 \times 10^{-5}$). Similarly, the white group has a significantly higher performance than the black group (*p-value* $= 3.1 \times 10^{-8}$), and the asian group has a significantly higher performance than the black group (*p-value* $= 6.7 \times 10^{-5}$), as shown in figure 5.8. We also monitor the performance of the intersectional groups, as shown in figure 5.7, and we find a significant difference between the groups of white males and Asian males, where the performance is higher for the former (*p-value* $= 5.98 \times 10^{-10}$). Similarly, the white males group has a significantly higher performance than the black males group (*p-value* $= 1.6 \times 10^{-11}$). We notice that the disparity between the asian and black groups disappears when investigating only males. Furthermore, the white females group scores significantly higher than both the asian (*p-value* $= 0.0001$) and black groups (*p-value* $= 1.9 \times 10^{-6}$). We also find disparities between ethnic groups for different segmentation classes, which we report in the appendix.

5.2.2 Test Set 2

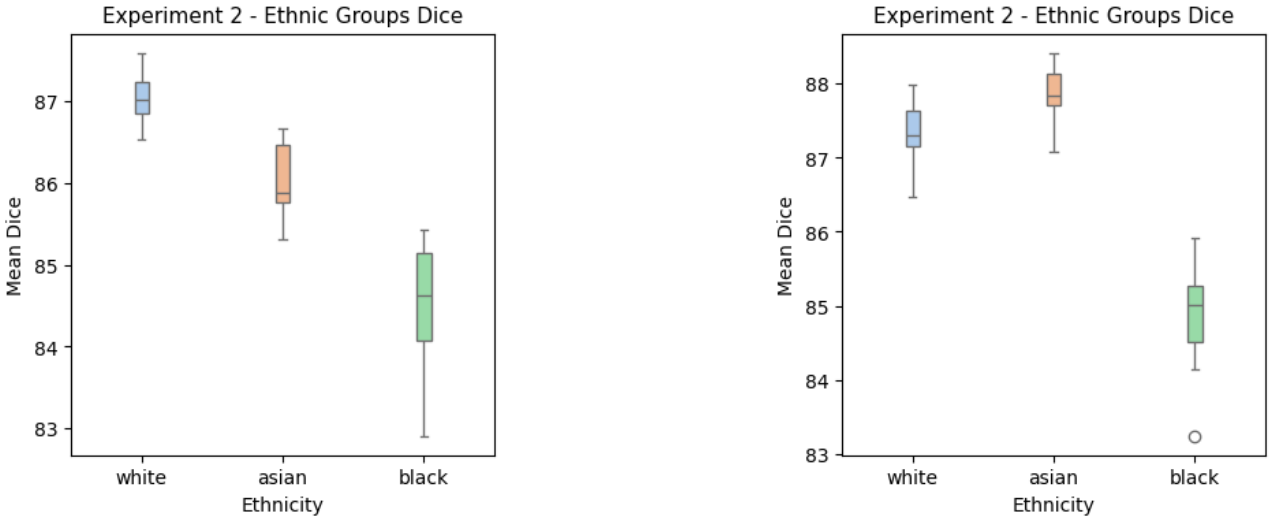
Sex Groups

For test set 2, there is a significant disparity between the male and female groups, where males score significantly higher (*p-value* $= 1.7 \times 10^{-7}$). Additionally, there are significant disparities between males and females within the ethnic groups, where males score higher than females for the asian (*p-value* $= 2.2 \times 10^{-9}$) and black (*p-value* $= 2.9 \times 10^{-5}$) groups. We also find disparities in segmentation labels, which we report in the appendix.

Ethnic Groups

We find that there is a significant difference between the white and asian groups, where the asian group performs better (*p-value* $= 0.012$), which is contrary to the previous test set. Additionally, we find that the performance for the black group is significantly lower than both the white (*p-value* $= 1.2 \times 10^{-7}$) and asian (*p-value* $= 5.2 \times 10^{-9}$) groups. However, the black males group performs significantly better than the white males (*p-value* $= 0.01$). The asian males group has significantly higher performance than the white males (*p-value* $= 1.2 \times 10^{-6}$) and the black males group (*p-value* $= .0005$). However, we find no significant difference for females between ethnic groups, i.e., asian females vs. white females. It is also

the case that this model performs better for the asian group for all the segmentation classes, which we report in the appendix.



(a) Test Set 1 **(b)** Test Set 2

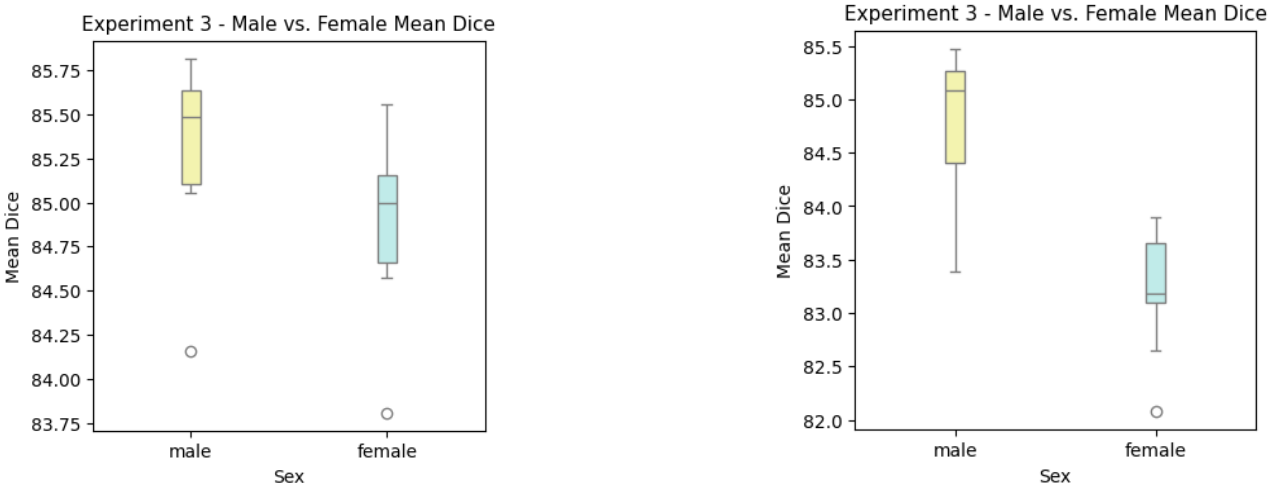
Figure 5.8 Experiment 2 - White vs. Asian vs. Black

5.3 Experiment 3

5.3.1 Test Set 1 Results

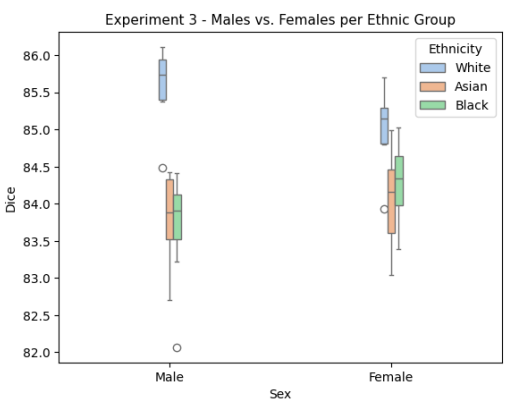
Sex Groups

For experiment 3, we find no significant difference between males and females, similar to experiment 2, as shown in figure 5.9. However, we find disparities only within the white ethnic group between males and females ($p\text{-value} = 0.01$) and no disparities within other ethnic groups, as shown in figure 5.10. We also find disparities between males and females for all segmentation classes, which we report in the appendix.

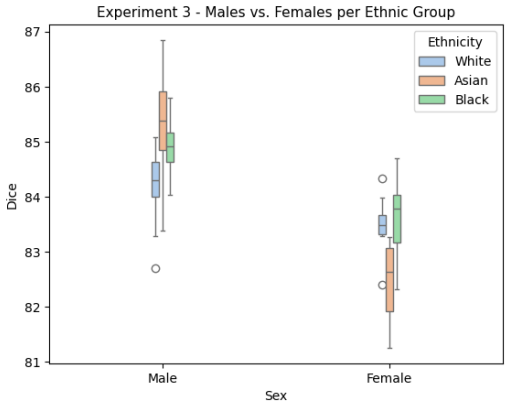


(a) Test Set 1 **(b)** Test Set 2

Figure 5.9 Experiment 3 - Males vs. Females



(a) Test Set 1



(b) Test Set 2

Figure 5.10 Experiment 3 - Males vs. Females

Ethnic Groups

We find statistically significant differences where the white group scores significantly higher than the asian group (*p-value* $= 0.0001$) and the black group (*p-value* $= 9.1 \times 10^{-5}$). Furthermore, the asian group scores significantly higher than the black group (*p-value* $= 0.03$), as shown in figure 5.11. Moreover, the White males group scores higher than asian males (*p-value* $= 5.6 \times 10^{-7}$) and black males (*p-value* $= 1.0 \times 10^{-6}$). Similarly, White females score significantly higher than asian females (*p-value* $= 0.0008$) and black females (*p-value* $= 0.005$), as shown in figure 5.10.

5.3.2 Test Set 2

Sex Groups

For test set 2, we find that males perform significantly better than females (*p-value* $= 1.8 \times 10^{-5}$). We also find that males perform significantly better than females within white ethnic group (*p-value* $= 0.02$), asian ethnic group (*p-value* $= 1.1 \times 10^{-6}$), and black ethnic group (*p-value* $= 0.0002$). Moreover, we report the disparities between segmentation classes in the appendix.

Ethnic Groups

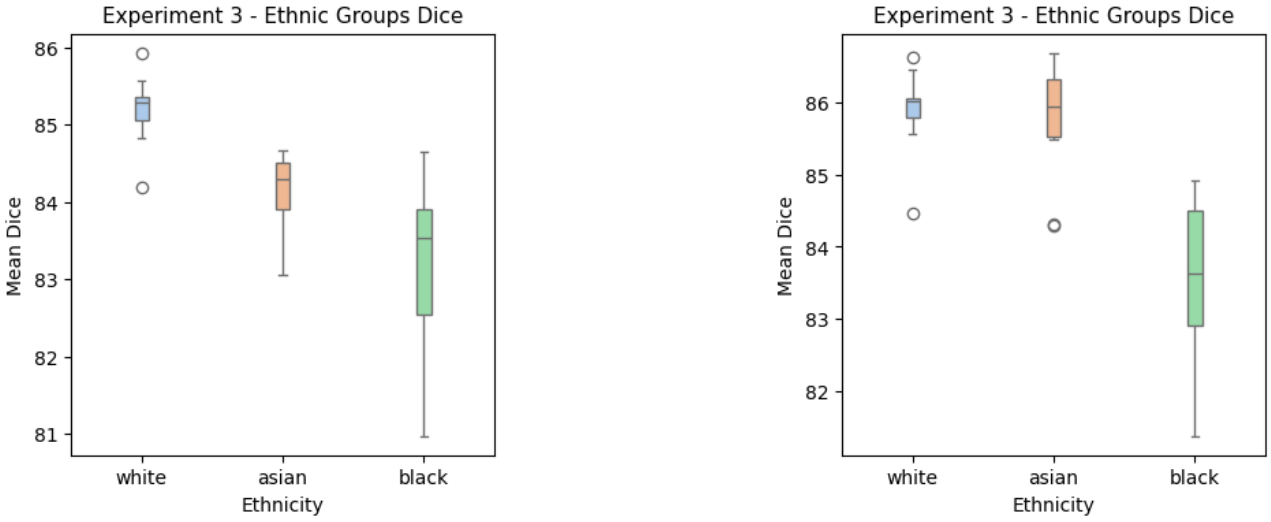
The black group has a significantly lower score than the white group (*p-value* $= 1.8 \times 10^{-5}$) and the asian group (*p-value* $= 0.0001$). However, black males perform significantly better than white males (*p-value* $= 0.02$). Furthermore, asian males significantly outperform white males (*p-value* $= 0.0001$). Moreover, we find no disparities between asians and black males. On the other hand, the asian females group scores significantly lower than white females (*p-value* $= 0.002$) and black females (*p-value* $= 0.002$). Differences between segmentation classes are also reported in the appendix.

