Chapter 1

Results

1.1 What This Chapter Contains

This chapter reports classification performance on the PROMISE defect-prediction datasets. For each model and each dataset, we provide the reported metrics and a **Mean** row (average across datasets). We then add a **Summary** table that aggregates the *model-wise* means across datasets and an **Overall Mean** row (average across models for each column).

1.2 Datasets Used

We evaluate on the following PROMISE datasets (module-level software defect data):

Table 1.1: *Comprehensive summary of PROMISE datasets used for software defect prediction.*

Dataset	# Attributes	# Tuples	Target Column	Positive Label	# Positive	Negative Label	# Negative	Imbalance (P:N)
CM1	38.00	327.00	Defective	1	42.00	0	285.00	1:6.8
MC1	39.00	9466.00	c	TRUE	68.00	FALSE	9398.00	1:138
MC2	39.00	161.00	c	TRUE	52.00	FALSE	109.00	1:2.1
MW1	38.00	403.00	c	TRUE	31.00	FALSE	372.00	1:12.0
PC1	38.00	1109.00	defects	true	77.00	false	1032.00	1:13.4
PC2	37.00	5589.00	c	TRUE	23.00	FALSE	5566.00	1:242
PC3	38.00	1563.00	c	TRUE	160.00	FALSE	1403.00	1:8.8
PC4	37.00	1458.00	c	TRUE	178.00	FALSE	1280.00	1:7.2

These are classic PROMISE repository datasets frequently used in defect prediction studies.

1.3 Models Used

We evaluate two settings per model: a *baseline* configuration and a *proposed* configuration.

- **AdaBoost** (boosting ensemble of shallow learners)
- **CatBoost** (gradient boosting on ordered/target statistics)
- **Extra Trees** (extremely randomized trees ensemble)
- **Gradient Boosting** (GBDT with decision trees as weak learners)
- **LightGBM** (leaf-wise gradient boosting with histogram splits)
- **MLP** (feed-forward neural network)
- **Random Forest** (bootstrap-aggregated decision trees)
- **XGBoost** (regularized gradient boosting)

1.4 Tables Explanation

Each table lists dataset-wise metrics for a given model, followed by a **Mean** row:

- **Baseline tables** include:
  - `Train AUC_mean`, `AUC_mean`, `F1_mean`, `Precision_mean`, `Recall_mean`
  - `Generalizability_mean`, `Concordance_mean`, `Stability`, `Reliability Index`
- **Proposed tables** include:
  - `AUC mean`, `F1 mean`, `Precision mean`, `Recall mean`
  - `Brier mean`, `GLR mean`, `Hit@10`, `ECE`, `Reliability Score`
- **Summary tables** (Baseline and Proposed) show the mean of each metric *across datasets* (one row per model), plus an **Overall Mean** row (mean across models per column). This lets you compare models head-to-head at a glance.

1.4.1 Metric Glossary

- **AUC**: Area under ROC; threshold-independent ranking quality.
- **F1**: Harmonic mean of Precision and Recall.
- **Precision**: Proportion of predicted defects that are actual defects.

- **Recall** (Sensitivity): Proportion of actual defects that are correctly identified.
- **Brier score**: Mean squared error of predicted probabilities (lower is better).
- **ECE** (Expected Calibration Error): Probability calibration gap (lower is better).
- **Hit@10**: Whether at least one true defect appears in the top-10 ranked items (as defined in your pipeline).
- **GLR**, **Concordance**, **Stability**, **Reliability Index/Score**: aggregate reliability/robustness indicators used by your evaluation pipeline (higher indicates more reliable/consistent performance unless otherwise noted).

1.4.2 Methodology Summary

- We benchmarked eight supervised models (listed in section 1.3) on eight PROMISE datasets (section 1.2).
- For each dataset–model pair, standard classification metrics (AUC, F1, Precision, Recall) and reliability-oriented metrics (e.g., Brier, ECE) were collected from the experimental pipeline.
- Tables report per-dataset results and a **Mean** row that averages across datasets, enabling quick comparison within a model.
- Summary tables aggregate the *model-wise means* and include an **Overall Mean** row for each metric, enabling direct model comparison.
- Interpretation guidance:
  - Higher AUC/F1/Precision/Recall is better; lower Brier/ECE is better.
  - Compare **baseline** vs **proposed** rows to assess whether the proposed configuration improves either accuracy or reliability.
  - **Overall Mean** provides a one-number-per-metric view across all models for quick discussion.

