BIOS226 - Topic 5 - Supervised Learning (Part 3)

How To Fail

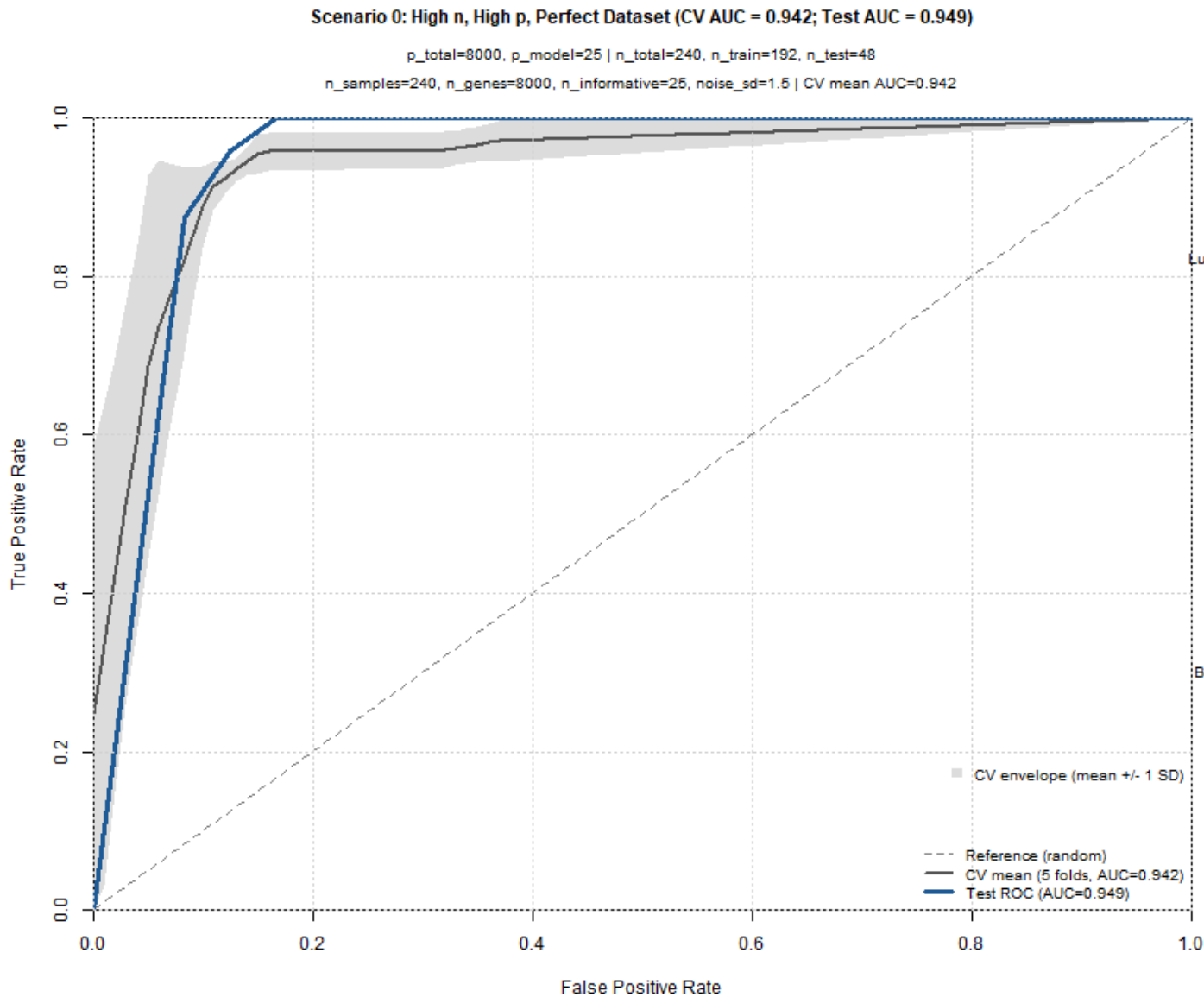
Dr. Robert Treharne