5.4 Experiment 4

5.4.1 Test Set 1 Results

Sex Groups

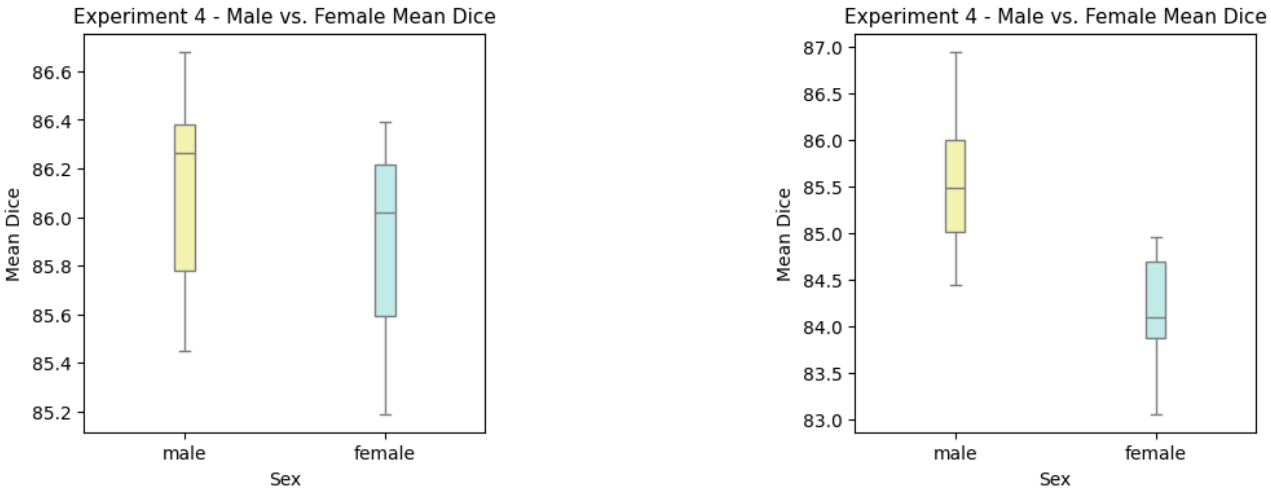
There is not sufficient evidence (*p-value* $= 0.3$) to conclude a statistically significant difference between males as shown in figure 5.12. When we analyze the performance between sex groups within each ethnic group, we find only significant disparity between black males and black females (*p-value* $= 0.04$) as shown in figure 5.13. Moreover, we find a significant difference between males and females for RV (*p-value* $=$



(a) Test Set 1 **(b)** Test Set 2

Figure 5.11 Experiment 3 - White vs. Asian vs. Black

0.04) where the score is higher for females. For LV (*p-value* $= 0.02$), and for MYO (*p-value* $= 5.2 \times 10^{-5}$), the score is significantly higher for males.



(a) Test Set 1 **(b)** Test Set 2

Figure 5.12 Experiment 4 - Males vs. Females

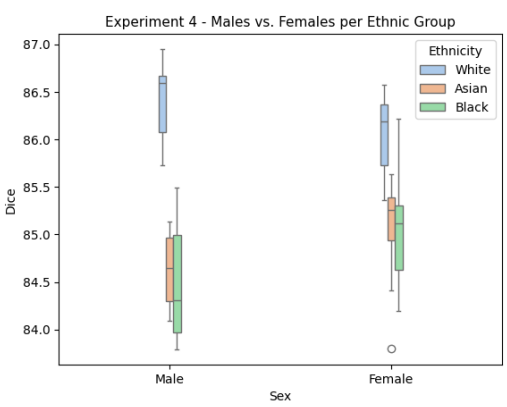
Ethnic Groups

For the first test set, we find significant differences between all groups as shown in figure 5.14. White group scores significantly higher than both Asian (*p-value* $= 0.0005$) and black (*p-value* $= 0.5 \times 10^{-5}$) groups. Furthermore, the asian group scores significantly higher than the black group (*p-value* $= 0.005$).

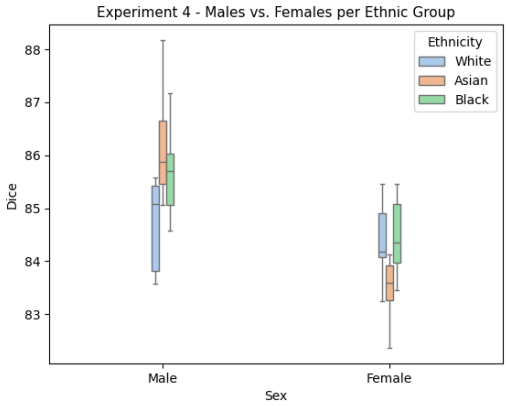
5.4.2 Test Set 2

Sex Groups

There is a statistically significant difference (*p-value* $= 0.0004$) between males and females where the model performs better for males. Furthermore, we find that the model performs better for males on LV



(a) Test Set 1



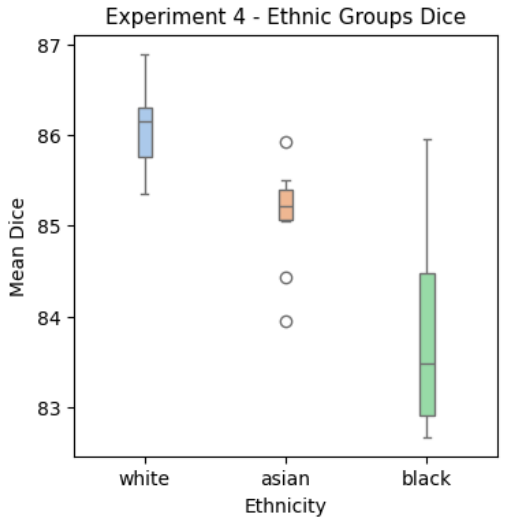
(b) Test Set 2

Figure 5.13 Experiment 4 - Males vs. Females

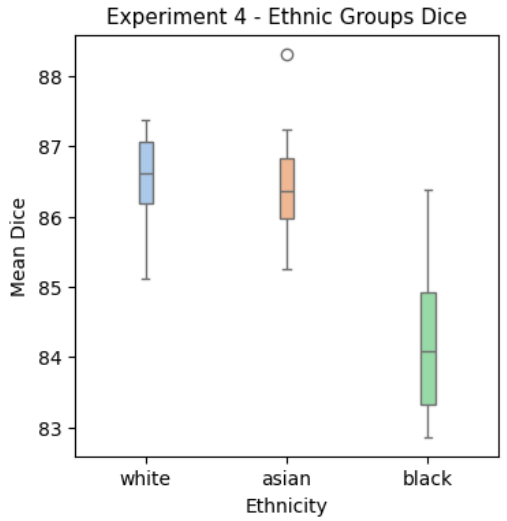
label (*p-value* $= 0.05$) and MYO (*p-value* $= 3.2 \times 10^{-7}$). However, there is no significant difference on the RV label (*p-value* $= 0.09$). For within ethnic group analysis, we find statistical difference between males and females for the asian ethnic group (*p-value* $= 4.4 \times 10^{-7}$) where males score higher. Furthermore, for the black ethnic groups, males score significantly higher than females (*p-value* $= 0.003$). However, we find no significant disparity between white males and females (*p-value* $= 0.5$).

Ethnic Groups

For the second test set, we only find that the black group scores significantly lower than both the white (*p-value* $= 3.3 \times 10^{-5}$) and asian (*p-value* $= 7.6 \times 10^{-5}$) groups. We find that white males group scores significantly lower than asian males (*p-value* $= 0.002$) and black males (*p-value* $= 0.02$) groups. We also find that asian females group scores significantly lower than white females group (*p-value* $= 0.003$) and black females group (*p-value* $= 0.003$).



(a) Test Set 1



(b) Test Set 2

Figure 5.14 Experiment 4 - White vs. Asian vs. Black

5.5 Model Comparison

We observe that the performance of model 2 is significantly higher than that of models 3 and 4 for almost all groups, intersectional groups, and segmentation classes on both test sets, as illustrated in Figures 5.15 through 5.17 and Tables 5.2 through 5.5. More detailed results are provided in the appendix.

Table 5.2 Mean and STD for Performance Test Set 1

Experiment	Overall Dice		Male Dice		Female Dice	
	Mean	STD	Mean	STD	Mean	STD
2	86.936213	0.311012	87.072816	0.258205	86.790966	0.391927
3	85.107979	0.467943	85.321747	0.490795	84.880683	0.485110
4	86.010615	0.395133	86.120725	0.414139	85.893536	0.445419

Table 5.3 Mean and STD for Performance Test Set 1

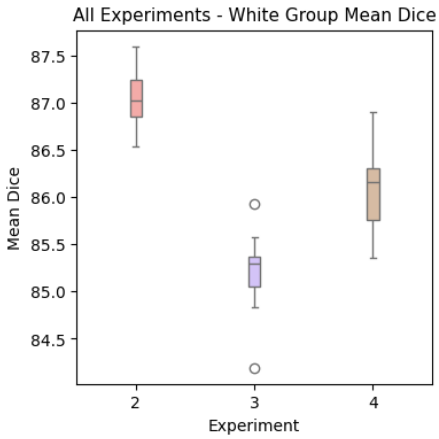
Experiment	White Dice		Asian Dice		Black Dice	
	Mean	STD	Mean	STD	Mean	STD
2	87.054328	0.356610	86.024801	0.471000	84.518008	0.794094
3	85.205564	0.464298	84.153583	0.517695	83.163051	1.203064
4	86.093194	0.477574	85.112742	0.555520	83.826883	1.123384

Table 5.4 Mean and STD for Performance Test Set 2

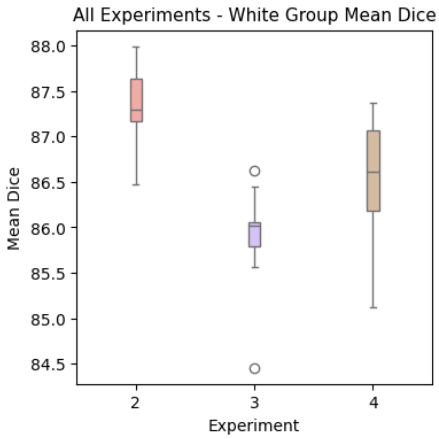
Experiment	Overall Dice		Male Dice		Female Dice	
	Mean	STD	Mean	STD	Mean	STD
2	85.805336	0.303834	86.429530	0.327011	85.156794	0.365431
3	84.035898	0.589093	84.809898	0.657889	83.231706	0.559140
4	84.853542	0.606976	85.509540	0.774845	84.171955	0.589603

Table 5.5 Mean and STD for Performance Test Set 1

Experiment	White Dice		Asian Dice		Black Dice	
	Mean	STD	Mean	STD	Mean	STD
2	87.325989	0.439230	87.861624	0.418177	84.877692	0.810649
3	85.881849	0.588996	85.717701	0.831268	83.502822	1.161620
4	86.516501	0.710226	86.488872	0.862828	84.190818	1.135954

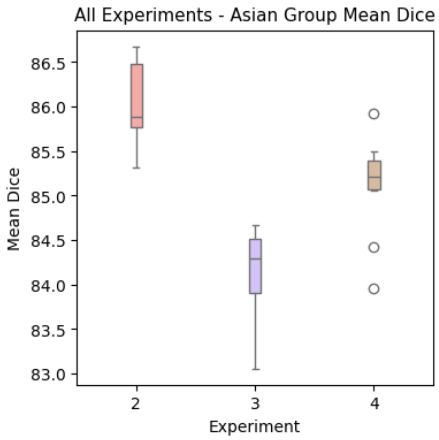


(a) Test Set 1

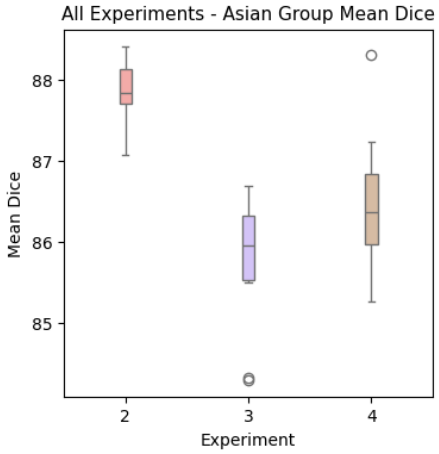


(b) Test Set 2

Figure 5.15 All Experiments - Overall Mean White Group Dice

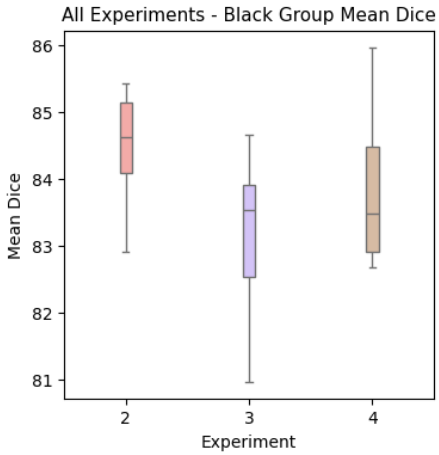


(a) Test Set 1

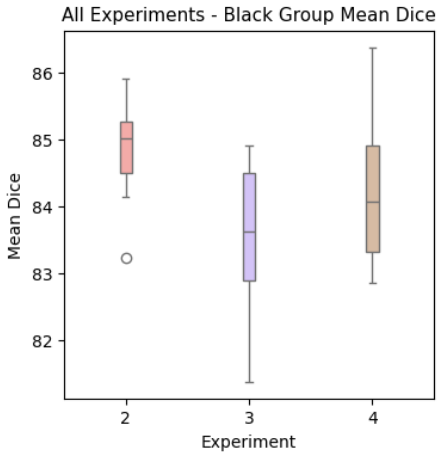


(b) Test Set 2

Figure 5.16 All Experiments - Overall Mean Asian Group Dice



(a) Test Set 1



(b) Test Set 2

Figure 5.17 All Experiments - Overall Mean Black Group Dice

6 Discussion

Train Set Balancing Effect

In our experiments, we attempted to mitigate representation bias by balancing the groups in the training sets. However, we found that this does not necessarily lead to a fair model. This behavior could not be explained solely by balancing the train and test data for sex or ethnicity, or both. For example, we expected models 2 and 4 to perform similarly for sex groups since they are trained on sex-balanced train sets, however, that was not the case. Moreover, we expected models 3 and 4 to perform similarly for ethnic groups since they are trained on balanced train sets for them, however, that was not the case. This finding is in line with the findings of Lee et al. [45] where they found that accuracy parity does not necessarily occur when groups are balanced.

