

UNIVERSITÀ DEGLI STUDI DI CASSINO E DEL LAZIO MERIDIONALE
CORSO DI DOTTORATO IN METODI, MODELLI e TECNOLOGIE PER L'INGEGNERIA
DIPARTIMENTO DI INGEGNERIA ELETTRICA E DELL'INFORMAZIONE



From Side-Channel Profiling to Adaptive Information Control

A Learning-Driven Framework for Intelligent Side-Channel Defense

Vincenzo Rega

vincenzo.rega@unicas.it

05/06/2026

TUTORS

Prof. Mario Molinara

Prof. Fabrizio Marignetti

COORDINATOR

Prof. Fabrizio Marignetti

UNIVERSITÀ DEGLI STUDI DI CASSINO E DEL LAZIO MERIDIONALE
CORSO DI DOTTORATO IN METODI, MODELLI E TECNOLOGIE PER
L'INGEGNERIA

Date: **05/06/2026**

Author: **Vincenzo Rega**

Title: **From Side-Channel Profiling to Adaptive Information Control
Learning-Driven Framework for Intelligent Side-Channel Defense**

Department: **DIPARTIMENTO DI INGEGNERIA ELETTRICA E
DELL'INFORMAZIONE**

Degree: **PHILOSOPHIAE DOCTOR**

Permission is herewith granted to university to circulate and to have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions.

Signature of Author

THE AUTHOR RESERVES OTHER PUBLICATION RIGHTS, AND NEITHER THE THESIS NOR EXTENSIVE EXTRACTS FROM IT MAY BE PRINTED OR OTHERWISE REPRODUCED WITHOUT THE AUTHOR'S WRITTEN PERMISSION.

THE AUTHOR ATTESTS THAT PERMISSION HAS BEEN OBTAINED FOR THE USE OF ANY COPYRIGHTED MATERIAL APPEARING IN THIS THESIS (OTHER THAN BRIEF EXCERPTS REQUIRING ONLY PROPER ACKNOWLEDGEMENT IN SCHOLARLY WRITING) AND THAT ALL SUCH USE IS CLEARLY ACKNOWLEDGED.

Dedication

*To Giuseppe and Martina,
the light of my eyes,
my strength in difficult days,
and the deepest reason behind every step forward.*

*To my grandfather Vincenzo,
who continues to protect me
from the most beautiful star in the sky.*

Acknowledgements

I would like to express my deepest gratitude to my supervisors, Fabrizio and Mario, for their guidance and support throughout my Ph.D. journey. In particular, I wish to thank Mario, whose patience, leadership, and constant encouragement have been invaluable. He guided me not only as a supervisor, but as a true leader does with his team, offering wise advice and steady support at every stage. A significant part of this achievement is undoubtedly due to his dedication, and I will always be grateful for his trust and mentorship.

I would also like to sincerely thank Nico for entrusting me with an important part of his work and for giving me the opportunity to contribute meaningfully to the research activities of this doctorate. Without his confidence and support, I would not have been able to achieve the level of quality reached in this work.

My gratitude also goes to Andrea, and especially to Luca, for their valuable advice, technical support, and constructive discussions during the research activities. Their contributions have been fundamental in shaping both the technical and scientific aspects of this work.

Last, but certainly not least, I would like to thank my wife, Tiziana. Her patience, strength, and understanding made this journey possible. She endured the stress, the anxiety, and the countless days spent away, often carrying the weight of family responsibilities on her own. Thank you for waiting for me, for believing in me, and for being the steady foundation behind this achievement.

Contents

Acknowledgements	iv
List of figures	xi
List of tables	xv
Introduction	1
1 Theoretical Background on Side-Channel Analysis and Machine Learning	4
1.1 Side-Channel Attacks	4
1.1.1 Classification of Side-Channel Attacks	5
1.1.2 Electromagnetic Side-Channel Analysis	6
1.1.3 Multi-Channel and Cross-Layer Leakage	6
1.1.4 Threat Models and Practical Considerations	6
1.2 Leakage Models and Measurement Setups	6
1.2.1 Analytical Leakage Models	7
1.2.2 Black-Box and Data-Driven Leakage Models	7
1.2.3 Measurement Setups as Part of the Leakage Model	7
1.2.4 Noise, Variability, and Non-Stationarity	7
1.2.5 Implications for Feature Representation and Adaptive Processing	8
1.3 Side-Channel Countermeasures and Defense Strategies	8
1.3.1 Hardware-Level Countermeasures	9
1.3.2 Software and Algorithmic Countermeasures	9
1.3.3 Informational Countermeasures and Obfuscation	10
1.3.4 Learning-Based and Adaptive Defense Approaches	10
1.3.5 Motivation for Reinforcement Learning-Based Adaptive Defense	10
1.4 Machine Learning and Deep Learning in Side-Channel Analysis	11
1.4.1 Profiling Attacks and Key Recovery	11
1.4.2 Behavioral Profiling of Embedded and IoT Systems	11
1.4.3 Anomaly Detection and Physical-Layer Intrusion Detection	11
1.4.4 Noise-Aware Feature Extraction and Denoising	12
1.4.5 Feature Selection, Temporal Modeling, and Decision Strategies	12
1.4.6 From Static Learning to Adaptive Defense	13
1.5 Conclusion	13

2	State of the Art in AI-Based Side-Channel Profiling and Security	15
2.1	Scope and Positioning of the State of the Art	15
2.2	Classical Side-Channel Attacks and Profiling Foundations	16
2.2.1	Power and Electromagnetic Analysis	16
2.2.2	Leakage Models and Hypothesis Testing	16
2.2.3	Operational Constraints and Practical Limitations	16
2.2.4	From Analytical Profiling to Data-Driven Approaches	17
2.3	AI-Based Side-Channel Profiling: From Static Classification to Adaptive Inference	17
2.3.1	Supervised Learning for Side-Channel Profiling	17
2.3.2	Temporal Modeling and Sequential Inference	18
2.3.3	Unsupervised Learning and Anomaly Detection	18
2.3.4	Noise, Generalization, and the Limits of Static Models	18
2.3.5	Realistic Scenarios Motivating Adaptive Defense	18
2.3.6	Toward Adaptive and Control-Oriented Learning	19
2.4	Noise, Environmental Variability, and the Limits of Static Side-Channel Models	19
2.5	Adaptive and Learning-Based Countermeasures in the Literature	20
2.5.1	Adaptive Preprocessing and Noise Mitigation	20
2.5.2	Learning-Based Obfuscation and Defensive Machine Learning	21
2.5.3	Adaptive Monitoring and Physical-Layer Intrusion Detection	21
2.5.4	Limitations of Existing Adaptive Approaches	21
2.5.5	Positioning Reinforcement Learning as a Unifying Framework	21
2.6	Realistic Threat Scenarios Motivating Adaptive Side-Channel Defense	22
2.6.1	Public and Semi-Public Embedded Systems	22
2.6.2	Smart City and Urban Infrastructures	22
2.6.3	Financial and Transaction-Oriented Devices	22
2.6.4	Remote and Air-Gapped Scenarios	23
2.6.5	Implications for Adaptive Defense Design	23
2.7	Discussion and Open Challenges	23
2.7.1	Generalization and Robustness Across Conditions	23
2.7.2	Static Pipelines in a Dynamic Leakage Landscape	24
2.7.3	Noise as a First-Class Design Constraint	24
2.7.4	Temporal Reasoning Beyond Window-Level Classification	24
2.7.5	Defensive Learning and the Limits of Static Countermeasures	24
2.7.6	Toward Closed-Loop and Adaptive Side-Channel Defense	24
2.7.7	Positioning of This Thesis	24
3	Experimental Setup and Methodological Framework	26
3.1	Hardware Platform and Acquisition Environment	27
3.2	Signal Acquisition and Preprocessing Pipeline	28
3.2.1	Removal of Background Noise Segment	29
3.2.2	Temporal Segmentation and Overlapping Windows	29
3.2.3	Energy-Based Aggregation via RMS	32

3.2.4	Normalization and Robust Scaling	32
3.2.5	Summary of Preprocessing Steps	35
3.2.6	Feature Set Description	35
3.2.7	Feature Selection Strategies and Inference Constraints	37
3.2.8	Pipeline Overview	40
3.2.9	Noise Robustness and Stability of Feature Relevance	40
3.3	From Static Pipelines to Adaptive Control: Summary and Transition	43
3.3.1	Achievements of Static Learning-Based Pipelines	43
3.3.2	Structural Limits of Static Robustness	43
3.3.3	Temporal Aggregation Is Not Adaptation	44
3.3.4	Noise and Variability as First-Class System Properties	44
3.3.5	Reframing Side-Channel Processing as a Control Problem	44
3.3.6	Toward Adaptive and Learning-Driven Side-Channel Defense	44
3.3.7	Transition to the Next Chapter	45
4	Results and Performance Evaluation	46
4.1	Evaluation Protocol	46
4.2	Overall Classification Performance	47
4.3	Scenario-Specific Analysis	47
4.4	Impact of Feature Selection	48
4.5	Robustness to Noise and Variability	48
4.5.1	Discussion	54
5	Side-Channel Leakage as an Information Control Problem	56
5.1	Toward a Feature-Aware Approach in Noisy Environments	56
5.2	Experimental Comparison Framework	57
5.3	Objectives of the Subsequent Analysis	57
5.4	Analysis of Selected Features under Clean Conditions	57
5.5	Analysis of Selected Features and Clean-Condition Performance	59
5.6	Classification Robustness with Respect to Dataset Dimensionality	60
5.7	Simulation of Noisy Operating Environments	62
5.8	Noise Model and Signal-to-Noise Ratio (SNR)	63
5.9	Qualitative Impact of Noise on Raw Side-Channel Signals	64
5.9.1	Current Signal: Robust vs. Fragile Scenarios	64
5.9.2	Electromagnetic Signal: Increased Sensitivity to Noise	65
5.10	Impact of Additive Noise on Classification Performance	66
5.11	Clean-to-Noisy Generalization and Domain Shift Effects	68
5.12	Mitigating Domain Shift via Noise-Aware Training	69
5.13	Feature Optimization under Noisy Conditions	70
5.14	Feature Optimization under Noisy Conditions	71
5.14.1	Feature Importance Drift across SNR Levels	71
5.14.2	Geometric Drift in Feature Space	72

5.15	Feature-Aware Analysis of Noise Impact	73
5.16	Feature-Weighted Classification under Noisy Conditions	75
6	Reinforcement Learning for Adaptive Feature-Aware Side-Channel Analysis	79
6.1	Motivation and Problem Formulation	79
6.2	Reinforcement Learning Background (Contextualized)	80
6.2.1	Markov Decision Process Formulation	80
6.2.2	Policy, Value Functions, and Optimality	80
6.2.3	Reinforcement Learning as Adaptive Control	81
6.2.4	State Representation	81
6.2.5	Action Space	82
6.2.6	Reward Design	82
6.2.7	Role of Reinforcement Learning in This Thesis	82
6.3	Feature-Aware Adaptive Denoising: From Deterministic to Reinforcement Learning	82
6.3.1	Static Preprocessing and Its Limitations	82
6.3.2	Deterministic Adaptive Preprocessing	83
6.3.3	Limitations of Deterministic Adaptation	83
6.3.4	Reinforcement Learning as a Generalization of Adaptive Control	83
6.3.5	Validation of the Adaptivity Hypothesis	83
6.4	Reinforcement Learning Architecture and Agent Placement	84
6.4.1	Overall Pipeline Description	84
6.4.2	Rationale for Feature-Level Agent Placement	84
6.4.3	Feature-Aware Control and Interpretability	85
6.4.4	Decoupling Adaptation from Classification	85
6.4.5	Comparison with End-to-End Learning Approaches	85
6.4.6	Summary	86
6.5	State, Action, and Reward Design	86
6.5.1	State Representation	86
6.5.2	Action Space	86
6.5.3	Reward Function	87
6.5.4	Design Considerations	87
6.6	Experimental Evaluation	87
6.6.1	Baseline and Light Action Set	87
6.6.2	RL Feature-Aware Selection with Light Action Set	88
6.6.3	Extension of the Action Space: Feature Weighting	91
6.6.4	Strong Reinforcement Learning Actions for Feature-Aware Adaptive Control	92
6.6.5	Strong RL Actions: Experimental Evaluation and Classifier-Aware Behavior	94
6.7	Threat Model and Side-Channel Defense Perspective	97
6.7.1	Defense Strategies in Side-Channel Attacks	97
6.7.2	Threat Model and Assumptions	98
6.7.3	Reinforcement Learning as Informational Defense	99
6.8	Proactive Reinforcement Learning-Based Side-Channel Defense	100

