AAMI Report – Group Project

Sondre Sigstad Wikberg (Other Group Members: XXXXXX, YYYYYY)
University College London, London, UK

1 Introduction

This project implements and compares extensions of a baseline deep learning model for optic disc and cup segmentation using the Chákṣu database, which contains 1,345 colour fundus images from Indian subjects captured on three devices: 810 Remidio (resolution 2448×3264), 95 Forus (2048×1536), and 104 Bosch (1920×1440) images in the training set, with 264, 31, and 41 images respectively in the test set [1]. Each image is annotated by five ophthalmologists with pixel-wise segmentations of the optic disc (annotation intensity 127), optic cup (255), and background (0), along with glaucoma assessments. These annotations are aggregated via pixel-wise mean, median, majority vote, and the STAPLE algorithm [2][1]. The dataset also includes individual and aggregated measurements of disc and cup area, width, height, and derived cup-to-disc ratios. Predefined training and test splits are provided by the dataset authors.

1.1 Organisation and Group Logistics

We designed a baseline U-Net architecture for segmentation and implemented a series of extensions to address various challenges. These included oversampling underrepresented devices to address class imbalance; a multi-task model predicting segmentation, majority (across experts) glaucoma status, and mean (across experts) cup-to-disc ratios; GradCAM applied to the final encoding layer for explainability; and uncertainty estimation using Monte Carlo dropout in the same layer. To reduce bias, a model using random image augmentations during training was also trained. Finally, a combined "mixture" model incorporated augmentations, oversampling, multi-task learning, and uncertainty estimation.

XXXXXXXX implemented the baseline architecture and YYYYYYYY developed preprocessing tools for augmentation and oversampling and contributed the foundation for uncertainty estimation. The author of this report implemented the multi-task and mixture models and handled GPU-based model training. All group members jointly developed the dataset class.

Imbalance and Bias There were differing views in the group on defining imbalance and bias. This author understood imbalance as arising from unequal sample sizes across devices or glaucoma classes, and bias as the presence of training-specific features that may not generalise (i.e., the bias in the bias-variance trade-off). Others viewed bias as the performance drop when training and testing on different devices (domain shift). The author acknowledges XXXXXXX for designing models to evaluate this cross-device performance gap. While bias in test results may stem from imbalance, this author sees imbalance as requiring balanced sampling, and bias as requiring augmentation to reduce overfitting.

[1] STAPLE is a probabilistic method commonly used in medical imaging to combine expert segmentations

2 Methods

2.1 Label Selection, Preprocessing, and Training Procedure

We selected the STAPLE-combined segmentations as ground truth to leverage consensus across experts while reducing memory usage. In a custom dataset class, all images and STAPLE labels were first resized such that the shorter side was scaled to 256 pixels using bilinear interpolation, with the longer side scaled proportionally to preserve aspect ratio. Interpolation was also applied to the STAPLE labels, as they are continuous-valued probabilistic maps ranging from 0 to 255. The resized images and labels were then centre-cropped to 256×256. Finally, the STAPLE label was discretised into three classes: background (BG), disc, and cup. Each pixel was assigned to one of these based on intensity: BG if $c_{i,j} \in [0, 64)$, disc if $c_{i,j} \in [64, 192)$, and cup if $c_{i,j} \in [192, 255]$.

All models were defined and trained using PyTorch 2.6.0 with CUDA 12.4 on an NVIDIA A100 via Google Colab. Training and validation were performed over 10 epochs on a fixed 80/20 further split of the 1009-sample Chákṣu training set. The resulting training set comprised 642 Remidio, 88 Bosch, and 77 Forus images. At each epoch, predictions were made per pixel by selecting the class with the highest activation. Dice scores for each class were averaged over the validation set, and model weights from the epoch with the lowest average dice (across all three classes) were saved. Multi-class cross-entropy loss was used, with class weights set as the normalised inverse pixel-wise frequency to counter background pixel imbalance. A batch size of 4 was used throughout.

2.2 Baseline Architecture

The baseline model follows a U-Net architecture consisting of four encoding and four decoding levels, connected via skip connections. In the encoder, spatial resolution is halved at each level using max-pooling, while the number of feature channels increases progressively to 512. The first encoding block uses 5×5 convolutional kernels to capture larger context early on, while subsequent layers use 3×3 kernels. Each block contains two convolutional layers followed by batch normalization and LeakyReLU activations. The bottleneck uses transposed convolutions to begin upsampling, and dropout is applied to its input for regularization. In the decoder, spatial dimensions are doubled at each level, while feature channels progressively decrease. The segmentation map is generated by a 1×1 convolution, producing logits for the three classes.

2.3 Contributions by Other Group Members

To address class imbalance, an augmentation model assigned each training sample a weight inversely proportional to the number of samples from its device type. Sampling was then performed with replacement using these weights, resulting in approximately equal representation of each device in each epoch when training on the baseline architecture.

