Early Detection of Cracking in Grapes under Field Conditions using Hyperspectral Imaging and Machine Learning

By Yovel Gani

Supervised by:

Dr. Alon Sela, Prof. Victor Alchanatis, Prof. Irad Ben-Gal

August 2026

A thesis submitted toward the degree of Master of Science (M.Sc.)

Abstract

Grape cracking is a severe physiological disorder that leads to substantial economic losses in table-grape production by reducing marketable yield and increasing susceptibility to secondary infections such as sour rot. Early, non-destructive detection is therefore critical for enabling timely management interventions. This study investigates the feasibility of hyperspectral imaging (HSI; VNIR 400–1000 nm) combined with machine learning for crack detection under field conditions, considering both early- and late-stage acquisitions (initial/minimal vs. advanced cracking). All spectral analyses were restricted to 450–925 nm and employed per-spectrum Standard Normal Variate (SNV) normalization.

A multi-stage analytical framework was developed, including pixel-level classification, whole-image crack detection, and wavelength selection. Pixel-level spectral signatures of healthy and cracked berries were manually annotated and used to train an XGBoost classifier, achieving high discrimination performance (accuracy $\approx 97\%$, AUC ≈ 0.99) and confirming the existence of a distinct spectral separation between tissue states.

Direct application of pixel-level predictions to field scenes resulted in background-induced misclassification. To address this limitation, a whole-image crack-detection pipeline was developed, aggregating pixel-level CRACK probabilities using spatial post-processing (thresholding, morphological filtering, and patch/blob-based decision rules). The system was evaluated using a spatially disjoint row-based calibration–test split and reported with emphasis on the Cracked class. On the held-out test row, the pipeline achieved $\mathrm{F1}_{\mathrm{CRACK}} = 0.704$ under early-stage acquisition and $\mathrm{F1}_{\mathrm{CRACK}} = 0.818$ under late-stage acquisition.

Wavelength selection experiments demonstrated that comparable detection performance can be retained using reduced spectral-band subsets, indicating the potential for compact sensing configurations suitable for field deployment. Overall, the results establish the feasibility of HSI-based crack detection under realistic vineyard conditions and provide a methodological foundation for operational, non-destructive monitoring tools.

Contents

1 Experimental Results **7**

1.1 Overview of Experimental Results . . . 7

1.1.1 Evaluation Baselines . . . 8

1.2 Spectral Separability and Informative Wavelengths . . . 9

1.2.1 Mean Spectral Signatures of Healthy and Cracked Tissue . . . 9

1.2.2 Spectral Separability Across Wavelength Regions . . . 10

1.3 Pixel-Level Binary Classification . . . 11

1.4 Pixel-Level Multi-Class Classification . . . 12

1.5 Autoencoder-Based Anomaly Detection . . . 13

1.6 Whole-Image Crack Detection . . . 15

1.7 Wavelength Selection via Backward Feature Selection . . . 18

1.7.1 BFS Trajectories and Subset Selection . . . 18

1.7.2 End-to-End Wavelength-Selection Results . . . 20

1.7.3 BFS Wavelength-Selection Stability Analysis . . . 21

A Supplementary Results **29**

A.1 Cross-Validation Statistics and Full Precision Tables . . . 29

A.1.1 Pixel-Level Classification: Binary and Multi-Class . . . 29

A.1.2 Autoencoder-Based Anomaly Detection . . . 29

A.2 Regional VIS–NIR Reflectance Summaries . . . 29

A.3 Binary Pixel-Level Classification (Balanced) . . . 32

A.4 Multi-Class Pixel-Level Classification (Balanced) . . . 33

A.5 Autoencoder Anomaly Detection (Balanced) . . . 34

A.5.1 One-Class Binary Formulation . . . 34

A.5.2 Crack-Only Training Formulation . . . 34

A.5.3 Non-Crack Training Formulation . . . 35

A.6 Wavelength Selection: Generalization and Performance Analysis . . . 36

A.6.1 BFS Thresholds (Full Precision) . . . 36

A.6.2 Wavelength Results (Full Precision) . . . 36

A.6.3 Generalization Analysis . . . 36

A.6.4 Wavelength Overlap Across Seeds 36

List of Figures

1.1 Mean spectral signatures (± standard deviation) for healthy and cracked tissue across the VIS–NIR range. 9

1.2 Wavelength-wise spectral separability metrics between healthy and cracked tissue, including absolute mean difference, Cohen's d, and Fisher score. . 10

1.3 BFS trajectories for the whole-image pipeline: CRACK-class PR–AUC (primary objective) and CRACK-class F1 versus the number of retained wavelengths (n; full baseline n=159). 18

1.4 BFS threshold-marker visualization on CRACK metrics. Markers indicate the maximum and the first n where performance drops more than 0.5% and 1.0% below that maximum. 19

1.5 Distribution and popularity of the top-30 selected wavelengths across five varying-split seeds. 21

1.6 Wavelength popularity at different bin sizes (1, 5, 10, 20 nm) for the top-30 subsets. 22

1.7 Stability heatmap for the top-30 BFS subsets across five seeds (objective: PR–AUC$_{\text{CRACK}}$). Each cell indicates how many seeds selected a wavelength within that 5 nm bin. 23

1.8 Wavelength selection frequency along the spectrum (2 nm sliding window) for the top-30 subsets. 24

1.9 Distribution and popularity of the top-11 selected wavelengths across five varying-split seeds. 25

1.10 Wavelength popularity at different bin sizes (1, 5, 10, 20 nm) for the top-11 subsets. 26

1.11 Stability heatmap for the top-11 BFS subsets across five seeds (objective: PR–AUC$_{\text{CRACK}}$; 0.5% drop setting). Each cell indicates how many seeds selected a wavelength within that 5 nm bin. 27

1.12 Wavelength selection frequency along the spectrum (2 nm sliding window) for the top-11 subsets. 28

A.1 Regional mean reflectance (±1 SD) and absolute mean class difference for the VIS and NIR ranges. 31

List of Tables

1.1 Pixel-level classification metrics (unbalanced, binary setting). Truncated mean values. Full cross-validation statistics in Appendix A.1. 11
1.2 Pixel-level classification metrics (unbalanced, multi-class). Truncated mean values. Full cross-validation statistics in Appendix A.1. 12
1.3 Pixel-level autoencoder anomaly detection (binary setting). Truncated mean values. Full statistics in Appendix A.1. 13
1.4 Pixel-level autoencoder anomaly detection (crack-only training). Truncated mean values. Full statistics in Appendix A.1. 14
1.5 Pixel-level autoencoder anomaly detection (all non-CRACK training). Truncated mean values. Full statistics in Appendix A.1. 14
1.6 Dataset composition for full-image crack-detection experiments under early- and late-stage acquisition conditions (initial/minimal vs. advanced cracking). Row 1 was used for calibration, while Row 2 served as an independent test set. 15
1.7 Full-image crack detection performance for early- and late-stage acquisition conditions. F1 (Cracked) is the primary metric; the Cracked class is positive. 16
1.8 Full-image crack detection using autoencoder-based anomaly detection. F1 (Cracked) is the primary metric; the Cracked class is positive. 17
1.9 Summary of BFS wavelength-count thresholds for CRACK metrics. (Full precision in Appendix A.6.1) 18
1.10 Full-image crack-detection performance for different wavelength subsets under early- and late-stage conditions. (Full precision in Appendix A.6.2) 20

A.1 Pixel-level classification metrics (unbalanced training set, binary). Values are reported as mean ± standard deviation across folds where available. Precision/recall and ROC–AUC refer to the *Cracked* class (positive class); F1 is the weighted F1-score. 29
A.2 Pixel-level classification metrics (**unbalanced** training set, multi-class). Values are reported as mean ± standard deviation across folds. 30

A.3 Pixel-level autoencoder anomaly detection: binary setting (non-cracked as normal class). Values are reported as mean ± standard deviation across LOGO folds. 30
A.4 Pixel-level autoencoder anomaly detection: trained on CRACK pixels (multi-class setting). Values are mean across LOGO folds. 30
A.5 Pixel-level autoencoder anomaly detection: trained on all non-CRACK classes (multi-class setting). Single group-wise train–test split. 30
A.6 Summary statistics of reflectance values across spectral regions for healthy and cracked tissue. 30
A.7 Regional spectral separability metrics between healthy and cracked tissue. Mean values are averaged across wavelengths within each range, while peak values indicate the maximum separability and the corresponding wavelength. 31
A.8 Pixel-level classification metrics (balanced training set). Values are reported as mean ± standard deviation across folds where available. Precision/recall and ROC–AUC refer to the *Cracked* class (positive class); F1 is the weighted F1-score. 32
A.9 Pixel-level classification metrics (**balanced** training set). Values are reported as mean ± standard deviation across folds. 33
A.10 Pixel-level anomaly detection performance using autoencoders trained on non-cracked samples (balanced sampling). Values are reported as mean ± standard deviation across folds. Accuracy is reported overall, while Precision/Recall/F1/ROC–AUC/PR–AUC correspond to the *Cracked* class (positive class). 34
A.11 Pixel-level anomaly detection performance using autoencoders trained exclusively on *Cracked* pixels and evaluated in a multi-class setting (balanced sampling). Values are reported as mean ± standard deviation across LOGO folds. Accuracy is reported overall, while Precision/Recall/F1/ROC–AUC/PR–AUC correspond to the *Cracked* class (positive class). 34
A.12 Pixel-level anomaly detection performance using autoencoders trained on all semantic classes except *Cracked* and evaluated in a multi-class setting (balanced sampling). Values are reported for the held-out test split. Accuracy is reported overall, while Precision/Recall/F1/ROC–AUC/PR–AUC correspond to the *Cracked* class (positive class). 35
A.13 Summary of BFS wavelength-count thresholds for CRACK metrics (full precision). 36
A.14 Full-image crack-detection performance for different wavelength subsets under early- and late-stage conditions (full precision). 36
A.15 Model Generalization Analysis: Comparison between Calibration and Test MCC for wavelength subsets. 37

A.16 Pairwise Jaccard similarity of top-30 wavelengths across seeds (varying split). Mean Jaccard = 0.493. 37
A.17 Pairwise Jaccard similarity of top-11 wavelengths across seeds (varying split). Mean Jaccard = 0.398. 37

Chapter 1

Experimental Results

1.1 Overview of Experimental Results

This chapter presents the experimental evaluation of the hyperspectral crack-detection framework. The findings are structured to systematically address the research hypotheses formulated in Chapter 3, progressing from fundamental spectral characterization to operational field-level detection. Specifically, the chapter is divided into four main sections:

1. Validation of the physical and spectral distinguishability between healthy and cracked tissues.
2. Pixel-level classification benchmarking.
3. Whole-image spatial aggregation and detection under early- and late-stage field conditions.
4. Wavelength selection analysis to determine the minimal spectral bands required for robust inference.