6.8.1	Fundamental Assumptions	100
6.8.2	Defense by Execution Control Rather Than Signal Manipulation	100
6.8.3	Roles and Information Flow in the Proposed Model	101
6.8.4	Informational Defense Objective	102
6.8.5	Adaptive Informational Control Architecture	102
6.8.6	Dual-Channel Interpretation: Defense vs. Attacker	103
6.8.7	Proactive Defense via Informational Asymmetry	104
6.8.8	Analysis of RL Action Usage in the Defensive Channel	104
6.8.9	Analysis of RL Action Usage in the Attacker-Observable Channel	105
6.8.10	Confusion Matrix Analysis	107
7	Results and Performance Evaluation	110
7.1	Summary of the Thesis	110
7.2	Original Scientific Contributions	111
7.3	Limitations and Future Work	113
	Bibliography	115
	Appendices	118
A	Detailed Experimental Parameters	120
A.1	Acquisition Parameters	120
A.2	Windowing and Segmentation	120
B	Feature Analysis and Supplementary Results	121
B.1	Complete Feature Set and Channel Contribution	121
B.2	Feature Selection and Classification with Random Forest	122
B.2.1	Recursive Feature Elimination (RFE)	122
B.2.2	Cross-Validation and Performance Evaluation	123
B.3	Noise Injection and SNR Modeling	123
B.4	Denoising Techniques on RMS Features: Case Study at SNR = 20 dB	123
B.4.1	Baseline RMS Classification under Noise	124
B.4.2	Moving Average Denoising	124
B.4.3	Savitzky–Golay Filtering	125
B.4.4	Wavelet-Based Denoising	125
B.4.5	Discussion	125
B.5	Extended Pipeline with Denoising and Time–Frequency Feature Extraction	126
B.5.1	Discussion and Interpretation of Results	126
B.6	The Role of Majority Voting	127
B.7	Supplementary Confusion Matrices	128
B.8	Window-Level vs Session-Level Inference	128
B.9	Methodological Implications	129
C	Algorithmic Details	130

C.1 Column-wise Majority Voting	131
D Implementation and Reproducibility Notes	133
D.1 CNN+LSTM Architecture	134
D.2 Role of the CNN+LSTM Model	134

List of Figures

1.1	Conceptual overview of a generic Side-Channel Attack (SCA). Physical leakages generated during device operation, including electromagnetic emissions, power fluctuations, and timing-related activity, are captured through external measurement equipment and analyzed to infer sensitive information. The attack pipeline typically involves signal acquisition, feature extraction, and information inference, enabling the reconstruction of cryptographic secrets, application behavior, or internal system activity without direct access to the device internals.	5
1.2	Conceptual overview of informational countermeasures and obfuscation strategies in side-channel analysis. In conventional side-channel attacks, observable physical leakages such as electromagnetic emissions, power fluctuations, and timing-related activity maintain strong statistical correlations with sensitive internal states, enabling high-confidence attacker inference. Informational obfuscation aims to disrupt these relationships by introducing adaptive perturbations, randomization, or activity shaping mechanisms that reduce the exploitable information content of observable signals. Rather than suppressing physical emissions entirely, the proposed defensive perspective weakens the attacker's inference capability while preserving normal system functionality, thereby increasing uncertainty, reducing model confidence, and degrading the effectiveness of learning-based attacks. . . .	9
1.3	Conceptual overview of a physical-layer Intrusion Detection System (IDS) based on side-channel analysis. Physical leakages such as electromagnetic emissions, power consumption, and timing-related activity are continuously acquired from the target device and used to learn models of normal system behavior through unsupervised or semi-supervised learning techniques. During runtime, newly observed traces are compared against the learned representation to compute anomaly scores and detect deviations associated with faults, cyberattacks, or unauthorized activity. By operating directly on hardware-level emissions, side-channel-based IDS provide non-intrusive and difficult-to-evade monitoring capabilities, complementing traditional software-based security mechanisms.	12
1.4	Conceptual comparison between static side-channel defense and reinforcement learning-based adaptive defense. In the static approach, physical leakages are processed through fixed preprocessing and feature extraction stages, making the resulting inference pipeline sensitive to noise, environmental variability, and adaptive attackers. In contrast, the proposed adaptive defense framework introduces a reinforcement learning agent that dynamically selects preprocessing, feature transformation, and representation-control actions according to the observed leakage conditions. By closing the loop between side-channel observations, policy decisions, and environmental feedback, the adaptive framework aims to preserve legitimate monitoring capabilities while degrading the exploitable information available to an attacker, thereby increasing uncertainty and improving robustness against non-stationary side-channel conditions.	14
3.1	Experimental setup for electromagnetic side-channel measurement. (1) Device Under Test (Raspberry Pi 4 Model B); (2) Magnetic Field Probe (ETS Lindgren 7406-901); (3) Oscilloscope (TiePie HS6); (4) Acquisition and control PC.	27
3.2	Raw current and magnetic field signals acquired at 1 MHz sampling frequency for Scenarios 1 to 4. Each acquisition lasts 40 seconds. The first 10 seconds correspond to background noise and are excluded from subsequent analysis.	28

3.3	Raw current and magnetic field signals for Scenarios 5 to 8. The observable increase in magnetic activity reflects emissions induced by the execution of application-level workloads. As in the previous scenarios, the first 10 seconds represent background noise and are removed prior to preprocessing.	29
3.4	Raw current and magnetic field signals for Scenarios 1 to 4. The first 10 seconds correspond to background electromagnetic noise acquired during idle system operation, while the subsequent interval reflects application-induced activity.	30
3.5	Raw current and magnetic field signals for Scenarios 5 to 8. As in the previous scenarios, the initial 10 seconds capture background noise, followed by a marked increase in electromagnetic activity associated with application execution.	30
3.6	RMS feature extraction using 100 ms windows with 50% overlap. RMS values are computed independently for current and magnetic field signals.	31
3.7	RMS time series for current and electromagnetic signals computed over 100 ms windows with 50% overlap. Compared to raw signals, RMS aggregation produces smoother and more interpretable temporal patterns, emphasizing workload-induced variations while reducing high-frequency noise.	33
3.8	Comparison of classification accuracy for different normalization techniques, with and without RMS preprocessing. Robust Scaler achieves the best performance when combined with RMS aggregation.	33
3.9	Overall feature importance obtained using a Random Forest classifier trained on the full feature set. Feature relevance is unevenly distributed across descriptors and channels, highlighting heterogeneous contributions to scenario discrimination.	38
3.10	Comparison of Random Forest classification accuracy under different feature selection strategies. Scenario-specific feature importance achieves the highest accuracy but is not applicable at inference time.	38
3.11	Comparison of feature importance between the full model and the reduced model obtained via Recursive Feature Elimination (RFE). The selected subset preserves the most discriminative descriptors while reducing dimensionality.	39
3.12	Confusion matrix obtained using Random Forest with the RFE-selected feature subset. High diagonal dominance confirms strong discriminative capability under inference-time constraints	40
3.13	Overview of the proposed side-channel processing pipeline. Raw current and electromagnetic signals are segmented, aggregated via RMS, normalized using Robust Scaler, processed through feature extraction and RFE-based selection, and finally classified using supervised learning models.	41
4.1	Effect of Gaussian noise injection on the normalized RMS feature of the current signal. Each point represents aggregation over a 5-second interval. Increasing noise levels induce progressively larger fluctuations and reduced temporal stability.	50
4.2	Effect of Gaussian noise injection on the normalized RMS feature of the magnetic field signal. Despite increasing perturbation levels, the signal exhibits higher stability compared to the current channel.	51
4.3	Confusion matrix at 10% Gaussian noise. Average accuracy: 88.33%.	52
4.4	Confusion matrix at 20% Gaussian noise. Average accuracy: 79.58%.	53
4.5	Confusion matrix at 30% Gaussian noise. Average accuracy: 80.42%.	54
4.6	Confusion matrix at 50% Gaussian noise. Average accuracy: 77.08%.	55
5.1	Random Forest feature importance ranking (top 20) obtained under clean conditions. . . .	58
5.2	Random Forest feature importance ranking under clean conditions. The plot reports the top 20 most influential features extracted from both current (ch1) and electromagnetic (ch2) channels, including time-domain and frequency-domain descriptors.	60
5.3	Window-level confusion matrix (%) obtained under clean conditions using the selected feature set and a Random Forest classifier. Rows represent true scenarios, while columns indicate predicted scenarios.	61
5.4	PCA projection of RMS-based features extracted under clean conditions. Each point corresponds to a window-level feature vector, with colors indicating different application scenarios. The first two principal components explain over 90% of the total variance. . . .	61

5.5	Current signal under clean and noisy conditions (SNR = 20 dB, 10 dB, 5 dB) for Scenarios 1–4. Noise primarily affects short-term fluctuations, while global signal structure remains preserved.	64
5.6	Current signal under clean and noisy conditions (SNR = 20 dB, 10 dB, 5 dB) for Scenarios 5–8. Noise significantly alters both local fluctuations and global signal structure. . . .	65
5.7	Electromagnetic signal under clean and noisy conditions (SNR = 20 dB, 10 dB, 5 dB) for Scenarios 1–4. Noise increases dispersion but preserves coarse statistical structure.	65
5.8	Electromagnetic signal under clean and noisy conditions (SNR = 20 dB, 10 dB, 5 dB) for Scenarios 5–8. Noise rapidly masks both local and global signal characteristics.	66
5.9	Window-level confusion matrix obtained under noisy–noisy conditions with SNR = 20 dB	67
5.10	Window-level confusion matrix obtained under noisy–noisy conditions with SNR = 10 dB	67
5.11	Signal-level confusion matrix in the clean $\rightarrow$ noisy setting with SNR = 20 dB	68
5.12	Signal-level confusion matrix for the mixed clean + noisy training setup (SNR = 20 dB), tested on noisy signals.	70
5.13	Feature importance drift across different SNR levels.	72
5.14	PCA geometry drift across different SNR levels.	73
5.15	Feature-aware pipeline with adaptive feature weighting.	75
5.16	Confusion matrix (%) obtained using feature-weighted representation under noisy conditions.	76
6.1	General reinforcement learning framework. At each time step, the agent observes the state s_t of the environment, selects an action a_t , and receives a reward r_t . This feedback loop allows the agent to update its policy over time with the objective of maximizing the expected cumulative reward.	81
6.2	Feature-aware reinforcement learning pipeline for adaptive side-channel inference. Raw electromagnetic or current traces are first subjected to controlled noise injection and pre-processing through windowing and RMS computation. Features are then extracted and passed to a Reinforcement Learning agent, which performs adaptive representation control before the final classification stage.	84
6.3	Window-level confusion matrix obtained under the static supervised baseline at SNR = 20 dB. The Random Forest classifier is trained using a mixed dataset of clean and noisy signals and evaluated exclusively on noisy data.	88
6.4	Window-level confusion matrix obtained using the RL feature-aware selector with light action set (no-op, moving average, Savitzky–Golay), evaluated on noisy data at SNR = 20 dB with 30 files per scenario	89
6.5	Window-level confusion matrix obtained using the RL feature-aware selector with light action set, evaluated on noisy data at SNR = 20 dB with 20 files per scenario	90
6.6	Window-level confusion matrix obtained using the RL feature-aware selector with extended action space including feature weighting, evaluated at SNR = 20 dB with 20 files per scenario	91
6.7	Window-level confusion matrix obtained using the RL feature-aware selector combined with a Random Forest classifier at SNR = 20 dB	95
6.8	Window-level confusion matrix obtained using the RL feature-aware selector combined with a CNN+LSTM classifier at SNR = 20 dB	96
6.9	Security-oriented reinterpretation of the proposed Reinforcement Learning framework for adaptive side-channel protection. The upper section illustrates the feature-aware adaptive control pipeline, where an RL agent dynamically adjusts feature representations under noisy conditions to maintain robust inference performance. The lower section reframes the same mechanism within a defensive threat model, highlighting the transition from traditional signal-level countermeasures toward representation-level protection. Instead of suppressing physical leakage directly, the proposed approach aims to control and distort the information exposed to learning-based adversaries, thereby reducing the exploitability of side-channel observations while preserving operational functionality.	97
6.10	Conceptual model of adaptive informational defense based on reinforcement learning. The physical side-channel remains observable by any passive observer, while an RL agent adaptively controls the execution behavior of the system to destabilize the mapping between observable leakage and sensitive information.	101

