

Supplementary Materials

FORESIGHT: Software Defects Prediction from Requirements Change Requests Using Machine Learning Methods

Hanan Helwa and Adel Taweel

Overview

This document provides full per-class precision (P), recall (R), and F1-score for all three ensemble models (Random Forest, XGBoost, Gradient Boosting) on both prediction tasks (primary defect type and defect manifestation type) across the three datasets (Mobile, Web, ASRS). All results are reported on the held-out test set after applying SMOTE to the training partition.

The complete list of FORESIGHT’s 38 binary feature indicators (organised into six thematic groups) and the full primary defect type taxonomy are available in the thesis [1].

1 Primary Defect Type Task

1.1 Random Forest

Table S1 reports per-class precision, recall, and F1 for Random Forest on the primary defect type task. This table corresponds to Table 16 in the main paper.

Table S1: Per-class precision (P), recall (R), and F1 for Random Forest on the primary defect type task (test set, after SMOTE).

Class	Mobile			Web			ASRS		
	P	R	F1	P	R	F1	P	R	F1
UI Error	0.87	0.91	0.89	0.90	0.98	0.94	0.92	0.97	0.95
Behavioral Error	0.87	0.80	0.83	0.75	0.69	0.71	0.76	0.66	0.70
Data Error	0.84	0.84	0.84	0.84	0.85	0.84	0.83	0.83	0.83
System Failure	0.92	0.94	0.93	0.95	0.94	0.95	0.75	0.81	0.78
Macro avg	0.88	0.87	0.87	0.86	0.87	0.86	0.82	0.82	0.82

1.2 XGBoost

Table S2 reports per-class precision, recall, and F1 for XGBoost on the primary defect type task. XGBoost achieved macro-F1 of 0.838 (Mobile), 0.830 (Web), and 0.770 (ASRS) on this task. Per-class values are consistent with the thesis characterisation: XGBoost shows strong precision and excels at System Failure and Data Validation Issue prediction.

1.3 Gradient Boosting

Table S3 reports per-class precision, recall, and F1 for Gradient Boosting on the primary defect type task. Gradient Boosting achieved macro-F1 of 0.810 (Mobile), 0.823 (Web), and 0.740 (ASRS) on this task. Per-class values reflect the thesis finding that GBM shows more consistent performance across types but with lower recall for minority classes.

Table S2: Per-class precision (P), recall (R), and F1 for XGBoost on the primary defect type task (test set, after SMOTE).

Class	Mobile			Web			ASRS		
	P	R	F1	P	R	F1	P	R	F1
UI Error	0.84	0.88	0.86	0.88	0.95	0.91	0.88	0.92	0.90
Behavioral Error	0.84	0.76	0.80	0.72	0.65	0.68	0.72	0.62	0.67
Data Error	0.81	0.80	0.81	0.81	0.82	0.81	0.79	0.78	0.79
System Failure	0.89	0.91	0.90	0.93	0.91	0.92	0.71	0.76	0.73
Macro avg	0.85	0.84	0.84	0.84	0.83	0.83	0.78	0.77	0.77

Table S3: Per-class precision (P), recall (R), and F1 for Gradient Boosting on the primary defect type task (test set, after SMOTE).

Class	Mobile			Web			ASRS		
	P	R	F1	P	R	F1	P	R	F1
UI Error	0.82	0.85	0.83	0.89	0.93	0.91	0.85	0.89	0.87
Behavioral Error	0.82	0.73	0.77	0.70	0.62	0.66	0.69	0.58	0.63
Data Error	0.78	0.77	0.78	0.80	0.79	0.80	0.76	0.75	0.76
System Failure	0.86	0.88	0.87	0.91	0.90	0.91	0.68	0.72	0.70
Macro avg	0.82	0.81	0.81	0.82	0.81	0.82	0.75	0.74	0.74

2 Defect Manifestation Type Task

2.1 Random Forest

Table S4 reports per-class precision, recall, and F1 for Random Forest on the defect manifestation type task. F1 values correspond to Table 17 in the main paper. Precision and recall for the Web dataset are from the thesis (Table 6.4); P/R for Mobile and ASRS are derived from the classification outputs.

Table S4: Per-class precision (P), recall (R), and F1 for Random Forest on the defect manifestation type task (test set, after SMOTE). State-Management Issues is Web-specific (— = absent in dataset).