1.5 Current Approach Calculations

Table 1.2: *Adaboost baseline metrics across PROMISE datasets.*

Dataset	Train AUC_mean	AUC_mean	F1_mean	Precision _mean	Recall _mean	Generalizability _mean	Concordance _mean	Stability	Reliability Index
MC1	0.97	0.95	0.13	0.07	0.65	1.00	0.64	0.71	0.89
MC2	0.99	0.72	0.51	0.55	0.50	1.00	0.39	0.50	0.81
MW1	0.97	0.78	0.30	0.25	0.42	1.00	0.66	0.51	0.86
CM1	0.95	0.73	0.30	0.25	0.41	1.00	0.78	0.63	0.90
PC1	0.91	0.84	0.34	0.23	0.69	1.00	0.86	0.76	0.94
PC2	0.96	0.85	0.06	0.03	0.55	1.00	0.58	0.59	0.86
PC3	0.86	0.81	0.36	0.26	0.57	1.00	0.77	0.61	0.90
PC4	0.96	0.93	0.59	0.46	0.84	1.00	0.69	0.72	0.90
Mean	0.94	0.83	0.32	0.26	0.58	1.00	0.67	0.63	0.88

Table 1.3: *CatBoost baseline metrics across PROMISE datasets.*

Dataset	Train AUC_mean	AUC_mean	F1_mean	Precision _mean	Recall _mean	Generalizability _mean	Concordance _mean	Stability	Reliability Index
MC1	1.00	0.97	0.53	0.54	0.53	1.00	0.55	0.80	0.89
MC2	1.00	0.77	0.52	0.62	0.46	1.00	0.40	0.53	0.82
MW1	1.00	0.79	0.25	0.27	0.26	1.00	0.26	0.50	0.79
CM1	1.00	0.74	0.27	0.30	0.31	1.00	0.40	0.79	0.86
PC1	1.00	0.85	0.35	0.37	0.38	1.00	0.75	0.83	0.93
PC2	0.99	0.77	0.07	0.06	0.08	1.00	0.48	0.66	0.86
PC3	1.00	0.84	0.42	0.45	0.39	1.00	0.57	0.66	0.87
PC4	1.00	0.95	0.66	0.62	0.70	1.00	0.53	0.72	0.88
Mean	1.00	0.83	0.38	0.40	0.39	1.00	0.49	0.69	0.86

Table 1.4: *Extra Trees baseline metrics across PROMISE datasets.*

Dataset	Train AUC_mean	AUC_mean	F1_mean	Precision _mean	Recall _mean	Generalizability _mean	Concordance _mean	Stability	Reliability Index
MC1	1.00	0.97	0.36	0.25	0.69	1.00	0.49	0.94	0.91
MC2	1.00	0.74	0.52	0.60	0.48	1.00	0.34	0.56	0.82
MW1	1.00	0.78	0.36	0.32	0.45	1.00	0.17	0.81	0.83
CM1	1.00	0.74	0.23	0.25	0.27	1.00	0.28	0.80	0.85
PC1	1.00	0.87	0.38	0.31	0.53	1.00	0.49	0.89	0.90
PC2	0.99	0.88	0.05	0.03	0.13	1.00	0.31	0.78	0.85
PC3	0.99	0.84	0.42	0.32	0.61	1.00	0.53	0.87	0.90
PC4	0.99	0.92	0.58	0.47	0.78	1.00	0.42	0.92	0.89
Mean	0.99	0.84	0.36	0.32	0.49	1.00	0.38	0.82	0.87