6.11	Reinforcement Learning-based proactive side-channel defense architecture. The RL controller operates on the feature representation, receiving feedback from classifier decisions and uncertainty metrics, and selecting adaptive transformations without modifying the classifier or the physical signal	103
6.12	Distribution of actions selected by the Reinforcement Learning agent in the defensive channel. The prevalence of the No-op action highlights the conservative and selective nature of the learned control policy.	105
6.13	Distribution of actions selected by the Reinforcement Learning controller in the attacker-observable channel. The prevalence of aggressive and structure-disrupting actions highlights the intentional degradation of informational content from the attacker's perspective	106
6.14	Window-level confusion matrix for the defensive channel. The matrix exhibits a strongly diagonal structure, indicating near-perfect class separability under adaptive RL control	107
6.15	Window-level confusion matrix for the attacker-observable channel. The absence of a diagonal structure indicates severe degradation of discriminative information	108
7.1	Conceptual illustration of <i>Adaptive Inferential Security (SIA)</i> . A reinforcement learning controller dynamically regulates the feature representations exposed through observable side-channel emissions, creating an inferential asymmetry between legitimate and adversarial observers. While authorized entities receive stable and informative representations suitable for reliable inference, adversarial observers are exposed to degraded, unstable, and low-confidence feature spaces. Rather than blocking signal access, the framework continuously adapts the informational structure of observable signals to control what can be inferred from system behavior under changing environmental and threat conditions.	112
B.1	Window-level confusion matrix obtained using RMS features extracted from raw signals corrupted with additive Gaussian noise at SNR = 20 dB (no denoising applied).	124
B.2	Extended processing pipeline for evaluating the combined impact of noise, denoising, and feature engineering. Denoising is applied to RMS features prior to time-frequency feature extraction.	126
B.3	Comparison between window-level and acquisition-level confusion matrices under noisy conditions (SNR = 20 dB). Temporal aggregation via majority voting restores perfect scenario-level classification despite window-level uncertainty.	127
B.4	Comparison between window-level and session-level inference. At the window level, each temporal segment is classified independently, producing local predictions that may fluctuate due to noise or partial loss of discriminative information. At the session level, the sequence of window-level predictions is aggregated through column-wise majority voting, producing a more stable and reliable acquisition-level decision.	129
C.1	Overview of the RMS-based preprocessing pipeline adopted for side-channel behavioral inference. Raw electromagnetic and current-consumption signals are processed through background removal, temporal segmentation, RMS computation, robust normalization, and feature extraction stages in order to generate compact and noise-resilient feature representations suitable for machine learning classification.	130
C.2	Column-wise majority voting strategy for acquisition-level behavioral inference. Temporal windows extracted from the side-channel acquisition are classified independently, producing a sequence of local predictions that may include unstable or noisy decisions. The majority voting mechanism aggregates these predictions and selects the most frequent class as the final acquisition-level scenario, improving robustness and reducing the impact of transient misclassifications.	132
D.1	Hybrid CNN+LSTM architecture adopted for side-channel behavioral inference. Sequences of feature vectors extracted from electromagnetic emissions, current consumption traces, RMS windows, and statistical descriptors are first processed by one-dimensional convolutional layers to capture local temporal patterns. The resulting high-level embeddings are then passed to an LSTM layer to model longer-term dependencies across consecutive windows. Dropout regularization and a final softmax layer produce scenario-level probability estimates.	133

List of Tables

5.1	Most noise-sensitive features.	74
5.2	Most noise-robust features.	74
5.3	Feature weights derived from noise-aware analysis.	75
6.1	Distribution of actions selected by the RL agent in the extended action-space configuration (20 files per scenario, SNR = 20 dB).	92
6.2	Action usage distribution for the RL feature-aware selector combined with a Random Forest classifier. The first two columns refer to a reduced configuration with 5 files per scenario, while the last two columns correspond to the full configuration with 20 files per scenario.	94
6.3	Action usage distribution for the RL feature-aware selector combined with a CNN+LSTM classifier.	96
B.1	Extracted features and corresponding signal domain.	122
B.2	Window-level accuracy obtained using moving average denoising on RMS features at SNR = 20 dB.	124
B.3	Window-level accuracy obtained using Savitzky–Golay filtering on RMS features at SNR = 20 dB (polynomial order = 2).	125
B.4	Window-level accuracy obtained using wavelet denoising on RMS features at SNR = 20 dB.	125
B.5	Window-level classification accuracy obtained with the extended pipeline under different denoising techniques (SNR = 20 dB).	126

Introduction

The rapid expansion of the Internet of Things (IoT) has profoundly transformed modern computing ecosystems, enabling pervasive connectivity across industrial control systems, healthcare infrastructures, smart cities, transportation networks, and other critical services. While this paradigm shift has delivered unprecedented levels of automation and efficiency, it has also significantly enlarged the attack surface of embedded and cyber-physical systems. IoT devices frequently operate under constrained computational resources, limited security mechanisms, and high degrees of physical accessibility, making them attractive targets for adversaries.

As the boundary between the digital and physical domains continues to blur, traditional software-centric security models prove insufficient to fully capture emerging threats. In particular, side-channel attacks represent a critical class of vulnerabilities that exploit unintended physical leakages generated during device operation. Processor microarchitectural activity, power consumption, and electromagnetic (EM) emissions may inadvertently disclose sensitive information, allowing adversaries to bypass cryptographic and logical protections even when robust algorithms are employed [27].

Among the various side channels, electromagnetic emissions are especially relevant in embedded and IoT environments. Electronic components unavoidably emit EM signals arising from clock activity, data transfers, and switching behavior. These emissions span a broad frequency spectrum, including low-frequency components—often in the order of a few megahertz—that are typical of microcontroller-based systems. Such low-frequency EM signals can propagate through obstacles and do not require line-of-sight access, enabling remote and non-invasive observation of device behavior [14].

The practical severity of EM side-channel leakage has been demonstrated in several real-world attacks. Notably, the *Screaming Channels* attack showed that digital activity in mixed-signal chips can couple into radio-frequency circuits and be amplified by the device antenna, allowing cryptographic keys to be recovered from distances of several meters [7]. These findings highlight how modern embedded platforms may unintentionally broadcast sensitive information beyond their immediate physical boundaries.

Side-channel attacks (SCAs) differ fundamentally from conventional attacks that target algorithmic weaknesses or software vulnerabilities. Instead, SCAs exploit physical emissions such as power consumption, electromagnetic radiation, timing variations, or acoustic signals to infer secret data or operational behavior [22, 13, 27]. Depending on their modus operandi, SCAs can be passive—where emissions are merely observed—or active, where controlled perturbations such as voltage glitches or electromagnetic pulses are injected to induce faults and reveal internal states [3].

The non-intrusive nature and high information content of electromagnetic side-channel analysis make it particularly suitable for both offensive and defensive applications. While attackers can exploit EM leakage to recover cryptographic secrets, defenders can leverage the same signals to monitor device behavior and detect deviations from expected execution patterns. This dual-use characteristic positions side-channel analysis as a powerful tool not only for attacks, but also for security monitoring and system assurance.

Artificial Intelligence (AI), and in particular Machine Learning (ML) and Deep Learning (DL), has dramatically reshaped the landscape of side-channel analysis. From an adversarial perspective, deep neural networks have demonstrated superior performance in extracting sensitive information from noisy and high-dimensional traces, often outperforming classical statistical techniques and reducing the need for handcrafted feature engineering [26, 21, 53]. Convolutional Neural Networks (CNNs), for instance, have proven effective in profiling attacks even in the presence of countermeasures.

From a defensive perspective, Machine Learning (ML) and Deep Learning (DL) techniques enable behavioral profiling and runtime monitoring of embedded and IoT systems through side-channel observations. By training classifiers on electromagnetic (EM) emissions or power consumption traces, it becomes possible to discriminate among different software activities—such as cryptographic operations, communication tasks, or idle states—without requiring access to internal variables or firmware instrumentation.

Previous studies have demonstrated classification accuracies above 90% using traditional ML approaches such as Random Forests, and up to 96% with deep learning architectures on real-world IoT platforms [16]. These capabilities provide valuable support for integrity verification, anomaly and malware detection, as well as forensic analysis, while preserving the non-intrusive nature of the monitoring process.

Effective application of ML and DL to side-channel data requires careful preprocessing and feature engineering. Raw physical traces are noisy, high-dimensional, and often non-stationary. Typical preprocessing steps include filtering, temporal segmentation, normalization, and noise reduction. Feature extraction may involve time-domain statistics, frequency-domain descriptors obtained via Fourier analysis, or hybrid time–frequency representations such as wavelets [25]. Feature selection techniques, including Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA), further reduce dimensionality and enhance model robustness by retaining the most informative components [18].

This thesis investigates a dual-use paradigm of AI-enhanced side-channel analysis, focusing not on cryptographic key extraction, but on inferring the *runtime behavior* of embedded systems deployed in realistic and often publicly accessible environments. By capturing and analyzing electromagnetic and power side-channel signals, the proposed approach demonstrates that application-level activities can be identified accurately without altering device firmware or software [38, 39].

In real-world scenarios such as smart cities, transportation hubs, ATM machines, ticketing kiosks, and other public-facing infrastructures, embedded devices operate continuously in environments where adversaries can opportunistically deploy side-channel sensing equipment. In these contexts, attackers may exploit inexpensive and widely available technologies—such as compact electromagnetic probes or software-defined radios—to collect EM emissions from nearby devices without physical contact. Hardware-level countermeasures are often impractical in such deployments, while static software defenses tend to become predictable over time.

To address these limitations, this work advances a conceptual shift from static side-channel profiling toward *adaptive informational defense*. Rather than attempting to suppress physical emissions, the proposed approach aims to dynamically degrade the information content exploitable by an adversary. This is achieved by modeling side-channel defense as a control problem at the feature-representation level, where preprocessing and transformation strategies are adapted in response to observed leakage characteristics.

Within this framework, Reinforcement Learning (RL) is employed as the enabling mechanism for adaptive, feature-aware control. The RL agent learns to select defensive actions that destabilize adversarial inference while preserving system functionality and observability for legitimate monitoring. This methodological transition—from side-channel-based profiling to adaptive control of information leakage—constitutes a central contribution of this Ph.D. work and is particularly suited to non-stationary, real-world environments.

Beyond profiling and defense, side-channel signals also provide a foundation for physical-layer Intrusion Detection Systems (IDS) in IoT environments. By monitoring power consumption or EM emissions, it is possible to detect runtime anomalies indicative of cyberattacks or system compromise. Such approaches complement traditional software-based IDS and are inherently more resilient to evasion, as they operate at the hardware level [11].

The experimental validation presented in this thesis is performed on a Raspberry Pi 4 Model B platform under eight representative application-layer scenarios, including web browsers, email clients, and network protocols such as HTTP, HTTPS, FTP, and FTPS. The obtained results demonstrate classification accuracies exceeding 98%, even in the presence of noisy operating conditions, thereby confirming the effectiveness and robustness of side-channel-based monitoring and adaptive defense mechanisms in realistic runtime environments [38].

Overall, this work contributes to the development of a new class of physical-layer and informational security mechanisms for embedded and IoT systems. By combining side-channel sensing with intelligent inference and adaptive control, the proposed framework enhances system observability and enables non-invasive mitigation of side-channel threats where traditional approaches fall short. The remainder of this thesis is organized as follows:

- Chapter 1 introduces the theoretical foundations of side-channel analysis, leakage modeling, machine learning, and adaptive learning paradigms for physical-layer security;
- Chapter 2 reviews the state of the art in AI-driven side-channel profiling, electromagnetic leakage analysis, adaptive countermeasures, and learning-based defensive approaches in realistic IoT and embedded-system scenarios;

- Chapter 3 presents the experimental setup and the methodological framework adopted throughout the thesis, including hardware acquisition, preprocessing, RMS-based feature extraction, normalization, feature engineering, and robustness-oriented pipeline design;
- Chapter 4 discusses the baseline experimental results and evaluates the performance of the proposed profiling framework under both clean and noisy operating conditions;
- Chapter 5 investigates the impact of noise, domain shift, and feature instability on side-channel inference, introducing a feature-aware perspective on robustness and adaptive information processing;
- Chapter 6 introduces the proposed reinforcement learning–based framework for adaptive and feature-aware side-channel analysis, presenting the RL architecture, control strategy, adaptive defense model, and experimental validation under adversarial and noisy conditions;
- Finally, Chapter 7 summarizes the main scientific contributions of the thesis, discusses current limitations, and outlines future research directions in adaptive and intelligent side-channel defense.

Chapter 1

Theoretical Background on Side-Channel Analysis and Machine Learning

This chapter introduces the theoretical foundations necessary to understand the side-channel attack (SCA) landscape and the role of Machine Learning (ML) and Deep Learning (DL) techniques in the analysis and exploitation of physical information leakage. The aim is not only to provide a classical overview of side-channel principles, but also to establish a coherent conceptual framework that connects physical signal acquisition, information representation, learning-based inference, and adaptive decision-making.

In contrast to purely algorithmic security analyses, side-channel analysis operates at the intersection of hardware behavior, software execution, and environmental interactions. As a result, the information available to an observer is inherently indirect, noisy, and temporally structured. Understanding how sensitive information propagates from internal system states to observable physical signals requires a multidisciplinary perspective that encompasses leakage modeling, signal processing, statistical learning, and temporal reasoning. This chapter provides the theoretical tools needed to interpret such leakages and to reason about their informational content under realistic operating conditions.

A central objective of this chapter is to motivate the data representations and preprocessing strategies adopted throughout the thesis. Raw side-channel signals are high-dimensional, non-stationary, and strongly affected by noise sources originating from both the device and its environment. Consequently, meaningful inference cannot rely on instantaneous observations alone, but must instead exploit aggregation over time, feature extraction, and dimensionality reduction. These aspects are introduced from a theoretical standpoint to justify the use of window-based representations, statistical and spectral features, denoising techniques, and temporal decision strategies employed in the experimental framework.

Beyond classical profiling attacks, this chapter also establishes the conceptual foundations for interpreting side-channel analysis as a dynamic and sequential inference problem. In realistic scenarios, leakage characteristics evolve over time due to changes in workload, environmental conditions, and system state. Static models trained under fixed assumptions may therefore become suboptimal or ineffective. This observation motivates the transition from traditional, static profiling approaches toward adaptive learning paradigms capable of responding to non-stationary leakage conditions.