No Disparities between Groups but Intersectional Groups

The absence of disparities in the overall mean does not imply the absence of disparities in subgroups or intersectional groups. This was evident in the second experiment, where the overall mean showed no significant disparities between males and females. However, further investigation into sex groups within each ethnic group revealed significant disparities. The overall mean is balanced because the disparities within each ethnic group sometimes favor females and sometimes favor males, thereby balancing the overall mean. This underscores the importance of considering intersectional factors when attempting to address fairness in model performance.

Test Set Effect

We observed variations in model performance across different test sets, particularly in relation to sex and ethnic groups. Interestingly, the performance difference between males and females varied depending on the test set used. This discrepancy may be attributed to differences in sample composition between the two test sets. It's possible that certain samples in the test sets, particularly those pertaining to females, present greater challenges for segmentation inference. Additionally, we noted that model performance varied based on the representation of ethnic groups in the test set. For instance, models trained on datasets with a higher proportion of samples from the white ethnic group tended to perform better on the first test set, but their performance dropped on the second test set. This underscores the impact of test set composition on model performance and highlights the need for careful consideration when evaluating model fairness.

Improve Overall Performance or Reduce Disparities

Our analysis reveals that improving the overall performance of a model does not guarantee fairness across all protected groups, underscoring the importance of bias analysis. Interestingly, Model 2 consistently outperforms Models 3 and 4 across various groups and subgroups, possibly due to its larger training set, despite all models being trained for a similar duration. However, it's noteworthy that Model 3, despite being trained on a larger train set than Model 4, performs worse across all groups and subgroups. This suggests that Model 3 may have required additional training time to achieve optimal performance.

Limitations and Future Work

In this work, we investigated bias mitigation to a limited extent by balancing the protected groups in the training set and exploring semi-supervised learning as a potential approach to mitigate bias. However, there are numerous other approaches that could be explored.

We began integrating distribution alignment from ReMixMatch [10] into UniMatch [80] to introduce a fairness aspect into the objective function, as outlined in the appendix A.1. However, this component was not fully integrated or tested due to the time needed to adapt it from image classification to image segmentation. Further enhancement of this component could involve aligning the distributions of protected groups through multi-task learning, where the model is trained to predict both the segmentation and the protected group of the patient.

We notice that the model's performance for slices near the base or apex of the heart is worse than for median slices, as expected. Therefore, further preprocessing of the training set to remove slices that are not very useful for the segmentation task could improve the model's performance.

We only relied on a loose balancing of the labeled data. In our setup, we use 10% of the total slices as labeled images, meaning that each patient will have some of the slices labeled and the rest unlabeled. This could result in a set where the labeled images are balanced per protected group patient-wise but not necessarily balanced for slice-wise in the labeled set, which is why further investigation is needed.

We highlight the limitation of relying on specific test sets for analysis. Therefore, we suggest investigating a similar approach to Glocker et al. [31], where they employ strategic resampling with replacement to obtain a balanced test set. This could aid in understanding the effect of the test set on model performance and disparities between groups.

Due to time constraints, we were unable to analyze bias for ethnic subgroups such as Indian, Pakistani, Bangladeshi, and other Asian backgrounds within the Asian ethnic group. This analysis could have provided valuable insights into the model's behavior.

7 Conclusion

This work underscores the critical role of leveraging AI models in healthcare to enhance the tools available for healthcare providers, ultimately saving time and resources. However, AI models pose potential risks if not carefully audited for bias and fairness. This can result in disparate performance between protected groups, subgroups, or intersectional groups. In this study, we propose semi-supervised learning as a potential approach to mitigate bias in AI models, focusing on the task of cardiac MRI segmentation. To this end, we trained a semi-supervised model and conducted various experiments to probe the factors influencing fairness. Our investigation revealed that while semi-supervised learning or representation learning can yield impressive performance in low-data regimes, it alone is insufficient to ensure fairness. Disparities persisted between different groups, subgroups, and intersectional groups across two test sets. Additionally, balancing the training set did not always eliminate these disparities, and the absence of disparities between certain groups did not imply fairness when considering intersectional groups. Furthermore, we examined the impact of test set sampling strategies, shedding light on potential explanations for observed disparities.

Moreover, we provided insights into possible avenues for future research, including incorporating distribution alignment, further preprocessing of the training set, investigating the effects of balancing labeled images, and employing different test set sampling strategies. Ultimately, we stress the importance of incorporating bias analysis and mitigation strategies into AI models, especially in healthcare applications. We encourage leveraging available tools to mitigate discrepancies and minimize harm to patients.

A Appendix

A.1 UniMatch with Distribution Alignment

We propose to integrate Distribution Alignment from ReMixMatch [10] into the objective function of UniMatch [80] to promote fairness in the model. In Distribution Alignment, as depicted in Eq. (1), we obtain $\tilde{q}^w$, signifying normalization, and $\hat{\tilde{q}}^w = \text{argmax}(\tilde{q}^w)$. When combined with Eq. (2) and Eq. (3), the unsupervised loss term of UniMatch, as stated in Eq. (4), transforms as follows:

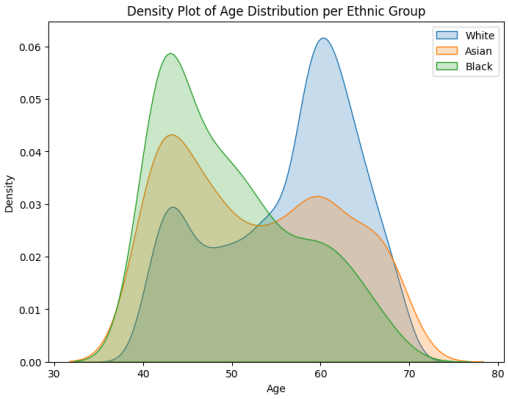
$$\frac{1}{\mu B}\sum 1(max(\tilde{q}^w) \geq \tau)(\lambda H(\hat{\tilde{q}}^w, p^{fp}) + \frac{\mu}{2}(H(\hat{\tilde{q}}^w, p^{s1}) + H(\hat{\tilde{q}}^w, p^{s2}))) \tag{A.1}$$

A.2 UK Biobank Dataset

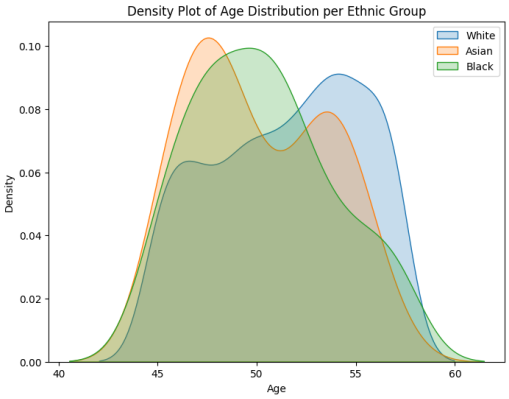
Code	Meaning
-3	Prefer not to answer
-1	Do not know
1	White
2	Mixed
3	Asian or Asian British
4	Black or Black British
5	Chinese
6	Other ethnic group
1001	British
1002	Irish
1003	Any other white background
2001	White and Black Caribbean
2002	White and Black African
2003	White and Asian
2004	Any other mixed background
3001	Indian
3002	Pakistani
3003	Bangladeshi
3004	Any other Asian background
4001	Caribbean
4002	African
4003	Any other Black background

