Balanced Sampling in Ensemble Methods for Unlearning

Miriam Bakija
McGill University
miriam.bakija@mail.mcgill.ca

Abstract

One technique for machine unlearning involves training several models on separate shards of the data, so when an unlearning request is made, retraining only has to be done on a fraction of the data. One concern raised about this technique is that it exacerbates unfairness because minority data points are even fewer in number when divided into several models. One approach to improving fairness in these ensemble models, previously unresearched in the context of unlearning, is balanced sampling. I show using empirical and mathematical methods that balanced sampling works to improve fairness when used in the unlearning ensemble framework while not sacrificing much performance, and that it can possibly improve the retraining cost of unlearning.

1 Introduction

One important pillar of privacy rights that has come to the forefront of discussions in recent years is the right for an individual to remove their data upon request. In machine learning (ML) this can pose difficulty because in order to remove an individual's impact, a retraining of the whole model is required. This can be computational and time inefficient, especially if the model is trained on a lot of data and/or the individual data is large. One general strategy for improving efficiency for retraining upon receiving an unlearning request is to split the data into shards, then train models on these individual shards and only aggregate the output at the end. This means only one model will be retrained on a smaller subsection of the data, and because of the linear relationship between training time and data size, speeds up retraining time. Specifically, Bourtoule et al. (2021) proposed the SISA framework for machine unlearning, which builds on the idea of separating the data into shards, but further reduces retraining time by splitting shards into smaller slices. In this paper I mainly focus on the idea of sharding but also discuss the impacts of slicing.

While the unlearning ensemble framework speeds up time for unlearning requests, concerns have arisen that it could weaken fairness. Particularly if the data is unbalanced in terms of sensitive attribute groups, by splitting the data up into shards, data points from minority groups are fewer so overfitting is more likely for minority groups than when trained in one model all together. One fairness technique that has not been explored in the context of the unlearning ensemble framework is balanced sampling. Balanced sampling combines Random Over Sampling (ROS) and Random Under Sampling (RUS), by having each shard contain an equal number of data points from each group. This is done by over sampling data points from minority groups so they appear in multiple shards, and under sampling individuals from majority groups, such that they appear in one or zero data shards.

In this paper I will first look at balanced sampling from a theoretical stand point, showing that under a few assumptions, the cost of retraining after unlearning with balanced sampling should equal or improve when compared to a baseline ensemble method without balanced sampling. Then I will empirically examine how balanced sampling affects fairness and performance. I find that overall balanced sampling improves fairness, especially when combined with other fairness techniques, while not significantly worsening performance. In this paper I will address the following research question: Is there a way to achieve fairness in the unlearning ensemble framework without worsening the cost of retraining or performance?

2 Related Work

2.1 Machine Unlearning and SISA

The first alarms being raised about the difficulty of removing data from ML algorithms were raised in

2015 (Cao and Yang, 2015), and with the growing prevalence ML in many contexts from healthcare to finance, the need for unlearning has become ever more important to insure individual privacy rights (Nguyen et al., 2022). The concept of machine unlearning is connected to other definitions of privacy within a ML context. For example epsilon differential privacy (Dwork, 2006) guarantees an individual that output from a model trained with their data and without will be no greater than some value epsilon, with the lower the epsilon the more privacy guaranteed. Exact unlearning on the other hand holds the guarantee that the model gives the same outputs as if it were not trained on an individuals data at all, a much stronger privacy guarantee than epsilon differential privacy (Kurmanji et al., 2024). The most obvious way to achieve this is to retrain the whole model from scratch without the individual's data, though this is time and resource expensive.

2.2 Fairness in Machine Unlearning

Unfairness caused by the unlearning ensemble framework, and ways to mitigate it has been researched, though mostly in the context of unfairness caused by class imbalance. Koch and Soll (2023) showed that even when using bias mitigation techniques like ROS, RUS, and in-processing regularization techniques, the SISA framework exacerbates performance disparities for minority classes. Yan et al. (2022) showed an improvement for SISA involving representative sampling and single-class shards can improve overall accuracy of the model as well as fairness in terms of performance for different classes. However both of these papers examine fairness in the context of class imbalance for classification tasks. To my knowledge, fairness for ensemble unlearning has not been studied in the context of non-classification tasks for imbalance in sensitive attribute groups.