Particular emphasis is placed on noise-aware analysis, temporal reasoning, and adaptive learning mechanisms, which collectively enable a shift from passive observation to active control of side-channel information. These concepts form the theoretical basis for the reinforcement learning-based adaptive defense strategies developed in the later chapters of this thesis, where side-channel processing is framed as a closed-loop control problem aimed at mitigating the information exploitable by an adversary.

1.1 Side-Channel Attacks

Side-channel attacks exploit indirect physical effects caused by the operation of electronic devices. Unlike traditional cryptanalytic attacks, which aim to compromise the mathematical strength of cryptographic algorithms, SCAs rely on observable phenomena that are statistically correlated with internal computations. Typical side channels include variations in power consumption [22], timing behavior [23], electromagnetic (EM) emissions [13], and, in some cases, acoustic or optical signals [14].

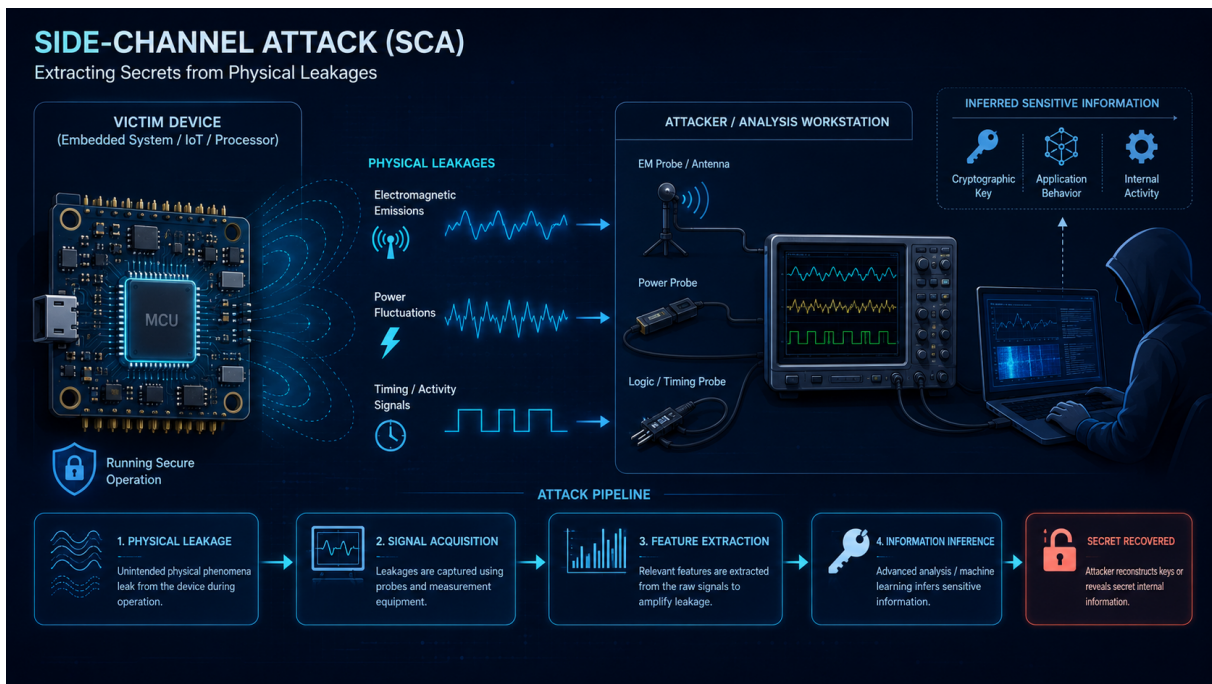


Figure 1.1: Conceptual overview of a generic Side-Channel Attack (SCA). Physical leakages generated during device operation, including electromagnetic emissions, power fluctuations, and timing-related activity, are captured through external measurement equipment and analyzed to infer sensitive information. The attack pipeline typically involves signal acquisition, feature extraction, and information inference, enabling the reconstruction of cryptographic secrets, application behavior, or internal system activity without direct access to the device internals.

These leakages originate from unavoidable interactions between hardware components, software execution, and the surrounding physical environment. Switching activity in logic gates, memory accesses, bus communications, and clock-driven operations all contribute to measurable physical emissions. As a result, even cryptographic primitives that are mathematically secure may become vulnerable when implemented on real hardware platforms [27, 47]. Side-channel analysis thus shifts the security focus from algorithms to implementations, exposing a gap between theoretical security guarantees and practical deployments.

1.1.1 Classification of Side-Channel Attacks

From an operational perspective, side-channel attacks are commonly classified into passive and active categories.

Passive SCAs observe physical emissions without altering the normal behavior of the target device. Classical examples include Simple Power Analysis (SPA) and Differential Power Analysis (DPA), where statistical correlations between measured traces and internal data values are exploited to recover secret information [22]. Passive attacks are particularly difficult to detect, as they do not introduce faults or observable disruptions in system operation.

Active SCAs, on the other hand, deliberately perturb the target system to induce faulty behavior. Techniques such as voltage glitches, clock manipulation, or electromagnetic fault injection can trigger computational errors that reveal sensitive intermediate states or bypass security checks [5, 3]. Active attacks often combine fault induction with side-channel observation, significantly increasing their effectiveness.

Beyond this basic dichotomy, side-channel attacks can also be classified according to the exploited physical channel (power, EM, timing), the required level of access (local, proximity-based, or remote), and the attack objective. While early work primarily focused on cryptographic key extraction, more recent studies have demonstrated that side-channel information can also be used to infer high-level system behavior, application execution, and protocol usage [46].

1.1.2 Electromagnetic Side-Channel Analysis

Among the various side channels, electromagnetic emissions are of particular relevance for embedded and IoT systems. EM side-channel analysis is inherently non-invasive and does not require electrical contact with the target device. Emissions originate from current flows in power delivery networks, clock lines, data buses, and memory interfaces, spanning a wide frequency spectrum.

Low-frequency EM emissions, typically in the range of a few megahertz to tens of megahertz, are especially significant in microcontroller-based and embedded platforms. These components often propagate through obstacles and enclosures with limited attenuation and do not require line-of-sight access, enabling proximity-based or even remote observation [14]. This characteristic makes EM side-channel analysis particularly attractive in real-world scenarios where physical access to the device is constrained but proximity is feasible.

The severity of EM leakage has been demonstrated in several practical attacks. The *Screaming Channels* attack showed that digital activity in mixed-signal chips can couple into radio-frequency circuitry and be unintentionally amplified by on-board antennas, allowing cryptographic keys to be recovered from distances of several meters [7]. Such results highlight that modern embedded platforms may inadvertently broadcast sensitive information beyond their immediate physical boundaries.

1.1.3 Multi-Channel and Cross-Layer Leakage

Modern devices rarely emit information through a single isolated channel. Instead, multiple side channels often coexist and interact. Multi-channel attacks exploit combined leakage from power, EM, and timing channels to improve inference accuracy and robustness [2]. Cross-layer leakage further arises when high-level software behavior influences low-level hardware activity in structured and repeatable ways.

In embedded and IoT systems, application-layer operations such as network communication, protocol handling, or data encryption induce characteristic execution patterns that manifest in physical emissions. As a result, side-channel analysis can extend beyond cryptographic attacks toward behavioral profiling, enabling the identification of running applications, services, or operational modes [38]. This shift significantly broadens the threat model, as attackers may no longer require precise knowledge of cryptographic implementations to extract valuable information.

1.1.4 Threat Models and Practical Considerations

The feasibility of side-channel attacks depends strongly on the assumed threat model. In laboratory settings, attackers often have full control over measurement conditions and repeated access to the target device. In contrast, real-world scenarios involve noise, variability, and limited access, particularly in IoT and cyber-physical systems deployed in uncontrolled environments.

Recent surveys emphasize that non-invasive side-channel attacks are increasingly relevant in realistic deployments, including public infrastructures and smart environments, where attackers may operate in proximity to devices using low-cost and portable measurement equipment [45]. In such contexts, noise and environmental interference do not eliminate side-channel leakage but rather motivate more sophisticated analysis techniques capable of extracting information under adverse conditions.

These considerations underscore that side-channel attacks should not be viewed solely as laboratory curiosities, but as a practical security concern for modern embedded systems. At the same time, they open the door to defensive applications that leverage the same physical signals for monitoring, profiling, and anomaly detection.

1.2 Leakage Models and Measurement Setups

Effective interpretation of side-channel signals requires the adoption of appropriate leakage models that describe how internal computational states and operations manifest as observable physical emissions. Leakage models provide an abstraction layer between the complex physical behavior of a device and the analytical or learning-based techniques used to extract information from measured traces. Rather than capturing the full physics of signal generation and propagation, leakage models aim to characterize the statistical relationship between sensitive variables and observable quantities.

From an information-theoretic perspective, a leakage model defines the mapping between internal states and external observations, implicitly determining what information can be inferred by an attacker or a monitoring system. The accuracy and generality of this model directly impact the effectiveness of both

attacks and defenses. Simplified models may enable tractable analysis, but they also introduce mismatches when applied to realistic systems operating under noise, variability, and environmental interference.

1.2.1 Analytical Leakage Models

Early work in side-channel analysis focused on analytical leakage models derived from simplified assumptions about hardware behavior. Among the most widely adopted are the **Hamming Weight** and **Hamming Distance** models, particularly in the context of power analysis [8]. The Hamming Weight model assumes that instantaneous power consumption is proportional to the number of bits set to one in a processed value, while the Hamming Distance model correlates power consumption with the number of bit transitions between consecutive states.

These models have proven effective for analyzing simple microcontroller architectures and cryptographic implementations with relatively deterministic execution patterns. However, their applicability diminishes as system complexity increases. Modern processors, embedded platforms, and IoT devices exhibit complex interactions between pipelines, caches, memory hierarchies, and peripheral activity, which violate the assumptions underlying simple analytical models.

To address these limitations, **linear and affine leakage models** have been introduced, particularly for electromagnetic and timing channels [41]. These models assume a linear relationship between the measured signal and a set of internal variables, enabling statistical analysis while offering greater flexibility than bit-level abstractions. Nevertheless, even linear models struggle to capture non-linearities, cross-coupling effects, and frequency-dependent behavior commonly observed in EM emissions.

1.2.2 Black-Box and Data-Driven Leakage Models

As analytical modeling becomes increasingly impractical for complex systems, **black-box leakage models** have gained prominence. These models make no explicit assumptions about the physical origin of leakage. Instead, they rely on data-driven techniques—most notably machine learning—to infer the relationship between observed traces and internal system behavior [18, 33].

Black-box models shift the focus from physical interpretation to empirical inference. From a practical standpoint, they are particularly well suited to heterogeneous IoT devices, system-on-chip platforms, and mixed-signal architectures, where leakage arises from multiple interacting sources. In these contexts, the leakage model is implicitly learned from data rather than analytically defined.

While powerful, black-box approaches introduce new challenges. Learned models may overfit specific operating conditions, devices, or noise characteristics, limiting their generalization. Moreover, the absence of an explicit physical model complicates interpretability and robustness analysis, especially under adversarial or non-stationary conditions. These limitations motivate the integration of domain knowledge, feature engineering, and adaptive learning strategies.

1.2.3 Measurement Setups as Part of the Leakage Model

Leakage models cannot be considered independently of the measurement setup used to acquire side-channel signals. In practice, the measurement chain—including sensors, acquisition hardware, synchronization mechanisms, and environmental conditions—acts as a transformation that shapes the observed leakage.

A typical side-channel measurement setup includes sensors such as magnetic field probes, current probes, or shunt resistors; high-speed acquisition hardware such as oscilloscopes or Analog-to-Digital Converters (ADCs); and synchronization mechanisms to ensure temporal alignment across traces. Shielding, grounding, and probe placement further influence signal quality and spectral characteristics [34].

From a modeling perspective, the measurement setup introduces filtering effects, noise sources, and bandwidth limitations that alter the observable signal. As a result, the effective leakage model seen by an analyst is a composition of the device’s intrinsic leakage and the measurement chain response. Ignoring this interaction can lead to inaccurate assumptions about signal structure and information content.

1.2.4 Noise, Variability, and Non-Stationarity

In controlled laboratory environments, measurement setups are optimized to minimize noise and variability. However, realistic deployments—particularly in embedded and IoT contexts—are inherently noisy

and non-stationary. Environmental interference, power supply fluctuations, concurrent software activity, and workload variability all contribute to time-varying leakage characteristics [45].

Noise and variability do not eliminate side-channel leakage, but they significantly affect its observability and statistical structure. Simple leakage models may fail under such conditions, while static preprocessing pipelines may become suboptimal as noise characteristics evolve. This observation has direct implications for both attack and defense strategies, as it challenges the assumption of stationary and repeatable leakage patterns.

1.2.5 Implications for Feature Representation and Adaptive Processing

The limitations of analytical models and the variability introduced by realistic measurement setups motivate the use of feature-based representations and adaptive processing strategies. By transforming raw traces into statistical, spectral, or time-frequency features, it becomes possible to capture invariant properties of leakage while mitigating sensitivity to noise and misalignment.