Table A.1 Ethnicity Codes

A.3 Experiments Train and Test Sets

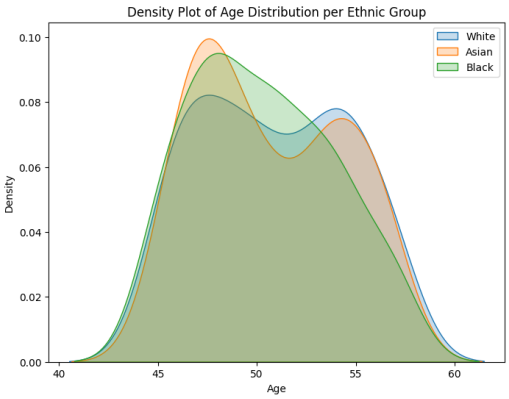


(a) Baseline - Age Distribution

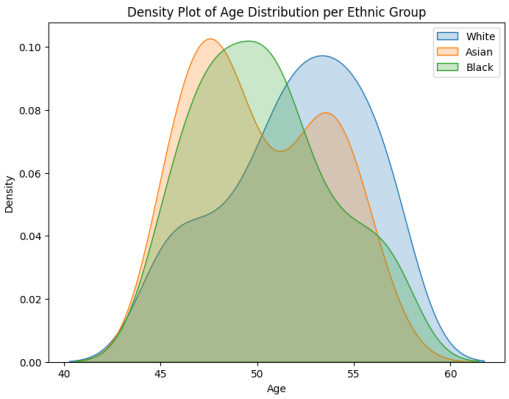


(b) Experiment 2 - Age Distribution

Figure A.1 Age Distribution

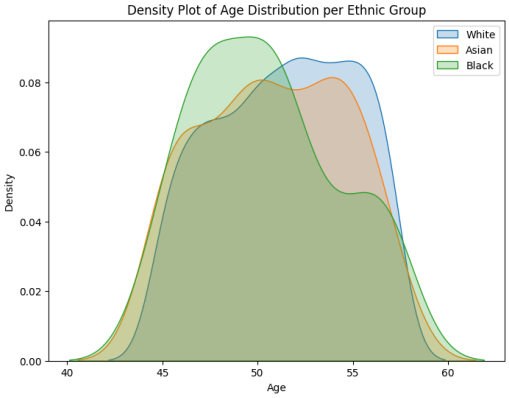


(a) Experiment 3 - Age Distribution

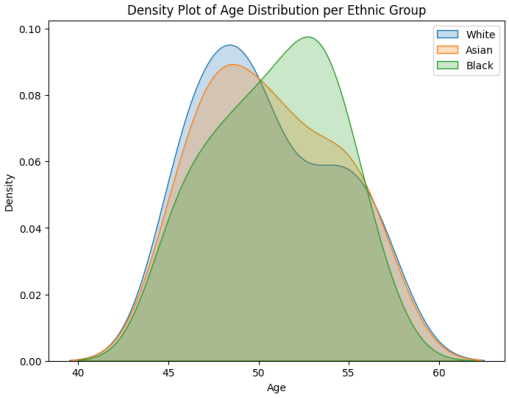


(b) Experiment 3 - Age Distribution

Figure A.2 Age Distribution



(a) Test Set 1 - Age Distribution



(b) Test Set 2 - Age Distribution

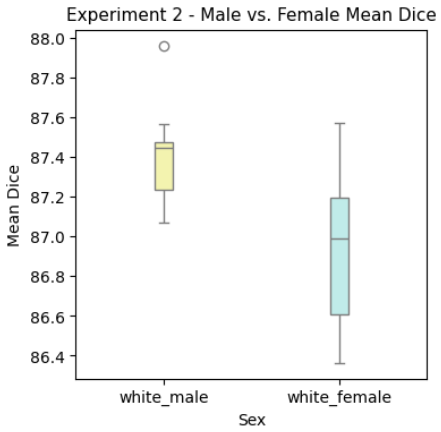
Figure A.3 Age Distribution

A.4 Experiment 1 - Results

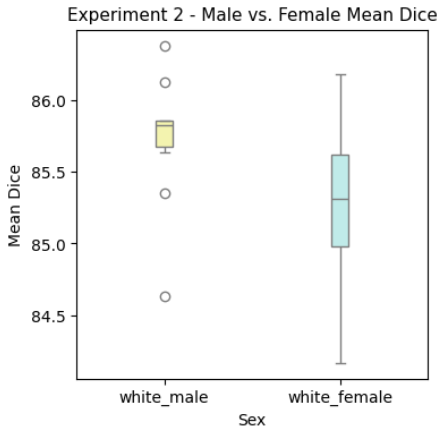
Table A.2 Model 1 Performance

Category	Test Set 1	Test Set 2
dice mean	92.194169	91.106028
dice lv	94.909766	94.143639
dice rv	90.823334	89.617494
dice myo	90.849408	89.556952
male	92.190938	91.555495
male lv	94.865803	94.228681
male rv	90.667754	89.932369
male myo	91.039257	90.505436
female	92.197605	90.639029
female lv	94.956512	94.055280
female rv	90.988760	89.290336
female myo	90.647544	88.571470
white	92.470077	91.986081
white rv	90.380295	90.942644
white lv	95.556949	95.321098
white myo	91.472988	89.694500
asian	91.719302	92.155741
asian rv	89.816690	90.530380
asian lv	94.682516	94.859992
asian myo	90.658700	91.076852
black	90.982253	91.406135
black rv	88.855189	89.692109
black lv	95.113475	95.241322
black myo	88.978094	89.284975
white male	92.397797	90.431073
asian male	91.350297	92.060546
black male	90.751102	92.080667
white female	92.376524	91.315370
asian female	91.212682	89.507563
black female	91.398476	90.879934