Presentation Note: For clarity and readability, the performance metrics reported in the main text and tables are presented as mean values truncated to three decimal places. Comprehensive cross-validation statistics, including standard deviations and fold-level data, are provided in Appendix A.1.

1.1.1 Evaluation Baselines

To ensure a rigorous and transparent evaluation, all experimental results are benchmarked against the following explicit baselines:

- **Pixel-Level Classification Baseline:** The reference model is **XGBoost** trained on the **full SNV-normalized hyperspectral signature** (450–925 nm, 159 contiguous bands), as defined in Section **??**. Binary-task performance is reported in the XGBoost rows of Table 1.1. Results for extended labeling and balancing regimes appear in Table 1.2, with balanced-training baselines detailed in Appendix A.

- **Whole-Image Crack-Detection Baseline:** System-level detection is evaluated using the spatial aggregation pipeline described in Section **??**. The baseline performance for this task relies on probability maps generated by the full-spectrum (159-band) XGBoost pixel-level model, establishing the upper bound for detection accuracy prior to any wavelength reduction. Early- and late-stage calibration and held-out test performances are summarized in Table 1.7.

- **Wavelength Selection (BFS) Reference:** The efficacy of Backward Feature Selection (BFS) is quantified by comparing the classification performance of reduced wavelength subsets directly against the full 159-band baseline. Elimination trajectories are illustrated in Figures 1.3 and 1.4, and whole-image test-set results at critical wavelength thresholds are reported in Table 1.10.

1.2 Spectral Separability and Informative Wavelengths

Prior to evaluating machine learning algorithms, a per-band spectral separability analysis was conducted to verify that healthy and cracked tissues exhibit physically distinguishable signatures, addressing Hypothesis 1. For this fundamental physical characterization, the relative reflectance spectra were analyzed *prior* to the application of per-pixel SNV normalization in order to preserve and quantify the absolute amplitude differences between the classes.

1.2.1 Mean Spectral Signatures of Healthy and Cracked Tissue

The mean spectral reflectance signatures ($\pm$ SD) for healthy and cracked grape tissues across the VIS–NIR range are presented in Figure 1.1. Cracked tissue consistently exhibits lower absolute reflectance across the full spectrum compared to intact healthy tissue. As visually evident in the plot, this divergence is exceptionally pronounced in the Near-Infrared (NIR) region, specifically between 700 nm and 850 nm.

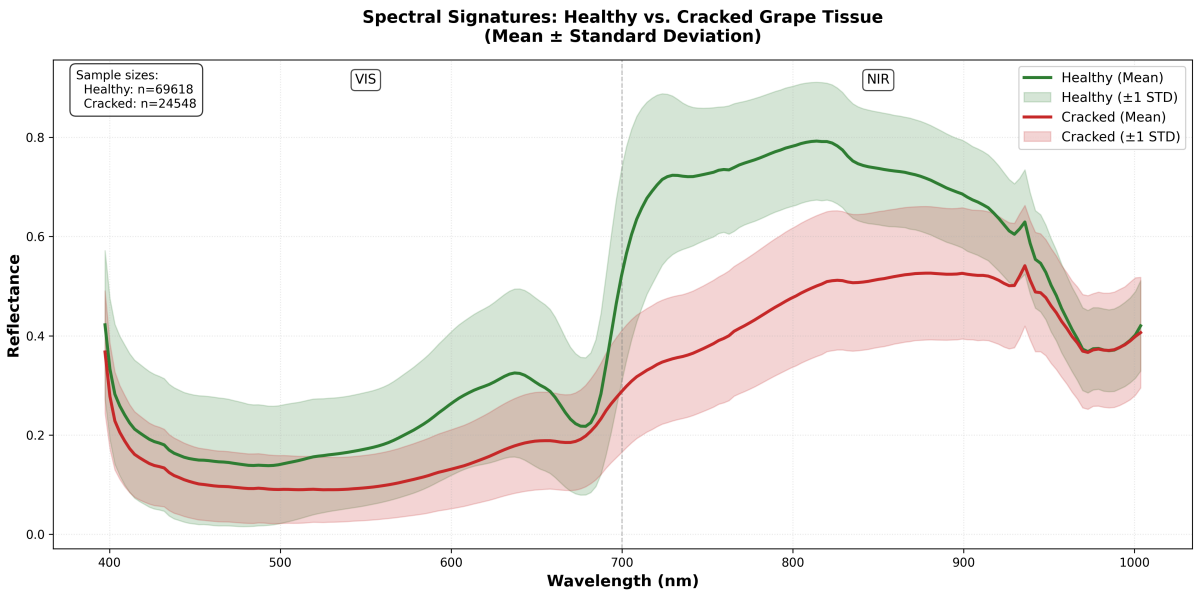


Figure 1.1: Mean spectral signatures ($\pm$ standard deviation) for healthy and cracked tissue across the VIS–NIR range.

1.2.2 Spectral Separability Across Wavelength Regions

To systematically quantify the discriminatory power of individual spectral bands, three complementary metrics were computed (Figure 1.2): **Absolute Mean Difference ($|\Delta\mu|$)** to capture the raw reflectance gap, **Cohen's d Cohen, 1988** to measure the standardized effect size, and the **Fisher Score Duda et al., 2001** to evaluate the ratio of inter-class separation to intra-class spread.

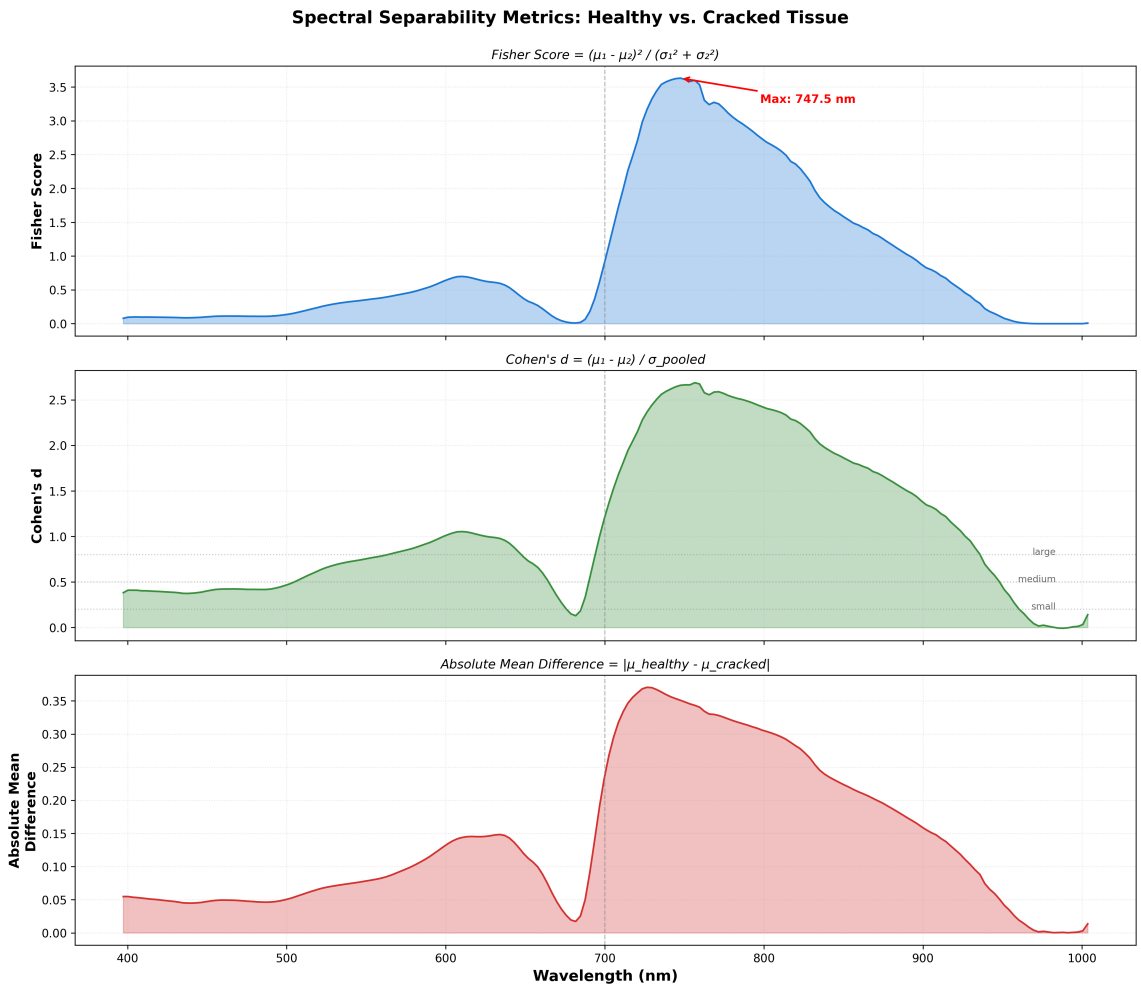


Figure 1.2: Wavelength-wise spectral separability metrics between healthy and cracked tissue, including absolute mean difference, Cohen's d, and Fisher score.

The separability metrics indicate that class separation is substantially higher in the NIR region, with peak discriminability consistently observed between 727–757 nm. As detailed in the regional summaries in Appendix A.2, healthy tissue reflectance exceeds that of cracked tissue by an average of 32% in the NIR and 37% in the VIS range. These clear physical differences confirm Hypothesis 1 and provide the basis for automated classification.

1.3 Pixel-Level Binary Classification

Supervised pixel-level classification addresses Hypothesis 2, following the LOGO cross-validation protocol described in Section **??**. PR–$\mathrm{AUC}_{\mathrm{CRACK}}$ is the primary metric.

Table 1.1 reports performance across LOGO folds (Section **??**; unbalanced class distribution). Balanced-training results are in Appendix A.

Table 1.1: Pixel-level classification metrics (unbalanced, binary setting). Truncated mean values. Full cross-validation statistics in Appendix A.1.

Model	Accuracy	Precision (Cracked)	Recall (Cracked)	Weighted F1	ROC–AUC (Cracked)
PLS-DA	0.992	0.999	0.968	0.992	1.000
Logistic Regression (L1)	0.996	0.989	0.994	0.996	0.999
SVM (RBF)	0.993	0.994	0.978	0.993	0.999
Random Forest	0.996	0.996	0.990	0.996	0.999
XGBoost	0.994	0.995	0.979	0.994	1.000

All five models achieve PR–$\mathrm{AUC}_{\mathrm{CRACK}} \geq 0.987$ (Table 1.1), supporting Hypothesis 2.

1.4 Pixel-Level Multi-Class Classification

Binary detection was extended to three labeling strategies (*crack_regular_rest*, *crack_vs_rest*, *multi_class*) as specified in Section **??**; PR–AUC$_{\text{CRACK}}$ is the primary metric. Balanced-training results are in Appendix A.

Table 1.2: Pixel-level classification metrics (unbalanced, multi-class). Truncated mean values. Full cross-validation statistics in Appendix A.1.