Moreover, the non-stationary nature of leakage suggests that preprocessing and feature extraction should not be treated as fixed design choices, but as adaptive components of the analysis pipeline. This perspective directly motivates noise-aware feature selection, temporal aggregation strategies, and learning-based control mechanisms, which are further explored in the subsequent sections and form the conceptual foundation for the reinforcement learning-based adaptive framework presented later in this thesis.

1.3 Side-Channel Countermeasures and Defense Strategies

The existence of side-channel leakage has motivated a broad spectrum of countermeasures aimed at mitigating the information that can be extracted from physical emissions. Unlike purely algorithmic vulnerabilities, which can often be addressed through formal analysis or cryptographic redesign, side-channel leakages are intrinsically tied to the physical realization of computing systems. They emerge from the interaction between hardware architecture, software execution, and environmental conditions, making their mitigation a fundamentally multidisciplinary challenge.

As a consequence, side-channel defenses cannot be confined to a single abstraction layer. Instead, existing strategies span multiple levels of the system stack and pursue different objectives. Some approaches aim to physically suppress or attenuate emissions at their source, while others seek to decorrelate observable signals from sensitive data or to reduce the exploitable information content available to an adversary. From a conceptual standpoint, these defenses can be broadly classified according to the layer at which they operate: hardware-level countermeasures, software and algorithmic techniques, and information-oriented or learning-based approaches.

Hardware-level countermeasures focus on limiting the generation or propagation of physical leakages, whereas software-based techniques attempt to disrupt data-dependent execution patterns through randomization or masking. More recently, information-oriented approaches have emerged, motivated by the increasing use of machine learning in side-channel attacks. These methods do not aim to eliminate physical emissions, but rather to interfere with the statistical structure exploited by an attacker’s inference model.

Despite their effectiveness in specific contexts, most existing countermeasures share a common limitation: they are static by design. Once deployed, their behavior remains fixed, implicitly assuming stationary leakage characteristics and a non-adaptive adversary. In realistic deployments, however, side-channel leakage is inherently non-stationary, influenced by workload variability, environmental noise, and adversarial adaptation. This mismatch motivates the exploration of adaptive defense mechanisms capable of dynamically adjusting mitigation strategies in response to evolving conditions.

In this perspective, adaptive and learning-based defenses represent a natural evolution of side-channel countermeasures, enabling closed-loop control over the information exposed through physical channels. The remainder of this section reviews existing defense categories, highlighting their strengths and limitations, and establishes the conceptual foundations for the reinforcement learning-based adaptive defense framework developed in the later chapters of this thesis.

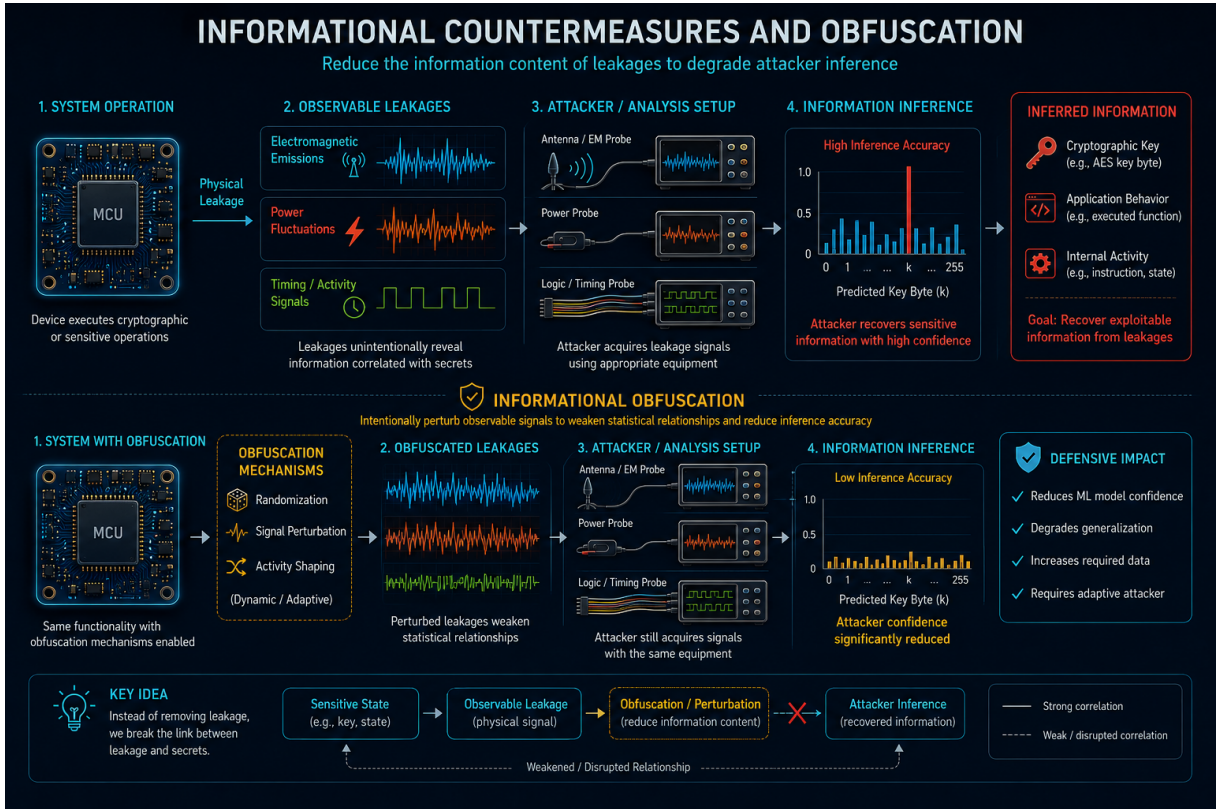


Figure 1.2: Conceptual overview of informational countermeasures and obfuscation strategies in side-channel analysis. In conventional side-channel attacks, observable physical leakages such as electromagnetic emissions, power fluctuations, and timing-related activity maintain strong statistical correlations with sensitive internal states, enabling high-confidence attacker inference. Informational obfuscation aims to disrupt these relationships by introducing adaptive perturbations, randomization, or activity shaping mechanisms that reduce the exploitable information content of observable signals. Rather than suppressing physical emissions entirely, the proposed defensive perspective weakens the attacker’s inference capability while preserving normal system functionality, thereby increasing uncertainty, reducing model confidence, and degrading the effectiveness of learning-based attacks.

1.3.1 Hardware-Level Countermeasures

Hardware-level countermeasures aim to suppress or attenuate physical emissions at their source. Common techniques include electromagnetic shielding, filtering of power supply lines, careful PCB layout design, and the use of differential signaling to reduce correlated emissions [44, 34]. In secure hardware designs, dedicated countermeasures such as dual-rail logic, current flattening, and random clocking are often employed to reduce data-dependent leakage.

While effective in controlled environments, hardware-based defenses present significant limitations in practical deployments. Shielding and filtering increase manufacturing costs and power consumption, while design-time countermeasures cannot be easily retrofitted into existing systems. Moreover, residual leakage often remains exploitable, particularly when attackers can average observations over time or combine multiple side channels [2].

In the context of IoT and embedded systems deployed in public or semi-public environments, hardware-level defenses are frequently impractical due to cost constraints, physical accessibility, and legacy deployments. These limitations motivate the exploration of complementary defense mechanisms at higher abstraction levels.

1.3.2 Software and Algorithmic Countermeasures

Software-based countermeasures aim to reduce the correlation between sensitive data and observable physical emissions through algorithmic transformations. Common techniques include instruction shuffling, random delays, dummy operations, and execution path randomization [28]. Noise injection strategies deliberately introduce random activity to obscure data-dependent patterns in power or EM traces.

At the cryptographic level, masking techniques split sensitive variables into multiple shares processed independently, reducing first-order leakage [8]. While theoretically sound, masking implementations are often complex, incur performance penalties, and may still leak information through higher-order effects or implementation flaws [41].

A key limitation of software-level defenses is their static nature. Once deployed, the defense behavior becomes predictable, allowing attackers to adapt their analysis strategies accordingly. Furthermore, excessive randomization or noise injection may degrade system performance or interfere with normal operation, limiting their applicability in real-time or resource-constrained environments.

1.3.3 Informational Countermeasures and Obfuscation

Beyond suppressing physical emissions, an alternative defensive perspective focuses on degrading the *information content* available to an attacker. Informational countermeasures do not attempt to eliminate leakage, but rather to reduce its exploitability by disrupting the statistical relationships between observed signals and sensitive system states.

Recent work has explored obfuscation strategies that deliberately perturb observable signals in a manner that frustrates machine learning-based attackers [29]. These approaches draw inspiration from adversarial machine learning, where carefully crafted perturbations can significantly reduce model confidence without requiring large signal modifications.

Such informational defenses are particularly attractive in side-channel contexts, where attackers increasingly rely on ML and DL models trained on large datasets. By targeting the feature space exploited by these models, defenders can reduce inference accuracy even when physical emissions remain observable.

However, existing obfuscation-based defenses are typically static or heuristic, relying on fixed perturbation strategies that may lose effectiveness as attacker models evolve.

1.3.4 Learning-Based and Adaptive Defense Approaches

The growing use of ML and DL in side-channel attacks naturally motivates the application of learning techniques for defense. Learning-based defenses aim to model the leakage process itself and dynamically adapt mitigation strategies based on observed signal characteristics.

Early learning-based approaches primarily focus on improving detection or classification accuracy for defensive monitoring, such as physical-layer intrusion detection systems based on power or EM analysis [11]. More recent work has begun exploring defensive ML techniques that actively manipulate observable signals to reduce attacker success [29].

Despite these advances, most learning-based defenses remain static once trained and do not explicitly account for the sequential and non-stationary nature of side-channel leakage. In realistic environments, leakage characteristics evolve over time due to workload changes, environmental noise, and adversarial adaptation. Static defenses may therefore become suboptimal or ineffective.

1.3.5 Motivation for Reinforcement Learning-Based Adaptive Defense

Reinforcement Learning (RL) provides a natural framework to address the limitations of static countermeasures by framing side-channel defense as a sequential decision-making problem. In this paradigm, the defender is modeled as an agent that observes side-channel characteristics and selects actions—such as denoising strategies, feature transformations, or obfuscation parameters—to optimize a long-term objective.

RL enables two complementary defensive roles. First, it supports *noise-aware compensation*, dynamically adapting preprocessing and feature extraction strategies to maximize legitimate observability under varying noise conditions. Second, it enables *adaptive informational defense*, where the agent actively degrades the consistency and reliability of features exploitable by an attacker, without requiring hardware modifications or intrusive software changes.

Unlike static defenses, an RL-based approach can continuously adapt to changing conditions and adversarial strategies, making it particularly suitable for embedded and IoT systems deployed in uncontrolled environments. To the best of the author’s knowledge, the use of reinforcement learning as a closed-loop control mechanism for adaptive side-channel defense has not been systematically investigated in prior literature, representing a key contribution of this thesis.

1.4 Machine Learning and Deep Learning in Side-Channel Analysis

The increasing complexity of modern computing systems and the intrinsic limitations of analytical leakage models have positioned Machine Learning (ML) and Deep Learning (DL) as central tools in contemporary side-channel analysis. Unlike classical statistical techniques, learning-based approaches enable automated feature extraction, robust classification, and scalable analysis of high-dimensional and noisy physical traces [25, 33]. This paradigm shift has fundamentally changed both the offensive and defensive landscape of side-channel research.

From a conceptual perspective, ML and DL transform side-channel analysis from a manually engineered correlation problem into a data-driven inference task. Rather than relying on explicitly defined leakage assumptions, learning-based models infer discriminative patterns directly from observed signals, making them particularly suitable for complex embedded platforms and heterogeneous IoT devices.

1.4.1 Profiling Attacks and Key Recovery

One of the earliest and most influential applications of ML in side-channel analysis is cryptographic key recovery through profiling attacks. In this setting, an adversary collects labeled side-channel traces from a reference device operating under known keys or inputs and trains a supervised model to associate physical leakage patterns with sensitive intermediate values [47]. During the attack phase, the trained model is applied to previously unseen traces to infer secret information.

Early profiling approaches employed classical classifiers such as Support Vector Machines, k -Nearest Neighbors, and Random Forests, demonstrating that machine learning could outperform traditional template attacks in noisy or imperfectly aligned scenarios. The introduction of deep learning further amplified this effect. Convolutional Neural Networks (CNNs), in particular, have proven effective at learning hierarchical representations directly from raw or minimally processed traces, significantly reducing the need for manual feature engineering [26, 21, 53].

These results highlight a key insight: side-channel leakage often contains exploitable structure even when traditional assumptions about alignment or noise are violated. This observation has profound implications beyond cryptography, as it suggests that physical emissions encode rich behavioral information that can be extracted through learning-based inference.

1.4.2 Behavioral Profiling of Embedded and IoT Systems

Beyond cryptographic key extraction, machine learning enables the inference of high-level operational behavior from side-channel signals. Behavioral profiling aims to identify which application, protocol, or functional mode is currently executed by a device by analyzing power consumption or electromagnetic emissions.