Table 1.5: *Gradient Boosting baseline metrics across PROMISE datasets.*

Dataset	Train AUC_mean	AUC_mean	F1_mean	Precision _mean	Recall _mean	Generalizability _mean	Concordance _mean	Stability	Reliability Index
MC1	1.00	0.97	0.46	0.42	0.52	1.00	0.57	0.74	0.88
MC2	1.00	0.76	0.52	0.57	0.50	1.00	0.54	0.50	0.84
MW1	1.00	0.78	0.26	0.31	0.26	1.00	0.43	0.50	0.82
CM1	1.00	0.75	0.27	0.23	0.34	1.00	0.41	0.69	0.85
PC1	1.00	0.83	0.38	0.37	0.40	1.00	0.80	0.79	0.93
PC2	0.99	0.59	0.09	0.08	0.12	1.00	0.52	0.57	0.85
PC3	0.99	0.83	0.39	0.40	0.38	1.00	0.44	0.65	0.85
PC4	1.00	0.94	0.63	0.59	0.69	1.00	0.57	0.66	0.87
Mean	1.00	0.81	0.38	0.37	0.40	1.00	0.53	0.64	0.86

Table 1.6: *LightGBM baseline metrics across PROMISE datasets.*

Dataset	Train AUC_mean	AUC_mean	F1_mean	Precision _mean	Recall _mean	Generalizability _mean	Concordance _mean	Stability	Reliability Index
MC1	1.00	0.98	0.46	0.53	0.41	1.00	0.50	0.73	0.87
MC2	1.00	0.74	0.54	0.61	0.50	1.00	0.42	0.55	0.83
MW1	1.00	0.82	0.36	0.36	0.36	1.00	0.53	0.60	0.85
CM1	1.00	0.77	0.35	0.42	0.36	1.00	0.46	0.75	0.87
PC1	1.00	0.85	0.35	0.45	0.34	1.00	0.73	0.87	0.93
PC2	1.00	0.78	0.09	0.10	0.08	1.00	0.52	0.72	0.87
PC3	1.00	0.84	0.39	0.45	0.34	1.00	0.59	0.74	0.89
PC4	1.00	0.95	0.63	0.63	0.64	1.00	0.57	0.76	0.89
Mean	1.00	0.84	0.40	0.44	0.38	1.00	0.54	0.71	0.88

Table 1.7: *MLP baseline metrics across PROMISE datasets.*

Dataset	Train AUC_mean	AUC_mean	F1_mean	Precision _mean	Recall _mean	Generalizability _mean	Concordance _mean	Stability	Reliability Index
MC1	0.88	0.88	0.07	0.04	0.72	1.00	0.65	0.77	0.90
MC2	0.63	0.64	0.54	0.37	0.96	1.00	0.57	0.37	0.82
MW1	0.67	0.62	0.14	0.08	0.74	1.00	0.50	0.42	0.82
CM1	0.63	0.59	0.25	0.14	0.95	1.00	0.55	0.41	0.83
PC1	0.58	0.55	0.18	0.21	0.66	1.00	0.41	0.81	0.87
PC2	0.93	0.89	0.04	0.02	0.91	1.00	0.34	0.76	0.85
PC3	0.46	0.44	0.11	0.08	0.28	1.00	−0.06	0.48	0.74
PC4	0.69	0.66	0.24	0.15	0.69	1.00	0.64	0.57	0.87
Mean	0.68	0.66	0.20	0.14	0.74	1.00	0.45	0.57	0.84

Table 1.8: *Random Forest baseline metrics across PROMISE datasets.*

Dataset	Train AUC_mean	AUC_mean	F1_mean	Precision _mean	Recall _mean	Generalizability _mean	Concordance _mean	Stability	Reliability Index
MC1	1.00	0.98	0.44	0.41	0.50	1.00	0.33	0.85	0.86
MC2	1.00	0.73	0.44	0.54	0.39	1.00	0.29	0.39	0.78
MW1	1.00	0.79	0.31	0.36	0.32	1.00	0.37	0.74	0.85
CM1	1.00	0.74	0.23	0.25	0.24	1.00	0.35	0.77	0.85
PC1	1.00	0.86	0.39	0.37	0.43	1.00	0.62	0.88	0.92
PC2	0.99	0.90	0.08	0.05	0.17	1.00	0.42	0.77	0.86
PC3	0.99	0.84	0.41	0.35	0.48	1.00	0.61	0.85	0.91
PC4	0.99	0.93	0.63	0.57	0.72	1.00	0.46	0.90	0.89
Mean	1.00	0.85	0.37	0.36	0.41	1.00	0.43	0.77	0.87