Label setup	Model	Accuracy	Balanced Acc.	Precision (Cracked)	Recall (Cracked)	Macro-F1	ROC–AUC	PR–AUC
crack_regular_rest								
	PLS-DA	0.508	0.687	0.882	0.753	0.581	0.997	0.863
	Logistic Regression (L1)	0.853	0.868	0.965	0.966	0.801	0.999	0.983
	Random Forest	0.917	0.917	0.894	0.939	0.890	0.997	0.962
	XGBoost	0.921	0.916	0.980	0.949	0.883	0.999	0.990
	MLP (Small)	0.917	0.916	0.953	0.972	0.875	0.999	0.991
crack_vs_rest								
	PLS-DA	0.512	0.694	0.884	0.762	0.589	0.997	0.866
	Logistic Regression (L1)	0.857	0.870	0.966	0.965	0.804	0.999	0.984
	Random Forest	0.914	0.915	0.887	0.936	0.885	0.997	0.961
	XGBoost	0.920	0.915	0.980	0.950	0.882	0.999	0.990
	MLP (Small)	0.916	0.915	0.952	0.973	0.874	0.999	0.991
multi_class								
	PLS-DA	0.967	0.978	0.720	0.992	0.901	0.997	0.967
	Logistic Regression (L1)	0.997	0.995	0.967	0.997	0.992	0.999	0.997
	Random Forest	0.996	0.994	0.967	0.996	0.990	0.999	0.997
	XGBoost	0.997	0.994	0.981	0.990	0.992	0.999	0.997
	MLP (Small)	0.990	0.989	0.924	0.986	0.969	0.996	0.989

The *multi_class* formulation yields the highest PR–AUC$_{\text{CRACK}}$ (≥ 0.967); LR, RF, and XGBoost all exceed 0.99 (Table 1.2).

1.5 Autoencoder-Based Anomaly Detection

In the one-class binary setting, the autoencoder achieved PR–AUC$_{\text{CRACK}}$ ranging from 0.925 to 0.970 (Table 1.3).

When trained on CRACK pixels as the normal class, PR–AUC$_{\text{CRACK}} \approx 1.0$ (Table 1.4), indicating trivial reconstruction of CRACK signatures. Training on all non-CRACK classes with CRACK defined as the anomaly yielded PR–AUC$_{\text{CRACK}} \approx 0.01$ (Table 1.5).

Supervised models outperform autoencoders for CRACK detection. Additional autoencoder formulations are reported in Appendix A.1.

Table 1.3: Pixel-level autoencoder anomaly detection (binary setting). Truncated mean values. Full statistics in Appendix A.1.

Architecture	Accuracy	Precision (Cracked)	Recall (Cracked)	F1 (Cracked)	ROC–AUC (Cracked)	PR–AUC (Cracked)
Autoencoder (64–32–16)	0.896	0.822	0.997	0.901	0.981	0.955
Autoencoder (128–64–32)	0.859	0.754	0.997	0.862	0.967	0.925
Autoencoder (64–32–8)	0.892	0.812	0.997	0.896	0.979	0.951

Table 1.4: Pixel-level autoencoder anomaly detection (crack-only training). Truncated mean values. Full statistics in Appendix A.1.

Architecture	Accuracy	Precision (non-CRACK)	Recall (non-CRACK)	F1 (non-CRACK)	ROC–AUC (non-CRACK)	PR–AUC (non-CRACK)
Autoencoder (64–32–16)	0.966	0.999	0.966	0.982	0.985	0.999
Autoencoder (128–64–32)	0.986	0.999	0.986	0.993	0.984	0.999
Autoencoder (64–32–8)	0.962	0.999	0.963	0.980	0.983	0.999

Table 1.5: Pixel-level autoencoder anomaly detection (all non-CRACK training). Truncated mean values. Full statistics in Appendix A.1.

Architecture	Accuracy	Precision (Cracked)	Recall (Cracked)	F1 (Cracked)	ROC–AUC (Cracked)	PR–AUC (Cracked)
Autoencoder (64–32–16)	0.996	0.009	0.111	0.017	0.457	0.008
Autoencoder (128–64–32)	0.998	0.011	0.012	0.012	0.451	0.008
Autoencoder (64–32–8)	0.998	0.020	0.031	0.025	0.454	0.011

Overall, autoencoder-based formulations yielded lower PR–AUC$_{\text{CRACK}}$ than the supervised XGBoost baseline and were therefore not adopted for subsequent pixel-level or whole-image detection stages.

1.6 Whole-Image Crack Detection

Using the protocol described in Section **??**, whole-image detection was assessed under early- and late-stage conditions, addressing Hypothesis 3. The primary metric is $\mathbf{F1_{CRACK}}$. The baseline for whole-image comparison is the full-spectrum XGBoost aggregation pipeline (159 bands, 450–925 nm) without wavelength selection, as defined in Section 1.1.1. Dataset composition is in Table 1.6; supervised and autoencoder results are in Tables 1.7 and 1.8, respectively.

Table 1.6: Dataset composition for full-image crack-detection experiments under early- and late-stage acquisition conditions (initial/minimal vs. advanced cracking). Row 1 was used for calibration, while Row 2 served as an independent test set.

Row	Stage	Split	Total	Unique	Pos	Neg
Row 1	Early	Calibration	172	60	32	140
Row 1	Late	Calibration	160	60	35	125
Row 2	Early	Test	60	60	32	28
Row 2	Late	Test	60	60	34	26

Table 1.7: Full-image crack detection performance for early- and late-stage acquisition conditions. F1 (Cracked) is the primary metric; the Cracked class is positive.

Stage	Split	TP	FP	TN	FN	Acc	Balanced Accuracy	Precision (Cracked)	Recall (Cracked)	F1 (Cracked)	F2 (Cracked)	Specificity (Healthy)	NPV (Healthy)	MCC
Late	Calibration	34	5	120	1	0.963	0.966	0.872	0.971	0.919	0.950	0.960	0.992	0.897
Late	Test	27	5	21	7	0.800	0.801	0.844	0.794	0.818	0.804	0.808	0.750	0.598
Early	Calibration	23	6	134	9	0.913	0.838	0.793	0.719	0.754	0.732	0.957	0.937	0.703
Early	Test	19	3	25	13	0.733	0.743	0.864	0.594	0.704	0.633	0.893	0.658	0.504

On the independent test set (Table 1.7), under late-stage conditions the supervised pipeline achieves $\mathrm{F1}_{\mathrm{CRACK}} = 0.818$. Under early-stage conditions $\mathrm{F1}_{\mathrm{CRACK}} = 0.704$.

Table 1.8: Full-image crack detection using autoencoder-based anomaly detection. F1 (Cracked) is the primary metric; the Cracked class is positive.

Stage	Split	TP	FP	TN	FN	Acc	Balanced Accuracy	Precision (Cracked)	Recall (Cracked)	F1 (Cracked)	F2 (Cracked)	Specificity (Healthy)	NPV (Healthy)	MCC
Late	Calibration	26	7	118	9	0.900	0.843	0.788	0.743	0.765	0.751	0.944	0.929	0.702
Late	Test	22	7	19	12	0.683	0.689	0.759	0.647	0.698	0.667	0.731	0.613	0.375
Early	Calibration	19	17	123	13	0.826	0.736	0.528	0.594	0.559	0.579	0.879	0.904	0.452
Early	Test	13	13	15	19	0.467	0.471	0.500	0.406	0.448	0.422	0.536	0.441	−0.058

On the independent test set (Table 1.8), the autoencoder-based pipeline yields lower $\mathrm{F1}_{\mathrm{CRACK}}$ than the supervised classifier in both stages (Late: $\mathrm{F1}_{\mathrm{CRACK}} = 0.698$ vs. $\mathrm{F1}_{\mathrm{CRACK}} = 0.818$; Early: $\mathrm{F1}_{\mathrm{CRACK}} = 0.448$ vs. $\mathrm{F1}_{\mathrm{CRACK}} = 0.704$).

The effect of wavelength selection on the whole-image pipeline is evaluated in the following section using backward feature selection.

1.7 Wavelength Selection via Backward Feature Selection

BFS was applied using the protocol described in Section **??**, optimizing PR–AUC$_{\text{CRACK}}$ on the whole-image pipeline. The full-spectrum baseline is 159 wavelengths (450–925 nm).

1.7.1 BFS Trajectories and Subset Selection

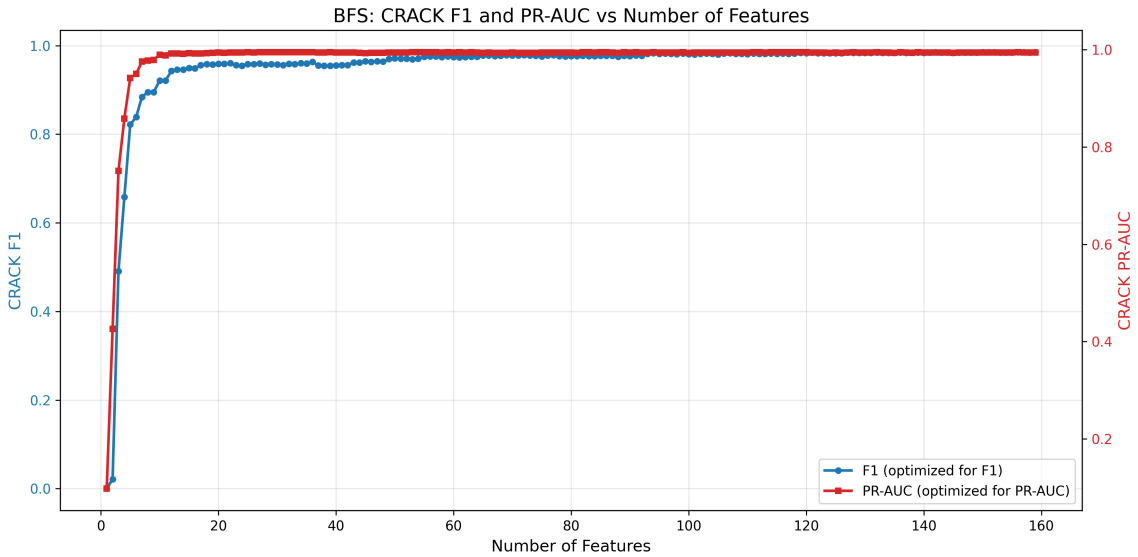
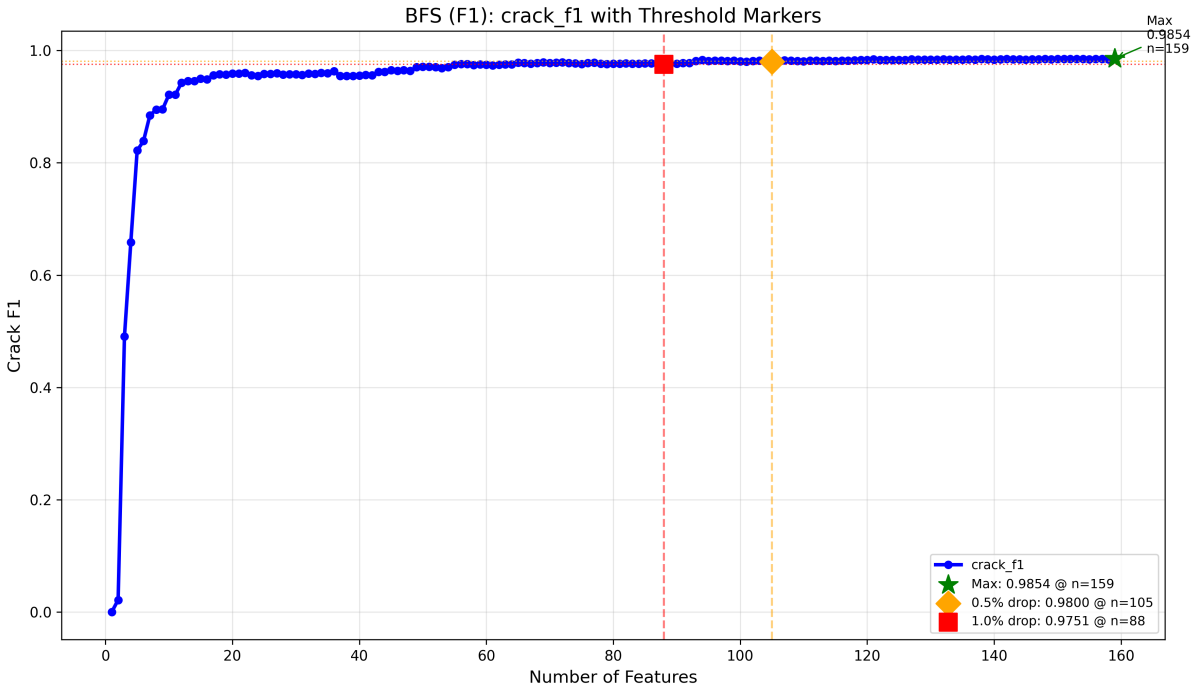