2.3 Balanced Sampling

Balanced sampling is a technique to improve fairness in ensemble learning that has been utilized in many different areas. Gu and Song (2015) proposed balanced sampling as a way to improve fairness in ensemble models using Ada-Boost. It is utilized commonly as as a strategy for fairness in credit-scoring algorithms (Qian et al., 2021; Moscato et al., 2021), but remains unexplored in other areas.

3 Speed-up of Balanced Learning

To show how balanced sampling affects the speedup of unlearning requests I will make the same assumption made by Bourtoule et al. (2021), that training time is directly proportional to dataset size.

Let N be the size of the dataset, M the number of shards, K the number of sensitive attribute groups, and $[k_1; k_2 \cdots k_1 \ldots k_k]$ be the set of percentages for sensitive attribute groups, such that $\sum_{i=0}^{k} k_i = 1$.

Under balanced sampling, each shard needs to contain $\frac{N}{M \cdot K}$ data points from group, and over M shards, we need $\frac{N}{K}$ data points from each group. A group has $k_i N$ data points, so $\frac{N}{K} \div k_1 N = \frac{1}{k_i K}$ is the maximum number of shards a data point from group i will need to appear in. Note that if $k_i > \frac{1}{K}$, data points from group i will appear in 0 or 1 shards.

Here a strict condition for balanced sampling I will enforce from now on is that $\min(k_i) \cdot K \geq \frac{1}{M}$, meaning no data point will appear in a given shard more than once. This is essential as 1) if a data point appears in a shard more than once the model is more likely to overfit on that group, meaning fairness for the smallest minority group decreases and 2) in order to extend speedup calculations to the SISA architecture, we want to hold that the speedup for slicing within shards is the same with and without balanced sampling, which is only true when each data point is in a shard at most once.

3.1 Sequential Unlearning Requests

First I will consider the case of sequential unlearning requests, where each unlearning request is received separately. Under the assumption that after an unlearning request the size of the shard remains relatively unchanged, the expected amount of retraining time is directly proportional to the number of shards containing the data point. Without balanced sampling, each data point appears in exactly one shard, so the retrained data is $\frac{N}{M}$. This is more complex with balanced sampling, however the expected amount of data to retrain on remains the same. We have a k_i probability of retraining $\frac{1}{k_i k}$ shards so the expected number of shards retrained after one unlearning request is:

$$E[c] = \sum_{i=0}^{k} k_i \left(\frac{1}{k_i K} \right) = 1$$

so the expected amount of data to retrain on is still $\frac{N}{M}$. Notably though the lower bound decreases to 0 , and the upper bound increases to $\frac{1}{\min(R_1)\cdot K}$, but on average the cost of retraining should be the same with and without balanced sampling.

3.2 Batched Unlearning Requests

If unlearning requests are batched together the expected cost equation becomes more complicated. I explore this case in Appendix B. Overall I find evidence that for a batched unlearning request of size H, balanced learning either equals or reduces the cost of retraining

4 Methodology

4.1 Dataset

In order for the unlearning ensemble framework to be desirable in a given context two conditions need to be fulfilled. The first of these is that the cost of training needs to be expensive enough to warrant a potential loss in performance. The second is that the context should require some need for privacy and where machine unlearning may be required. I chose to use the FairSeg dataset Tian et al., 2023 because the context fulfills these two conditions. The FairSeg dataset contains images of ocular scans from 10,000 patients, as well as information about several sensitive attributes including Race and Gender. The dataset is benchmarked on the task of medical segmentation, where the model receives human annotated medical images that outline two anatomical parts of the eye, the cup and the rim, and it aims to replicate these annotations. To compare fairness techniques I chose to focus on Race as a sensitive attribute because out of all the sensitive attributes in the dataset, it had the biggest disparity in performance when no fairness techniques were used, as seen in Tian et al.'s (2023) benchmark. The dataset is 9.19% Asian, 14.73% Black, and 76.08% White.

Medical segmentation is an important task, as it is often a the first step in machine learning models that are designed to diagnose individuals based on medical images. Specifically, ocular scans are used to diagnose diseases such as glaucoma, and worse performance on the medical segmentation task can lead to worse accuracy on disease prediction. This highlights the importance of achieving demographic parity or equally high performance for all sensitive attribute groups in the medical segmentation task. Further, medical segmentation is a task that is computationally expensive to train on because it makes predictions on each individual pixel in the image. Finally the option to remove your data in the context of medicine is valuable to give patients more control over their privacy (Kaissis et al., 2020).