In embedded and IoT systems, software activities such as network communication, encryption, or user interaction induce characteristic execution patterns that propagate to the physical layer. Learning-based classifiers trained on side-channel features can exploit these patterns to distinguish between browsing, communication, idle, or computation-intensive states with high accuracy [16, 38].

This shift from key-centric attacks to behavioral inference significantly broadens the scope of side-channel analysis. Rather than targeting isolated cryptographic routines, attackers and defenders alike can reason about system-level behavior, application execution, and protocol usage. From a defensive perspective, this capability enables non-intrusive monitoring and integrity verification without requiring access to internal system states or modifications to firmware.

1.4.3 Anomaly Detection and Physical-Layer Intrusion Detection

Unsupervised and semi-supervised learning methods extend side-channel analysis to anomaly detection and runtime monitoring. In this context, models are trained on traces collected during normal operation and used to identify deviations indicative of faults, attacks, or unauthorized behavior [51].

Autoencoders, clustering algorithms, and density-based methods learn compact representations of nominal side-channel behavior. At runtime, anomalies are detected by evaluating reconstruction errors or distance metrics between observed traces and learned models [11]. When deployed continuously, such approaches form the basis of physical-layer Intrusion Detection Systems (IDS), which monitor hardware-level behavior rather than software logs.

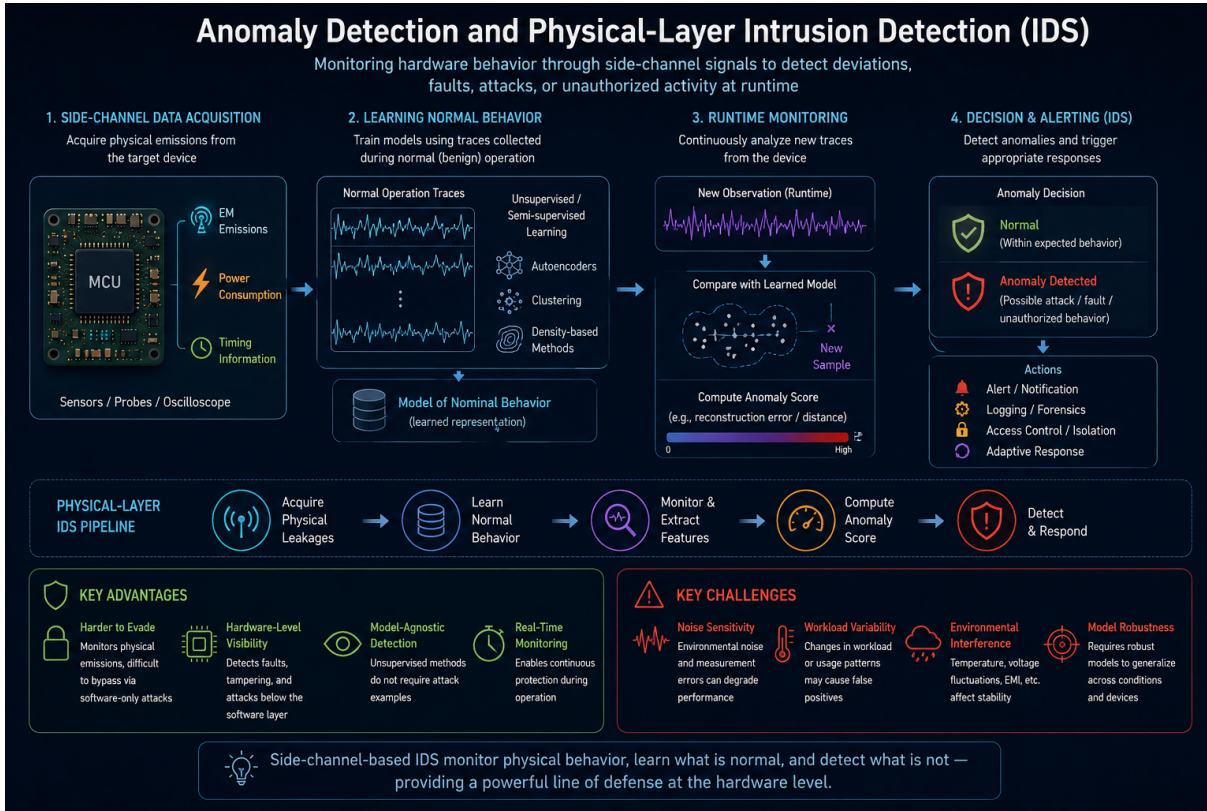


Figure 1.3: Conceptual overview of a physical-layer Intrusion Detection System (IDS) based on side-channel analysis. Physical leakages such as electromagnetic emissions, power consumption, and timing-related activity are continuously acquired from the target device and used to learn models of normal system behavior through unsupervised or semi-supervised learning techniques. During runtime, newly observed traces are compared against the learned representation to compute anomaly scores and detect deviations associated with faults, cyberattacks, or unauthorized activity. By operating directly on hardware-level emissions, side-channel-based IDS provide non-intrusive and difficult-to-evade monitoring capabilities, complementing traditional software-based security mechanisms.

A key advantage of side-channel-based IDS lies in their resilience to evasion. Since they operate on physical emissions, they are inherently more difficult to bypass through purely software-level manipulation. However, their effectiveness critically depends on robustness to noise, workload variability, and environmental interference.

1.4.4 Noise-Aware Feature Extraction and Denoising

Raw side-channel signals are inherently noisy, non-stationary, and high-dimensional. Noise originates from environmental interference, concurrent software activity, power supply fluctuations, and measurement artifacts. Effective preprocessing and denoising are therefore essential to enable reliable learning-based inference, particularly in realistic deployments.

Classical signal processing techniques such as moving averages, Savitzky–Golay filtering, and wavelet-based denoising have long been employed to suppress high-frequency noise while preserving informative components [28, 43]. These methods exploit assumptions about signal smoothness and spectral locality, providing a first level of noise mitigation.

More recent work has demonstrated that denoising can be tightly integrated into learning-based pipelines. Deep learning-based denoising and feature fusion methods, including encoder–decoder architectures, have shown improved robustness and classification performance under low signal-to-noise ratio (SNR) conditions [19]. These approaches treat denoising as a learnable transformation rather than a fixed preprocessing step, enabling adaptation to complex and evolving noise characteristics.

1.4.5 Feature Selection, Temporal Modeling, and Decision Strategies

Feature extraction alone is insufficient to guarantee robust inference. Feature selection techniques further refine representations by removing redundant or irrelevant dimensions, reducing model complexity

and improving generalization. Recursive Feature Elimination (RFE) [15] and information-theoretic approaches based on mutual information [54] are widely used to identify the most discriminative features in side-channel datasets.

Equally critical is the temporal dimension of side-channel inference. Individual samples or extremely short windows are dominated by noise and transient effects, rarely conveying sufficient semantic information to characterize system behavior. Window-based representations mitigate this limitation by aggregating information over time, but reliable inference often requires reasoning over sequences of windows rather than isolated observations.

This observation is particularly relevant when decisions are made at different temporal granularities, ranging from window-level inference to sequence-level behavioral classification. In practice, a single decision window may provide limited or ambiguous evidence, while consistent patterns emerging across multiple windows can reveal stable behavioral signatures.

Temporal aggregation strategies, such as majority voting or confidence-based fusion, exploit the consistency of predictions over time to compensate for local misclassifications [24].

From a methodological standpoint, these considerations imply that side-channel inference pipelines should explicitly account for temporal structure and decision hierarchy. Rather than optimizing classifiers solely at the window level, effective systems must reason about how local predictions accumulate over time to support robust, higher-level decisions. This perspective naturally extends toward adaptive strategies in which temporal aggregation, feature relevance, and noise mitigation are dynamically adjusted based on observed signal conditions.

1.4.6 From Static Learning to Adaptive Defense

While the majority of existing studies focus on improving attack effectiveness or static profiling accuracy, recent work has begun exploring learning-based defensive strategies. Approaches based on adversarial obfuscation and defensive machine learning aim to perturb observable signals in order to frustrate attacker models [29]. These methods demonstrate that learning-driven defense is feasible, but they typically rely on fixed or heuristic perturbation strategies.

The inherent non-stationarity of side-channel leakage motivates a more general and adaptive approach. Reinforcement Learning (RL) offers a principled framework to model side-channel defense as a sequential decision-making problem. In this paradigm, preprocessing, feature extraction, and representation strategies are no longer static design choices, but controllable actions optimized over time.

Within this framework, RL can serve a dual role. First, it enables noise-aware compensation by dynamically adapting denoising and feature extraction strategies to changing environmental conditions. Second, it supports adaptive informational defense by actively degrading the stability and reliability of features exploitable by an adversary.

It is worth noting that the objective of the proposed framework is not to guarantee information-theoretic security, but to actively reduce the exploitable information available to an adversary under realistic physical, computational, and operational constraints. Rather than attempting to eliminate side-channel leakage altogether—which is often infeasible in practice—the proposed approach aims to control and distort the statistical structure of observable signals in a way that limits adversarial inference.

Unlike static countermeasures, an RL-based defense can continuously adapt to evolving noise conditions, workload variability, and potential attacker strategies. To the best of the author’s knowledge, such a closed-loop, reinforcement learning–based formulation for adaptive side-channel defense has not been systematically explored in prior literature, motivating the contributions developed in the subsequent chapters of this thesis.

1.5 Conclusion

This chapter presented the theoretical foundations required to frame side-channel analysis as an information-driven inference problem and to understand the role of machine learning and deep learning in extracting, modeling, and exploiting physical leakage. Rather than treating side-channel emissions as isolated signal artifacts, the chapter emphasized their interpretation as structured, noisy, and temporally correlated manifestations of internal system behavior.

The discussion of leakage models and measurement setups highlighted the limitations of analytical assumptions in realistic environments, where noise, variability, and non-stationarity fundamentally shape observable signals.

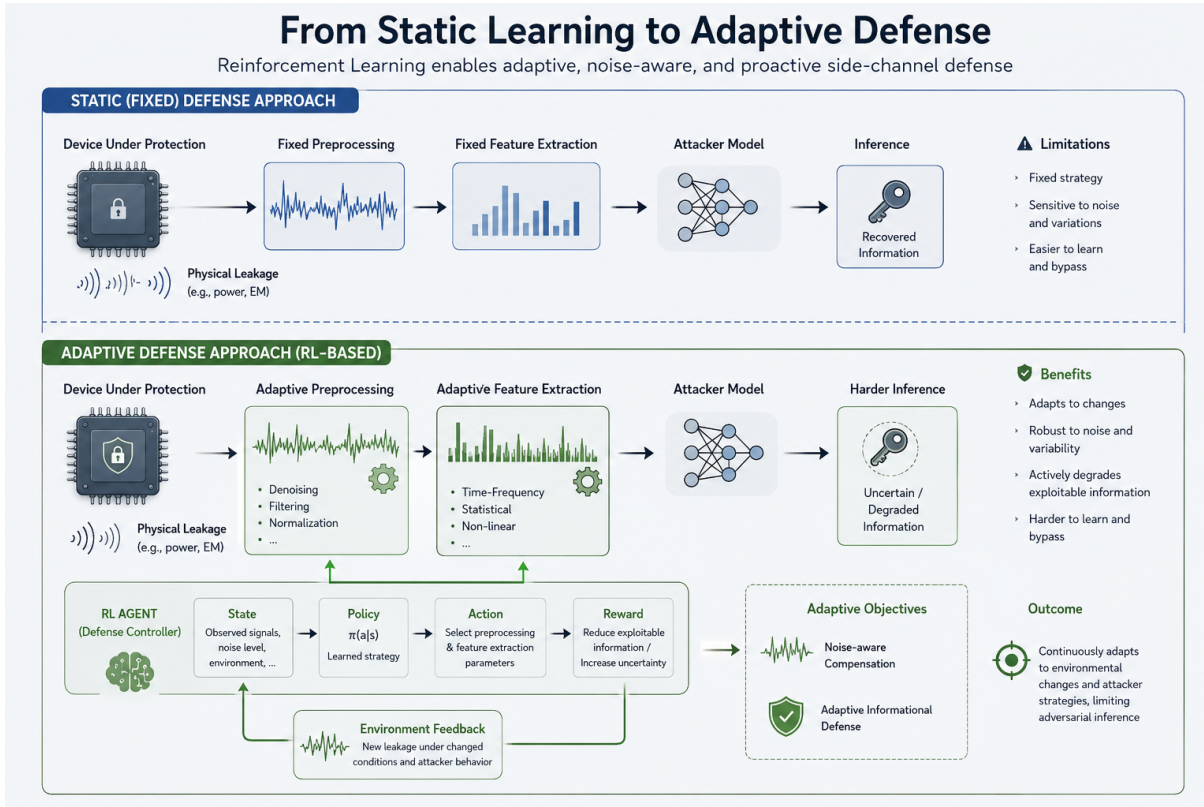


Figure 1.4: Conceptual comparison between static side-channel defense and reinforcement learning-based adaptive defense. In the static approach, physical leakages are processed through fixed preprocessing and feature extraction stages, making the resulting inference pipeline sensitive to noise, environmental variability, and adaptive attackers. In contrast, the proposed adaptive defense framework introduces a reinforcement learning agent that dynamically selects preprocessing, feature transformation, and representation-control actions according to the observed leakage conditions. By closing the loop between side-channel observations, policy decisions, and environmental feedback, the adaptive framework aims to preserve legitimate monitoring capabilities while degrading the exploitable information available to an attacker, thereby increasing uncertainty and improving robustness against non-stationary side-channel conditions.