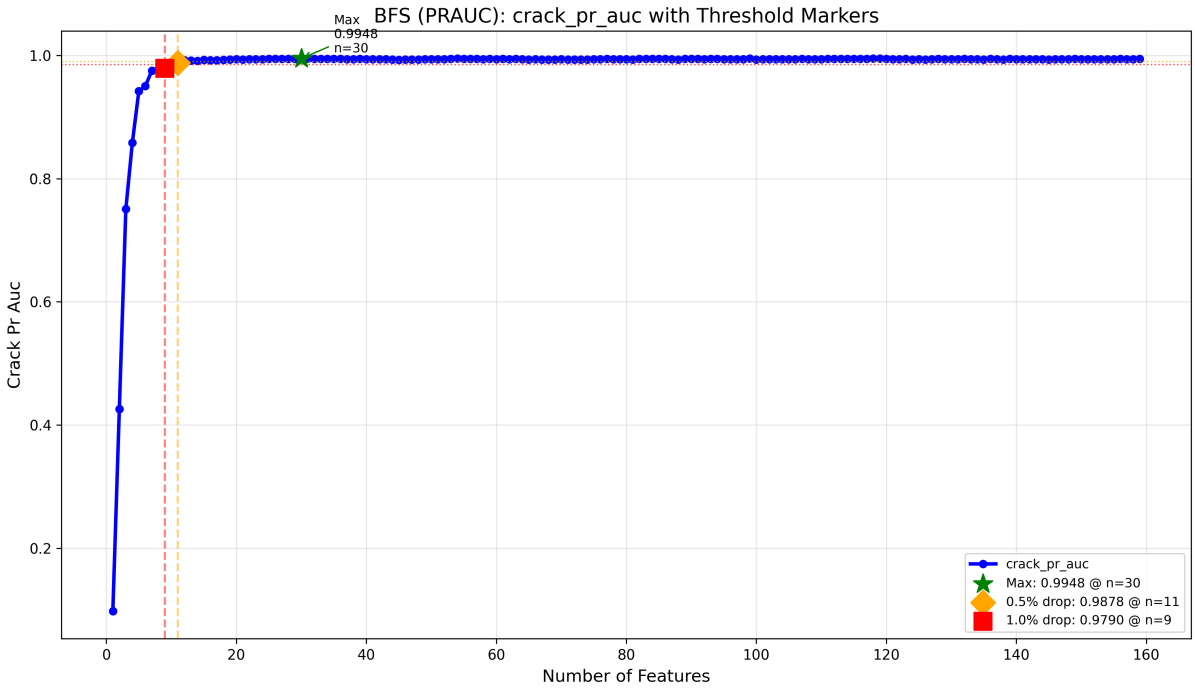
Figure 1.3: BFS trajectories for the whole-image pipeline: CRACK-class PR–AUC (primary objective) and CRACK-class F1 versus the number of retained wavelengths (n; full baseline n=159).

Table 1.9: Summary of BFS wavelength-count thresholds for CRACK metrics. (Full precision in Appendix A.6.1)

Metric	Max score	$n_{\max}$	0.5% drop: n	score	1.0% drop: n	score
CRACK PR–AUC	0.994	30	11	0.987	9	0.979
CRACK F1	0.985	159	105	0.980	88	0.975



(a) CRACK-class F1 vs. number of wavelengths (n), with markers at the maximum and at 0.5%/1.0% drop points.



(b) CRACK-class PR–AUC (primary BFS objective) vs. number of wavelengths (n), with the same threshold markers.

Figure 1.4: BFS threshold-marker visualization on CRACK metrics. Markers indicate the maximum and the first n where performance drops more than 0.5% and 1.0% below that maximum.

1.7.2 End-to-End Wavelength-Selection Results

Table 1.10 reports whole-image detection metrics for representative wavelength subsets, evaluated separately for early- and late-stage acquisitions on the held-out test set.

Table 1.10: Full-image crack-detection performance for different wavelength subsets under early- and late-stage conditions. (Full precision in Appendix A.6.2)

Stage	# Wavelengths	Accuracy	Balanced Accuracy	Precision (Cracked)	Recall (Cracked)	F1 (Cracked)	MCC
Early	9	0.783	0.790	0.880	0.687	0.771	0.587
Late	9	0.800	0.805	0.866	0.764	0.812	0.605
Early	11	0.750	0.757	0.840	0.656	0.736	0.519
Late	11	0.816	0.811	0.828	0.852	0.840	0.625
Early	30	**0.816**	**0.821**	**0.888**	**0.750**	**0.813**	**0.644**
Late	30	**0.833**	**0.834**	**0.875**	0.823	0.848	**0.665**
Early	159	0.800	0.803	0.857	**0.750**	0.800	0.607
Late	159	**0.833**	0.825	0.833	**0.882**	**0.857**	0.659

The top-30 subset achieves $F1_{CRACK} = 0.8136$ (Early) and 0.8485 (Late), matching the 159-band baseline in Early stage (0.8000) and approaching it in Late stage (0.8571); the top-11 subset yields $F1_{CRACK} = 0.7368$ (Early) and 0.8406 (Late). Additional metrics and generalization-gap analysis are provided in Appendix A.6.

1.7.3 BFS Wavelength-Selection Stability Analysis

Cross-seed wavelength overlap statistics are reported in Appendix A.6.

Wavelength Popularity and Stability – Top 30

The top-30 wavelength subsets demonstrate strong cross-seed consistency in their selection patterns, with dominant spectral regions remaining stable under coarser binning schemes. The top-30 subset exhibits substantially higher cross-seed stability compared to smaller subsets such as the top-11. The consistently selected regions are predominantly located in the near-infrared (NIR) portion of the spectrum. Figures 1.5–1.8 provide comprehensive visualization of this cross-seed wavelength stability.
Exact wavelength identities and cross-seed overlap statistics are provided in Appendix A.6.

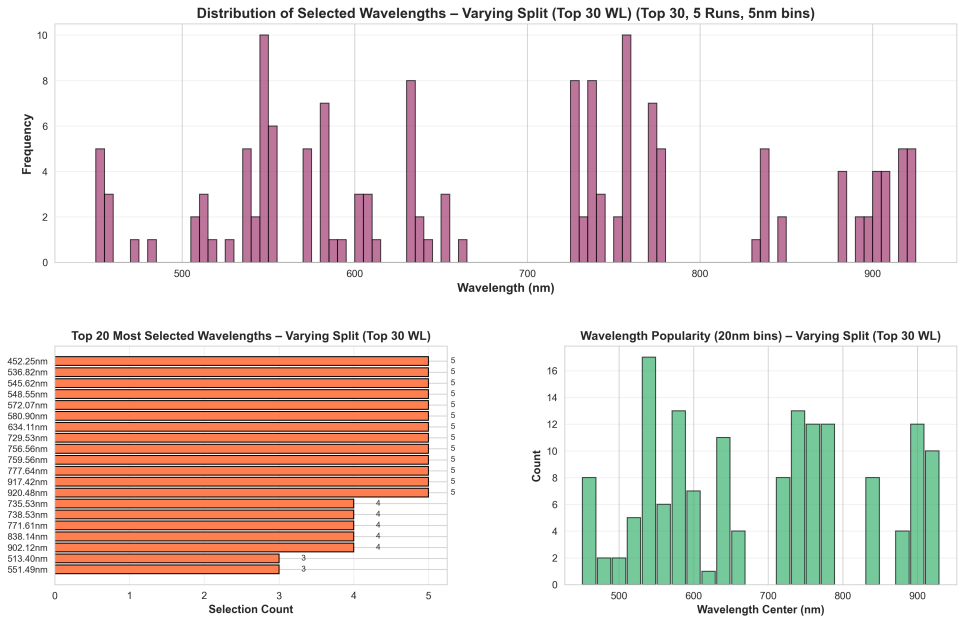


Figure 1.5: Distribution and popularity of the top-30 selected wavelengths across five varying-split seeds.