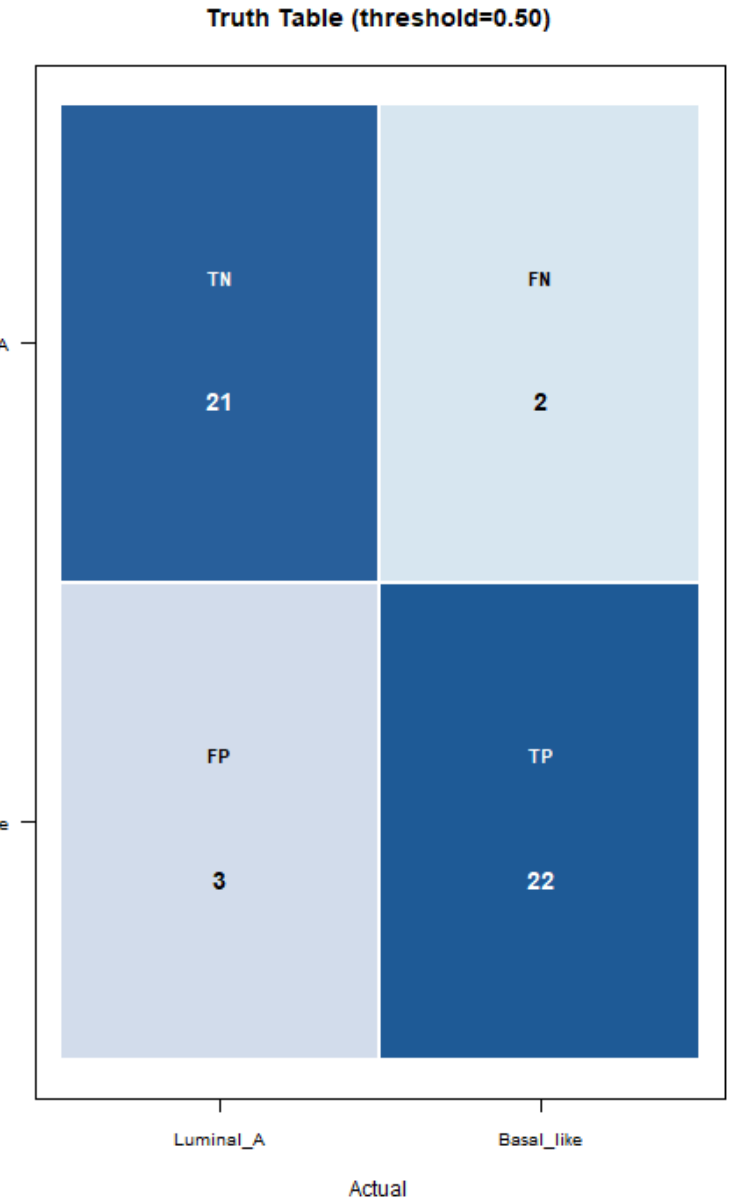
Overview

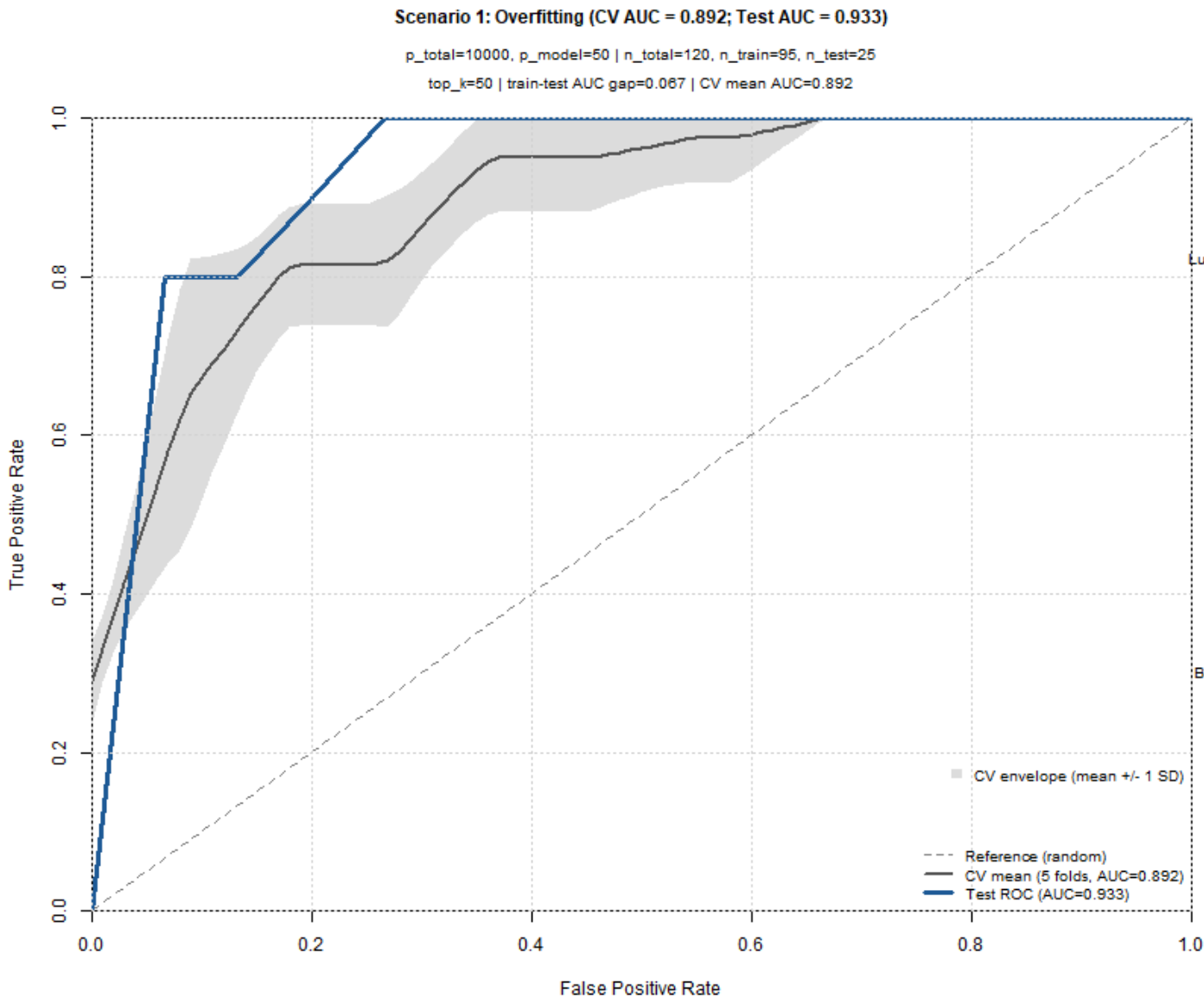
In supervised learning, models rarely fail loudly. They fail silently.