4.2 Training details

In order to understand balanced sampling's effects on group fairness and accuracy, I run four ensemble models under different conditions. The first is a base, which I call reg, where data is randomly split up into shards. The next is batch sampling, which employs the strategy of oversampling minority groups and under sampling majority groups to create equal representation shards. Also, starting from the saved weights at the 20th epoch from the two previous models, I fine tuned over 5 epochs using the fairness regularizer FEBs Tian et al., 2023. Each ensemble model consists of 10 identical models, each trained using the architecture and pretrained weights of the Segment Anything Model over 25 total epochs, and aggregated using prediction vector aggregation (see Appendix D for more information on the Segment Anything Model and FEBs). I used the same hyperparameters as the benchmark from Tian et al. (2023), with the exception of batch size and epochs, which I discuss in the results section.

5 Results

The four ensemble models are compared against the benchmark monolith models to show how using the ensemble unlearning framework affects performance. Dice score is a performance metric that measures pixel-wise correct predictions and is typically used to measure performance on medical segmentation tasks (Azad et al., 2023). ES-Dice (Tian et al., 2023)is a fairness adjusted Dice score which takes the original Dice score and reduces it if there is large disparity in group performances. The benchmark monolith models were trained for 100 epochs on a batch size of 24, while the ensemble models were trained each for 25 epochs on a batch size of 12. The cost of training a monolith model is very computationally and time expensive, which on the one hand emphasizes that the cost of retraining for unlearning is high, but meant that I could not recreate the monolith model for the same batch size and number of epochs as the ensemble models, due to limited resources. In terms

of batch size because it only differs by a magnitude of 2, the difference in performance shouldn't change that much. The one way I consider it might change is increasing the effectiveness for smaller batch of the FEBs regularizer, as it regularizes over batches. In order to see how the reduced number of epochs affected performance, in Figure 1 (Appendix A) I graphed performance over the first 25 for the ensemble models. Overall Rim Dice starts to plateau after 15 epochs, while Cup Dice steadily climbs, though the Rim Dice score is boosted by introducing FEBs in the last 5 epochs. By epoch 25, comparing to the monolith model's 100th epoch, for Cup Dice there is only a very small performance gap, which could probably close over more epochs. As for Rim Dice, there is a noticeable gap between the monolith model Dice and the ensemble models without FEBs, but when including FEBs this gap gets much smaller.For the fairness adjusted Cup ES-Dice score, we see that balanced sampling well outperforms the ensemble model without it, and is about on par with the monolith. Interestingly Rim ES-Dice is highest with balanced sampling and FEBs slightly even over the monolith model. Overall the results from Figure 1 suggest that the combination of balanced sampling and FEBs performs the best in terms of fairness for ensemble models, and is about on par with performance of the monolith models. In appendix C tables 1 and 2 show full results from the last epoch for all racial groups. The main takeaway from these tables is that the order of performance between the groups does not significantly change, except that Black Rim Dice is higher than Asian Rim Dice in the ensemble models, and vice versa in the monolith models.

6 Limitations and Future Directions

6.1 Privacy and Fairness

One important concern to consider when using balanced sampling is that by oversampling from minority groups, this could potentially make members of the minority groups more susceptible to membership inference attacks, decreasing their privacy. While using ensemble methods can reduce success of membership attacks because of the aggregation step (Chen et al., 2021), it is nonetheless essential to ensure privacy is not decreased as this is exactly counter to the goal of unlearning. A future path of research could explore how balanced sampling affects success of membership attacks for minority and majority groups, and if differential privacy techniques could equalize privacy risks for all groups.

6.2 Unlearning Distribution and Fairness

Another consideration that must be made is that so far we have assumed that the distribution of unlearning requests is random, but this is likely not the case. Koch and Soll (2023) argues that unlearning requests are usually not evenly distributed and provides evidence that young and/or rich individuals have more privacy awareness, and are more likely to remove their data. Overall correlations between unlearning likelihood and sensitive attribute groups, specifically ones considered in balanced sampling, raise ethical questions about equal access to privacy that should be considered within the context of the situations unlearning is used in.

6.3 Robustness and Unlearning

An additional aspect to consider is how the robustness of the model is affected by balanced sampling. (Qian et al., 2023) found that the SISA framework was more vulnerable than monolith models to robustness attacks without defense mechanisms. The success of these robustness attacks could increase under balanced sampling because some data points appear multiple times, so future research could examine how balanced sampling affects robustness.