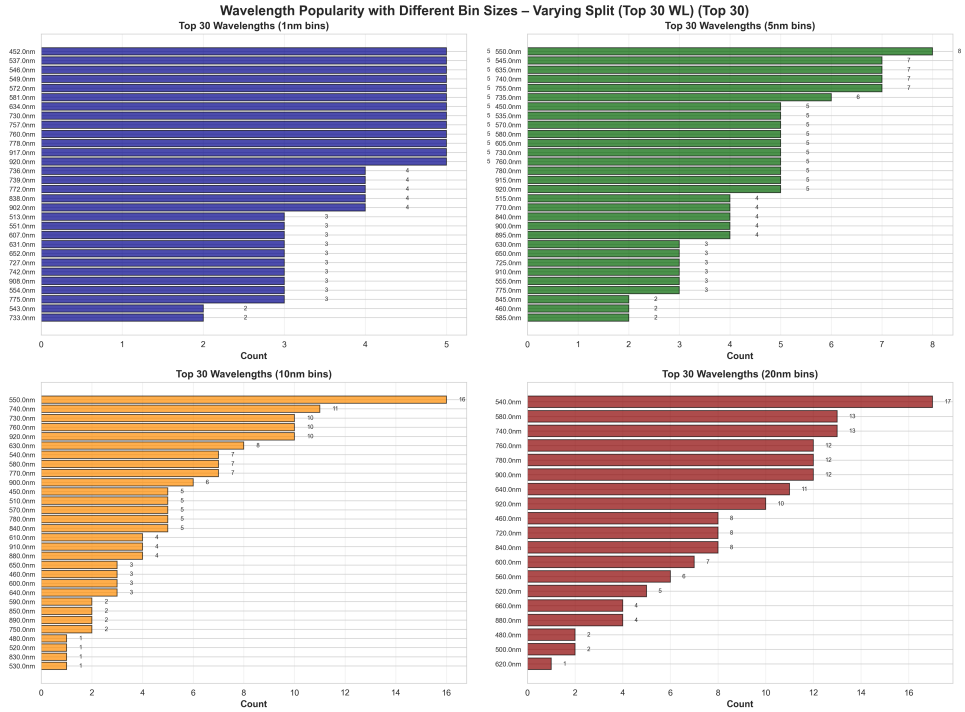


Figure 1.6: Wavelength popularity at different bin sizes (1, 5, 10, 20 nm) for the top-30 subsets.

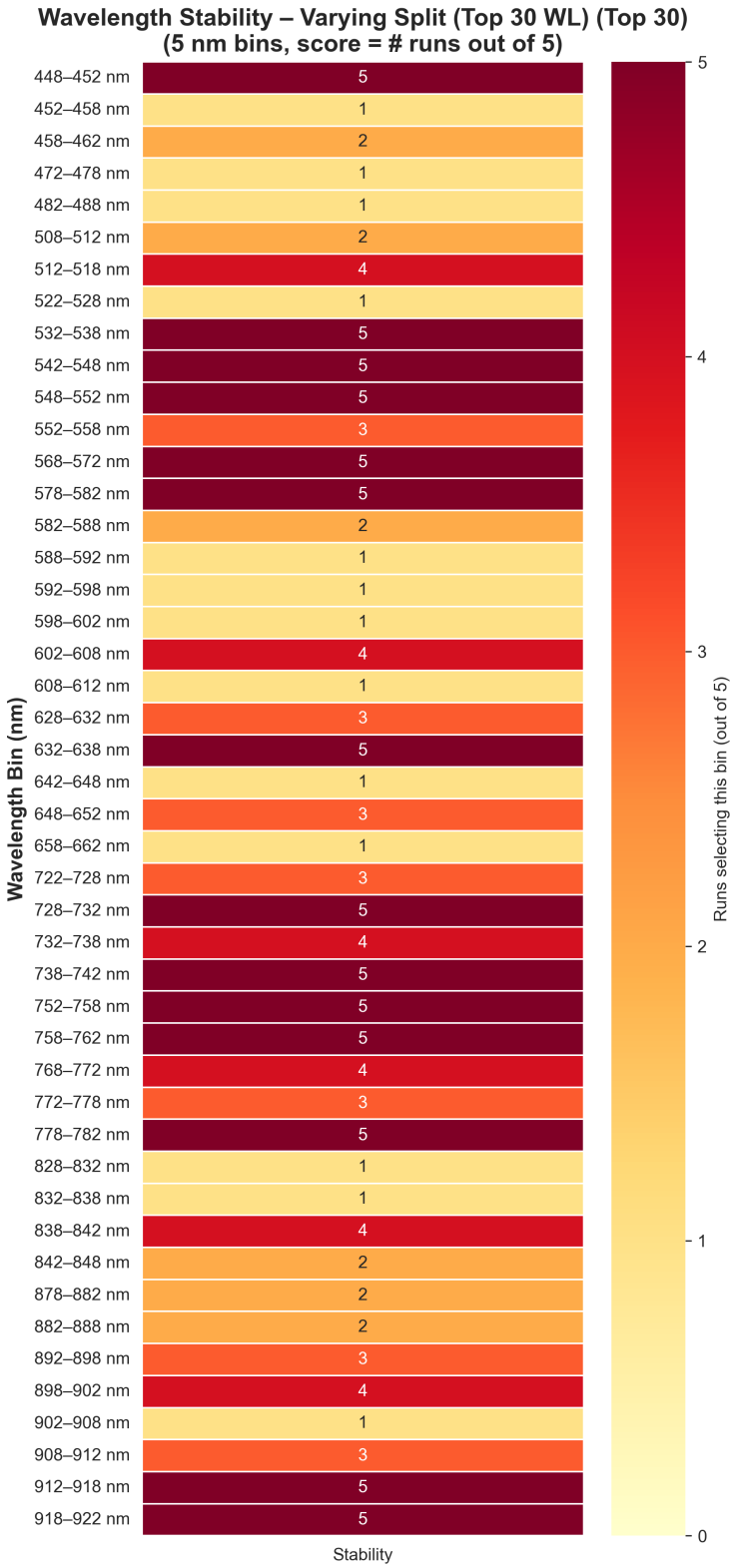


Figure 1.7: Stability heatmap for the top-30 BFS subsets across five seeds (objective: PR–AUC$_{\text{CRACK}}$). Each cell indicates how many seeds selected a wavelength within that 5 nm bin.

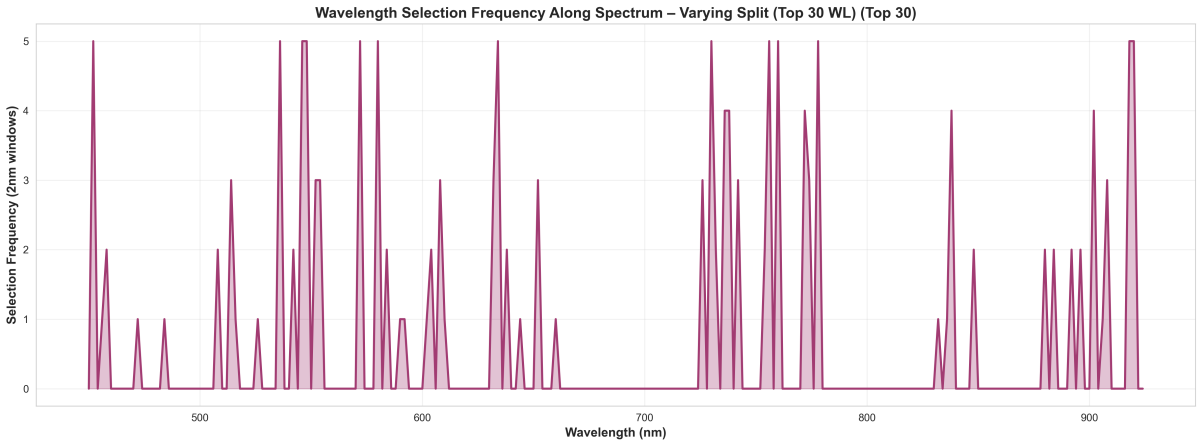


Figure 1.8: Wavelength selection frequency along the spectrum (2 nm sliding window) for the top-30 subsets.

Wavelength Popularity and Stability – Top 11

The top-11 wavelength subsets (within 1% of peak PR–AUC$_{\text{CRACK}}$) exhibit lower cross-seed stability than the top-30 subset, a consequence of the stronger dimensionality reduction. Nevertheless, consistent spectral regions remain identifiable across seeds (Figures 1.9–1.12). Detailed overlap statistics are reported in Appendix A.6.

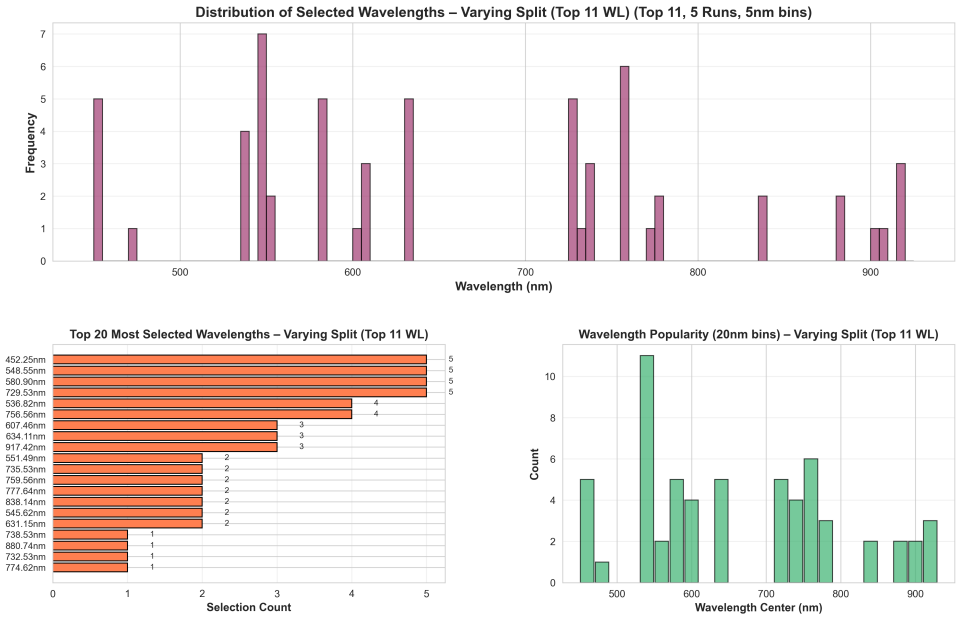


Figure 1.9: Distribution and popularity of the top-11 selected wavelengths across five varying-split seeds.

