Weighted Loss on MLP Prediction on Conway's Game of Life

Zhang Yiran

28 November 2025

1 Motivation

In previous experiments, the MLP model trained with standard (unweighted) binary cross-entropy achieved high overall accuracy ($\sim 90\%$) but consistently under-predicted the minority class (alive cells). This reflected a strong imbalance in the dataset ($N_{\text{dead}} \gg N_{\text{alive}}$). To address this, we evaluated whether adding a class weight to the positive class improves minority-class detection.

2 Experiment Design

2.1 Basic Settings

For every combination of

- density $p \in \{0.2, 0.4, 0.6\}$,
- regime $\in \{\text{early}, \text{mid}, \text{late}\}$,
- patch size $\in \{3 \times 3, 5 \times 5\}$,

All models were trained for 20 epochs with early stopping and evaluated on an independent test set (3 random seeds merged). In total, 36 trained models were compared.

2.2 Loss Functions

All models are trained using binary cross-entropy with logits. The **unweighted** formulation corresponds to the standard BCE:

$$\mathcal{L}_{\text{unweighted}} = -\Big(y \log \sigma(x) + (1-y)\log(1-\sigma(x))\Big),$$

where x is the model output logit and $y \in \{0,1\}$ is the ground-truth cell state.

Weighted BCE. To address the strong class imbalance ($N_{\text{dead}} \gg N_{\text{alive}}$), we apply a positive-class weight to penalize false negatives more heavily:

$$\mathcal{L}_{\text{weighted}} = -\Big(w_{\text{pos}}\, y \log \sigma(x) + (1-y)\log(1-\sigma(x))\Big),$$

where the positive-class weight is defined as

$$w_{\text{pos}} = \min\left(8, \sqrt{\frac{N_{\text{dead}}}{N_{\text{alive}}}}\right).$$

This weighting amplifies the cost of misclassifying alive cells (false negatives), forcing the model to pay more attention to the minority class. In all experiments, the negative-class weight is fixed at 1.

3 Overall Findings

Weighted loss significantly changes the model's behavior:

- **Recall increases dramatically** (average improvement +0.30 $\sim$ +0.38)
- **F1-score consistently improves** (+0.04 to +0.11)
- **Precision decreases**, as expected when the model predicts more alive cells
- **Accuracy decreases slightly** (1%–3%)

Thus, weighted loss addresses the systematic under-detection of alive cells and yields a better-balanced classifier.