7 Conclusion

In this paper I explore the potential to use balanced sampling to improve fairness in an ensemble unlearning framework. I find theoretical evidence that the cost of retraining for unlearning requests should either remain the same or decrease under balanced sampling compared to random selection. I also find empirical evidence that balanced sampling improves group fairness without sacrificing significant performance. Overall balanced sampling in an ensemble unlearning framework has the potential to be a useful technique for improving sensitive attribute group fairness, but should not be implemented in any real context until further research is conducted on the possible risks mentioned in the limitations section.

8 Key Learnings

In this project I learned methods for research in computer science, and how to read and understand academic papers in the field of Responsible AI.

I also learned how to outline and write my first computer science project paper following standard guidelines. In terms of specific educational knowledge I gained, I learned more about the intersection of privacy and fairness in machine learning, and pre-processing and in-processing fairness techniques.

References

Reza Azad, Moein Heidary, Kadir Yilmaz, Michael Hüttemann, Sanaz Karimijafarbigloo, Yuli Wu, Anke Schmeink, and Dorit Merhof. 2023. Loss functions in the era of semantic segmentation: A survey and outlook. *arXiv preprint arXiv:2312.05391*.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159. IEEE.

Yinzhi Cao and Junfeng Yang. 2015. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pages 463–480. IEEE.

Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang. 2021. When machine unlearning jeopardizes privacy. In *Proceedings of the 2021 ACM SIGSAC conference on computer and communications security*, pages 896–911.

Cynthia Dwork. 2006. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer.

Hong Gu and Tao Song. 2015. Balanced sampling method for imbalanced big data using adaboost. In *International Conference on Bioinformatics Models, Methods and Algorithms*, volume 2, pages 189–194. SCITEPRESS.

Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. 2020. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6):305–311.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026.

Korbinian Koch and Marcus Soll. 2023. No matter how you slice it: Machine unlearning with sisa comes at the expense of minority classes. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 622–637. IEEE.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. 2024. Towards unbounded machine unlearning. *Advances in Neural Information Processing Systems*, 36.

Vincenzo Moscato, Antonio Picariello, and Giancarlo Sperlí. 2021. A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications*, 165:113986.

Thanh Tam Nguyen, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. 2022. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*.

Hongyi Qian, Shen Zhang, Baohui Wang, Lei Peng, Songfeng Gao, and You Song. 2021. A comparative study on machine learning models combining with outlier detection and balanced sampling methods for credit scoring. *arXiv preprint arXiv:2112.13196*.

Wei Qian, Chenxu Zhao, Wei Le, Meiyi Ma, and Mengdi Huai. 2023. Towards understanding and enhancing robustness of deep learning models against malicious unlearning attacks. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1932–1942.

Yu Tian, Min Shi, Yan Luo, Ava Kouhana, Tobias Elze, and Mengyu Wang. 2023. Fairseg: A large-scale medical image segmentation dataset for fairness learning with fair error-bound scaling. *arXiv preprint arXiv:2311.02189*.

Haonan Yan, Xiaoguang Li, Ziyao Guo, Hui Li, Fenghua Li, and Xiaodong Lin. 2022. Arcane: An efficient architecture for exact machine unlearning. In *IJCAI*, volume 6, page 19.