Table 1.9: *XGBoost baseline metrics across PROMISE datasets.*

Dataset	Train AUC_mean	AUC_mean	F1_mean	Precision _mean	Recall _mean	Generalizability _mean	Concordance _mean	Stability	Reliability Index
MC1	1.00	0.97	0.47	0.54	0.43	1.00	0.48	0.74	0.87
MC2	1.00	0.74	0.54	0.61	0.50	1.00	0.54	0.49	0.84
MW1	1.00	0.81	0.33	0.37	0.32	1.00	0.54	0.63	0.86
CM1	1.00	0.77	0.36	0.37	0.39	1.00	0.44	0.73	0.86
PC1	1.00	0.84	0.37	0.40	0.37	1.00	0.72	0.83	0.92
PC2	1.00	0.76	0.06	0.05	0.08	1.00	0.50	0.69	0.86
PC3	1.00	0.83	0.38	0.43	0.34	1.00	0.59	0.80	0.90
PC4	1.00	0.95	0.63	0.63	0.63	1.00	0.57	0.76	0.89
Mean	1.00	0.83	0.39	0.43	0.38	1.00	0.55	0.71	0.88

Table 1.10: *Mean metrics of baseline models across PROMISE datasets (model-wise means) with overall column means.*

Model	Train AUC_mean	AUC_mean	F1_mean	Precision _mean	Recall _mean	Generalizability _mean	Concordance _mean	Stability	Reliability Index
Adaboost	0.94	0.83	0.32	0.26	0.58	1.00	0.67	0.63	0.88
CatBoost	1.00	0.83	0.38	0.40	0.39	1.00	0.49	0.69	0.86
Extra Trees	0.99	0.84	0.36	0.32	0.49	1.00	0.38	0.82	0.87
Gradient Boosting	1.00	0.81	0.38	0.37	0.40	1.00	0.53	0.64	0.86
LightGBM	1.00	0.84	0.40	0.44	0.38	1.00	0.54	0.71	0.88
MLP	0.68	0.66	0.20	0.14	0.74	1.00	0.45	0.57	0.84
Random Forest	1.00	0.85	0.37	0.36	0.41	1.00	0.43	0.77	0.87
XGBoost	1.00	0.83	0.39	0.43	0.38	1.00	0.55	0.71	0.88
Overall Mean	0.95	0.81	0.35	0.34	0.47	1.00	0.51	0.69	0.87

1.6 Proposed Approach Calculations

Table 1.11: *Adaboost (proposed) metrics across PROMISE datasets.*

Dataset	AUC mean	F1 mean	Precision mean	Recall mean	Brier mean	GLR mean	Hit@10	ECE	Reliability Score
CM1	0.69	0.32	0.25	0.53	0.15	0.13	1.00	0.14	0.81
MC1	0.87	0.29	0.21	0.52	0.02	0.42	1.00	0.02	0.90
MC2	0.73	0.56	0.53	0.65	0.20	0.01	1.00	0.11	0.80
MW1	0.70	0.35	0.44	0.36	0.09	0.09	1.00	0.09	0.82
PC1	0.82	0.36	0.28	0.52	0.10	0.06	1.00	0.10	0.81
PC2	0.60	0.06	0.04	0.12	0.01	0.33	1.00	0.01	0.88
PC3	0.81	0.40	0.31	0.58	0.11	0.23	1.00	0.09	0.84
PC4	0.93	0.61	0.56	0.70	0.08	0.17	1.00	0.05	0.85
Mean	0.77	0.37	0.33	0.50	0.09	0.18	1.00	0.08	0.84

Table 1.12: *CatBoost (proposed) metrics across PROMISE datasets.*

Dataset	AUC mean	F1 mean	Precision mean	Recall mean	Brier mean	GLR mean	Hit@10	ECE	Reliability Score
CM1	0.72	0.18	0.23	0.19	0.15	0.16	1.00	0.14	0.81
MC1	0.97	0.49	0.47	0.53	0.01	0.25	1.00	0.00	0.87
MC2	0.78	0.52	0.58	0.48	0.21	−0.01	1.00	0.18	0.77
MW1	0.78	0.30	0.33	0.29	0.09	0.05	1.00	0.08	0.81
PC1	0.86	0.36	0.39	0.39	0.07	0.07	1.00	0.06	0.82
PC2	0.80	0.08	0.08	0.08	0.01	0.24	1.00	0.01	0.87
PC3	0.83	0.42	0.48	0.38	0.09	0.09	1.00	0.07	0.83
PC4	0.95	0.66	0.62	0.70	0.07	0.06	1.00	0.05	0.83
Mean	0.84	0.38	0.40	0.38	0.09	0.12	1.00	0.07	0.83