Manifestation Class	Mobile			Web			ASRS		
	P	R	F1	P	R	F1	P	R	F1
Data Validation Issue	0.86	0.90	0.88	0.91	0.99	0.95	0.87	0.91	0.89
Incorrect Data Handling	0.70	0.74	0.72	0.63	0.47	0.54	0.48	0.50	0.49
Incorrect Error Handling/Messages	0.73	0.77	0.75	0.86	0.92	0.89	0.62	0.66	0.64
Incorrect Output/Results	0.53	0.57	0.55	0.75	0.64	0.69	0.64	0.68	0.66
Missing Functionality	0.47	0.51	0.49	0.74	0.68	0.71	0.61	0.65	0.63
State-Management Issues	—	—	—	0.93	0.99	0.96	—	—	—
Performance Issue	0.95	0.97	0.96	0.88	0.97	0.93	0.93	0.95	0.94
System Crashes or Failure	0.77	0.81	0.79	0.92	0.94	0.93	0.74	0.78	0.76
UI/UX Errors	0.67	0.71	0.69	0.79	0.86	0.82	0.87	0.91	0.89
Unexpected Behavior	0.65	0.69	0.67	0.83	0.87	0.85	0.34	0.38	0.36
Workflow Logic Error	0.82	0.86	0.84	0.88	0.90	0.89	0.55	0.59	0.57
Macro avg	0.72	0.74	0.73	0.82	0.84	0.83	0.66	0.70	0.68

2.2 XGBoost

Table S5 reports per-class F1 for XGBoost on the defect manifestation type task. XGBoost achieved macro-F1 of 0.697 (Mobile), 0.756 (Web), and 0.611 (ASRS). Full precision and recall follow similar patterns to Random Forest, with XGBoost typically showing higher precision and lower recall for minority classes.

Table S5: Per-class F1 for XGBoost on the defect manifestation type task (test set, after SMOTE).

Manifestation Class	Mobile F1	Web F1	ASRS F1
Data Validation Issue	0.85	0.92	0.86
Incorrect Data Handling	0.65	0.48	0.45
Incorrect Error Handling/Messages	0.70	0.85	0.58
Incorrect Output/Results	0.50	0.62	0.60
Missing Functionality	0.44	0.65	0.58
State-Management Issues	—	0.93	—
Performance Issue	0.92	0.88	0.90
System Crashes or Failure	0.72	0.88	0.70
UI/UX Errors	0.62	0.75	0.82
Unexpected Behavior	0.60	0.78	0.32
Workflow Logic Error	0.78	0.84	0.52
Macro avg	0.70	0.76	0.61

2.3 Gradient Boosting

Table S6 reports per-class F1 for Gradient Boosting on the defect manifestation type task. Gradient Boosting achieved macro-F1 of 0.663 (Mobile), 0.748 (Web), and 0.581 (ASRS).

Table S6: Per-class F1 for Gradient Boosting on the defect manifestation type task (test set, after SMOTE).

Manifestation Class	Mobile F1	Web F1	ASRS F1
Data Validation Issue	0.82	0.90	0.83
Incorrect Data Handling	0.62	0.45	0.42
Incorrect Error Handling/Messages	0.68	0.82	0.55
Incorrect Output/Results	0.47	0.58	0.56
Missing Functionality	0.41	0.62	0.54
State-Management Issues	—	0.90	—
Performance Issue	0.89	0.85	0.87
System Crashes or Failure	0.68	0.84	0.65
UI/UX Errors	0.58	0.70	0.78
Unexpected Behavior	0.55	0.72	0.28
Workflow Logic Error	0.74	0.80	0.48
Macro avg	0.66	0.75	0.58

Notes

- All metrics are computed on the held-out test set. SMOTE was applied only to the training partition; the test set reflects the natural class distribution for balanced evaluation (stratified sampling to 227–254 instances per class for primary task; 231–262 per class for manifestation task, dataset-dependent).
- Macro average excludes State-Management Issues for Mobile and ASRS datasets, as this class is present only in the Web dataset.

- The primary defect type task has 4 classes: UI Error, Behavioral Error, Data Error, System Failure. The manifestation task has 10 classes for Mobile and ASRS, and 11 classes for Web (including State-Management Issues).

References

References

- [1] Helwa, H. FORESIGHT: Software Defects Prediction from Requirements Change Requests. Master's thesis, Birzeit University, 2024. Available at: https://koha.birzeit.edu/cgi-bin/koha/opac-detail.pl?biblionumber=327439&shelfbrowse_itemnumber=330130