Weekly Meeting with Dr. Hannah

Hasan Shaikh

Quantitative Imaging Research and
Artificial Intelligence Lab (QIRAIL)

The devil is in the detail

1. Cohort Split vs. 5-Fold Stratified Cross-Validation

In this reproducibility study of the CNN-based prognosis model, we followed both evaluation strategies as outlined in the original paper, using the **Canada** and **MAASTRO** datasets.

✅ **Cohort Split (CI 95%):**

- **Training Set** → `HGJ` and `CHUS`
- **Validation Set** → `HMR` and `CHUM`
- **Test Set** → Entire **MAASTRO** dataset (`maastro_c/` folder)

✅ **5-Fold Stratified Cross-Validation:**

We also performed **5-fold stratified cross-validation** using the **Canada** dataset to ensure class balance (event vs. non-event) across folds.

2. Early Stopping Epoch vs. Best Epoch Performance

- The training pipeline uses an **early stopping mechanism**, which halts training when the **training AUC** reaches a defined threshold (e.g., `0.95`).
- However, this "***does not guarantee* the best validation or testing AUC is achieved at that specified epoch**". For example:

  👉 In the **Distant Metastasis (DM)** task, early stopping triggered at **epoch 708** (when training AUC reached 0.95), but the **best validation AUC** was actually observed at **epoch 689**.

- Therefore, it's important to **review AUCs across all epochs** to identify the model's best generalization point.

What Actually Triggers Model Saving?

```
if store_model.get(MODEL_PATH) and \
  metrics[VALIDATION][ROC][AUC] > best_val_auc and \
  metrics[VALIDATION][ROC][AUC] > store_model.get(THRESHOLD, 0) and \
  abs(metrics[VALIDATION][ROC][AUC] - metrics[TRAIN_METRICS][ROC][AUC]) < store_model.get(MAX_DIFFERENCE, 1):
```

This means the model will be saved **only if all of these conditions are met**:

- ✅ A valid path is given to save the model (`MODEL_PATH`)
- ✅ **Validation ROC AUC is higher than any previous epoch**
- ✅ The new validation AUC is **above a threshold** (usually 0 if not explicitly set)
- ✅ The **difference between training AUC and validation AUC is not too large** (typically < 1)

What Happens When a Model is Saved?

```
save_model(
    store_model[MODEL_PATH],
    epoch,
    model,
    optimizer,
    metrics[TRAIN_METRICS][LOSS],
    model_id=store_model[MODEL_ID] or str(type(model))
)
best_val_auc = metrics[VALIDATION][ROC][AUC]
```

This will:

- Save the model to disk using the epoch number in the filename (e.g., `model_dm_947.pth.tar`)
- Update the `best_val_auc`

So the ***"last model saved"*** is the one that had the **best validation AUC** during the run.

3. Log Files for AUC Tracking

All results for each outcome have been saved under the `log/` directory for transparency and further analysis:

- `log_dm.txt`
- `log_lrf.txt`
- `log_os.txt`

Cohort-split

- `log_dm_5-fold_cv_fold {0,1,2,3,4}.txt`
- `log_lrf_5-fold_cv_fold {0,1,2,3,4}.txt`
- `log_os_5-fold_cv_fold {0,1,2,3,4}.txt`

5-fold stratified cross validation

Each file contains AUC values for training, validation, and testing across all epochs, which allowed me to identify the true performance peaks.

Table 1: Comparative performance (AUCs) for different outcomes of the reproduced HNC-CNN studies and our result

Event	Paper Result		Our Result		Our Result with our Dataset	
	Cohort split (CI 95%)	5-fold CV	Cohort split (CI 95%)	5-fold CV	Cohort split (CI 95%)	5-fold CV
Distant Metastasis (2 years)						
Training	0.91 [0.84, 0.96]	0.87 (0.84–0.92)	**0.81**	**0.87**	-	-
Validation	0.89 [0.81, 0.96]	0.86 (0.77–0.96)	**0.84**	**0.85**	-	-
Testing	0.89 [0.79, 0.98]	0.83 (0.76–0.90)	**0.81**	**0.74**	-	-
Locoregional failure (2 years)						
Training	0.76 [0.64, 0.88]	0.77 (0.72–0.86)	**0.71**	**0.81**	-	-
Validation	0.77 [0.58, 0.92]	0.76 (0.72–0.84)	**0.72**	**0.80**	-	-
Testing	0.45 [0.32, 0.57]	0.53 (0.48–0.59)	**0.49**	**0.57**	-	-
Overall survival (4 years)						
Training	0.84 [0.75, 0.92]	0.82 (0.68–0.94)	**0.75**	**0.78**	-	-
Validation	0.80 [0.66, 0.91]	0.77 (0.62–0.96)	**0.77**	**0.79**	-	-
Testing	0.67 [0.57, 0.77]	0.63 (0.57–0.72)	**0.67**	**0.75**	-	-

For DM:

- Cohort-Split: 947 epoch
- Cross-Validation: 2-fold, 603 epoch

For LRF:

- Cohort-Split: 1359 epoch
- Cross-Validation: 3-fold, 619 epoch

For OS:

- Cohort-Split:
- Cross-Validation: 1-fold, 923 epoch

** Minimum AUC threshold for the validation set change from 0.75 to 0.70*