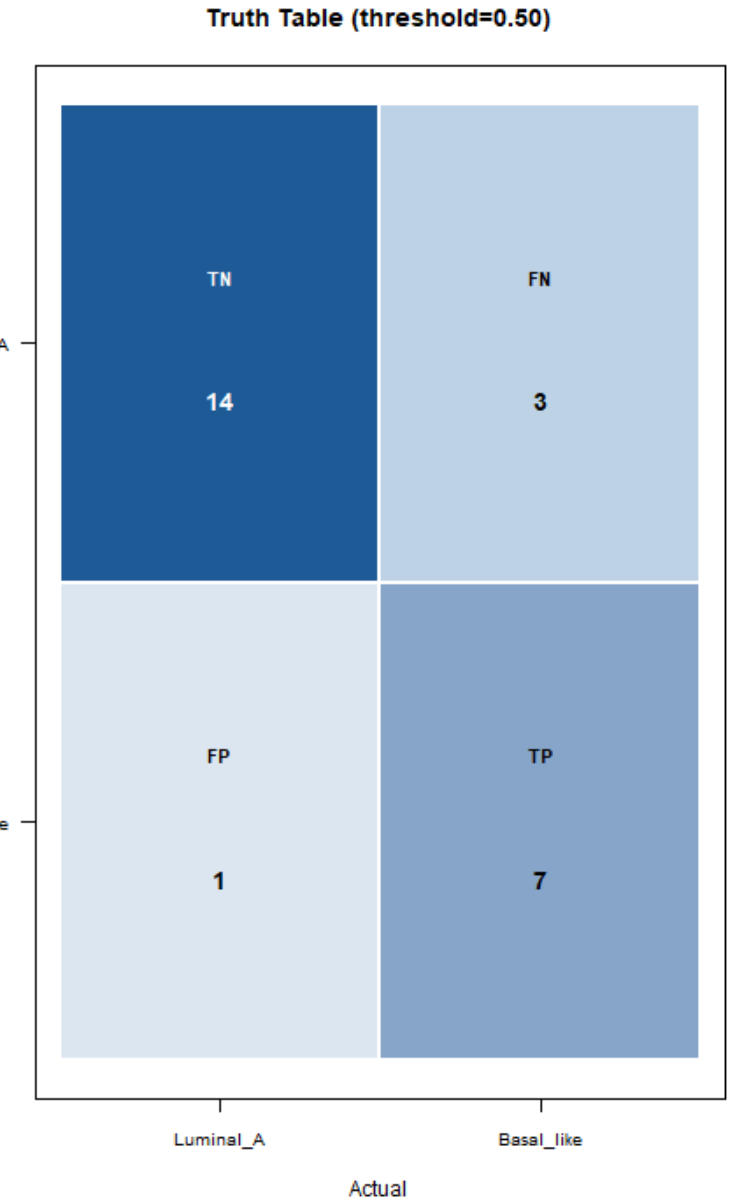
Common failure modes:

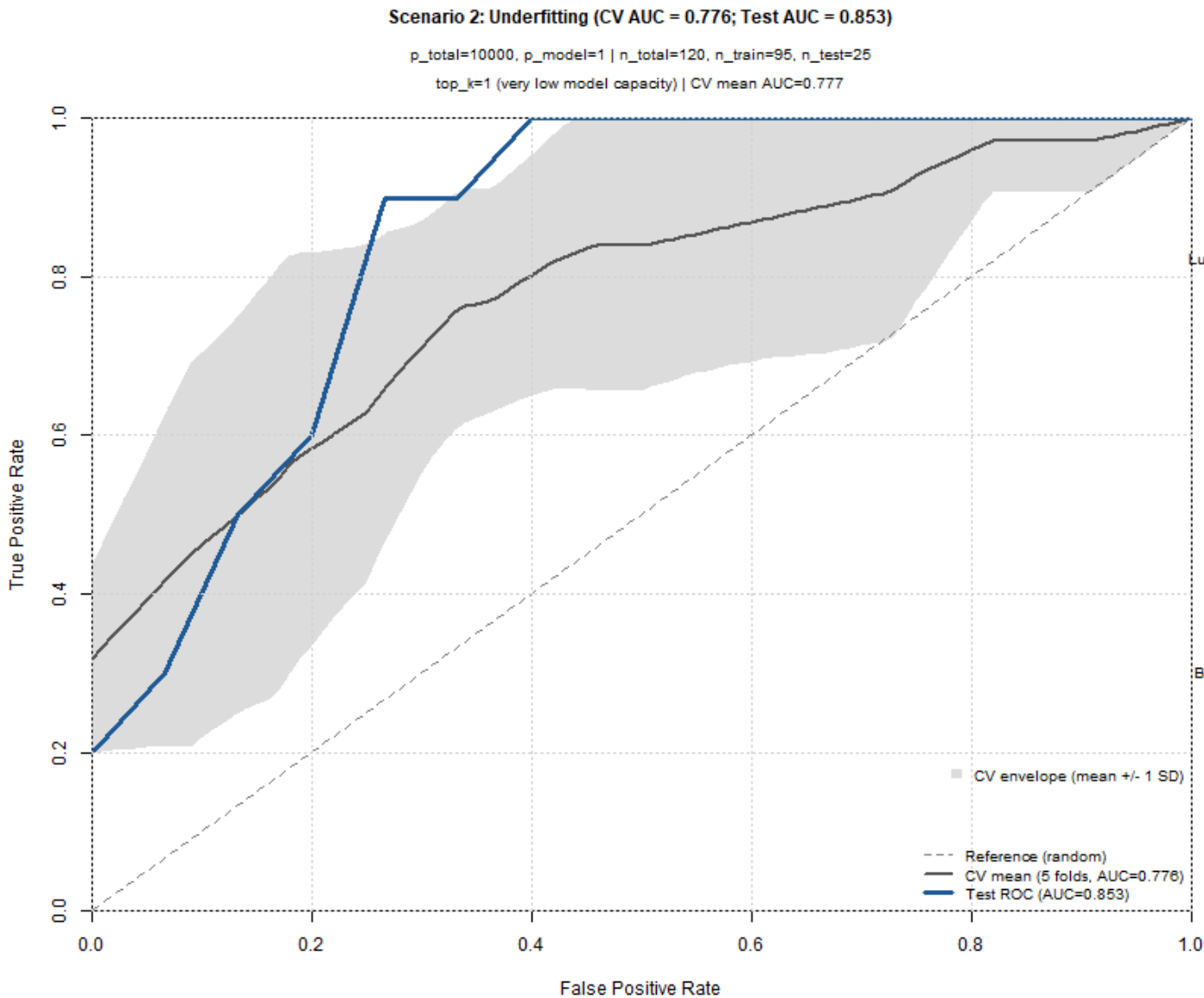
- Overfitting (learning noise instead of signal)
- Underfitting (model too simple)
- Data leakage (information from test data enters training)
- Ignoring class imbalance
- Choosing the wrong threshold for the clinical context
- Evaluating on training data only