A Appendix

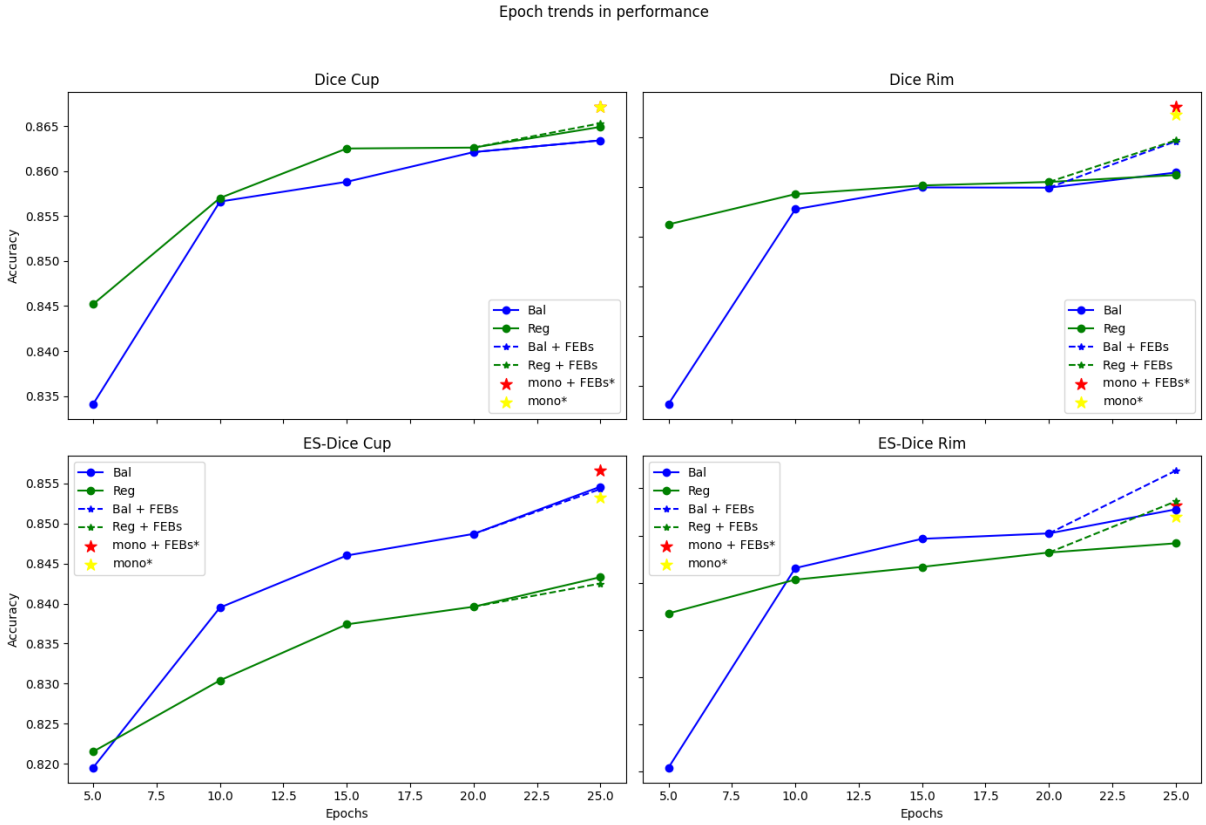

Figure 1: Graphs of how selected performance scores change over epochs, compared to monolith models

B Appendix

B.1 Batched Unlearning Requests

Lets say we receive $H \quad i.i.d.$ unlearning requests. We want to know the expected number of shards affected, and how many times each shard is affected to calculate cost of retraining. We will work under the assumption again that $\frac{N}{M} - H \approx \frac{N}{M}$, that is the number of unlearning requests is trivial in comparison to the size of a shard. In this case the cost of retaining is $C = \frac{N}{M} \cdot p$, where p is the number of shards retained.

$$E[c] = \frac{N}{M} E[p]$$

To understand how balanced sampling affects $E[p]$, lets look first at the case where $H = 2$. The expected number of shards to retrain is the expected number of shards affected by each individual data point, minus the expected number of shards that contain both data points.

$$E\left[p \middle| H = 2\right] = E\left[p_1\right] + E\left[p_2\right] - E\left[\frac{P_1 \cdot P_2}{M}\right]$$

from sequential learning requests I show that $E\left[p \middle| H = 1\right] = 1$ with and with out balanced sampling. Next we need to find what $E\left[p^2\right]$ is

$$E\left[p^2\right] = \sum_{i=0}^{K} k_i \left(\frac{1}{k_i K}\right)^2 E\left[p^2\right] = \frac{1}{K} \sum_{i=0}^{K} \frac{1}{k_i}$$

