BIOS226 – Supervised Learning (Part 3)

Student Companion: How To Fail

This companion document explains the scenarios shown in the lecture slides (Part 3: How To Fail).

The purpose of this section is to help you understand how supervised learning models fail — often silently — and how to recognise those failures using ROC curves and confusion matrices.

Why Study Failure?

In biological modelling, the greatest danger is not that a model performs poorly.

The real danger is that a model appears to perform well — but is fundamentally flawed.

In high-dimensional data (10,000 genes, 120 patients), it is very easy to build models that look impressive but are unreliable.

Each scenario in the slides shows a different failure mode.

Scenario 0: High n, High p, Clean Signal (The Ideal Case)

This scenario represents a healthy modelling pipeline.

Key characteristics:

- Adequate number of samples
- Reasonable number of selected genes
- True biological signal present
- No data leakage
- Balanced classes

What you observe:

- Cross-validation AUC similar to Test AUC
- ROC curve is smooth and high
- Small gap between CV and test performance
- Confusion matrix shows sensible error rates

Interpretation:

The model generalises well. It has learned genuine structure rather than memorising noise.

Scenario 1: Overfitting

Overfitting occurs when the model learns noise specific to the training data.

Typical causes:

- Too many features relative to number of samples (p » n)
- Weak regularisation
- Highly flexible model structure

What you observe:

- Strong performance during cross-validation
- Noticeable instability across folds
- Potential difference between CV and test metrics

Conceptually:

The model memorises patterns unique to the training set rather than capturing generalisable biology.

This is extremely common in omics datasets.

Scenario 2: Underfitting

Underfitting happens when the model is too simple to capture real signal.

Possible causes:

- Too few informative features selected
- Model overly constrained
- Signal weak relative to noise

What you observe:

- ROC curve closer to diagonal
- Lower AUC values
- Poor performance on both CV and test sets

Conceptually:

The model lacks sufficient capacity to separate Luminal_A from Basal_like tumors.

Underfitting is less deceptive than overfitting — but still clinically useless.

Scenario 3: Wrong Labels in Training

This scenario simulates mislabelled samples.

Examples in real research:

- Data entry errors
- Misclassified tumor subtype
- Incorrect clinical annotation

What you observe:

- ROC curve near diagonal or worse
- Very low AUC
- Confusion matrix dominated by errors

Key lesson:

Supervised learning depends entirely on accurate labels.

If labels are incorrect, the model will confidently learn incorrect biology.

Scenario 4: Feature Selection Leakage

This is one of the most dangerous failure modes.

Feature selection was performed before splitting the data.

This means the test data influenced which genes were selected.

What you observe:

- Extremely high CV AUC
- Extremely high Test AUC
- Performance appears unrealistically perfect

Why this is dangerous:

The test set is no longer independent.

The model has indirectly "seen the answers."

Leakage produces artificially inflated performance.

Scenario 5: Ignoring Class Imbalance

In this case, one subtype heavily dominates the dataset (e.g., 90% Luminal_A, 10% Basal_like).

What you observe:

- CV AUC may look reasonable
- Test AUC may drop sharply
- Confusion matrix may show zero true positives for Basal_like

Why this happens:

The model learns to predict the majority class.

If it predicts Luminal_A for everyone, accuracy remains high but Basal_like tumors are never detected.

Accuracy becomes misleading in imbalanced datasets.

This is especially dangerous in medical contexts.

Recognising Failure Using ROC Curves

ROC curves help identify patterns of failure:

Overfitting: - Large variance across folds - Instability between CV and test

Underfitting: - Curve close to diagonal - AUC near 0.5

Leakage: - Suspiciously perfect performance

Label errors: - Performance worse than random

Class imbalance: - ROC may hide poor minority detection - Confusion matrix becomes essential

---

Why Confusion Matrices Matter

ROC curves describe discrimination across all thresholds.

Confusion matrices describe consequences at one chosen threshold.

In clinical settings:

- False negatives may deny necessary treatment
- False positives may cause unnecessary intervention

Model evaluation must consider clinical consequences, not just AUC.

The Meta-Lesson

All of these failures share one theme:

The model appears objective and mathematical.

But modelling is a human-designed pipeline.

Failure happens when:

- We split incorrectly
- We leak information
- We mis-handle imbalance
- We misunderstand metrics
- We report only favourable numbers

Good modelling is not about maximising AUC.

It is about preventing self-deception.

Final Reflection Questions

When evaluating any supervised model, ask:

1. Was feature selection done after splitting?
2. Were scaling parameters learned on training data only?
3. Is there a large gap between CV and test performance?
4. Are the labels trustworthy?
5. Is class balance appropriate?
6. Does the chosen threshold match the clinical goal?

If you cannot answer these confidently, the model may be failing silently.