Reproducibility Study

CNN-Based Prognosis for Head and Neck Cancer

Progress Update | December 2025

By:
Hasan Shaikh
Quantitative Imaging Research and Artificial Intelligence Lab (QIRAIL)

Under the guidance:
Dr. Hannah Mary Thomas T
Scientist, BMIU
CMC, Vellore

Study Overview

Paper Being Reproduced

Mateus et al. (2023) - "Image based prognosis in head and neck cancer using convolutional neural networks"

Three Clinical Outcomes

- Distant Metastasis (2y)
- Locoregional Recurrence (2y)
- Overall Survival (4y)

Two Model Configurations

- Imaging-only (CNN on CT)
- Imaging + Clinical data

Objective

Reproduce the paper's methodology on the Maastro dataset, then apply validated approach to our CMC dataset for external validation

Canada dataset (training) → Maastro (testing) | Our work: Reproduce + CMC validation

Reproduction Results - Imaging Only

✓ Successfully reproduced original results

Table 1. Comparative AUC performance for three clinical outcomes in head and neck cancer across the original published study (**"Paper Results"**), our reproduced models on the same dataset (**"Our Results"**)

Events	Cohort	Paper Results (CNN)	Our Results (CNN)
		Cohort split (CI 95%) / 5-fold CV	**Cohort split (CI 95%) / 5-fold CV**
Distant Metastasis (2y)	Train	0.91 [0.84, 0.96] / 0.87 (0.84–0.92)	**0.85 [0.75, 0.93] / 0.88 (0.83–0.93)**
	Val	0.89 [0.81, 0.96] / 0.86 (0.77–0.96)	**0.87 [0.73, 0.98] / 0.86 (0.81–0.95)**
	Test	0.89 [0.79, 0.98] / 0.83 (0.76–0.90)	**0.87 [0.67, 0.99] / 0.79 (0.74–0.85)**
Locoregional failure (2y)	Train	0.76 [0.64, 0.88] / 0.77 (0.72–0.86)	**0.71 [0.57, 0.84] / 0.78 (0.71–0.81)**
	Val	0.77 [0.58, 0.92] / 0.76 (0.72–0.84)	**0.72 [0.53, 0.88] / 0.79 (0.74–0.82)**
	Test	0.45 [0.32, 0.57] / 0.53 (0.48–0.59)	**0.49 [0.36, 0.62] / 0.56 (0.54–0.57)**
Overall survival (4y)	Train	0.84 [0.75, 0.92] / 0.82 (0.68–0.94)	**0.75 [0.61, 0.86] / 0.79 (0.77–0.81)**
	Val	0.80 [0.66, 0.91] / 0.77 (0.62–0.96)	**0.77 [0.62, 0.90] / 0.79 (0.78–0.80)**
	Test	0.67 [0.57, 0.77] / 0.63 (0.57–0.72)	**0.67 [0.56, 0.76] / 0.71 (0.67–0.75)**

Reproduction Results - With Clinical Data

✓ Reproduced both CNN and ANN architectures

Table 2. Comparative performance (AUCs) including clinical data for predicting multiple clinical outcomes in head and neck cancer across the original study results (***"Paper Results"***), our reproduced CNN and ANN models (***"Our Results"***)

Events	Cohort	Paper Results (CNN)	Our Results (CNN)
		Cohort split (CI 95%) / 5-fold CV	**Cohort split (CI 95%) / 5-fold CV**
Distant Metastasis (2y)	Train	0.91 [0.86, 0.95] / 0.88 (0.81–0.93)	**0.90 [0.84, 0.95] / 0.89 (0.85–0.93)**
	Val	0.89 [0.79, 0.98] / 0.88 (0.81–0.93)	**0.86 [0.68, 0.98] / 0.87 (0.84–0.89)**
	Test	0.93 [0.86, 0.99] / 0.88 (0.86–0.90)	**0.92 [0.86, 0.98] / 0.87 (0.85–0.88)**
Locoregional failure (2y)	Train	0.84 [0.76, 0.93] / 0.77 (0.62–0.87)	**0.75 [0.63, 0.86] / 0.55 (0.55–0.55)**
	Val	0.70 [0.54, 0.84] / 0.72 (0.60–0.84)	**0.70 [0.52, 0.85] / 0.70 (0.70–0.70)**
	Test	0.59 [0.47, 0.70] / 0.57 (0.53–0.60)	**0.57 [0.44, 0.68] / 0.40 (0.40–0.40)**
Overall survival (4y)	Train	0.74 [0.64, 0.84] / 0.83 (0.74–0.94)	**0.74 [0.64, 0.84] / 0.81 (0.79–0.82)**
	Val	0.74 [0.58, 0.86] / 0.81 (0.73–0.93)	**0.72 [0.58, 0.86] / 0.81 (0.78–0.83)**
	Test	0.69 [0.59, 0.79] / 0.68 (0.63–0.71)	**0.69 [0.59, 0.79] / 0.60 (0.57–0.64)**

CMC Dataset - Preprocessing Complete

Preprocessing Success

134/138

97.10% success rate

Pipeline Steps

- DICOM → NIFTI conversion
- GTV structure identification
- RAS orientation
- Slice selection
- PNG generation

Quality Validation

- Tumor centering validated
- Consistent HU windows
- Matches Maastro dimensions
- Ready for training

Failures (4): Missing GTV structures (2) | Corrupted DICOM (2)

Work Completed - Summary

Phase 1: Maastro Reproduction

Imaging-only and imaging+clinical models. Both CNN & ANN. Cohort split & 5-fold CV validated.

Phase 2: CMC Preparation

Pipeline adapted. 292/296 patients processed. Ready for validation.

Key Questions for Discussion

1. Model Training Strategy

Should we train models from scratch or use model weights for reproducing work?

2. Cross-Validation Reporting

How exactly did you calculate the 5-fold CV results? Should we simply average the 5 AUC values and report range, or bootstrap within each fold?

3. CMC Data Integration

For external validation on CMC: Should we train from scratch or use a trained model? Any specific considerations?