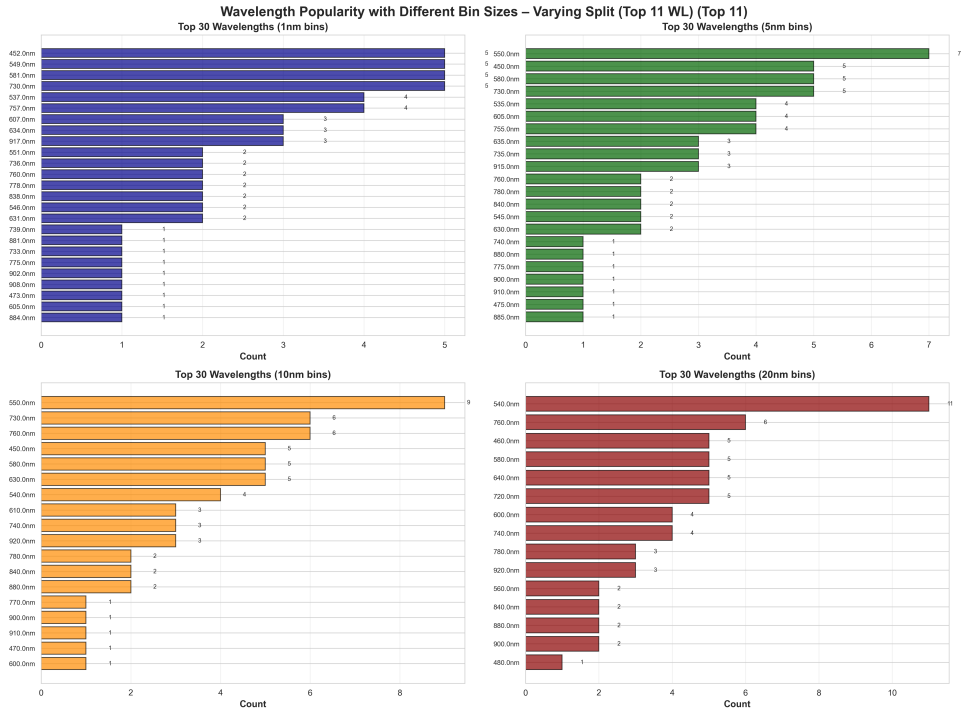


Figure 1.10: Wavelength popularity at different bin sizes (1, 5, 10, 20 nm) for the top-11 subsets.

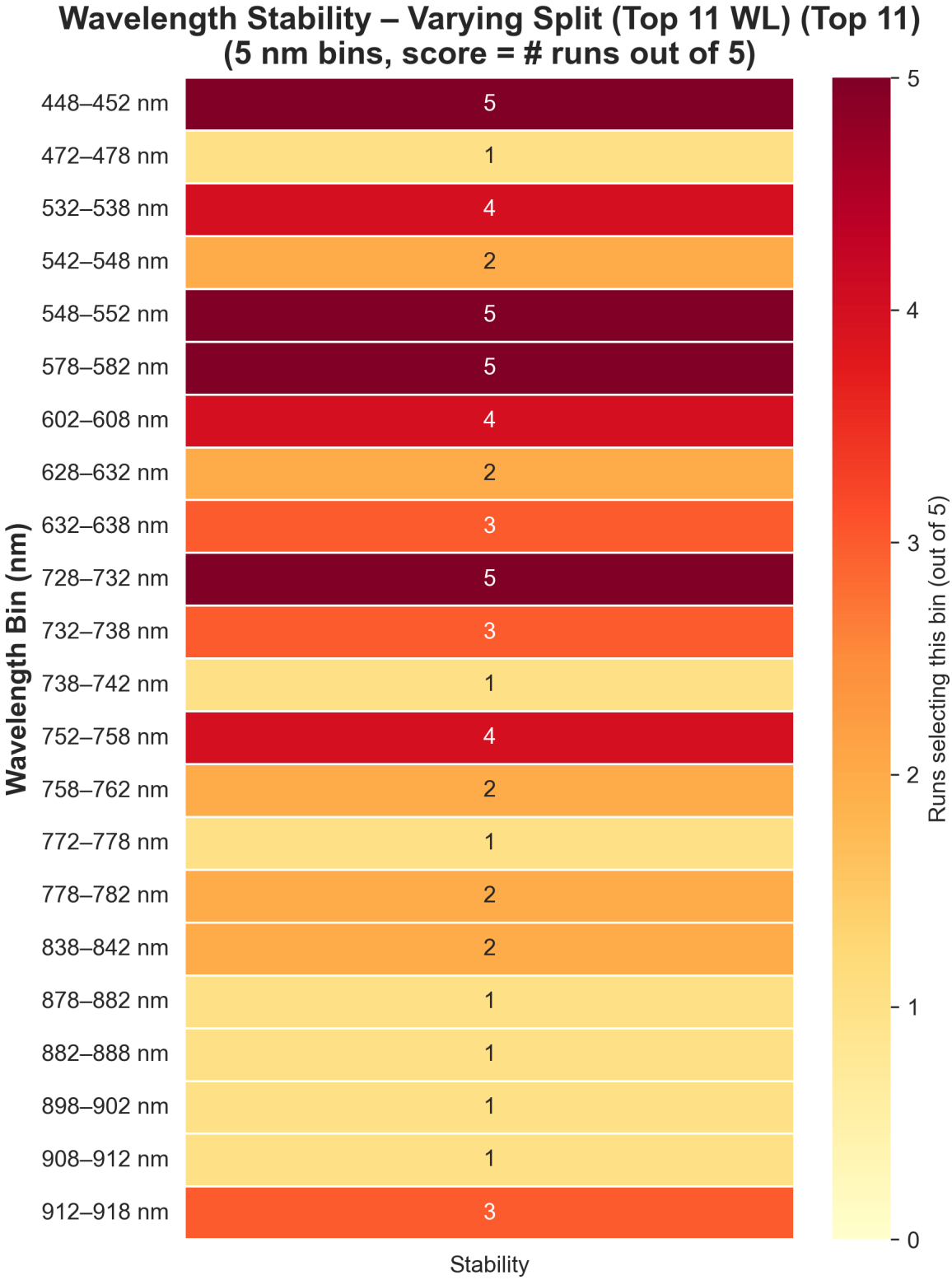


Figure 1.11: Stability heatmap for the top-11 BFS subsets across five seeds (objective: PR–AUC$_{\text{CRACK}}$; 0.5% drop setting). Each cell indicates how many seeds selected a wavelength within that 5 nm bin.

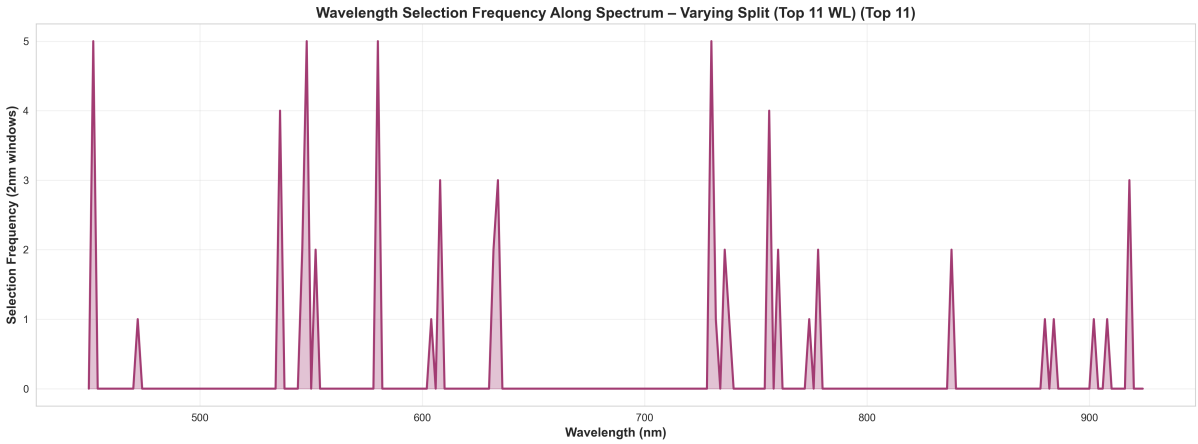


Figure 1.12: Wavelength selection frequency along the spectrum (2 nm sliding window) for the top-11 subsets.

Appendix A

Supplementary Results

A.1 Cross-Validation Statistics and Full Precision Tables

This appendix contains full cross-validation results with complete precision and standard deviations for all experiments discussed in Chapter 1. The main text reports truncated mean values for clarity; unreduced values are provided below for reproducibility and detailed analysis.

A.1.1 Pixel-Level Classification: Binary and Multi-Class

Table A.1: Pixel-level classification metrics (unbalanced training set, binary). Values are reported as mean ± standard deviation across folds where available. Precision/recall and ROC–AUC refer to the *Cracked* class (positive class); F1 is the weighted F1-score.

Model	Accuracy	Precision (Cracked)	Recall (Cracked)	Weighted F1	ROC–AUC (Cracked)
PLS-DA	0.9927 ± 0.0168	0.9999	0.9682	0.9924	1.0000 ± 0.0000
Logistic Regression (L1)	0.9962 ± 0.0074	0.9890	0.9948	0.9963	0.9996 ± 0.0012
SVM (RBF)	0.9938 ± 0.0094	0.9943	0.9787	0.9938	0.9997 ± 0.0010
Random Forest	0.9968 ± 0.0055	0.9964	0.9904	0.9968	0.9990 ± 0.0035
XGBoost	0.9947 ± 0.0103	0.9953	0.9798	0.9946	1.0000 ± 0.0000

A.1.2 Autoencoder-Based Anomaly Detection

A.2 Regional VIS–NIR Reflectance Summaries

Cracked tissue shows lower reflectance than healthy tissue across the full VIS–NIR range, with the divergence most pronounced beyond 700 nm. Reflectance values were aggregated over two broad spectral regions (VIS and NIR); healthy tissue exceeds cracked tissue by 32% in the NIR (0.660 vs. 0.450) and 37% in the VIS (0.218 vs. 0.137).

Table A.2: Pixel-level classification metrics (**unbalanced** training set, multi-class). Values are reported as mean ± standard deviation across folds.