To mitigate bias and encourage generalisation, an augmentation model applied random augmentations during training: horizontal flips (probability 0.5),

vertical flips (0.2), and gamma adjustments (0.3, range [0.6, 1.4]) during training on the baseline architecture.

Explainability was introduced using Grad-CAM. Functionality was built to generate a class activation map (CAM) by computing gradients of the class-specific outputs of a model with respect to the final decoding feature maps, globally averaging them, and combining them with the activations. The resulting CAMs could be upsampled and overlaid on the original image to visualise influential regions for each class.

Finally, a Monte Carlo dropout function was used to estimate model uncertainty. By retaining dropout layers at test time and averaging softmax outputs across a given number of forward passes, majority vote segmentation and predictive entropy could be computed to highlight regions of low model confidence.

2.4 Multitask Model

A clear imbalance was observed between subjects with and without glaucoma in the training set. Additionally, the mean cup-to-disc area (ACDR), vertical (VCDR), and horizontal (HCDR) ratios were significantly higher in glaucoma cases, as shown in Table 1, with significant differences confirmed via Welch's t-tests. Since these ratios reflect anatomical features that segmentation is expected to capture, they were used as auxiliary regression targets in a multitask extension of the baseline model. The multitask model fed globally pooled bottleneck

Ratio Type	**All Subjects (n=807)**	**Glaucoma (n=102)**	**Normal (n=705)**	**T-test**
ACDR	0.217 ± 0.095	0.341 ± 0.084	0.199 ± 0.082	$p = 3.85 \times 10^{-32}$
VCDR	0.461 ± 0.097	0.590 ± 0.066	0.442 ± 0.086	$p = 1.72 \times 10^{-45}$
HCDR	0.450 ± 0.106	0.570 ± 0.083	0.433 ± 0.097	$p = 6.72 \times 10^{-32}$

Table 1. Mean ± standard deviation of ACDR, VCDR, and HCDR in training set . Welch's t-test p-values compare the glaucoma and normal groups.

features into a regression head predicting the three cup-to-disc ratios. These regression outputs, after gradient detachment (to avoid the classifier dominating the regressor), were concatenated with the pooled features and passed to a classification head for glaucoma status. The regression and classification outputs were further projected back into the decoder and injected as 64 spatial maps to support the segmentation task. Importantly, the weights of both the regressor and classifier were updated not only through their respective loss terms, but also indirectly through gradients from the segmentation loss via this decoder injection. The total loss combined three terms: multi-class cross-entropy for segmentation, mean squared error (MSE) for the regression targets, and positively weighted binary cross-entropy for glaucoma classification (with the positive weight equal to the ratio of negative to positive training samples). As the initial validation losses for regression and classification were an order of magnitude smaller than that for segmentation, the final loss weighting was set as: $5 \cdot \mathcal{L}_{\text{seg}} + 1 \cdot \mathcal{L}_{\text{cls}} + 5 \cdot \mathcal{L}_{\text{reg}}$.

2.5 Mixture Model

The final mixture model was designed to jointly address bias, imbalance, and task complexity. It extended the multitask model—predicting both segmentation

and auxiliary glaucoma targets—while also incorporating both random augmentations (as described above) and a modified sampling strategy. Each sample was weighted by the mean of two inverse-frequency weights: one based on its device type and the other on its glaucoma status, ensuring balanced sampling across both dimensions. Sampling based on Glaucoma status was introduced to account for imbalance in results of the oversampling model, which generally had the best test-results among the earlier models. Finally, uncertainty estimation and majority vote per pixel using Monte Carlo dropout can be applied at test time, as dropout was retained in the final encoder layer.

3 Results

Table 2 presents Dice and HD95 (Hausdorff 95% score) metrics for each model, averaged over the full test set (336 samples), stratified by glaucoma status (51 positive), and over the Forus samples only. Dice provides a measure of spatial overlap between predicted and ground truth segmentations, while HD95 reflects the extent of the worst misclassifications by measuring the 95th percentile of boundary distances. This makes HD95 useful for identifying outlier errors and boundary inaccuracies. The Forus-specific results reflect performance on the least represented device (31 samples), offering insight into domain generalisation. Bold values indicate the best score per metric. The oversampling model achieved the most best-in-column scores (14), followed by the mixture model (8). Oversampling consistently reduced HD95 and improved Dice for the disc relative to the baseline, suggesting both more accurate and more spatially coherent predictions.