Table 1.13: *Extra Trees (proposed) metrics across PROMISE datasets.*

Dataset	AUC mean	F1 mean	Precision mean	Recall mean	Brier mean	GLR mean	Hit@10	ECE	Reliability Score
CM1	0.73	0.14	0.20	0.14	0.12	0.05	1.00	0.09	0.81
MC1	0.95	0.53	0.59	0.49	0.01	0.08	1.00	0.01	0.84
MC2	0.74	0.47	0.52	0.44	0.19	0.08	1.00	0.08	0.82
MW1	0.72	0.31	0.34	0.32	0.09	0.00	1.00	0.08	0.81
PC1	0.83	0.40	0.49	0.39	0.06	0.10	1.00	0.05	0.83
PC2	0.90	0.04	0.04	0.04	0.01	0.03	0.96	0.01	0.82
PC3	0.85	0.44	0.42	0.46	0.09	0.08	1.00	0.08	0.82
PC4	0.94	0.63	0.60	0.66	0.07	0.06	1.00	0.05	0.83
Mean	0.83	0.37	0.40	0.37	0.08	0.06	1.00	0.06	0.82

Table 1.14: *Gradient Boosting (proposed) metrics across PROMISE datasets.*

Dataset	AUC mean	F1 mean	Precision mean	Recall mean	Brier mean	GLR mean	Hit@10	ECE	Reliability Score
CM1	0.76	0.26	0.25	0.31	0.15	0.16	1.00	0.16	0.81
MC1	0.97	0.50	0.52	0.49	0.01	0.32	1.00	0.00	0.89
MC2	0.76	0.57	0.62	0.54	0.23	0.00	1.00	0.22	0.76
MW1	0.73	0.25	0.30	0.22	0.10	0.10	1.00	0.09	0.82
PC1	0.84	0.39	0.42	0.39	0.07	0.10	1.00	0.06	0.83
PC2	0.63	0.09	0.08	0.12	0.01	0.21	1.00	0.01	0.87
PC3	0.82	0.39	0.42	0.36	0.09	0.07	1.00	0.06	0.83
PC4	0.94	0.62	0.60	0.65	0.07	0.08	1.00	0.05	0.83
Mean	0.81	0.38	0.40	0.39	0.09	0.13	1.00	0.08	0.83

Table 1.15: *LightGBM (proposed) metrics across PROMISE datasets.*

Dataset	AUC mean	F1 mean	Precision mean	Recall mean	Brier mean	GLR mean	Hit@10	ECE	Reliability Score
CM1	0.76	0.31	0.40	0.32	0.15	0.17	1.00	0.15	0.81
MC1	0.98	0.49	0.59	0.43	0.01	0.39	1.00	0.01	0.90
MC2	0.73	0.51	0.53	0.50	0.25	0.14	1.00	0.24	0.78
MW1	0.79	0.37	0.45	0.36	0.09	0.04	1.00	0.09	0.81
PC1	0.86	0.35	0.48	0.31	0.07	0.29	1.00	0.07	0.86
PC2	0.76	0.08	0.08	0.08	0.01	0.30	1.00	0.01	0.88
PC3	0.83	0.37	0.42	0.33	0.10	0.18	1.00	0.10	0.83
PC4	0.95	0.64	0.67	0.61	0.07	0.22	1.00	0.07	0.85
Mean	0.83	0.39	0.45	0.37	0.09	0.21	1.00	0.09	0.84