A.5 Experiment 2 - Results

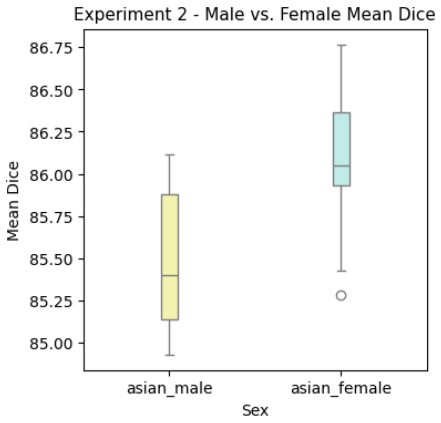


(a) Test Set 1

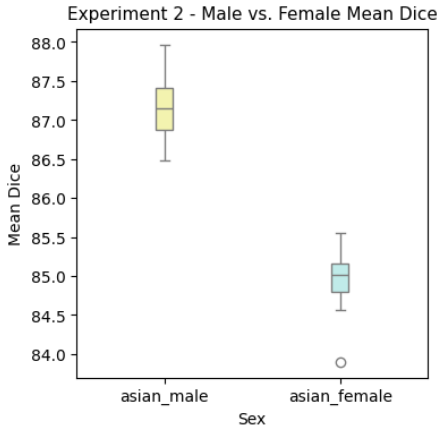


(b) Test Set 2

Figure A.4 Experiment 2 - White Males vs. Females

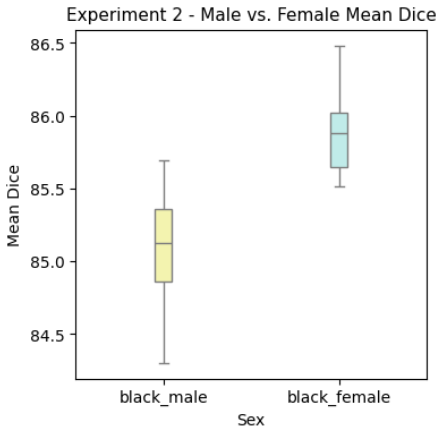


(a) Test Set 1

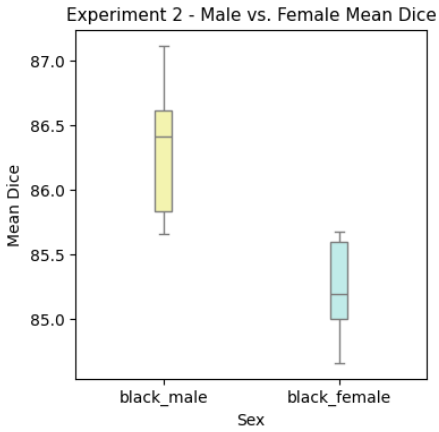


(b) Test Set 2

Figure A.5 Experiment 2 - Asian Males vs. Females

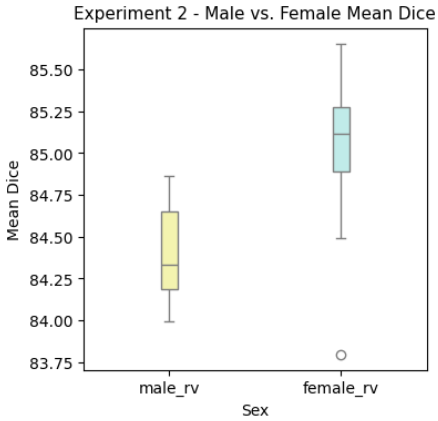


(a) Test Set 1

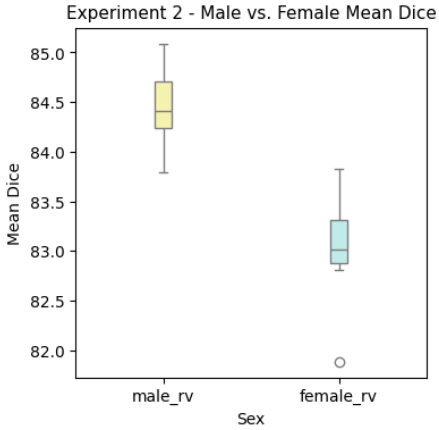


(b) Test Set 2

Figure A.6 Experiment 2 - Black Males vs. Females

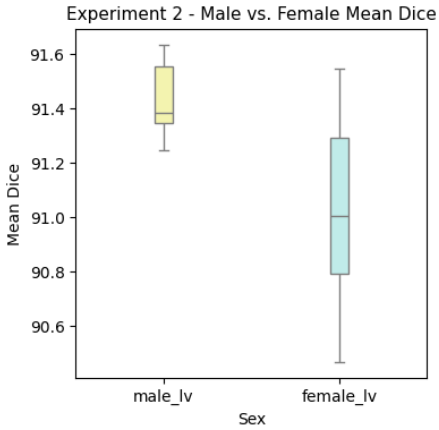


(a) Test Set 1

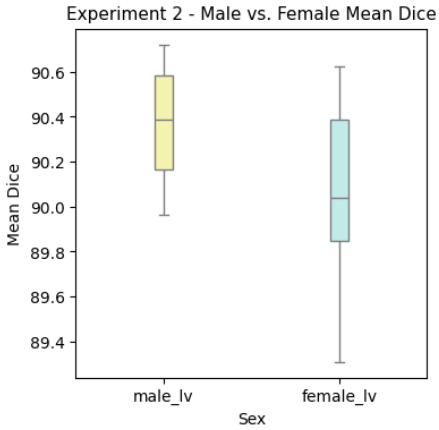


(b) Test Set 2

Figure A.7 Experiment 2 - Males vs. Females RV

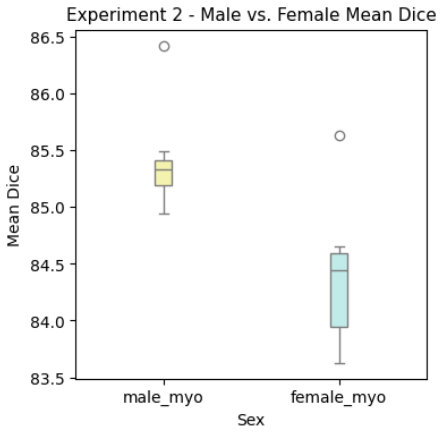


(a) Test Set 1

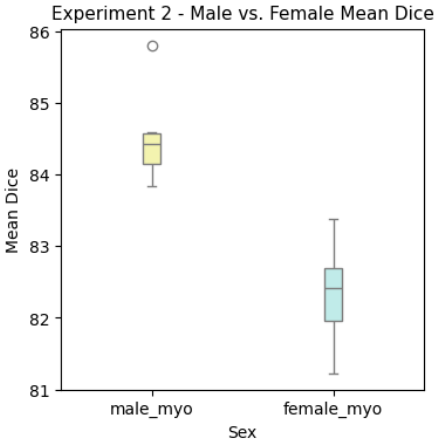


(b) Test Set 2

Figure A.8 Experiment 2 - Males vs. Females LV

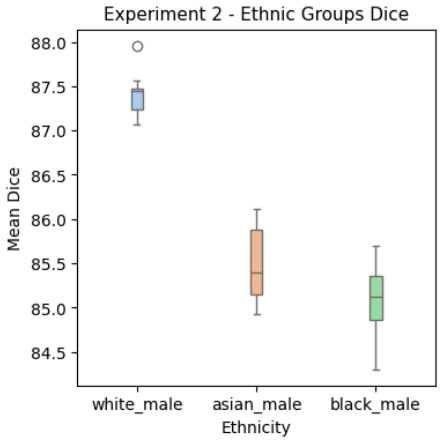


(a) Test Set 1

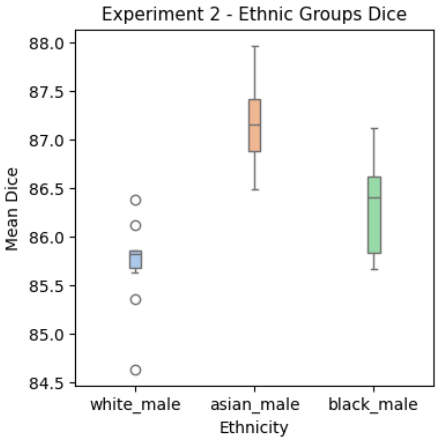


(b) Test Set 2

Figure A.9 Experiment 2 - Males vs. Females MYO

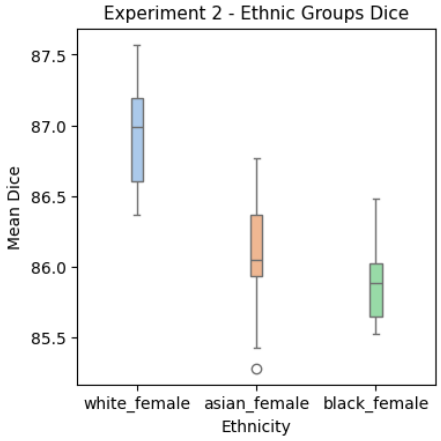


(a) Test Set 1

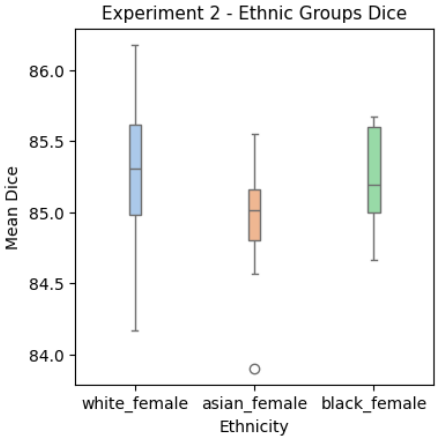


(b) Test Set 2

Figure A.10 Experiment 2 - White vs. Asian vs. Black Males

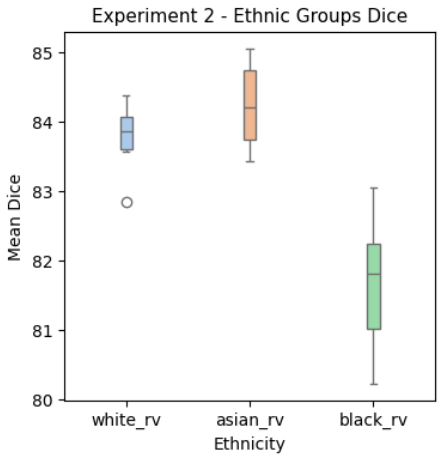


(a) Test Set 1

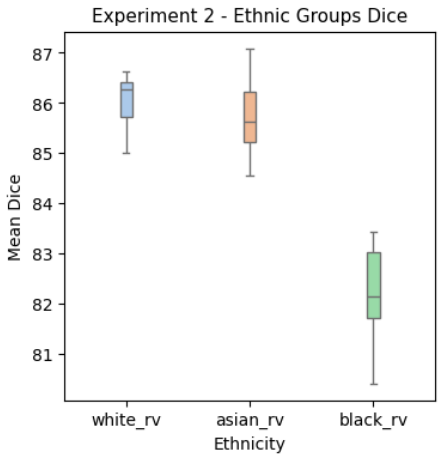


(b) Test Set 2

Figure A.11 Experiment 2 - White vs. Asian vs. Black Females

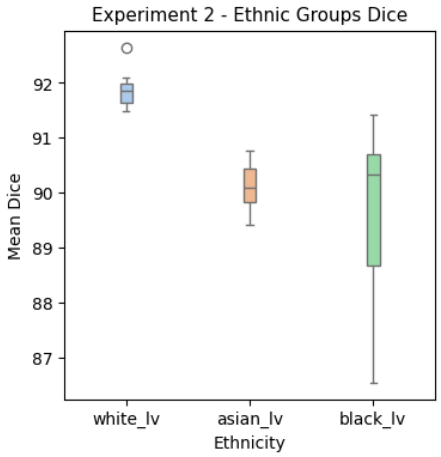


(a) Test Set 1

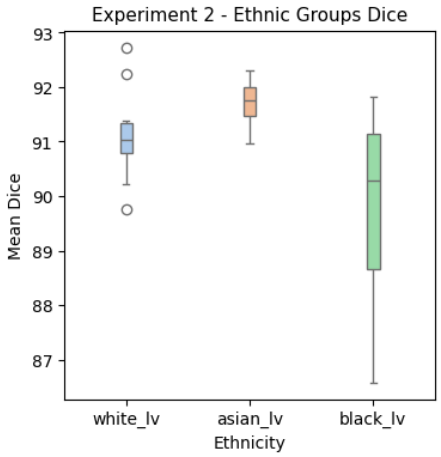


(b) Test Set 2

Figure A.12 Experiment 2 - White vs. Asian vs. Black RV

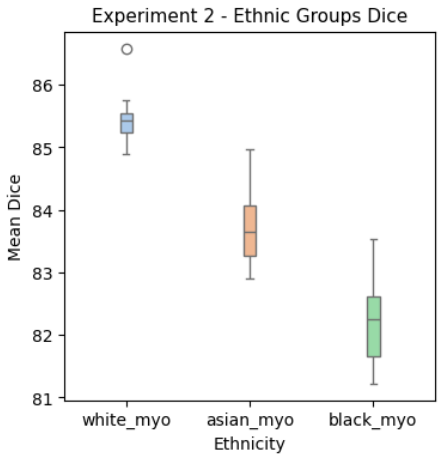


(a) Test Set 1

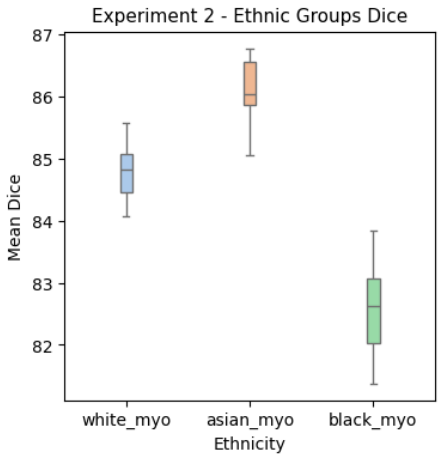


(b) Test Set 2

Figure A.13 Experiment 2 - White vs. Asian vs. Black LV



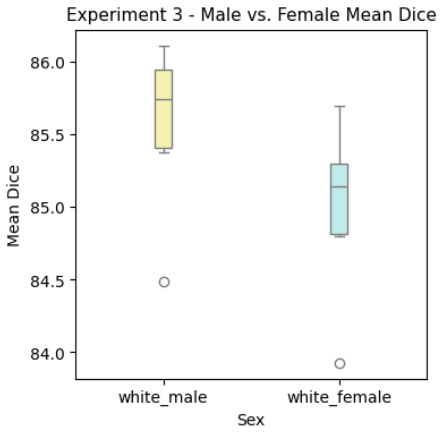
(a) Test Set 1



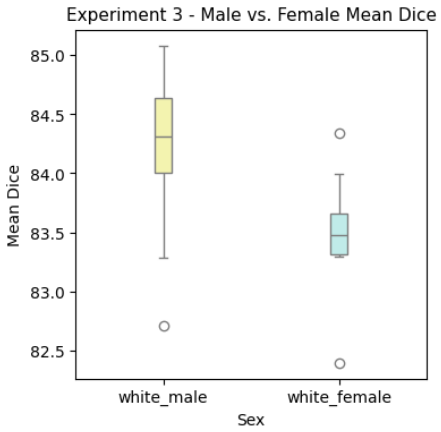
(b) Test Set 2

Figure A.14 Experiment 2 - White vs. Asian vs. Black MYO

A.6 Experiment 3 - Results

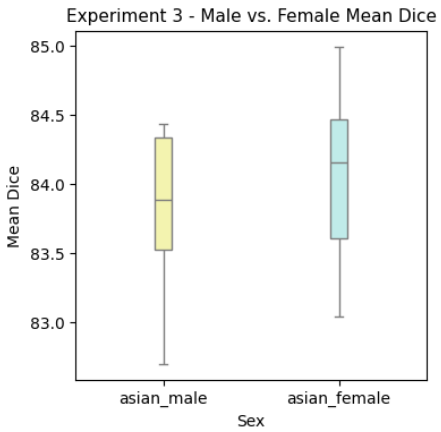


(a) Test Set 1

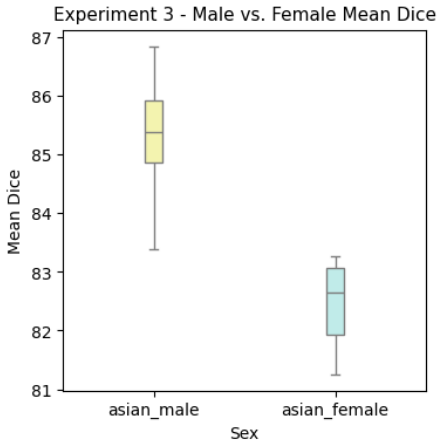


(b) Test Set 2

Figure A.15 Experiment 3 - White Males vs. Females

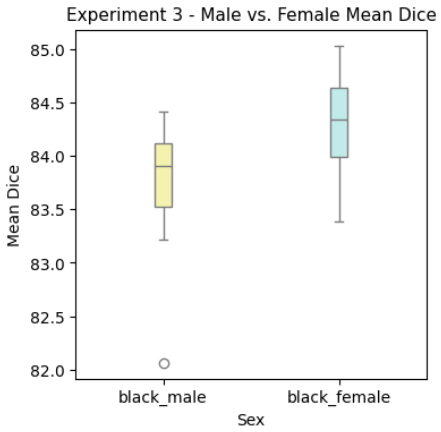


(a) Test Set 1

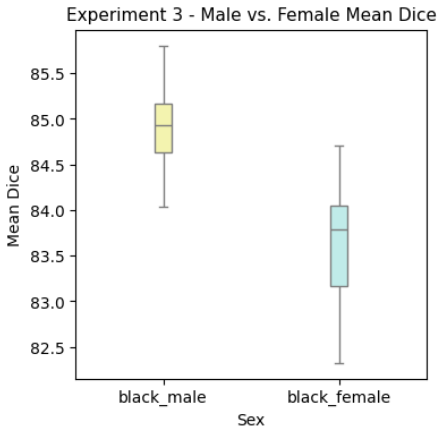


(b) Test Set 2

Figure A.16 Experiment 3 - Asian Males vs. Females

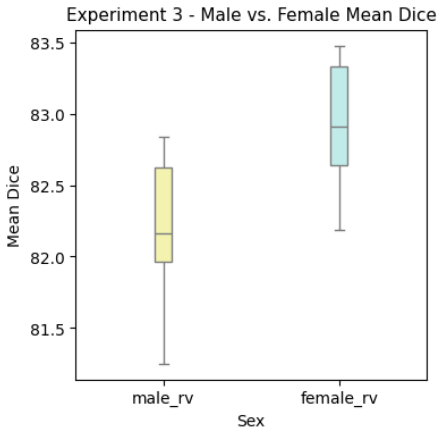


(a) Test Set 1

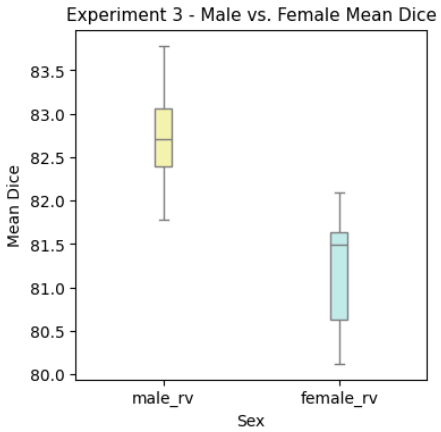


(b) Test Set 2

Figure A.17 Experiment 3 - Black Males vs. Females

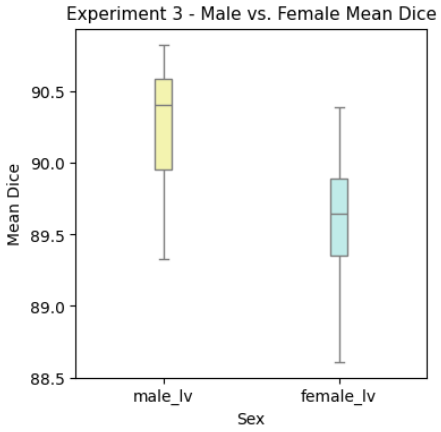


(a) Test Set 1

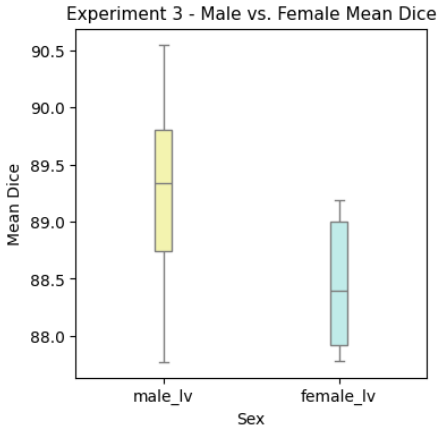


(b) Test Set 2

Figure A.18 Experiment 3 - Males vs. Females RV

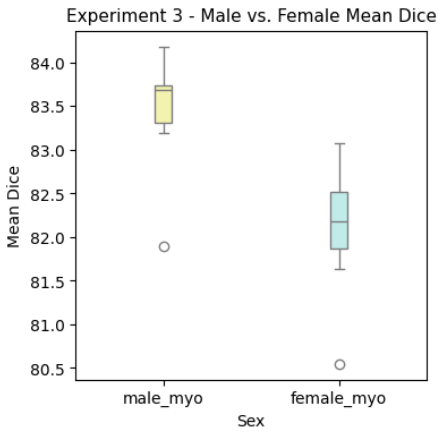


(a) Test Set 1

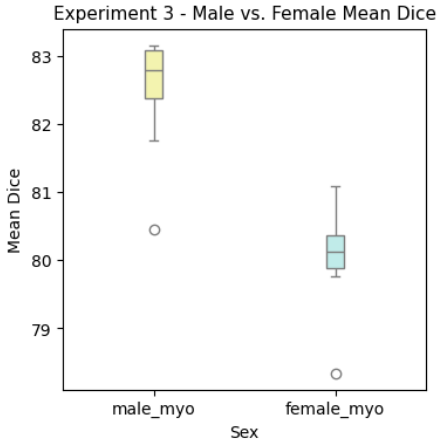


(b) Test Set 2

Figure A.19 Experiment 3 - Males vs. Females LV

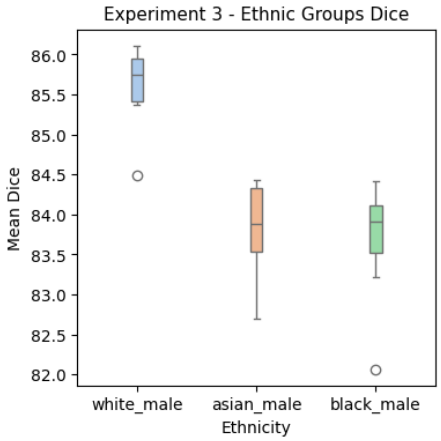


(a) Test Set 1

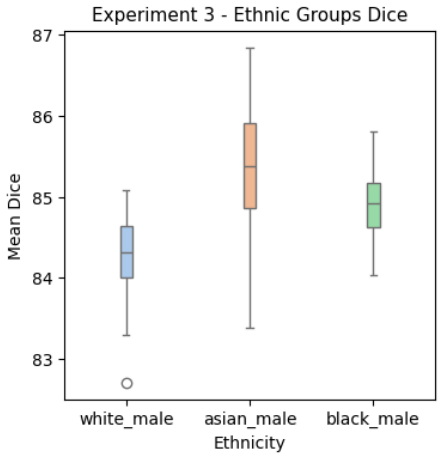


(b) Test Set 2

Figure A.20 Experiment 3 - Males vs. Females MYO

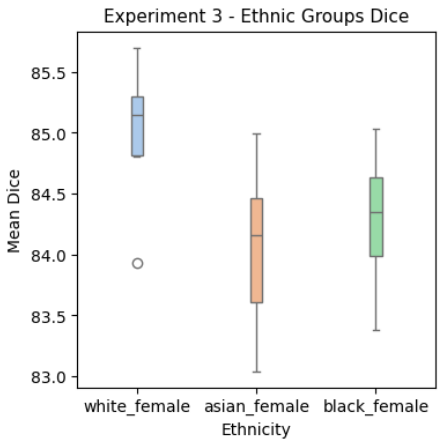


(a) Test Set 1

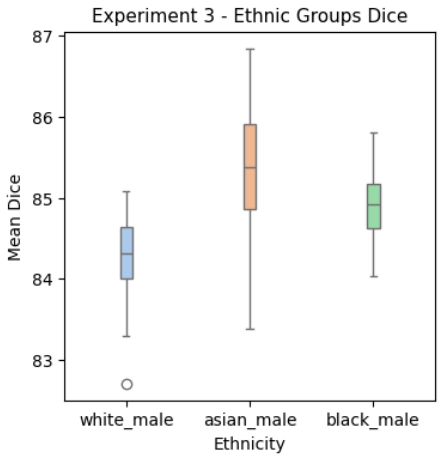


(b) Test Set 2

Figure A.21 Experiment 3 - White vs. Asian vs. Black Males

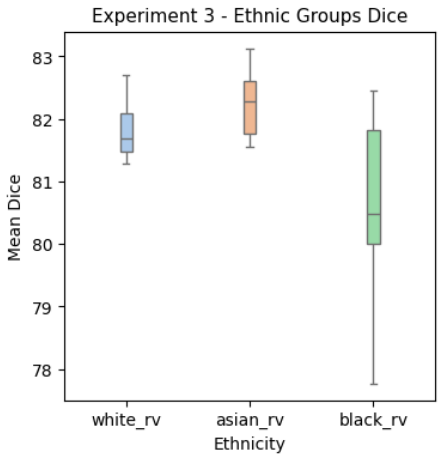


(a) Test Set 1

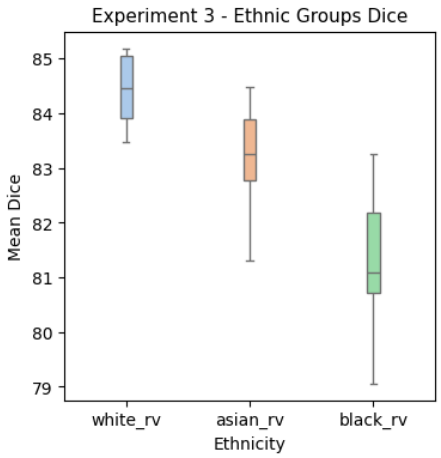


(b) Test Set 2

Figure A.22 Experiment 3 - White vs. Asian vs. Black Females

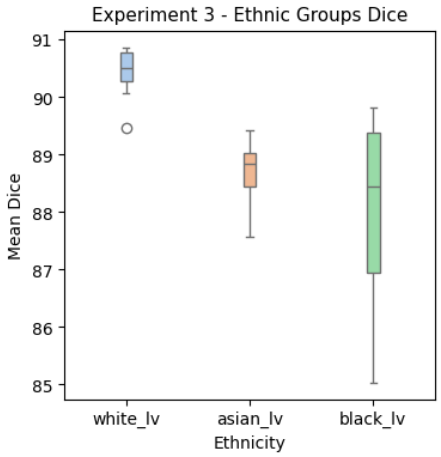


(a) Test Set 1

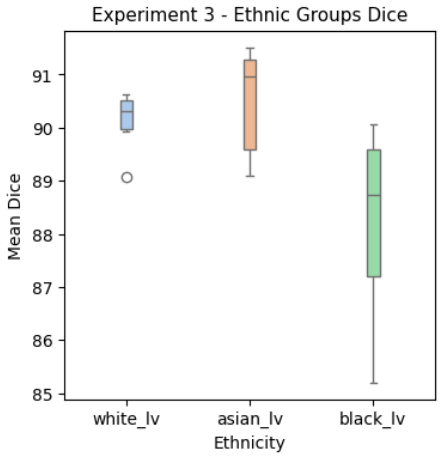


(b) Test Set 2

Figure A.23 Experiment 3 - White vs. Asian vs. Black RV

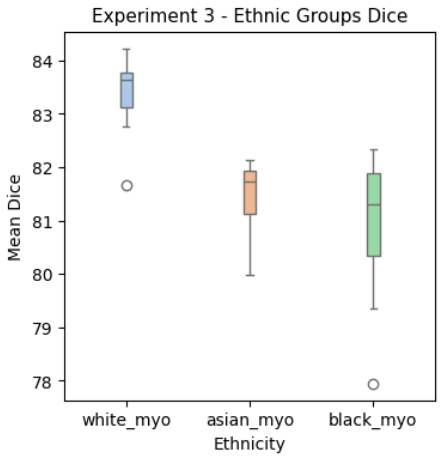


(a) Test Set 1

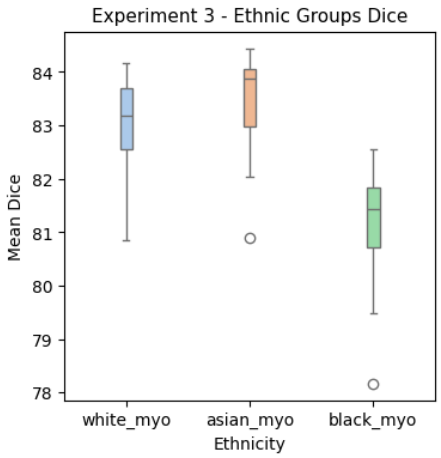


(b) Test Set 2

Figure A.24 Experiment 3 - White vs. Asian vs. Black LV



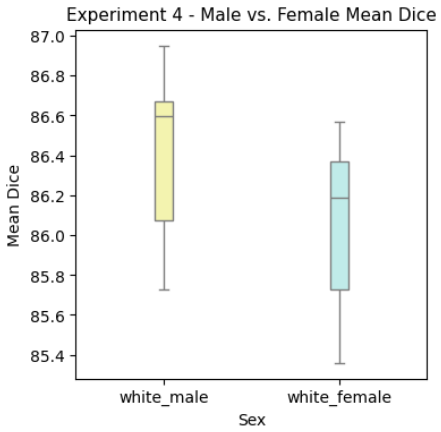
(a) Test Set 1



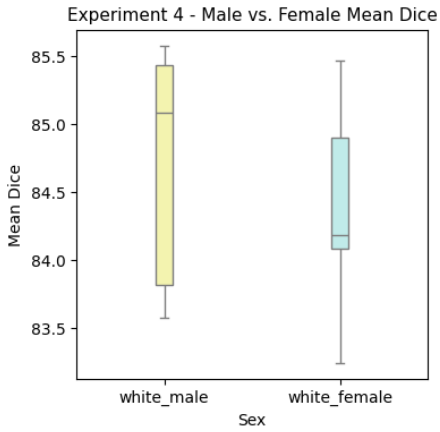
(b) Test Set 2

Figure A.25 Experiment 3 - White vs. Asian vs. Black MYO

A.7 Experiment 4 - Results

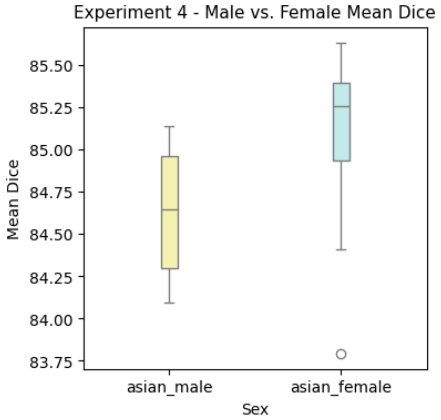


(a) Test Set 1

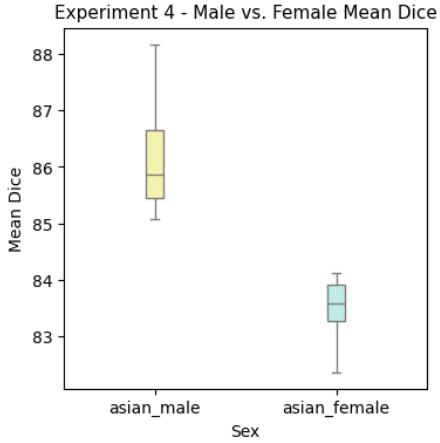


(b) Test Set 2

Figure A.26 Experiment 4 - White Males vs. Females

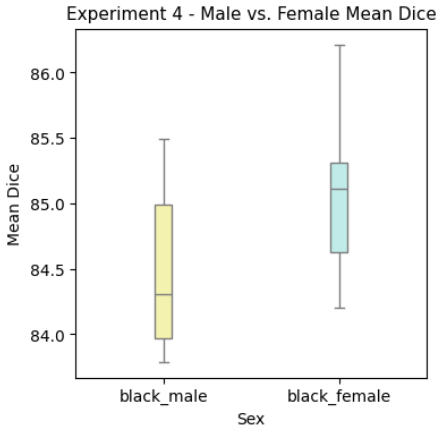


(a) Test Set 1

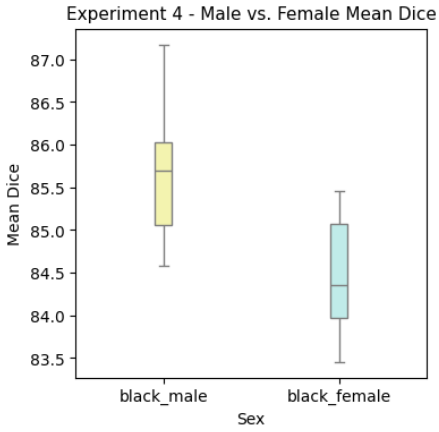


(b) Test Set 2

Figure A.27 Experiment 4 - Asian Males vs. Females

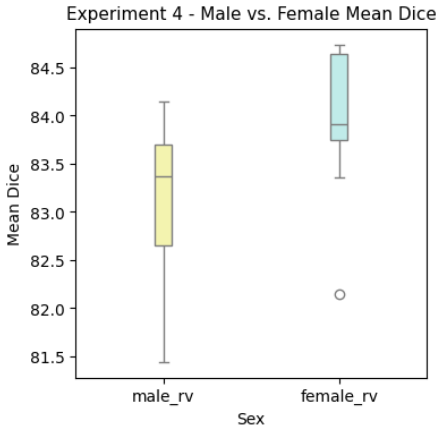


(a) Test Set 1

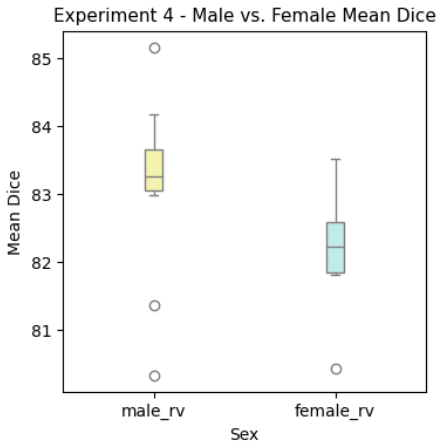


(b) Test Set 2

Figure A.28 Experiment 4 - Black Males vs. Females

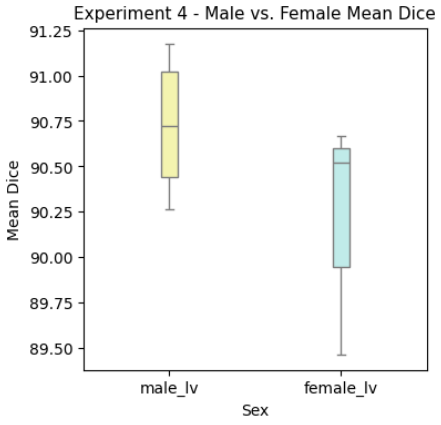


(a) Test Set 1

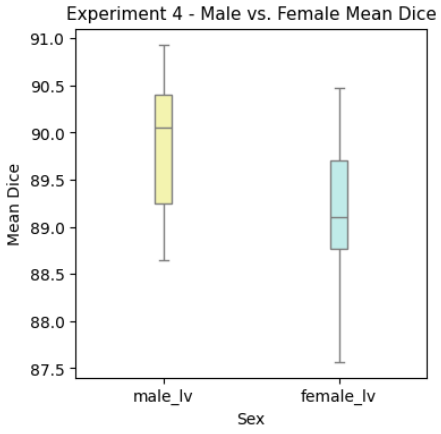


(b) Test Set 2