All Subjects

Model	BG	Disc	Cup	BG	Disc	Cup
Baseline	0.997	0.812	**0.899**	14.25	23.35	7.15
Oversampled	**0.998**	**0.852**	0.894	**9.71**	**11.56**	7.04
Multi-task	0.997	0.811	0.872	15.56	30.30	12.77
Uncertainty	**0.998**	0.831	0.897	14.72	27.03	8.72
Augmented	0.993	0.763	0.810	30.77	38.25	37.56
Mixture	**0.998**	0.839	0.897	12.35	16.89	**6.80**

Glaucoma Only

Model	BG	Disc	Cup	BG	Disc	Cup
Baseline	0.997	0.796	**0.929**	**11.57**	**13.52**	6.36
Oversampled	**0.998**	**0.820**	0.906	15.17	19.68	10.64
Multi-task	0.997	0.807	0.912	15.03	20.12	14.16
Uncertainty	**0.998**	0.812	0.920	14.80	25.12	**5.52**
Augmented	0.993	0.744	0.834	29.92	31.06	28.37
Mixture	**0.998**	0.814	0.909	15.74	17.23	5.94

Normal Only

Model	BG	Disc	Cup	BG	Disc	Cup
Baseline	0.997	0.815	0.893	14.72	25.11	7.29
Oversampled	**0.998**	**0.858**	0.892	**8.73**	**10.11**	**6.40**
Multi-task	0.997	0.812	0.865	15.65	32.12	12.52
Uncertainty	**0.998**	0.834	0.893	14.71	27.37	9.29
Augmented	0.993	0.767	0.806	30.92	39.53	39.21
Mixture	**0.998**	0.843	**0.895**	11.74	16.82	6.96

Forus Only

Model	BG	Disc	Cup	BG	Disc	Cup
Baseline	0.997	0.856	0.905	20.23	39.25	12.86
Oversampled	**0.999**	**0.891**	0.887	**5.58**	10.51	14.41
Multi-task	0.998	0.852	0.891	15.22	63.31	19.94
Uncertainty	0.998	0.860	0.892	27.25	60.05	20.96
Augmented	0.991	0.753	0.809	39.82	76.83	61.60
Mixture	0.998	0.882	**0.896**	11.83	**12.21**	**10.83**

Table 2. Dice Score and pixel-wise HD95 (lower is better) across groups and models. Uncertainty had 5 forward passes; metrics calculated from majority argmax class.

Figure 1 (left) shows segmentation loss curves over training epochs on the validation set. Monitoring segmentation loss helps assess the stability and con-

vergence of model training, as lower and smoother loss indicates effective learning of spatial structure. A sharp rise at epoch six is seen for the baseline and multi-task models, suggesting instability or overfitting, potentially due to interactions with auxiliary tasks or insufficient regularisation. The mixture model shows the most stable and consistent decrease, indicating better optimisation. On the right, segmentation, regression, and classification losses for the multi-task and mixture models are compared. Classification loss is substantially lower for the mixture model, suggesting that shared representation learning and balanced sampling help it generalise more effectively across task heads.

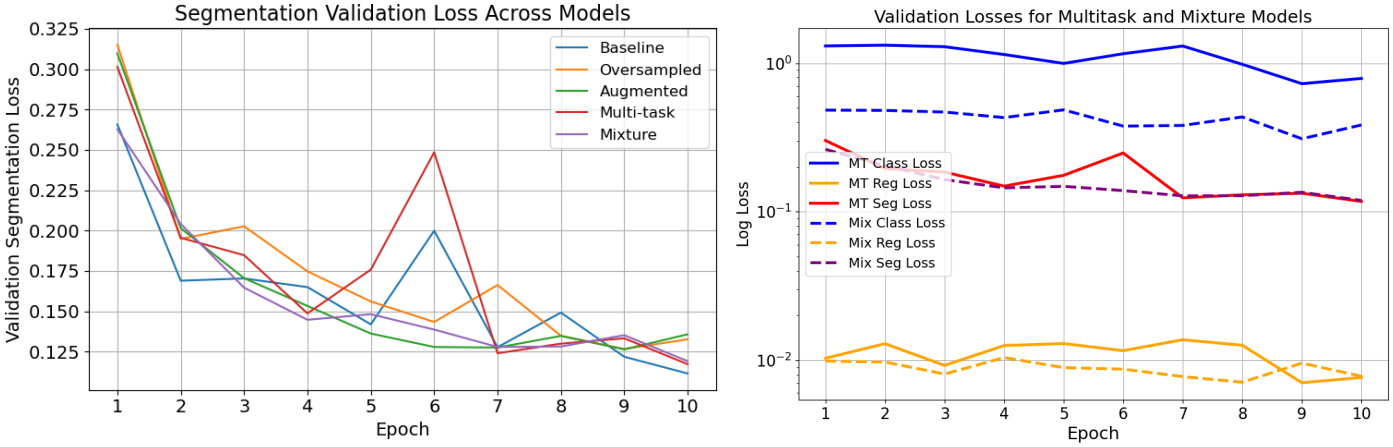

Fig. 1. Validation losses across epochs for different models. Left: Segmentation loss only. Right: Classifier, Reg., and Segmentation loss for multi-task and mixture models

Figure 2 illustrates the segmentation outputs for a Bosch sample in the test set. The oversampling model produces the most coherent output, while the augmentation model shows scattered false positives. Monte Carlo dropout-based entropy maps (five forward passes) highlight high uncertainty around anatomical boundaries, but also in some regions away from the disc and cup.