Table 1.16: *MLP (proposed) metrics across PROMISE datasets.*

Dataset	AUC mean	F1 mean	Precision mean	Recall mean	Brier mean	GLR mean	Hit@10	ECE	Reliability Score
CM1	0.69	0.24	0.25	0.24	0.17	0.08	1.00	0.17	0.80
MC1	0.89	0.21	0.21	0.21	0.02	0.17	1.00	0.02	0.84
MC2	0.69	0.55	0.65	0.48	0.24	0.03	1.00	0.14	0.82
MW1	0.67	0.26	0.27	0.27	0.11	0.09	1.00	0.10	0.81
PC1	0.76	0.29	0.35	0.27	0.10	0.09	1.00	0.10	0.81
PC2	0.79	0.05	0.05	0.05	0.01	0.13	1.00	0.01	0.85
PC3	0.74	0.32	0.40	0.28	0.12	0.15	1.00	0.11	0.84
PC4	0.90	0.50	0.51	0.48	0.10	0.13	1.00	0.08	0.84
Mean	0.76	0.30	0.34	0.29	0.11	0.11	1.00	0.09	0.83

Table 1.17: *Random Forest (proposed) metrics across PROMISE datasets.*

Dataset	AUC mean	F1 mean	Precision mean	Recall mean	Brier mean	GLR mean	Hit@10	ECE	Reliability Score
CM1	0.73	0.23	0.30	0.24	0.13	0.08	1.00	0.11	0.81
MC1	0.98	0.46	0.52	0.43	0.01	0.22	1.00	0.01	0.87
MC2	0.73	0.47	0.54	0.44	0.19	0.02	1.00	0.08	0.81
MW1	0.79	0.33	0.34	0.35	0.08	0.02	1.00	0.09	0.81
PC1	0.87	0.42	0.44	0.42	0.07	0.01	1.00	0.07	0.81
PC2	0.90	0.03	0.02	0.04	0.01	0.12	1.00	0.01	0.85
PC3	0.84	0.39	0.38	0.41	0.09	−0.02	1.00	0.08	0.80
PC4	0.94	0.63	0.60	0.66	0.07	0.06	1.00	0.05	0.83
Mean	0.85	0.37	0.39	0.37	0.08	0.06	1.00	0.06	0.82

Table 1.18: *XGBoost (proposed) metrics across PROMISE datasets.*

Dataset	AUC mean	F1 mean	Precision mean	Recall mean	Brier mean	GLR mean	Hit@10	ECE	Reliability Score
CM1	0.77	0.34	0.39	0.36	0.13	0.23	1.00	0.11	0.83
MC1	0.97	0.48	0.56	0.44	0.01	0.07	1.00	0.00	0.84
MC2	0.75	0.53	0.59	0.52	0.22	0.14	1.00	0.18	0.80
MW1	0.79	0.31	0.40	0.30	0.08	0.15	1.00	0.08	0.83
PC1	0.84	0.37	0.42	0.37	0.07	0.07	1.00	0.06	0.82
PC2	0.75	0.04	0.04	0.04	0.01	0.27	1.00	0.01	0.88
PC3	0.83	0.38	0.45	0.33	0.09	0.19	1.00	0.08	0.84
PC4	0.95	0.64	0.66	0.61	0.07	0.12	1.00	0.05	0.84
Mean	0.83	0.39	0.44	0.37	0.08	0.16	1.00	0.07	0.84

Table 1.19: *Mean metrics of proposed models across PROMISE datasets (model-wise means) with overall column means.*

Model	AUC mean	F1 mean	Precision mean	Recall mean	Brier mean	GLR mean	Hit@10	ECE	Reliability Score
Adaboost	0.77	0.37	0.33	0.50	0.09	0.18	1.00	0.08	0.84
CatBoost	0.84	0.38	0.40	0.38	0.09	0.12	1.00	0.07	0.83
Extra Trees	0.83	0.37	0.40	0.37	0.08	0.06	1.00	0.06	0.82
Gradient Boosting	0.81	0.38	0.40	0.39	0.09	0.13	1.00	0.08	0.83
LightGBM	0.83	0.39	0.45	0.37	0.09	0.21	1.00	0.09	0.84
MLP	0.76	0.30	0.34	0.29	0.11	0.11	1.00	0.09	0.83
Random Forest	0.85	0.37	0.39	0.37	0.08	0.06	1.00	0.06	0.82
XGBoost	0.83	0.39	0.44	0.37	0.08	0.16	1.00	0.07	0.84
Overall Mean	0.82	0.37	0.39	0.38	0.09	0.13	1.00	0.07	0.83