Figure A.29 Experiment 4 - Males vs. Females RV

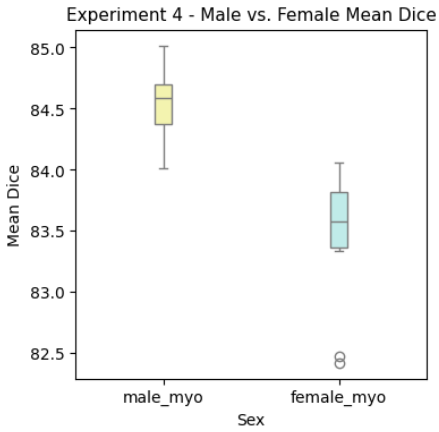


(a) Test Set 1

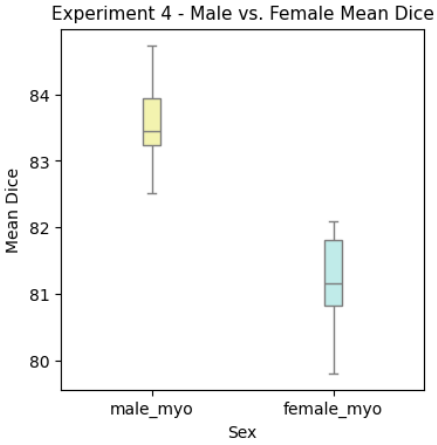


(b) Test Set 2

Figure A.30 Experiment 4 - Males vs. Females LV

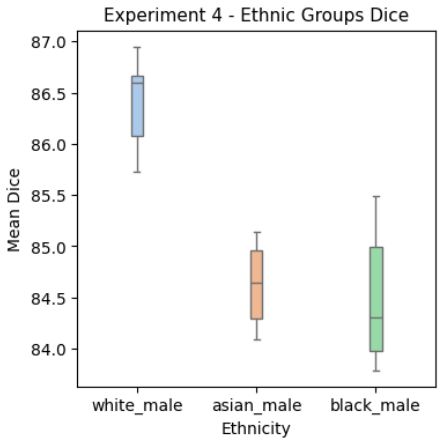


(a) Test Set 1

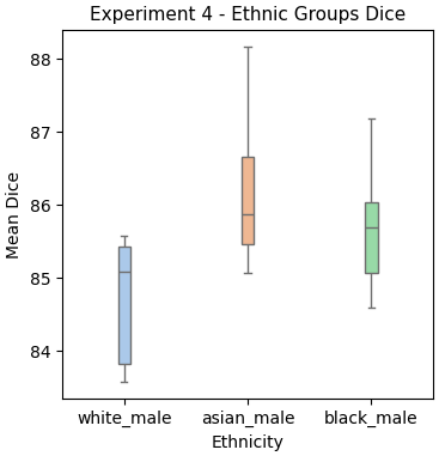


(b) Test Set 2

Figure A.31 Experiment 4 - Males vs. Females MYO

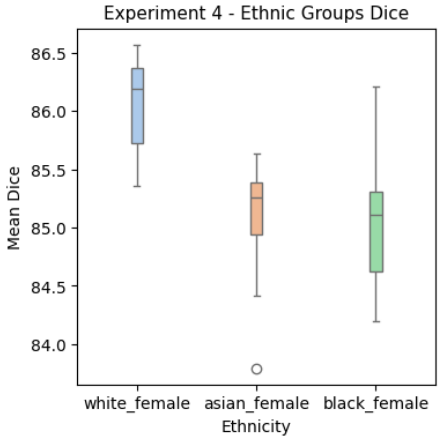


(a) Test Set 1

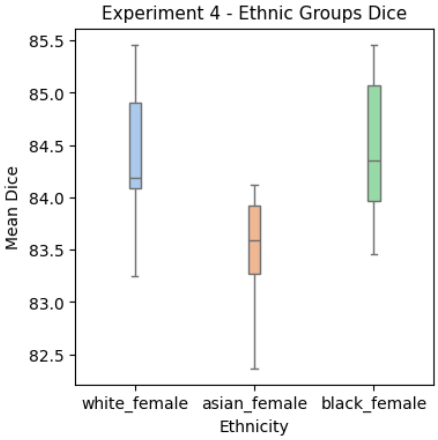


(b) Test Set 2

Figure A.32 Experiment 4 - White vs. Asian vs. Black Males

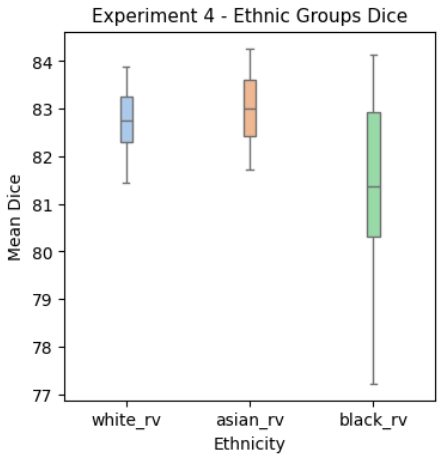


(a) Test Set 1

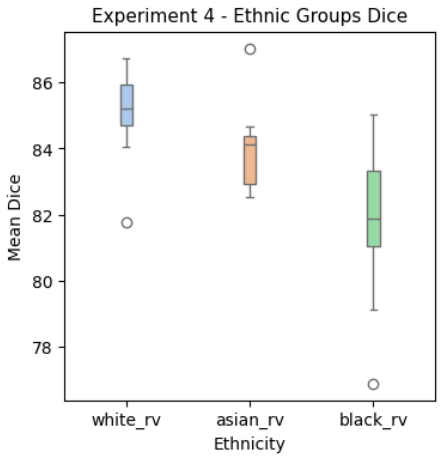


(b) Test Set 2

Figure A.33 Experiment 4 - White vs. Asian vs. Black Females

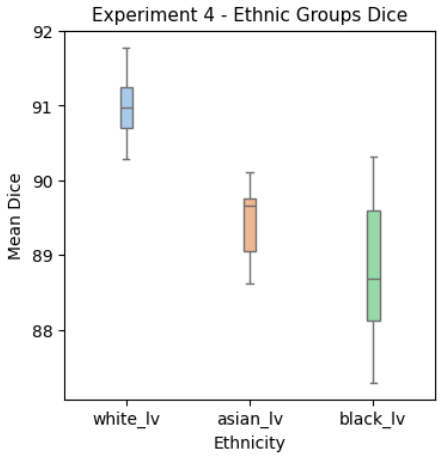


(a) Test Set 1

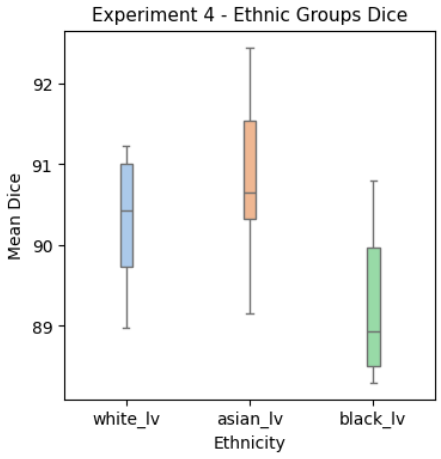


(b) Test Set 2

Figure A.34 Experiment 4 - White vs. Asian vs. Black RV

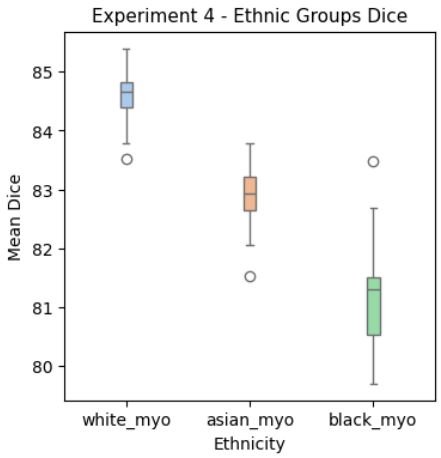


(a) Test Set 1

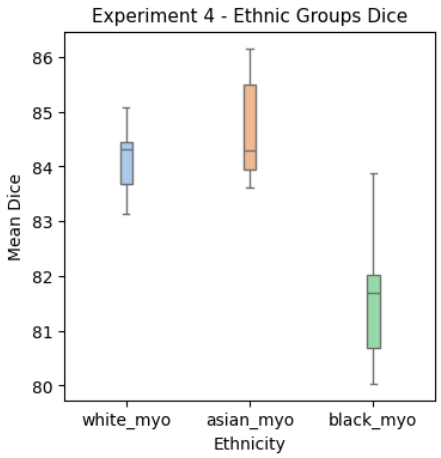


(b) Test Set 2

Figure A.35 Experiment 4 - White vs. Asian vs. Black LV



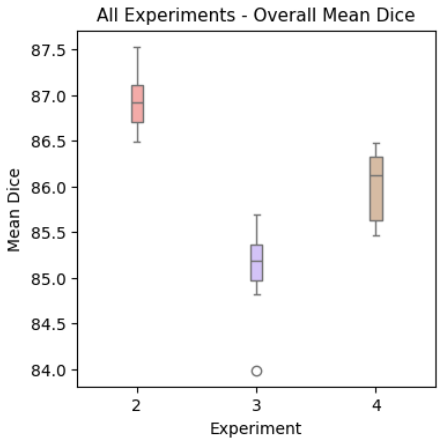
(a) Test Set 1



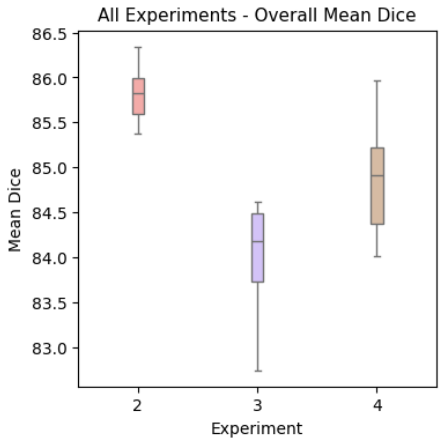
(b) Test Set 2

Figure A.36 Experiment 4 - White vs. Asian vs. Black MYO

A.8 Models Comparison



(a) Test Set 1



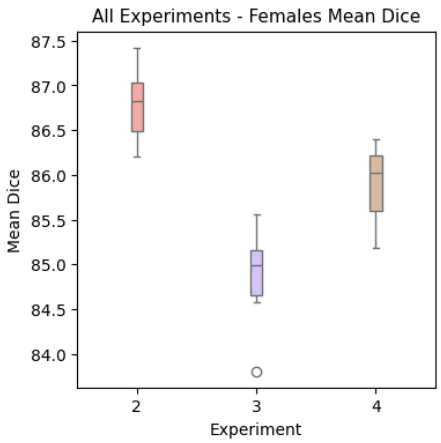
(b) Test Set 2

Figure A.37 All Experiments - Overall Mean Dice

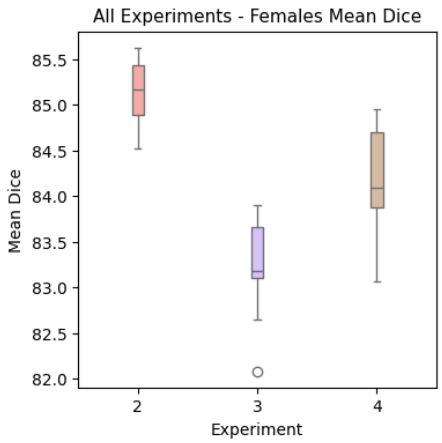
(a) Test Set 1

(b) Test Set 2

Figure A.38 All Experiments - Overall Mean Males

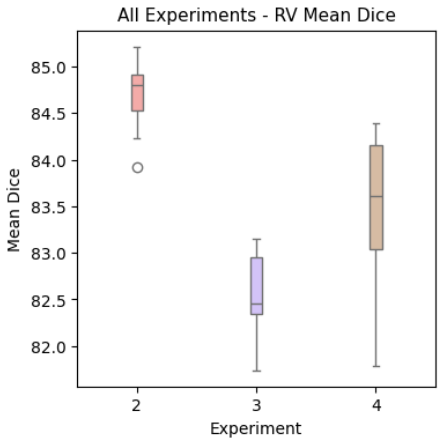


(a) Test Set 1

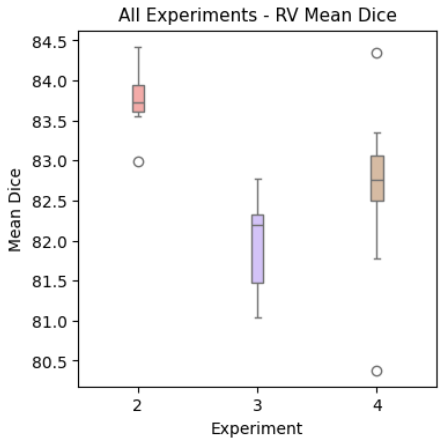


(b) Test Set 2

Figure A.39 All Experiments - Overall Mean Females Dice

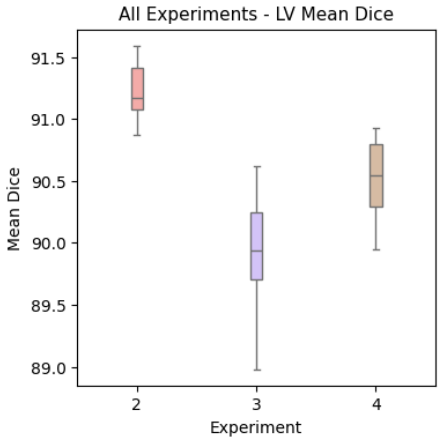


(a) Test Set 1

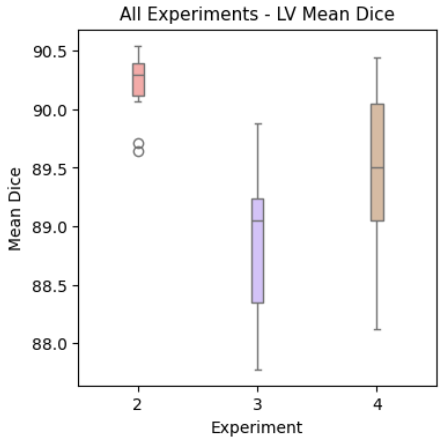


(b) Test Set 2

Figure A.40 All Experiments - RV Mean Dice



(a) Test Set 1



(b) Test Set 2

Figure A.41 All Experiments - LV Mean Dice

Bibliography

[1] Bharath Ambale-Venkatesh, Xiaoying Yang, Colin O Wu, Kiang Liu, W Gregory Hundley, Robyn McClelland, Antoinette S Gomes, Aaron R Folsom, Steven Shea, Eliseo Guallar, et al. Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis. *Circulation research*, 121(9):1092–1101, 2017.

[2] Shekoofeh Azizi, Laura Culp, Jan Freyberg, Basil Mustafa, Sebastien Baur, Simon Kornblith, Ting Chen, Patricia MacWilliams, S Sara Mahdavi, Ellery Wulczyn, et al. Robust and efficient medical imaging with self-supervision. *arXiv preprint arXiv:2205.09723*, 2022.

[3] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *Advances in neural information processing systems*, 27, 2014.

[4] Ricardo Baeza-Yates. Bias on the web. *Communications of the ACM*, 61(6):54–61, 2018.

[5] Yunhao Bai, Duowen Chen, Qingli Li, Wei Shen, and Yan Wang. Bidirectional copy-paste for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11514–11524, 2023.

[6] Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, Po-Chih Kuo, et al. Reading race: Ai recognises patient's racial identity in medical images. *arXiv preprint arXiv:2107.10356*, 2021.

[7] Hritam Basak and Zhaozheng Yin. Pseudo-label guided contrastive learning for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19786–19797, 2023.

[8] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

[9] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018.

[10] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019.

[11] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

[12] John Bridle, Anthony Heading, and David MacKay. Unsupervised classifiers, mutual information and'phantom targets. *Advances in neural information processing systems*, 4, 1991.

[13] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022.

[14] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

[15] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[16] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[17] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022.

[18] Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation. *arXiv preprint arXiv:2204.13258*, 2022.

[19] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022.

[20] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[21] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806, 2017.

[22] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 113–123, 2019.

[23] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.