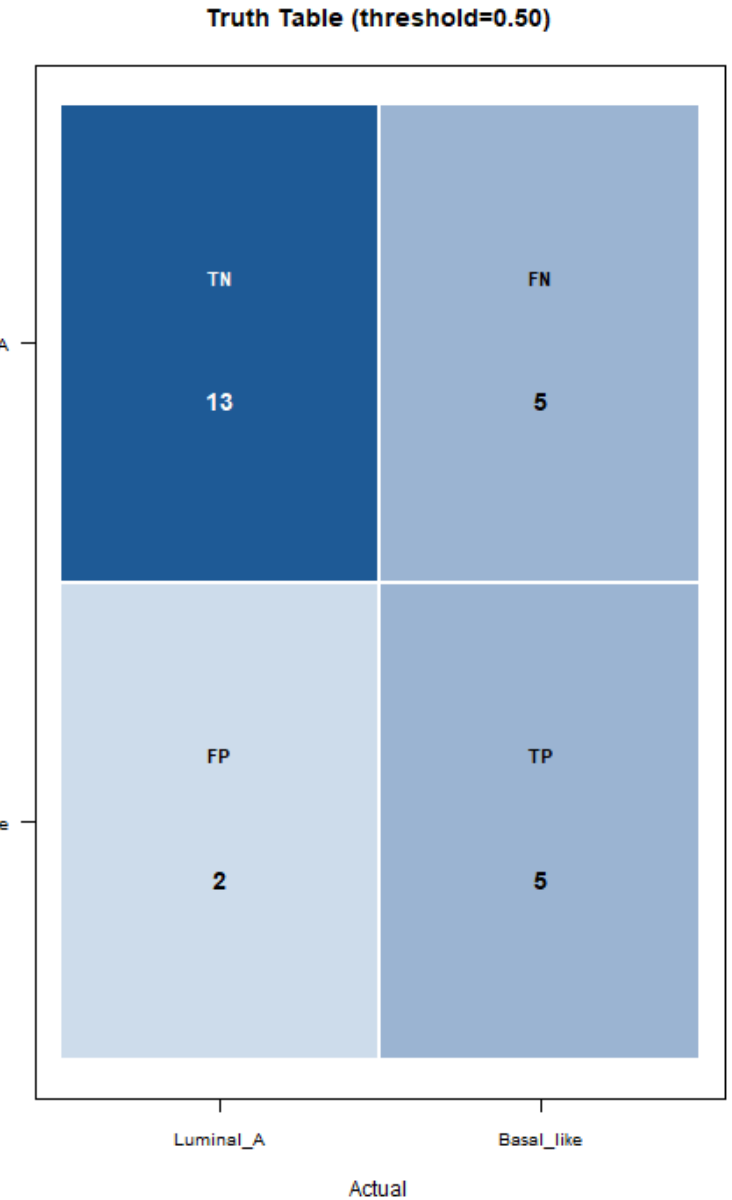

Scenario 0: High n, High p, Perfect Dataset (CV AUC = 0.942; Test AUC = 0.949)
p_total=8000, p_model=25 | n_total=240, n_train=192, n_test=48
n_samples=240, n_genes=8000, n_informative=25, noise_sd=1.5 | CV mean AUC=0.942
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
CV envelope (mean +/- 1 SD)
Reference (random)
CV mean (5 folds, AUC=0.942)
Test ROC (AUC=0.949)