Label setup	Model	Accuracy	Balanced Accuracy	Precision (Cracked)	Recall (Cracked)	Macro-F1	ROC–AUC (Cracked)	PR–AUC (Cracked)
crack_regular_rest								
	PLS-DA	0.5087 ± 0.0044	0.6879 ± 0.0239	0.8825 ± 0.1174	0.7537 ± 0.2385	0.5819 ± 0.0199	0.9977 ± 0.0049	0.8630 ± 0.1242
	Logistic Regression (L1)	0.8537 ± 0.0021	0.8683 ± 0.0075	0.9654 ± 0.0185	0.9663 ± 0.0748	0.8010 ± 0.0049	0.9993 ± 0.0021	0.9839 ± 0.0116
	Random Forest	0.9175 ± 0.0056	0.9179 ± 0.0116	0.8942 ± 0.0644	0.9394 ± 0.0701	0.8904 ± 0.0106	0.9975 ± 0.0037	0.9622 ± 0.0495
	XGBoost	0.9212 ± 0.0040	0.9162 ± 0.0094	0.9804 ± 0.0118	0.9493 ± 0.0894	0.8832 ± 0.0074	0.9992 ± 0.0017	0.9902 ± 0.0160
	MLP (Small)	0.9175 ± 0.0059	0.9166 ± 0.0122	0.9530 ± 0.0272	0.9729 ± 0.0398	0.8759 ± 0.0176	0.9994 ± 0.0017	0.9911 ± 0.0122
crack_vs_rest								
	PLS-DA	0.5120 ± 0.0047	0.6940 ± 0.0189	0.8843 ± 0.1117	0.7623 ± 0.2303	0.5893 ± 0.0164	0.9979 ± 0.0045	0.8669 ± 0.1186
	Logistic Regression (L1)	0.8574 ± 0.0022	0.8704 ± 0.0076	0.9669 ± 0.0190	0.9651 ± 0.0759	0.8049 ± 0.0050	0.9993 ± 0.0022	0.9847 ± 0.0114
	Random Forest	0.9147 ± 0.0060	0.9159 ± 0.0126	0.8871 ± 0.0701	0.9368 ± 0.0751	0.8857 ± 0.0116	0.9975 ± 0.0041	0.9610 ± 0.0535
	XGBoost	0.9200 ± 0.0043	0.9154 ± 0.0098	0.9809 ± 0.0120	0.9508 ± 0.0888	0.8825 ± 0.0078	0.9992 ± 0.0018	0.9900 ± 0.0165
	MLP (Small)	0.9163 ± 0.0062	0.9158 ± 0.0129	0.9521 ± 0.0282	0.9736 ± 0.0390	0.8741 ± 0.0188	0.9994 ± 0.0017	0.9912 ± 0.0122
multi_class								
	PLS-DA	0.9670 ± 0.0211	0.9787 ± 0.0121	0.7204 ± 0.1749	0.9927 ± 0.0127	0.9013 ± 0.0709	0.9978 ± 0.0025	0.9677 ± 0.0372
	Logistic Regression (L1)	0.9972 ± 0.0031	0.9959 ± 0.0097	0.9675 ± 0.0225	0.9977 ± 0.0078	0.9920 ± 0.0085	0.9996 ± 0.0016	0.9971 ± 0.0079
	Random Forest	0.9965 ± 0.0030	0.9940 ± 0.0120	0.9675 ± 0.0234	0.9969 ± 0.0090	0.9906 ± 0.0100	0.9994 ± 0.0026	0.9970 ± 0.0077
	XGBoost	0.9978 ± 0.0021	0.9946 ± 0.0103	0.9812 ± 0.0149	0.9908 ± 0.0206	0.9923 ± 0.0074	0.9995 ± 0.0021	0.9977 ± 0.0069
	MLP (Small)	0.9905 ± 0.0249	0.9890 ± 0.0224	0.9244 ± 0.1648	0.9866 ± 0.0259	0.9699 ± 0.0704	0.9965 ± 0.0083	0.9898 ± 0.0247

Table A.3: Pixel-level autoencoder anomaly detection: binary setting (non-cracked as normal class). Values are reported as mean ± standard deviation across LOGO folds.

Architecture	Accuracy	Precision (Cracked)	Recall (Cracked)	F1 (Cracked)	ROC–AUC (Cracked)	PR–AUC (Cracked)	Train Time [s]	Infer Time [s]
Autoencoder (64–32–16)	0.8969 ± 0.1646	0.8226 ± 0.1925	0.9971 ± 0.0016	0.9017 ± 0.1047	0.9814 ± 0.0327	0.9554 ± 0.0900	63.34 ± 2.80	0.043 ± 0.004
Autoencoder (128–64–32)	0.8597 ± 0.1813	0.7544 ± 0.2308	0.9974 ± 0.0015	0.8628 ± 0.1270	0.9673 ± 0.0469	0.9253 ± 0.1190	60.90 ± 1.48	0.057 ± 0.001
Autoencoder (64–32–8)	0.8923 ± 0.1648	0.8126 ± 0.1938	0.9978 ± 0.0013	0.8965 ± 0.1050	0.9799 ± 0.0334	0.9511 ± 0.0913	61.02 ± 1.46	0.041 ± 0.001

Table A.4: Pixel-level autoencoder anomaly detection: trained on CRACK pixels (multi-class setting). Values are mean across LOGO folds.

Architecture	Accuracy	Precision (non-CRACK)	Recall (non-CRACK)	F1 (non-CRACK)	ROC–AUC (non-CRACK)	PR–AUC (non-CRACK)	Train Time [s]	Infer Time [s]
Autoencoder (64–32–16)	0.9660	0.9995	0.9662	0.9826	0.9858	0.9999	26.96	0.148
Autoencoder (128–64–32)	0.9861	0.9991	0.9869	0.9930	0.9846	0.9999	27.12	0.207
Autoencoder (64–32–8)	0.9629	0.9994	0.9632	0.9809	0.9834	0.9999	26.20	0.155

Table A.5: Pixel-level autoencoder anomaly detection: trained on all non-CRACK classes (multi-class setting). Single group-wise train–test split.

Architecture	Accuracy	Precision (Cracked)	Recall (Cracked)	F1 (Cracked)	ROC–AUC (Cracked)	PR–AUC (Cracked)	Train Time [s]	Infer Time [s]
Autoencoder (64–32–16)	0.9969	0.0096	0.1110	0.0177	0.4575	0.0083	1429.22	0.388
Autoencoder (128–64–32)	0.9986	0.0118	0.0126	0.0122	0.4513	0.0084	999.73	0.255
Autoencoder (64–32–8)	0.9984	0.0207	0.0316	0.0250	0.4542	0.0110	916.85	0.141

Table A.6: Summary statistics of reflectance values across spectral regions for healthy and cracked tissue.

Spectral Region	Class	Mean Reflectance	Std. Dev.	Median
VIS (400–700 nm)	Healthy	0.218	0.128	0.186
	Cracked	0.137	0.076	0.123
NIR (700–1000 nm)	Healthy	0.660	0.114	0.630
	Cracked	0.450	0.121	0.425

Table A.7: Regional spectral separability metrics between healthy and cracked tissue. Mean values are averaged across wavelengths within each range, while peak values indicate the maximum separability and the corresponding wavelength.

Metric	VIS Range	NIR Range	Full Spectrum
Mean $\lvert\Delta\mu\rvert$ (Abs. Mean Diff.)	0.0810	0.2101	0.1422
Max $\lvert\Delta\mu\rvert$ (nm)	0.2325 (708.6)	0.3704 (726.5)	0.3704 (726.5)
Mean Cohen's d	0.6330	1.6217	1.1014
Max Cohen's d (nm)	1.1910 (699.5)	2.6877 (756.6)	2.6877 (756.6)
Mean Fisher Score	0.2895	1.6590	0.9436
Max Fisher Score (nm)	0.8821 (694.5)	3.6311 (747.5)	3.6311 (747.5)

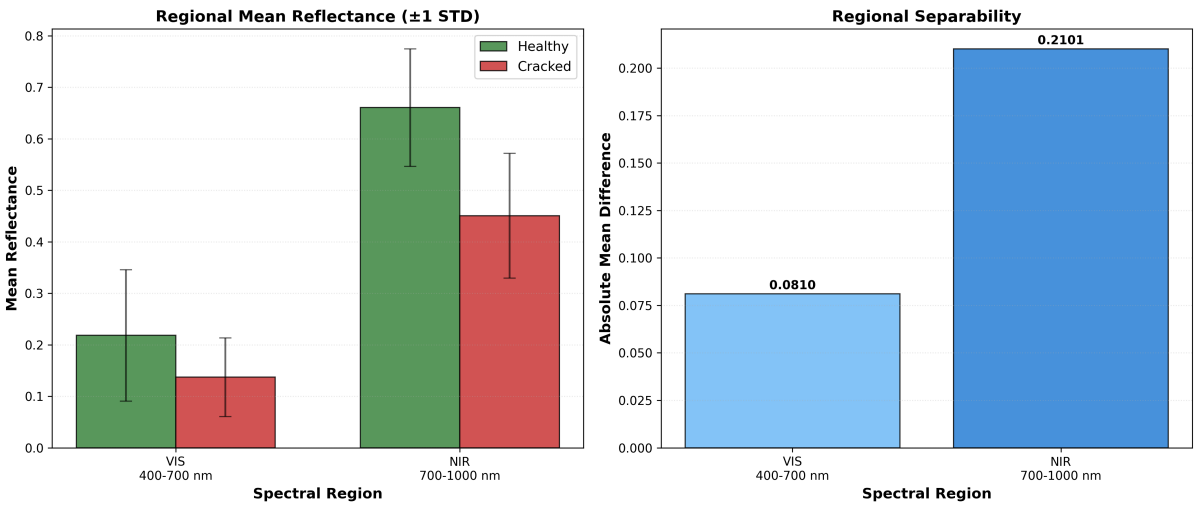


Figure A.1: Regional mean reflectance (±1 SD) and absolute mean class difference for the VIS and NIR ranges.

This appendix reports pixel-level classification and anomaly detection results obtained under class-balanced training, corresponding to the unbalanced experiments presented in the main Results chapter. Model rankings are consistent with those observed under unbalanced sampling.

A.3 Binary Pixel-Level Classification (Balanced)

Table A.8 reports binary pixel-level classification performance under balanced training.

Table A.8: Pixel-level classification metrics (balanced training set). Values are reported as mean ± standard deviation across folds where available. Precision/recall and ROC–AUC refer to the *Cracked* class (positive class); F1 is the weighted F1-score.

Model	Accuracy	Precision (Cracked)	Recall (Cracked)	Weighted F1	ROC–AUC (Cracked)
PLS-DA	0.9961 ± 0.0086	0.9998	0.9929	0.9961	1.0000 ± 0.0000
Logistic Regression (L1)	0.9930 ± 0.0148	0.9994	0.9861	0.9930	0.9999 ± 0.0002
SVM (RBF)	0.9876 ± 0.0202	0.9948	0.9771	0.9875	0.9998 ± 0.0007
Random Forest	0.9935 ± 0.0110	0.9970	0.9882	0.9934	0.9990 ± 0.0038
XGBoost	0.9934 ± 0.0107	0.9972	0.9882	0.9933	1.0000 ± 0.0001

A.4 Multi-Class Pixel-Level Classification (Balanced)

Table A.9 reports multi-class pixel-level classification performance under balanced training across all three labeling strategies.

Table A.9: Pixel-level classification metrics (**balanced** training set). Values are reported as mean ± standard deviation across folds.