[24] Rhodri H Davies, João B Augusto, Anish Bhuva, Hui Xue, Thomas A Treibel, Yang Ye, Rebecca K Hughes, Wenjia Bai, Clement Lau, Hunain Shiwani, et al. Precision measurement of cardiac structure and function in cardiovascular magnetic resonance using machine learning. *Journal of Cardiovascular Magnetic Resonance*, 24(1):16, 2022.

[25] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[26] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[27] Golnoosh Farnadi, Behrouz Babaki, and Lise Getoor. Fairness in relational domains. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 108–114, 2018.

[28] Sina Farsiu, Stephanie J Chiu, Rachelle V O'Connell, Francisco A Folgar, Eric Yuan, Joseph A Izatt, Cynthia A Toth, Age-Related Eye Disease Study 2 Ancillary Spectral Domain Optical Coherence Tomography Study Group, et al. Quantitative classification of eyes with and without intermediate age-related macular degeneration using optical coherence tomography. *Ophthalmology*, 121(1):162–172, 2014.

[29] Christopher N Ford, R Whitney Leet, LM Kipling, Mary K Rhee, Sandra L Jackson, Peter WF Wilson, Lawrence S Phillips, and Lisa R Staimez. Racial differences in performance of hba1c for the classification of diabetes and prediabetes among us adults of non-hispanic black and white race. *Diabetic Medicine*, 36(10):1234–1242, 2019.

[30] Michael Frick, Ingo Paetsch, Chiel den Harder, Marc Kouwenhoven, Harald Heese, Sebastian Dries, Bernhard Schnackenburg, Wendy de Kok, Rolf Gebker, Eckart Fleck, et al. Fully automatic geometry planning for cardiac mr imaging and reproducibility of functional cardiac parameters. *Journal of Magnetic Resonance Imaging*, 34(2):457–467, 2011.

[31] Ben Glocker, Charles Jones, Melanie Bernhardt, and Stefan Winzeck. Algorithmic encoding of protected characteristics in image-based models for disease detection. *arXiv preprint arXiv:2110.14755*, 2021.

[32] JW Goldfarb, J Cheng, and JJ Cao. Automatic optimal frequency adjustment for high field cardiac mr imaging via deep learning. In *CMR 2018–A Joint EuroCMR/SCMR Meeting Abstract Supplement*, pages 437–438, 2018.

[33] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17, 2004.

[34] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[35] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[36] Carmel Hayes, Devos Daniel, Xiaoguang Lu, Marie-Pierre Jolly, and Michaela Schmidt. Fully automatic planning of the long-axis views of the heart. *Journal of Cardiovascular Magnetic Resonance*, 15:1–2, 2013.

[37] Deborah Hellman. *When is discrimination wrong?* Harvard University Press, 2008.

[38] Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. Kiut: Knowledge-injected u-transformer for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19809–19818, 2023.

[39] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

[40] 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.

[41] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

[42] Ira Ktena, Olivia Wiles, Isabela Albuquerque, Sylvestre-Alvise Rebuffi, Ryutaro Tanno, Abhijit Guha Roy, Shekoofeh Azizi, Danielle Belgrave, Pushmeet Kohli, Taylan Cemgil, et al. Generative models improve fairness of medical classifiers under distribution shifts. *Nature Medicine*, pages 1–8, 2024.

[43] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

[44] Tiarna Lee, Esther Puyol-Antón, Bram Ruijsink, Keana Aitcheson, Miaojing Shi, and Andrew P King. An investigation into the impact of deep learning model choice on sex and race bias in cardiac mr segmentation. In *Workshop on Clinical Image-Based Procedures*, pages 215–224. Springer Nature Switzerland Cham, 2023.

[45] Tiarna Lee, Esther Puyol-Antón, Bram Ruijsink, Miaojing Shi, and Andrew P King. A systematic study of race and sex bias in cnn-based cardiac mr segmentation. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 233–244. Springer, 2022.

[46] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[47] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022.

[48] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[49] Xiangde Luo, Minhao Hu, Tao Song, Guotai Wang, and Shaoting Zhang. Semi-supervised medical image segmentation via cross teaching between cnn and transformer. In *International Conference on Medical Imaging with Deep Learning*, pages 820–833. PMLR, 2022.

[50] Geoffrey J McLachlan. Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis. *Journal of the American Statistical Association*, 70(350):365–369, 1975.

[51] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[52] Juzheng Miao, Cheng Chen, Furui Liu, Hao Wei, and Pheng-Ann Heng. Caussl: Causality-inspired semi-supervised learning for medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21426–21437, 2023.