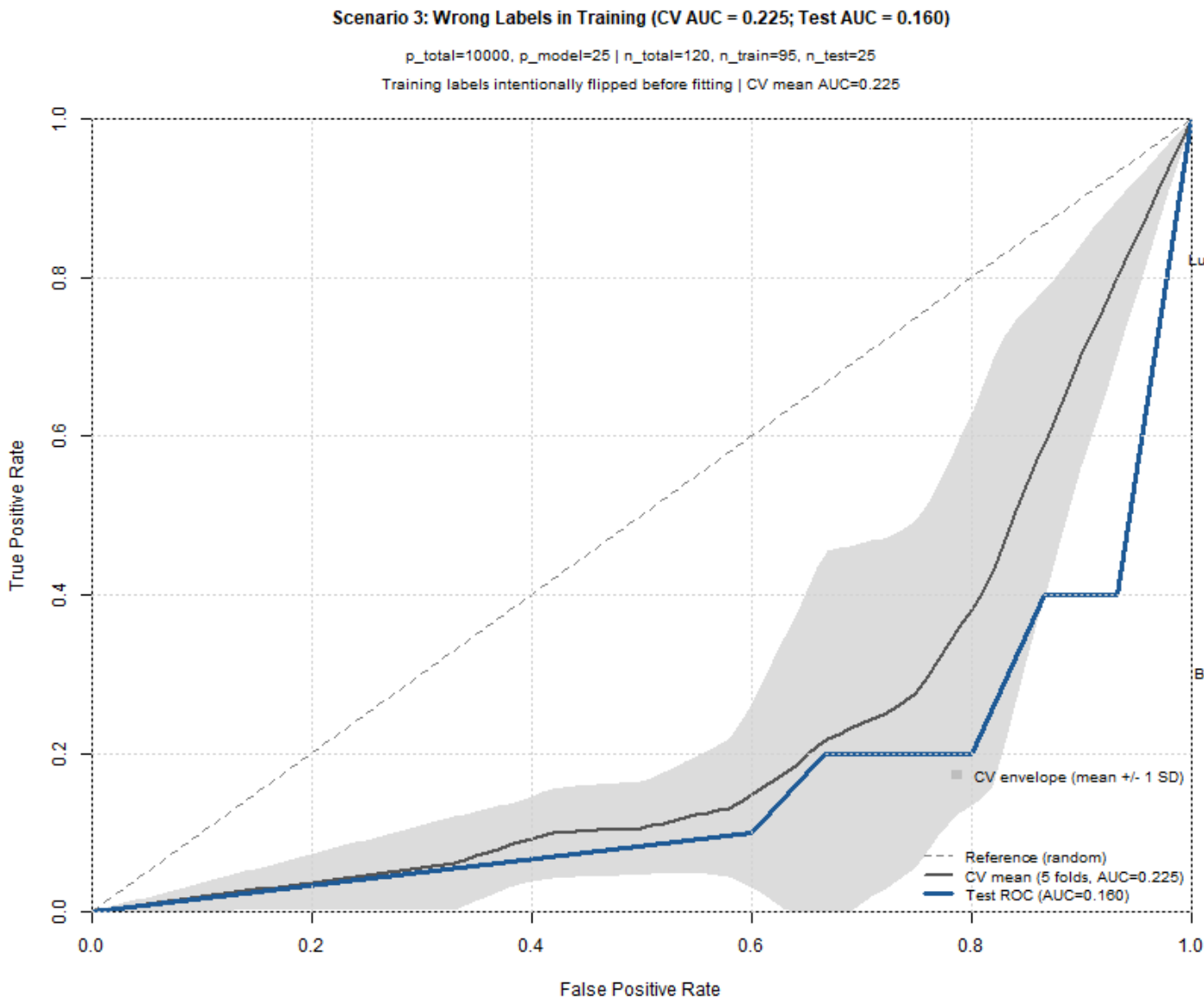

Truth Table (threshold=0.50)
TN
21
FN
2
FP
3
TP
22
Luminal_A
Basal_like
Luminal_A
Basal_like
Actual

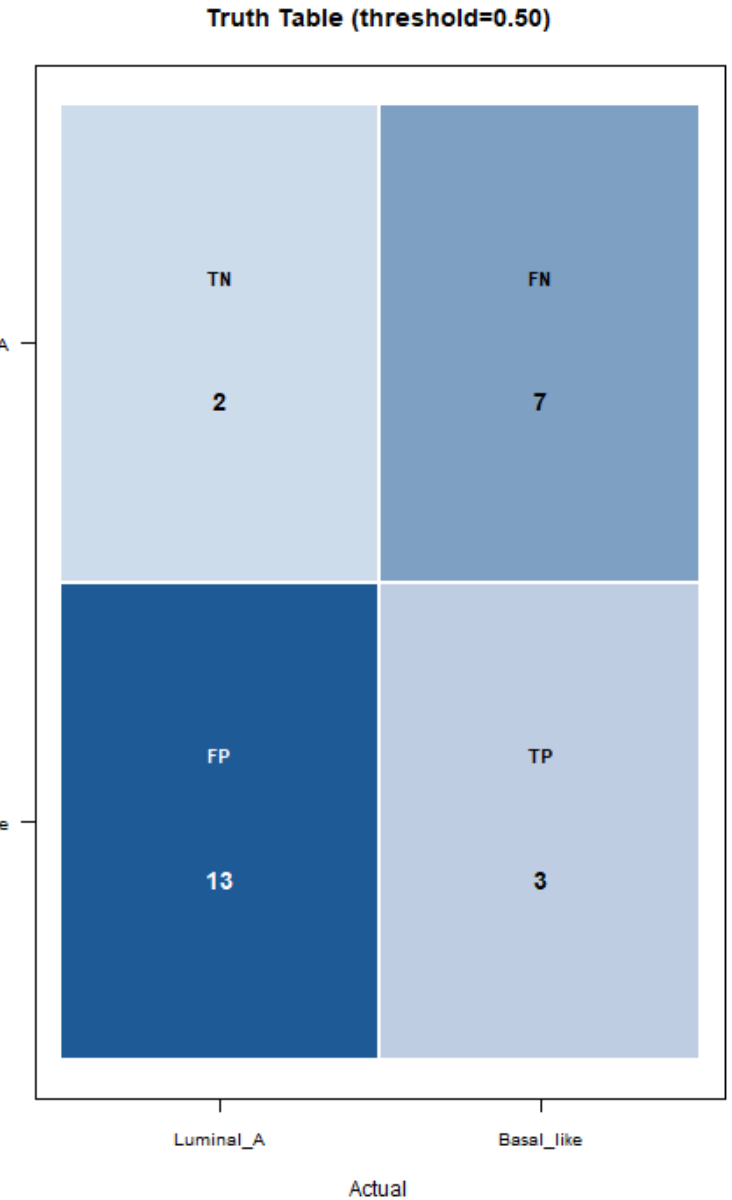

Scenario 1: Overfitting (CV AUC = 0.892; Test AUC = 0.933)
p_total=10000, p_model=50 | n_total=120, n_train=95, n_test=25
top_k=50 | train-test AUC gap=0.067 | CV mean AUC=0.892
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
CV envelope (mean +/- 1 SD)
Reference (random)
CV mean (5 folds, AUC=0.892)
Test ROC (AUC=0.933)