4 Representative Confusion Matrices

4.1 Unweighted Models

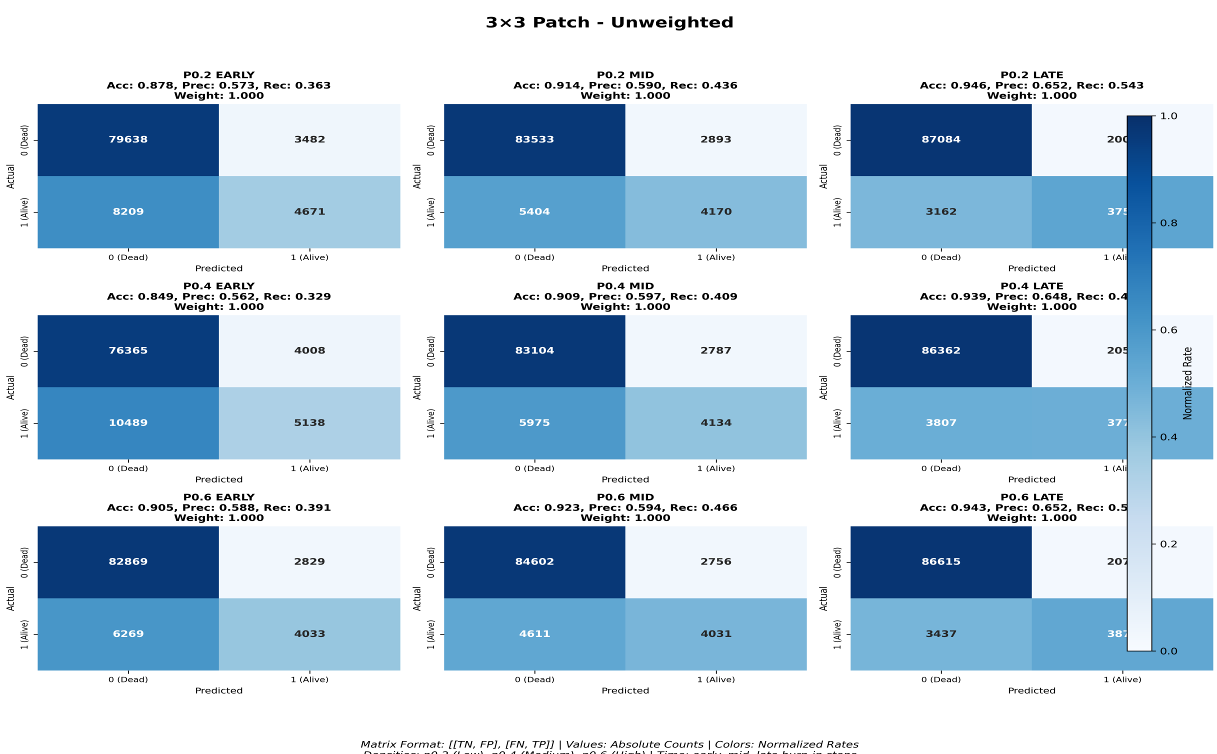

Figure 1: Confusion matrices for **3×3 unweighted** models across densities and regimes.

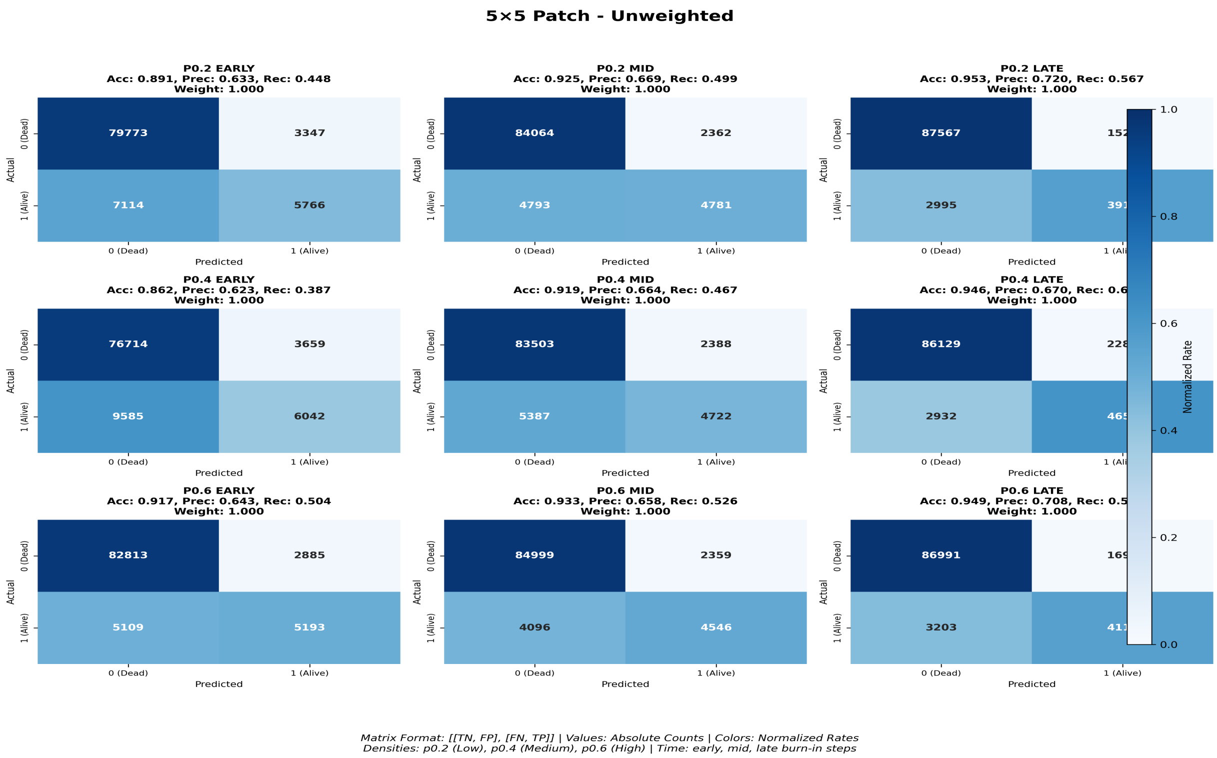

Figure 2: Confusion matrices for **5×5 unweighted** models across densities and regimes.