Label setup	Model	Accuracy	Balanced Accuracy	Precision (Cracked)	Recall (Cracked)	Macro-F1	ROC–AUC (Cracked)	PR–AUC (Cracked)
crack_regular_rest								
	PLS-DA	0.4771 ± 0.0088	0.7061 ± 0.0107	0.8433 ± 0.0910	0.9451 ± 0.1069	0.5875 ± 0.0090	0.9982 ± 0.0046	0.9500 ± 0.0899
	Logistic Regression (L1)	0.8629 ± 0.0033	0.8653 ± 0.0073	0.9865 ± 0.0067	0.9671 ± 0.0739	0.8097 ± 0.0054	0.9994 ± 0.0018	0.9874 ± 0.0090
	Random Forest	0.9277 ± 0.0046	0.9199 ± 0.0110	0.9101 ± 0.0533	0.9194 ± 0.1124	0.9047 ± 0.0087	0.9975 ± 0.0035	0.9725 ± 0.0438
	XGBoost	0.9242 ± 0.0026	0.9111 ± 0.0084	0.9932 ± 0.0042	0.9437 ± 0.0954	0.8839 ± 0.0067	0.9993 ± 0.0016	0.9919 ± 0.0175
	MLP (Small)	0.9219 ± 0.0053	0.9140 ± 0.0125	0.9831 ± 0.0123	0.9832 ± 0.0223	0.8869 ± 0.0167	0.9982 ± 0.0057	0.9934 ± 0.0097
crack_vs_rest								
	PLS-DA	0.4782 ± 0.0090	0.7078 ± 0.0106	0.8445 ± 0.0904	0.9455 ± 0.1063	0.5889 ± 0.0093	0.9982 ± 0.0046	0.9507 ± 0.0886
	Logistic Regression (L1)	0.8642 ± 0.0032	0.8662 ± 0.0074	0.9866 ± 0.0067	0.9674 ± 0.0738	0.8109 ± 0.0055	0.9994 ± 0.0018	0.9876 ± 0.0089
	Random Forest	0.9261 ± 0.0049	0.9191 ± 0.0113	0.9092 ± 0.0546	0.9192 ± 0.1128	0.9038 ± 0.0091	0.9974 ± 0.0036	0.9722 ± 0.0444
	XGBoost	0.9238 ± 0.0027	0.9110 ± 0.0085	0.9931 ± 0.0042	0.9439 ± 0.0954	0.8836 ± 0.0068	0.9993 ± 0.0016	0.9919 ± 0.0176
	MLP (Small)	0.9221 ± 0.0052	0.9143 ± 0.0126	0.9830 ± 0.0124	0.9833 ± 0.0222	0.8871 ± 0.0165	0.9982 ± 0.0057	0.9934 ± 0.0097
multi_class								
	PLS-DA	0.9666 ± 0.0210	0.9786 ± 0.0120	0.7179 ± 0.1759	0.9926 ± 0.0129	0.9010 ± 0.0707	0.9978 ± 0.0025	0.9674 ± 0.0375
	Logistic Regression (L1)	0.9973 ± 0.0031	0.9960 ± 0.0097	0.9677 ± 0.0224	0.9977 ± 0.0078	0.9921 ± 0.0085	0.9996 ± 0.0016	0.9971 ± 0.0079
	Random Forest	0.9966 ± 0.0030	0.9941 ± 0.0120	0.9676 ± 0.0233	0.9969 ± 0.0090	0.9907 ± 0.0100	0.9994 ± 0.0026	0.9970 ± 0.0077
	XGBoost	0.9979 ± 0.0021	0.9947 ± 0.0103	0.9813 ± 0.0149	0.9908 ± 0.0206	0.9924 ± 0.0074	0.9995 ± 0.0021	0.9977 ± 0.0069
	MLP (Small)	0.9906 ± 0.0249	0.9891 ± 0.0224	0.9246 ± 0.1647	0.9866 ± 0.0259	0.9700 ± 0.0704	0.9965 ± 0.0083	0.9898 ± 0.0247

A.5 Autoencoder Anomaly Detection (Balanced)

A.5.1 One-Class Binary Formulation

Table A.10 reports one-class autoencoder performance trained on non-cracked pixels under balanced sampling.

Table A.10: Pixel-level anomaly detection performance using autoencoders trained on non-cracked samples (balanced sampling). Values are reported as mean ± standard deviation across folds. Accuracy is reported overall, while Precision/Recall/F1/ROC–AUC/PR–AUC correspond to the *Cracked* class (positive class).

Architecture	Accuracy	Precision (Cracked)	Recall (Cracked)	F1 (Cracked)	ROC–AUC (Cracked)	PR–AUC (Cracked)	Train Time [s]	Infer Time [s]
Autoencoder (64–32–16)	0.9343 ± 0.0989	0.9122 ± 0.1348	0.9906 ± 0.0340	0.9437 ± 0.0820	0.9823 ± 0.0457	0.9702 ± 0.0592	65.35 ± 1.90	0.008 ± 0.002
Autoencoder (128–64–32)	0.9083 ± 0.1332	0.8771 ± 0.1576	0.9982 ± 0.0058	0.9256 ± 0.1016	0.9919 ± 0.0178	0.9833 ± 0.0357	58.49 ± 3.87	0.007 ± 0.001
Autoencoder (64–32–8)	0.9353 ± 0.1027	0.9074 ± 0.1338	0.9977 ± 0.0100	0.9449 ± 0.0833	0.9858 ± 0.0344	0.9757 ± 0.0501	52.04 ± 2.44	0.006 ± 0.002

A.5.2 Crack-Only Training Formulation

Table A.11 reports multi-class autoencoder performance trained exclusively on cracked pixels under balanced sampling.

Table A.11: Pixel-level anomaly detection performance using autoencoders trained exclusively on *Cracked* pixels and evaluated in a multi-class setting (balanced sampling). Values are reported as mean ± standard deviation across LOGO folds. Accuracy is reported overall, while Precision/Recall/F1/ROC–AUC/PR–AUC correspond to the *Cracked* class (positive class).

Architecture	Accuracy	Precision (Cracked)	Recall (Cracked)	F1 (Cracked)	ROC–AUC (Cracked)	PR–AUC (Cracked)	Train Time [s]	Infer Time [s]
Autoencoder (64–32–16)	0.9660 ± 0.0106	0.9995 ± 0.0008	0.9662 ± 0.0108	0.9826 ± 0.0055	0.9858 ± 0.0281	0.9999 ± 0.0001	61.06 ± 7.56	0.045 ± 0.013
Autoencoder (128–64–32)	0.9861 ± 0.0052	0.9991 ± 0.0017	0.9869 ± 0.0051	0.9930 ± 0.0027	0.9846 ± 0.0274	0.9999 ± 0.0001	63.03 ± 8.09	0.056 ± 0.004
Autoencoder (64–32–8)	0.9629 ± 0.0149	0.9994 ± 0.0011	0.9632 ± 0.0150	0.9809 ± 0.0079	0.9834 ± 0.0298	0.9999 ± 0.0002	59.80 ± 8.55	0.041 ± 0.003

A.5.3 Non-Crack Training Formulation

Table A.12 reports multi-class autoencoder performance trained on all classes except cracked under balanced sampling.

Table A.12: Pixel-level anomaly detection performance using autoencoders trained on all semantic classes except *Cracked* and evaluated in a multi-class setting (balanced sampling). Values are reported for the held-out test split. Accuracy is reported overall, while Precision/Recall/F1/ROC–AUC/PR–AUC correspond to the *Cracked* class (positive class).

Architecture	Accuracy	Precision (Cracked)	Recall (Cracked)	F1 (Cracked)	ROC–AUC (Cracked)	PR–AUC (Cracked)	Train Time [s]	Infer Time [s]
Autoencoder (64–32–16)	0.9791	0.0000	0.0000	0.0000	0.2472	0.0083	2567.17	0.381
Autoencoder (128–64–32)	0.9789	0.0000	0.0000	0.0000	0.2847	0.0084	2260.73	0.299
Autoencoder (64–32–8)	0.9781	0.0000	0.0000	0.0000	0.4513	0.0110	1853.00	0.227

A.6 Wavelength Selection: Generalization and Performance Analysis

This section provides supplementary metrics from the backward feature selection (BFS) analysis on the whole-image pipeline, including calibration-to-test generalization gaps and performance metrics across additional wavelength subsets.

A.6.1 BFS Thresholds (Full Precision)

Table A.13: Summary of BFS wavelength-count thresholds for CRACK metrics (full precision).

Metric	Max score	$n_{\max}$	0.5% drop: n	score	1.0% drop: n	score
CRACK PR–AUC	0.9948	30	11	0.9878	9	0.9790
CRACK F1	0.9854	159	105	0.9800	88	0.9751

A.6.2 Wavelength Results (Full Precision)

Table A.14: Full-image crack-detection performance for different wavelength subsets under early- and late-stage conditions (full precision).

Stage	# Wavelengths	Accuracy	Balanced Accuracy	Precision (Cracked)	Recall (Cracked)	F1 (Cracked)	MCC
Early	9	0.7833	0.7902	0.8800	0.6875	0.7719	0.5873
Late	9	0.8000	0.8054	0.8667	0.7647	0.8125	0.6054
Early	11	0.7500	0.7567	0.8400	0.6562	0.7368	0.5195
Late	11	0.8167	0.8111	0.8286	0.8529	0.8406	0.6254
Early	30	**0.8167**	**0.8214**	**0.8889**	**0.7500**	**0.8136**	**0.6447**
Late	30	**0.8333**	**0.8348**	**0.8750**	0.8235	0.8485	**0.6652**
Early	159	0.8000	0.8036	0.8571	**0.7500**	0.8000	0.6071
Late	159	**0.8333**	0.8258	0.8333	**0.8824**	**0.8571**	0.6591

A.6.3 Generalization Analysis

Calibration-to-test generalization gap. Table A.15 shows a consistent calibration-to-test gap across all wavelength subsets and both acquisition stages. Early-stage gaps range from 0.056 to 0.151; late-stage gaps are larger (0.212–0.291).

A.6.4 Wavelength Overlap Across Seeds

Tables A.16 and A.17 report the pairwise Jaccard similarity between the wavelength subsets selected by each seed for the top-30 and top-11 configurations, respectively.

Table A.15: Model Generalization Analysis: Comparison between Calibration and Test MCC for wavelength subsets.

Stage	# Features	MCC Calibration	MCC Test	**Gap (Δ)**
Early	9	0.7381	0.5873	0.1508
	30	0.7232	**0.6447**	**0.0785**
	159	0.6631	0.6071	0.0560
Late	9	0.8968	0.6054	0.2914
	30	0.8771	**0.6652**	**0.2119**
	159	0.8931	0.6591	0.2340

Table A.16: Pairwise Jaccard similarity of top-30 wavelengths across seeds (varying split). Mean Jaccard = 0.493.

	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5
Seed 1	—	0.409	0.590	0.722	0.442
Seed 2	0.409	—	0.378	0.409	0.476
Seed 3	0.590	0.378	—	0.590	0.442
Seed 4	0.722	0.409	0.590	—	0.476
Seed 5	0.442	0.476	0.442	0.476	—

Table A.17: Pairwise Jaccard similarity of top-11 wavelengths across seeds (varying split). Mean Jaccard = 0.398.

	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5
Seed 1	—	0.500	0.412	0.333	0.412
Seed 2	0.500	—	0.412	0.412	0.333
Seed 3	0.412	0.412	—	0.500	0.333
Seed 4	0.333	0.412	0.500	—	0.333
Seed 5	0.412	0.333	0.333	0.333	—

The top-30 subsets share 13 wavelengths across all five seeds (mean Jaccard = 0.493); the top-11 subsets share 4 wavelengths (mean Jaccard = 0.398). The cross-seed intersection at top-11 comprises 452.25, 548.55, 580.90, and 729.53 nm—all four also appear in the top-30 intersection, confirming consistent selection under both subset sizes.

Bibliography

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd). Lawrence Erlbaum Associates.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification* (2nd). Wiley-Interscience.