4 Discussion

While all models achieved high average Dice scores, their comparative performance reveals important trade-offs. The augmentation model, designed to improve generalisation, consistently underperformed—likely due to insufficient epochs to reconcile robustness to distortion with image structure learning. figure 1 confirms this, showing early stagnation and scattered predictions in figure 2.

In contrast, the oversampling model improved segmentation across most groups against the baseline, reducing HD95 and increasing Dice for the disc class. Although oversampling risks overfitting to minority devices, Table 2 suggests it instead helped mitigate imbalance-induced bias on the Forus samples, while also improving general performance. However, the oversampling model did

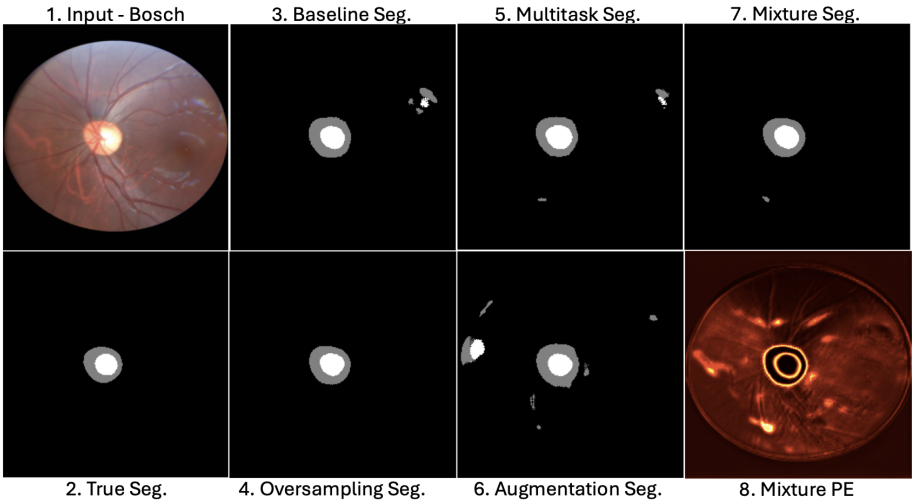

Fig. 2. 1: Fundus image from test set (Bosch device), 2: Ground truth STAPLE segmentation (thesholded a described above), 3-7: segmentations from indicated models generated by pixel-wise argmax on the classes, 8: Predictive Entropy on segmentations from mixture model (generated by five forward passes with Monte Carlo dropout).

not upsample glaucoma positives, and this is reflected in its similar performance to the baseline for these samples.

Multi-task and mixture models struggled with auxiliary outputs: classification and regression losses remained high throughout, relative to the size of the respective parameters and the relevant loss functions (Fig. 1, right). However, segmentation losses decreased well, which may have benefited from joint task learning and device-balanced sampling for the mixture model. Notably, the mixture model's Forus-only performance generally exceeded the full test set average, suggesting enhanced robustness to domain shift.

Finally, uncertainty model (majority of 5 fwd passes) improved Dice on BG and disc against baseline, but increased HD95, potentially due to dropout noise affecting boundary regions. The entropy map in figure 2 confirm higher uncertainty at anatomical edges and occasionally in irrelevant regions.

4.1 Future Work

Extending training to 30–50 epochs may allow models—especially those trained with augmentation or multiple tasks—to better adapt to transformations and optimise all loss heads. For the mixture model, reducing glaucoma oversampling could help avoid sacrificing device balance for class balance.

Architectural changes may also yield gains. Incorporating self-attention or transformer-based modules may help reduce false positives by enabling long-range spatial reasoning, potentially lowering HD95 and improving segmentation coherence. Further, augmentations targeting spatial invariance—e.g., cropping, zoom, elastic distortion—should be explored to enhance robustness to framing and resolution variability.

References

1. Harish Kumar, J.R., Seelamantula, C.S., Gagan, J.H., Kamath, Y.S., Kuzhuppilly, N.I.R., Vivekanand, U., Gupta, P., Patil, S.: Chákṣu: A glaucoma-specific fundus image database. Scientific Data 10, 70 (2023). https://pmc.ncbi.nlm.nih.gov/articles/PMC9898274/
2. Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous Truth and Performance Level Estimation (STAPLE): An Algorithm for the Validation of Image Segmentation. IEEE Transactions on Medical Imaging 23(7), 903–921 (2004). https://doi.org/10.1109/TMI.2004.828354