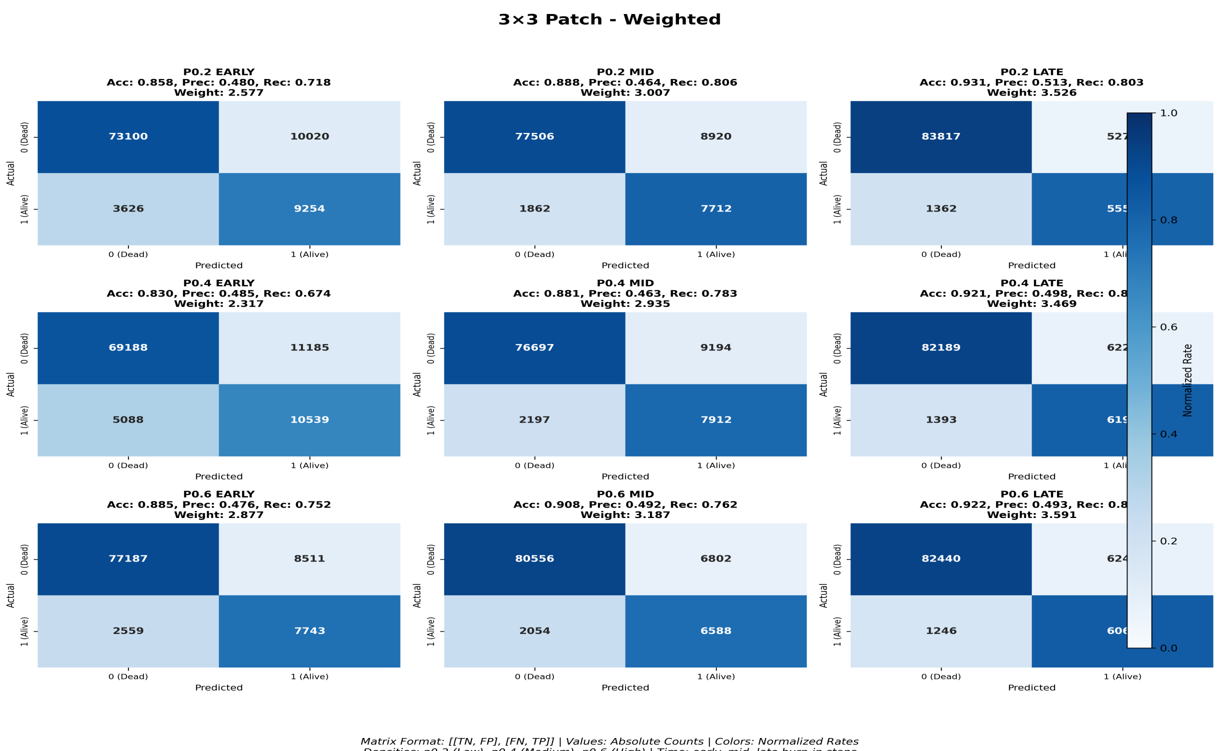

Figure 3: Confusion matrices for **3×3 weighted** models. Note the much higher true-positive counts (lower FN).