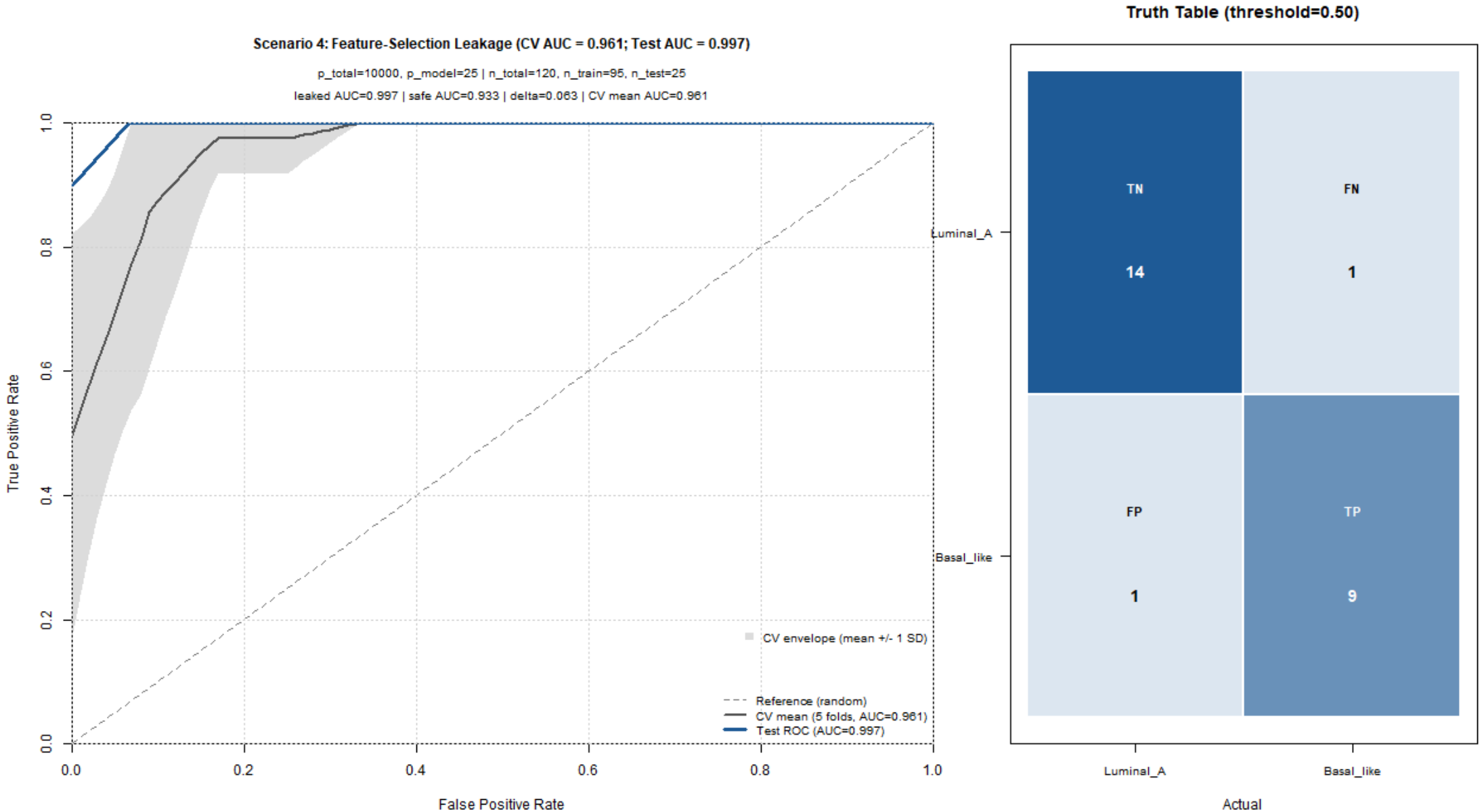

Truth Table (threshold=0.50)
TN
14
FN
3
FP
1
TP
7
Luminal_A
Basal_like
Luminal_A
Basal_like
Actual