These considerations motivated the adoption of feature-based representations, denoising techniques, and temporal aggregation strategies as essential components of robust side-channel analysis pipelines.

By reviewing learning-based approaches to key recovery, behavioral profiling, and anomaly detection, the chapter illustrated how ML and DL enable scalable inference across complex embedded and IoT platforms. At the same time, the inherent limitations of static models and fixed preprocessing pipelines were identified, particularly in the presence of evolving workloads, environmental interference, and adaptive adversaries.

Collectively, these insights establish a conceptual bridge from traditional, static side-channel profiling toward adaptive and learning-driven defense mechanisms. In particular, the combination of noise-aware processing, temporal reasoning, and feature-level control provides the foundation for formulating side-channel mitigation as a sequential decision-making problem. This perspective directly motivates the reinforcement learning-based adaptive framework introduced in the subsequent chapters, where side-channel analysis evolves from passive observation to active control of the information exposed through physical channels.

Chapter 2

State of the Art in AI-Based Side-Channel Profiling and Security

2.1 Scope and Positioning of the State of the Art

The purpose of this chapter is to review the current state of the art in Artificial Intelligence-based side-channel profiling and security, with a specific focus on learning-driven analysis, robustness under noise, and adaptive defense strategies. Rather than reiterating the fundamental principles of side-channel analysis introduced in Chapter 1, this chapter situates existing research within a broader methodological and technological landscape, identifying dominant trends, structural limitations, and open challenges that directly motivate the contributions of this thesis.

While Chapter 1 established the theoretical foundations of side-channel leakage, measurement abstractions, and learning-based inference, the present chapter concentrates on how these concepts have been instantiated in practice. In particular, emphasis is placed on works that employ Machine Learning (ML) and Deep Learning (DL) techniques for profiling, classification, anomaly detection, and security monitoring based on power and electromagnetic (EM) side-channel signals.

The scope of this review is intentionally focused on embedded and IoT-class systems, where low-frequency side-channel emissions are both practically observable and security-relevant. Such platforms are typically characterized by constrained computational resources, heterogeneous and time-varying workloads, and deployment in uncontrolled or semi-public environments. In these contexts, assumptions commonly adopted in laboratory settings—such as stationary noise conditions, stable execution patterns, or perfect trace alignment—are rarely satisfied. Consequently, robustness, generalization, and adaptability emerge as central requirements for both offensive and defensive side-channel methodologies.

A key theme emerging from the reviewed literature is the progressive shift from handcrafted feature extraction and analytical leakage modeling toward data-driven and black-box approaches. Supervised learning techniques have demonstrated remarkable effectiveness in cryptographic key recovery and behavioral profiling tasks, while unsupervised and semi-supervised models have enabled physical-layer anomaly detection and runtime monitoring. However, despite these advances, most existing approaches rely on static preprocessing pipelines, fixed feature representations, and offline-trained models, implicitly assuming that the statistical structure of side-channel leakage remains stable over time.

This chapter critically examines such assumptions by surveying recent contributions that explicitly address noise, variability, and environmental interference in side-channel analysis. Particular attention is devoted to techniques for denoising, feature selection, and temporal modeling, as well as to emerging attack scenarios that challenge traditional threat models, including proximity-based and remote electromagnetic attacks and cross-layer leakage exploitation.

Finally, the chapter introduces learning-based defensive perspectives, focusing on approaches that attempt to mitigate side-channel leakage or reduce its exploitability through obfuscation, monitoring, or adaptive signal processing. Although reinforcement learning has been widely adopted in control systems, signal processing, and cybersecurity more broadly, its application as a principled mechanism for adaptive side-channel defense remains largely unexplored. By systematically reviewing the existing literature, this chapter identifies this gap and establishes the conceptual and methodological motivation for the reinforcement learning-based adaptive framework developed in the subsequent chapters of this thesis.

2.2 Classical Side-Channel Attacks and Profiling Foundations

Early research on side-channel attacks established the fundamental feasibility of extracting sensitive information from physical emissions generated during device operation. These works demonstrated that side-channel leakage constitutes a practical and powerful attack vector even in the absence of software vulnerabilities or cryptographic weaknesses. Classical side-channel analysis primarily focused on cryptographic key recovery and provided the conceptual and methodological foundations upon which later learning-based approaches have been developed.

Traditional side-channel attacks exploit physical signals—most notably power consumption and electromagnetic (EM) emissions—that exhibit statistical dependence on internal computations. Rather than relying on direct signal interpretation, these techniques leverage the repeatability of leakage patterns across multiple executions. By correlating measured traces with hypotheses about internal data values or operations, an attacker can progressively amplify weak data-dependent effects and infer secret information.

2.2.1 Power and Electromagnetic Analysis

Power analysis attacks were among the first practical side-channel techniques to be demonstrated. Simple Power Analysis (SPA) exploits visually observable variations in power consumption corresponding to different operations or control-flow decisions. While conceptually simple, SPA can be effective when execution patterns are strongly data-dependent or when branching behavior differs across operations.

Differential Power Analysis (DPA) introduced a more systematic and statistically grounded methodology [22]. By collecting large sets of power traces and partitioning them according to hypothetical intermediate values, DPA enables the amplification of subtle data-dependent variations that are otherwise obscured by noise. Correlation Power Analysis (CPA) further refined this approach by explicitly modeling the relationship between predicted intermediate values and observed leakage using correlation metrics.

Electromagnetic side-channel analysis follows analogous principles but operates on radiated emissions rather than conducted power signals. EM analysis offers important practical advantages, including non-invasive measurement and spatial selectivity, as probes can be positioned to isolate specific components or subsystems. EM leakage typically spans a wide frequency range and reflects complex interactions among switching activity, clock signals, and bus communications [13].

Despite their effectiveness, classical power and EM attacks generally rely on carefully controlled measurement setups, precise synchronization, and handcrafted signal processing pipelines. As a result, their performance is highly sensitive to noise, misalignment, and environmental variability, limiting their robustness outside laboratory conditions.

2.2.2 Leakage Models and Hypothesis Testing

At the core of classical side-channel attacks lies the concept of a leakage model, which formalizes the assumed relationship between internal variables and observable signals. Widely adopted abstractions include the Hamming Weight and Hamming Distance models, particularly in power analysis [8]. These models enable the construction of hypotheses about sensitive intermediate values that can be statistically tested against measured traces.

Template attacks represent one of the most powerful classical profiling techniques. In a template attack, an adversary builds multivariate statistical models—typically Gaussian distributions—of side-channel leakage for each possible value of a sensitive intermediate. During the attack phase, likelihood evaluations are used to identify the most probable secret value [47].

While highly effective under ideal conditions, template attacks exhibit several structural limitations. They require large numbers of high-quality profiling traces, accurate trace alignment, and strong assumptions regarding leakage stationarity. Moreover, the dimensionality of modern side-channel traces and the presence of noise often render classical multivariate modeling impractical without aggressive dimensionality reduction.

2.2.3 Operational Constraints and Practical Limitations

Classical side-channel attacks are predominantly developed and evaluated under controlled laboratory conditions, where noise sources are minimized and measurement repeatability is ensured. In realistic

deployment scenarios—particularly for embedded and IoT systems—these assumptions rarely hold.

Environmental interference, concurrent software activity, power supply fluctuations, and device aging introduce variability that degrades the effectiveness of analytical leakage models. In addition, modern embedded platforms exhibit complex execution behavior involving multitasking, caching, and peripheral interactions, all of which contribute to side-channel signals in ways that are difficult to capture analytically.

Scalability represents another critical limitation of classical approaches. Handcrafted preprocessing pipelines, manual feature selection, and hypothesis-driven analysis do not scale well to large datasets, heterogeneous devices, or diverse workloads. Consequently, classical techniques struggle to generalize across platforms or to adapt to evolving leakage characteristics over time.

2.2.4 From Analytical Profiling to Data-Driven Approaches

The limitations of classical side-channel techniques have progressively motivated a transition toward data-driven methodologies. Rather than relying on simplified analytical leakage models, learning-based approaches aim to infer discriminative patterns directly from data, enabling automated handling of high-dimensional, noisy, and partially misaligned traces.

From a conceptual standpoint, classical profiling attacks can be viewed as early instances of supervised learning, where statistical models are trained to associate leakage patterns with sensitive values. However, their reliance on fixed leakage assumptions and handcrafted features constrains robustness and adaptability under realistic conditions.

The emergence of Machine Learning and Deep Learning techniques has fundamentally altered this landscape. Learning-based models can automatically extract relevant features, tolerate misalignment, and generalize across varying noise regimes. These capabilities have enabled more effective profiling and classification, but they also introduce new challenges related to generalization, robustness, and security in non-stationary environments.

This evolution marks a conceptual shift from hypothesis-driven analysis toward representation learning and data-centric inference. The following section builds upon this transition by reviewing state-of-the-art AI-driven side-channel profiling techniques, highlighting how modern learning models address the limitations of classical approaches while exposing new vulnerabilities that motivate adaptive and learning-based defense strategies.

2.3 AI-Based Side-Channel Profiling: From Static Classification to Adaptive Inference

The integration of Artificial Intelligence into side-channel analysis has marked a decisive transition from analytically driven attacks to data-centric inference frameworks. Unlike classical approaches, which rely on explicit leakage assumptions and handcrafted preprocessing pipelines, AI-based methods learn discriminative representations directly from measured signals. This shift has enabled side-channel analysis to scale to complex systems, tolerate noise and misalignment, and generalize across heterogeneous devices.

In the state of the art, AI-driven side-channel profiling is primarily framed as a supervised or unsupervised learning problem, where models are trained to associate physical emissions with internal system states, cryptographic secrets, or behavioral classes. While these approaches have significantly improved attack effectiveness, they also expose fundamental limitations of static learning pipelines when deployed in realistic, non-stationary environments. In most existing works, this formulation implicitly assumes a passive observation model, where the learner adapts to the signal but does not influence its generation.

2.3.1 Supervised Learning for Side-Channel Profiling

Supervised Machine Learning represents the most established AI paradigm in side-channel analysis. In profiling scenarios, labeled traces collected from a reference device are used to train classifiers capable of inferring sensitive information from unseen measurements. Early works demonstrated that traditional ML models—such as Support Vector Machines, Random Forests, and k -Nearest Neighbors—could outperform classical template attacks under noisy or imperfectly aligned conditions [25, 18].

The introduction of Deep Learning further expanded these capabilities. Convolutional Neural Networks (CNNs) have become the dominant architecture in AI-based side-channel attacks due to their ability to automatically learn hierarchical features from raw or minimally processed traces [26, 53]. CNN-based

models have shown remarkable robustness to desynchronization, masking countermeasures, and signal distortions, effectively reducing the reliance on domain-specific preprocessing.

From a conceptual standpoint, supervised profiling attacks reframe side-channel analysis as a pattern recognition problem, where leakage is treated as a structured signal embedding information about internal computation rather than as a noisy physical artifact. This abstraction allows models to exploit subtle correlations that may be inaccessible to analytical techniques, but it also introduces new dependencies on training data quality, representativeness, and stationarity.

2.3.2 Temporal Modeling and Sequential Inference

Side-channel signals are inherently temporal, reflecting the sequential execution of software and hardware operations. While early AI-based approaches treated traces as static vectors, recent work has emphasized the importance of temporal modeling. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) architectures have been applied to capture temporal dependencies in power and EM traces, enabling the classification of execution phases, workloads, and protocol behavior [50].

Temporal modeling is particularly relevant for behavioral profiling, where the objective extends beyond single-operation inference to the identification of sustained system activity. In such scenarios, the discriminative information emerges not from isolated samples, but from consistent temporal patterns observed across multiple windows. This observation reinforces the need for sequence-level reasoning and aggregation strategies, a theme that directly connects profiling accuracy to temporal decision hierarchies.

2.3.3 Unsupervised Learning and Anomaly Detection

Beyond supervised profiling, unsupervised and semi-supervised learning techniques have gained attention for side-channel-based monitoring and intrusion detection. Autoencoders, clustering algorithms, and density estimation models are trained on traces collected during nominal operation and used to detect deviations indicative of faults, malicious activity, or unauthorized software execution [51, 10].

These approaches are particularly attractive for security monitoring, as they do not require labeled attack data and can detect previously unseen behaviors. However, their effectiveness critically depends on the stability of the learned representation. Changes in noise conditions, workload characteristics, or environmental interference may cause benign deviations to be misclassified as anomalies, leading to high false-positive rates.