[53] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[54] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[55] Keiron O'shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

[56] Steffen E Petersen, Paul M Matthews, Jane M Francis, Matthew D Robson, Filip Zemrak, Redha Boubertakh, Alistair A Young, Sarah Hudson, Peter Weale, Steve Garratt, et al. Uk biobank's cardiovascular magnetic resonance protocol. *Journal of cardiovascular magnetic resonance*, 18(1):8, 2016.

[57] 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.

[58] 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.

[59] Peijie Qiu, Jin Yang, Sayantan Kumar, Soumyendu Sekhar Ghosh, and Aristeidis Sotiras. Agileformer: Spatially agile transformer unet for medical image segmentation. *arXiv preprint arXiv:2404.00122*, 2024.

[60] John Rawls. *Justice as fairness: A restatement*. Harvard University Press, 2001.

[61] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[62] David A Rosman, Jean Jacques Nshizirungu, Emmanuel Rudakemwa, Crispin Moshi, Jean de Dieu Tuyisenge, Etienne Uwimana, and Louise Kalisa. Imaging in the land of 1000 hills: Rwanda radiology country report. *Journal of Global Radiology*, 1(1), 2015.

[63] Henry Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371, 1965.

[64] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[65] 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.

[66] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

[67] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[68] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

[69] Harini Suresh and John V Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2(8):73, 2019.

[70] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[71] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[72] Georgia D Tourassi. Journey toward computer-aided diagnosis: role of image texture analysis. *Radiology*, 213(2):317–320, 1999.

[73] Athanasios Tragakis, Chaitanya Kaul, Roderick Murray-Smith, and Dirk Husmeier. The fully convolutional transformer for medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3660–3669, 2023.

[74] Martin Vallières, Alex Zwanenburg, Bodgan Badic, Catherine Cheze Le Rest, Dimitris Visvikis, and Mathieu Hatt. Responsible radiomics research for faster clinical translation, 2018.

[75] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[76] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022.

[77] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[78] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11558–11567, 2023.

[79] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268, 2020.

[80] Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7236–7246, 2023.

[81] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. St++: Make self-training work better for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4268–4277, 2022.

[82] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

[83] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

[84] Haoran Zhang, Natalie Dullerud, Karsten Roth, Lauren Oakden-Rayner, Stephen Pfohl, and Marzyeh Ghassemi. Improving the fairness of chest x-ray classifiers. In *Conference on health, inference, and learning*, pages 204–233. PMLR, 2022.

[85] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[86] Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Weidi Xie, and Yanfeng Wang. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*, 14(1):4542, 2023.

[87] Xingyu Zhang, Bharath Ambale-Venkatesh, David A Bluemke, Brett R Cowan, J Paul Finn, Alan H Kadish, Daniel C Lee, Joao AC Lima, William G Hundley, Avan Suinesiaputra, et al. Information maximizing component analysis of left ventricular remodeling due to myocardial infarction. *Journal of translational medicine*, 13:1–9, 2015.

[88] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Xiaoguang Han, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Volumetric medical image segmentation via a 3d transformer. *IEEE Transactions on Image Processing*, 2023.

[89] Dominik Zietlow, Michael Lohaus, Guha Balakrishnan, Matthäus Kleindessner, Francesco Locatello, Bernhard Schölkopf, and Chris Russell. Leveling down in computer vision: Pareto inefficiencies in fair deep classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10410–10421, 2022.

[90] Yongshuo Zong, Yongxin Yang, and Timothy Hospedales. Medfair: benchmarking fairness for medical imaging. *arXiv preprint arXiv:2210.01725*, 2022.