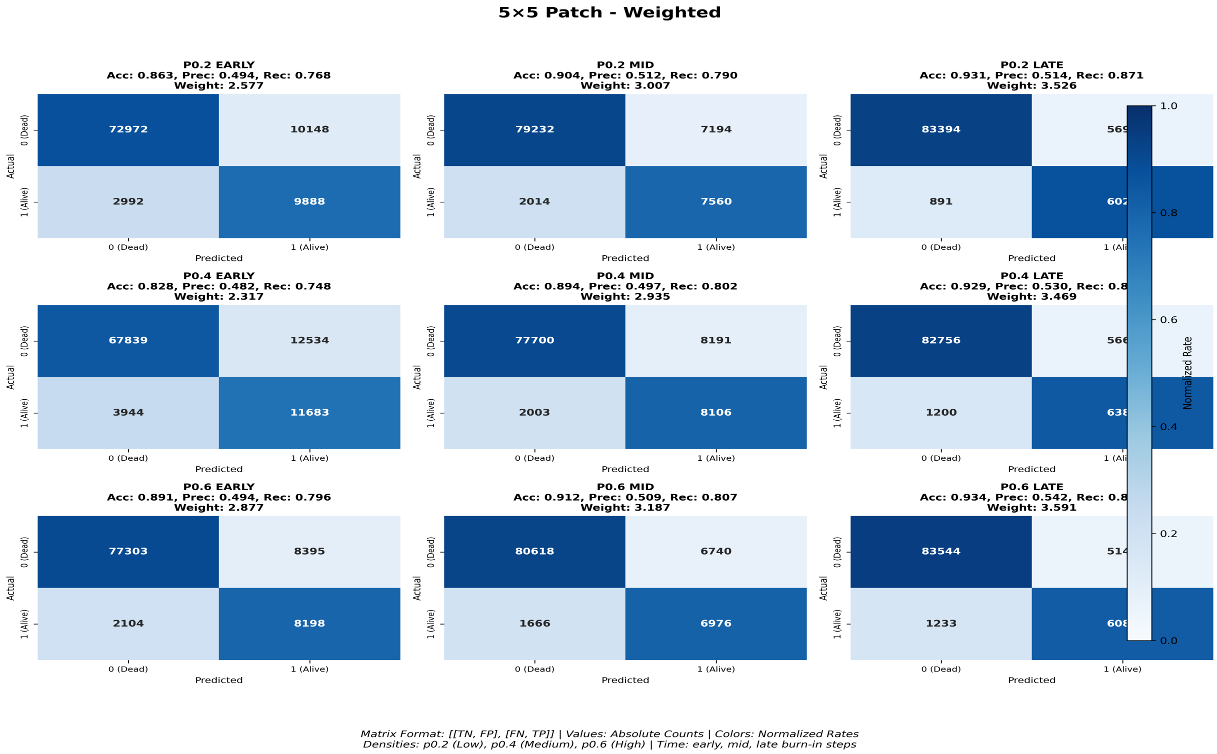

Figure 4: Confusion matrices for **5×5 weighted** models. FN decreases substantially compared to unweighted models.

4.2 Weighted Models

5 Quantitative Summary

5.1 Global Weighted vs Unweighted Performance

Table 1: Global performance averaged across all densities, regimes, and patches.

	Accuracy	Precision	Recall	F1
Unweighted	0.9098	0.6281	0.4840	0.5573
Weighted	0.8904	0.5025	0.7813	0.6150
Δ (W - UW)	-0.0194	-0.1256	$+0.2973$	$+0.0577$

5.2 Patch-Level Comparison

Table 2: Effect of weighted loss by patch size.

Patch Size	Accuracy	Precision	Recall	F1
3×3 Unweighted	0.9031	0.5780	0.3979	0.4628
3×3 Weighted	0.8783	0.4713	0.7632	0.5785
5×5 Unweighted	0.9165	0.6675	0.5698	0.6151
5×5 Weighted	0.9025	0.5336	0.7998	0.6496

6 Visual Summary of Weighted vs Unweighted

Figure 5: Comparison of accuracy, precision, recall, and F1 for weighted vs. unweighted training. Weighted loss significantly boosts recall and F1 at the cost of precision and small accuracy reductions.

7 Conclusion

Weighted loss provides a large improvement in detecting alive cells and consistently increases F1-score. Although accuracy and precision decrease slightly, the new models avoid the systematic failure mode of predicting almost all cells as dead.

This correction is essential before moving toward more advanced architectures (e.g., GeoGNN) or autoregressive rollouts, because the baseline classifier must reliably recognize minority alive states in the first place.