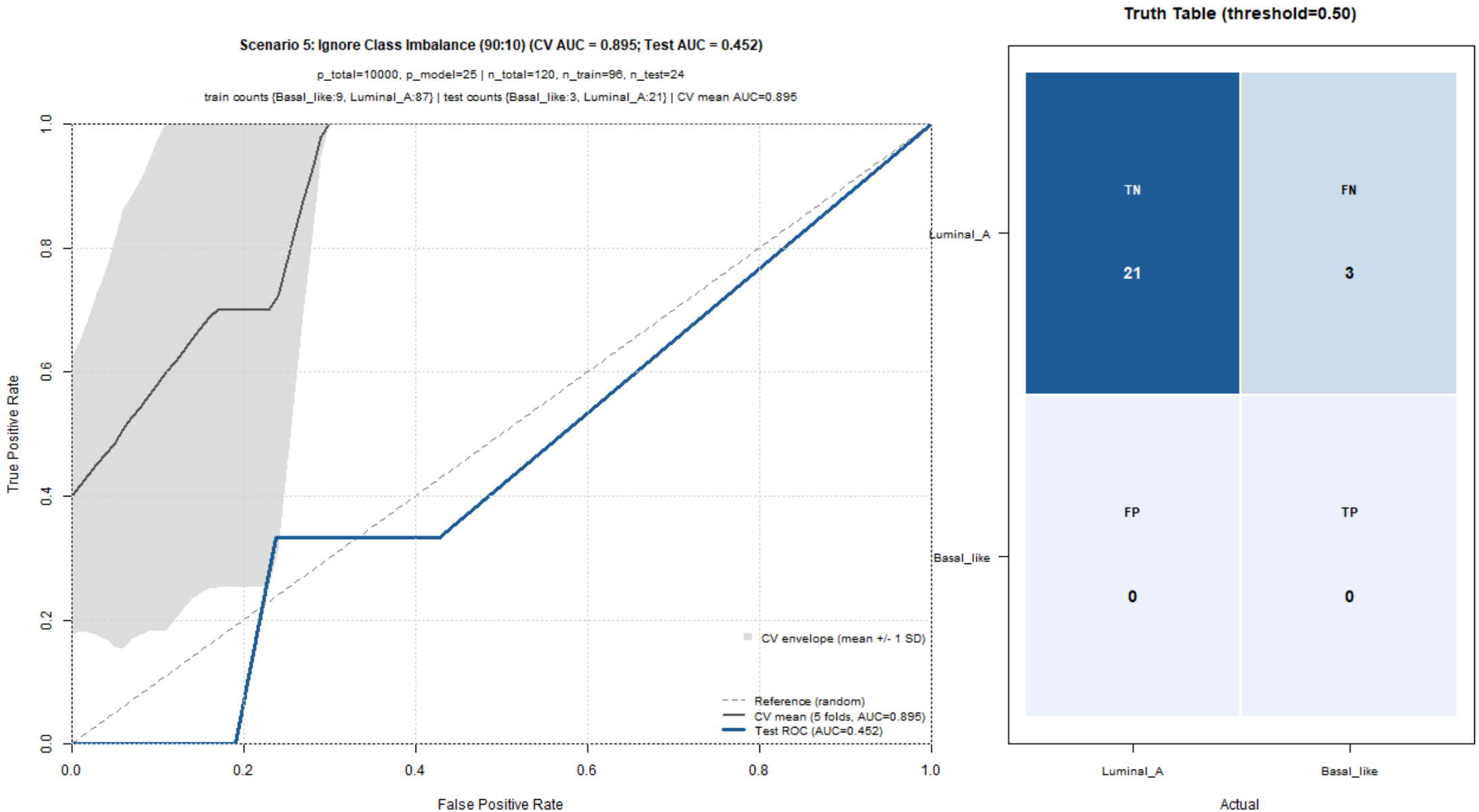

Scenario 2: Underfitting (CV AUC = 0.776; Test AUC = 0.853)
p_total=10000, p_model=1 | n_total=120, n_train=95, n_test=25
top_k=1 (very low model capacity) | CV mean AUC=0.777
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
CV envelope (mean +/- 1 SD)
Reference (random)
CV mean (5 folds, AUC=0.776)
Test ROC (AUC=0.853)


Truth Table (threshold=0.50)
TN
13
FN
5
FP
2
TP
5
Luminal_A
Basal_like
Luminal_A
Basal_like
Actual


Scenario 3: Wrong Labels in Training (CV AUC = 0.225; Test AUC = 0.160)
p_total=10000, p_model=25 | n_total=120, n_train=95, n_test=25
Training labels intentionally flipped before fitting | CV mean AUC=0.225
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
CV envelope (mean +/- 1 SD)
Reference (random)
CV mean (5 folds, AUC=0.225)
Test ROC (AUC=0.160)


Truth Table (threshold=0.50)
TN
2
FN
7
FP
13
TP
3
Luminal_A
Basal_like
Luminal_A
Basal_like
Actual


Scenario 4: Feature-Selection Leakage (CV AUC = 0.961; Test AUC = 0.997)
p_total=10000, p_model=25 | n_total=120, n_train=95, n_test=25
leaked AUC=0.997 | safe AUC=0.933 | delta=0.063 | CV mean AUC=0.961
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
CV envelope (mean +/- 1 SD)
Reference (random)
CV mean (5 folds, AUC=0.961)
Test ROC (AUC=0.997)
Truth Table (threshold=0.50)
TN
FN
14
1
FP
TP
1
9
Luminal_A
Basal_like
Actual


Scenario 5: Ignore Class Imbalance (90:10) (CV AUC = 0.895; Test AUC = 0.452)
p_total=10000, p_model=25 | n_total=120, n_train=96, n_test=24
train counts {Basal_like:9, Luminal_A:87} | test counts {Basal_like:3, Luminal_A:21} | CV mean AUC=0.895
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
CV envelope (mean +/- 1 SD)
Reference (random)
CV mean (5 folds, AUC=0.895)
Test ROC (AUC=0.452)
Truth Table (threshold=0.50)
TN
FN
21
3
FP
TP
0
0
Luminal_A
Basal_like
Actual

By the end of this topic, you should be able to:

- Explain the difference between exploratory and supervised learning in biological data
- Describe a leakage-safe supervised learning pipeline from raw data to final evaluation
- Interpret a confusion matrix and ROC curve in a clinical context
- Recognise common failure modes (overfitting, leakage, imbalance, label errors)
- Critically evaluate whether a model is genuinely generalisable or silently flawed