By the harmonic mean geometric mean inequality

$$\sum_{i=0}^{k} \frac{1}{k_i} \geq K$$

So the expected number of repeat shards depends on the distribution of groups in the population. But what we can infer from this is that the more imbalanced the classes are, the larger $\sum_{i=0}^{k} \frac{1}{k_1}$ is. In the ease with no balanced sampling, this is equivalent to $K = 1$ groups, where $k_1 = 1$

$$E\left[\frac{p_1 \cdot p_2}{M}\right] = \frac{1}{M}$$

Lets compare this to an example case for balanced sampling where $K = 3$, $k_1 = 0.7$, $k_2 = 0.2$, and $k_3 = 0.1$

$$E\left[\frac{p_1 \cdot p_2}{M}\right] \approx \frac{5.47}{M}$$

This means that for a batch size of 2 :

$$E[p|H = 2]_{normal} \geq E[p|H = 2]_{balancedsampling}$$

So the expected cost of retraining is the same or less with balanced sampling of size 2. This becomes more complex when we receive a batch size of H Using induction and the inclusion-exclusion principle I derive

$$E[p] = \sum_{i=1}^{H} (-1)^{i+1} Hi \cdot \frac{\sum_{j=0}^{K} \left(\frac{1}{k_j}\right)^{i-1}}{K^i M^{i-1}}$$

This expectation would grow unstably, especially for very small values of k_j as $\frac{1}{\min(k_j)}$ would grow at a much faster rate then K^{i-1}. This is when the strict bound of $\min\left(k_j\right) \cdot K \geq \frac{1}{M}$, so these terms grow at a reasonable rate. Because this is a complex expectation and hard to reduce I opted to code random experiments. with random choices for K, M and H in the following ranges

$$K = [2, 10], M = [3, 30] \quad H = [2, 2 \cdot M]$$

and creating random k_j which fulfilled the conditions. I graphed the ratio of $E[p]$ of normal against balanced sampling over 10,000 random episodes.

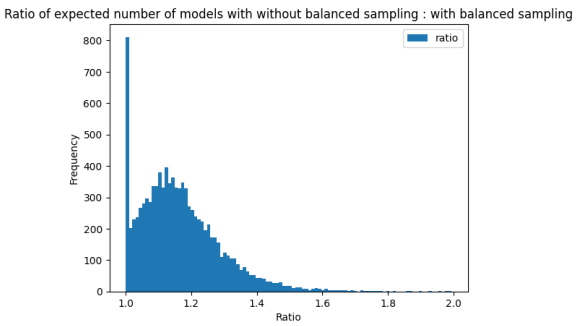

Figure 2: Ratio of Expected number of shards retrained with and without balanced sampling over 10,000 random episodes

As seen in figure 2, the minimum ratio is 1, meaning the expected number of shards to retrain, and therefore the expected cost is at least equal if not better with balanced sampling. Although this is not a strict proof that balanced sampling equals or reduces cost to retrain it provides promising evidence that could be tested empirically in future research. It is important also investigate how much the cost of retraining is reduced, and consider that in certain contexts where we want limit the upper bound of how much retraining we have to do, balanced sampling may not be a good choice.

C Appendix

Table 1: Optic Cup segmentation performance on FairSeg dataset

Method	Overall Dice	Es-Dice	Asian Dice	Black Dice	White Dice
Regular	0.8649	0.8433	0.8477	0.8594	0.8679
Balanced Sampling	0.8634	0.8546	0.8550	0.8626	0.8644
Regular + FEBS	0.8653	0.8425	0.8484	0.8584	0.8685
Balanced Sampling + FEBS	0.8634	0.8543	0.8556	0.8618	0.8646
Monolith	0.8671	0.8532	0.8568	0.8730	0.8670
Monolith + FEBS	0.8671	0.8566	0.8587	0.8708	0.8672

Table 2: Optic Rim segmentation performance on FairSeg dataset

Method	Overall Dice	Es-Dice	Asian Dice	Black Dice	White Dice
Regular	0.8049	0.7367	0.7569	0.7721	0.8166
Balanced Sampling	0.8059	0.7511	0.7672	0.7809	0.8151
Regular + FEBS	0.8189	0.7545	0.7784	0.7851	0.8300
Balanced Sampling + FEBS	0.8159	0.7675	0.7868	0.7924	0.8271
Monolith	0.8291	0.7478	0.7890	0.7758	0.8444
Monolith + FEBS	0.8323	0.7529	0.7952	0.7789	0.8473

D Appendix

D.1 Segment Anything Model

The segment anything model Kirillov et al., 2023 is a state of the art model that is very popular for medical segmentation tasks, though it is designed for any image segmentation.

D.2 FEBs

FEBs or fair error bounded scaling is a in-processing fairness technique introduced by Tian et al., 2023. It works by changing the making the weight of the loss for individuals heavier if they are in the group with the highest Dice loss. In the benchmark dataset it was compared to several other fairness techniques like adversarial fair representations, which both did well on different sub tasks.