2.3.4 Noise, Generalization, and the Limits of Static Models

Despite their success, AI-based side-channel models inherit fundamental limitations from their static training paradigm. Most models are trained offline under specific acquisition conditions and implicitly assume that the statistical properties of leakage remain stationary over time. In realistic deployments, this assumption rarely holds.

Low-frequency EM measurements, in particular, are highly sensitive to environmental noise, probe placement, and concurrent system activity. Variations in power supply quality, temperature, or software workload can significantly alter the observable signal distribution. As a result, models trained under one condition may degrade rapidly when deployed under another, even in the absence of adversarial manipulation.

Several studies have attempted to mitigate these issues through data augmentation, transfer learning, or robust feature engineering. While these techniques improve generalization to some extent, they remain fundamentally static: once trained, the model cannot actively adapt its processing strategy in response to changing conditions.

2.3.5 Realistic Scenarios Motivating Adaptive Defense

The limitations of static AI-based profiling become particularly evident in realistic deployment scenarios. Consider public or semi-public infrastructures such as ticketing machines, ATM terminals, smart kiosks, or embedded controllers deployed in transportation hubs. In such environments, attackers may operate in proximity to devices using portable EM sensors, exploiting low-frequency emissions without direct physical contact.

Recent works on long-range and indirect EM leakage, including attacks exploiting unintended antenna effects or memory-induced emissions, demonstrate that physical proximity constraints are increasingly

relaxed [7, 49]. These scenarios highlight that side-channel leakage cannot be assumed to be either static or confined to controlled environments.

From a defensive perspective, such settings impose strict constraints. Hardware retrofitting is often infeasible, software modifications must preserve functionality, and environmental noise cannot be eliminated. Defense mechanisms must therefore operate adaptively, responding to evolving conditions while maintaining acceptable performance.

2.3.6 Toward Adaptive and Control-Oriented Learning

The convergence of these challenges motivates a fundamental rethinking of AI-based side-channel defense. Rather than treating preprocessing, feature extraction, and classification as fixed stages, recent research trends suggest modeling side-channel analysis as a closed-loop system. In this perspective, signal processing and representation choices become controllable parameters that influence both observability and security.

Reinforcement Learning (RL) provides a natural framework for this formulation. By modeling side-channel processing as a sequential decision-making problem, an RL agent can dynamically select denoising strategies, feature representations, or aggregation policies based on observed signal characteristics. This enables continuous adaptation to noise, workload variability, and potential adversarial behavior.

Importantly, the objective of such an approach is not to achieve information-theoretic security, but to actively reduce the exploitable information available to an adversary under realistic physical and operational constraints. By controlling the statistical structure of observable signals, adaptive defense mechanisms can degrade the effectiveness of profiling attacks while preserving legitimate system functionality.

Although RL has been extensively studied in control systems and adaptive signal processing, its application as a defensive mechanism in side-channel contexts remains largely unexplored. This gap in the literature motivates the adaptive framework proposed in this thesis, which builds upon the limitations identified in state-of-the-art AI-based profiling and extends them toward noise-aware, control-oriented defense strategies.

2.4 Noise, Environmental Variability, and the Limits of Static Side-Channel Models

A fundamental limitation of many side-channel analysis techniques proposed in the literature lies in the implicit assumption of stationarity. Classical attack and profiling models are typically developed and evaluated under controlled laboratory conditions, where environmental noise, workload variability, and measurement uncertainty are carefully minimized. While such assumptions facilitate repeatability and benchmarking, they rarely hold in realistic deployment scenarios, particularly for embedded and IoT systems operating in open or semi-controlled environments.

In practice, side-channel signals are affected by multiple sources of noise and variability. Environmental electromagnetic interference, power supply fluctuations, temperature changes, and concurrent system activity introduce stochastic perturbations that alter the observable leakage patterns over time. Even minor variations in probe placement, device orientation, or surrounding electronic equipment can significantly affect the spectral and temporal characteristics of measured EM or power traces. As a result, side-channel leakage in real-world settings is inherently non-stationary and context-dependent.

Several studies have shown that noise does not eliminate side-channel leakage, but rather changes its statistical structure [45]. Features that are highly discriminative under clean conditions may become unreliable or misleading when signal-to-noise ratio (SNR) degrades. Conversely, information that is not exploitable in isolation may emerge when observations are aggregated over longer time horizons. This behavior highlights a critical distinction between instantaneous signal quality and long-term information content.

Pipelines designed under static assumptions struggle to preserve performance when the statistical structure of leakage evolves over time. Once configured, these pipelines implicitly assume that the noise characteristics and leakage structure remain constant. In realistic environments, however, noise statistics may drift over time, rendering previously optimal preprocessing choices suboptimal or even counterproductive. This phenomenon is particularly evident in low-frequency EM analysis, where interference from external sources overlaps with device-specific spectral components.

Noise-aware techniques have been proposed to mitigate these issues, including adaptive filtering, wavelet-based denoising, and learning-based noise suppression [28, 19]. While these methods improve robustness under certain conditions, they are typically applied in an offline or static manner. Parameters are tuned during a training phase and remain fixed during deployment, limiting their ability to respond to unforeseen changes in the operating environment.

From a modeling perspective, environmental variability challenges the core assumptions underlying many profiling attacks. Models trained on traces collected under specific conditions may generalize poorly when deployed in different contexts, devices, or noise regimes. This limitation has been extensively documented in cross-device and cross-environment evaluations, where classification accuracy degrades sharply when training and testing conditions differ [33]. Such observations underscore the fragility of static models in adversarial or uncontrolled settings.

Beyond attack scenarios, the same limitations apply to defensive applications such as side-channel-based monitoring and intrusion detection. A physical-layer IDS trained on nominal behavior under clean conditions may suffer from elevated false-positive or false-negative rates when noise characteristics evolve. In these cases, the inability to distinguish between benign environmental changes and genuine anomalies undermines the reliability of static detection systems.

These considerations motivate a shift in perspective: noise should not be treated merely as an external nuisance to be filtered out, but as a dynamic component of the side-channel ecosystem that must be explicitly modeled and managed. Effective side-channel analysis in realistic environments requires mechanisms capable of adapting preprocessing, feature representation, and decision strategies in response to observed signal conditions.

This observation naturally leads to the formulation of side-channel analysis as a sequential and adaptive problem rather than a static inference task. Instead of optimizing a fixed pipeline offline, the system must continuously assess the quality and structure of incoming signals and adjust its processing strategy accordingly. Such adaptability is essential not only to maintain performance under changing noise conditions, but also to respond to adversarial behaviors that deliberately attempt to exploit or manipulate side-channel characteristics.

In this context, adaptive learning paradigms—and reinforcement learning in particular—offer a principled framework to address the limitations of static side-channel models. By framing noise mitigation and information control as a closed-loop optimization problem, adaptive systems can dynamically balance observability, robustness, and security objectives over time. This perspective provides a critical conceptual bridge between traditional side-channel profiling and the adaptive defense mechanisms explored in the subsequent sections and chapters of this thesis.

2.5 Adaptive and Learning-Based Countermeasures in the Literature

The increasing effectiveness of machine learning-based side-channel attacks has naturally motivated the exploration of adaptive and learning-driven countermeasures. Unlike traditional static defenses, these approaches attempt to respond dynamically to evolving leakage patterns, noise conditions, or attacker capabilities. This section reviews the main directions explored in the literature, highlighting both their contributions and their limitations, and positioning the need for a more general adaptive control framework.

2.5.1 Adaptive Preprocessing and Noise Mitigation

Several works have investigated adaptive preprocessing techniques to improve robustness against noise and environmental variability. These approaches typically focus on dynamically adjusting filtering parameters, window sizes, or denoising thresholds based on observed signal characteristics. Wavelet-based denoising schemes with adaptive thresholding have been proposed to preserve informative components under varying noise levels, while smoothing techniques such as Savitzky–Golay filters have been tuned to balance noise suppression and signal distortion [28, 43].

More recently, learning-based denoising methods have been introduced, where neural networks are trained to suppress noise or enhance leakage-related components [19]. Although these methods improve performance under specific conditions, they are typically trained offline and applied statically at runtime. As a result, their adaptability is limited to the noise distributions observed during training, reducing effectiveness when operating conditions drift or adversarial interference is present.

2.5.2 Learning-Based Obfuscation and Defensive Machine Learning

Another emerging direction focuses on actively degrading the exploitable information available to an attacker rather than improving signal quality. Defensive machine learning techniques introduce controlled perturbations or obfuscation mechanisms that target the statistical structure exploited by side-channel attacks. Notably, adversarial obfuscation strategies have been proposed to reduce the effectiveness of ML-based architectural and microarchitectural side-channel attacks by injecting carefully crafted noise or modifying execution patterns [29].

These approaches demonstrate that it is possible to interfere with attacker inference models without eliminating physical leakage altogether. However, most existing solutions rely on fixed perturbation strategies or heuristics determined during design time. Once deployed, the defense behavior remains static, allowing attackers to adapt by retraining or adjusting their models. Furthermore, these techniques often focus on specific leakage channels or attack models, limiting their generality.

2.5.3 Adaptive Monitoring and Physical-Layer Intrusion Detection

Adaptive learning has also been explored in the context of physical-layer monitoring and intrusion detection. Side-channel-based IDS solutions employ ML or DL models to detect deviations from nominal behavior by analyzing power or EM traces [11, 17]. Some systems incorporate incremental learning or periodic retraining to account for workload changes and long-term drift.

While such approaches introduce a degree of adaptivity, they typically treat adaptation as an offline or semi-online process, rather than as a continuous decision-making problem. The monitoring system reacts to detected changes but does not actively control the observable leakage or preprocessing strategy. As a result, these systems remain reactive rather than proactive and do not explicitly optimize the trade-off between observability, robustness, and information exposure.

2.5.4 Limitations of Existing Adaptive Approaches

Across these research directions, a common limitation emerges: adaptivity is often local, heuristic, or confined to a single stage of the side-channel processing pipeline. Preprocessing parameters, feature representations, and decision strategies are rarely optimized jointly. Moreover, adaptation is typically driven by predefined rules or periodic retraining rather than by continuous interaction with the environment.

Most importantly, existing approaches do not explicitly model side-channel defense as a sequential control problem. Decisions are made independently at each stage, without considering their long-term impact on information leakage or adversarial inference. This fragmented view limits the ability of current methods to cope with non-stationary noise, evolving workloads, and adaptive attackers simultaneously.

2.5.5 Positioning Reinforcement Learning as a Unifying Framework

These observations suggest that side-channel analysis and defense should be reformulated not as isolated inference problems, but as sequential processes in which decisions taken at one stage influence future observability. The gaps identified in the literature suggest the need for a more principled and unified adaptive defense strategy. Reinforcement Learning (RL) provides a natural framework to address these challenges by modeling side-channel processing as a closed-loop interaction between a defender and its environment. In this formulation, the defender observes side-channel characteristics, selects actions affecting preprocessing, feature extraction, or obfuscation, and receives feedback based on performance objectives.

Unlike heuristic adaptation, RL enables optimization over long-term objectives and explicit trade-offs between competing goals, such as robustness to noise, preservation of legitimate observability, and reduction of exploitable information. Importantly, RL does not require an explicit model of the attacker or the leakage process, making it well suited to complex and partially observable environments.

While reinforcement learning has been widely applied in control, networking, and cybersecurity domains, its use as a feature-aware, noise-adaptive control mechanism for side-channel defense remains largely unexplored. This observation positions the approach developed in this thesis as a natural and necessary extension of existing adaptive countermeasures, bridging the gap between static profiling, heuristic adaptation, and fully autonomous defense strategies.

2.6 Realistic Threat Scenarios Motivating Adaptive Side-Channel Defense

While a significant portion of side-channel research has historically focused on laboratory-grade attacks and controlled measurement setups, recent advances demonstrate that physical leakage can be exploited in increasingly realistic and unconstrained environments. In such scenarios, the assumptions underlying static profiling and fixed countermeasures no longer hold. This section reviews representative real-world threat scenarios that motivate the need for adaptive, noise-aware, and learning-driven side-channel defense mechanisms.

2.6.1 Public and Semi-Public Embedded Systems

Many embedded systems operate in public or semi-public environments where physical proximity to the device is feasible without requiring privileged access. Examples include ticketing machines, self-service kiosks, access control terminals, industrial HMIs, and embedded controllers deployed in transportation infrastructures.

In these contexts, an attacker may deploy low-cost electromagnetic probes or antennas in close proximity to the target device, exploiting low-frequency EM emissions that propagate beyond device enclosures. Unlike laboratory attacks, such scenarios are characterized by:

- Uncontrolled and time-varying environmental noise,
- Limited measurement duration and imperfect synchronization,
- Dynamic workloads driven by user interaction.

Despite these constraints, prior work has shown that side-channel leakage remains exploitable when analyzed using robust feature extraction and machine learning techniques [46, 38]. Importantly, the non-stationary nature of these environments renders static preprocessing and fixed defense strategies insufficient, as leakage characteristics evolve continuously over time.

2.6.2 Smart City and Urban Infrastructures

Smart city deployments represent a particularly relevant threat model for modern side-channel analysis. Distributed embedded devices—such as traffic controllers, smart meters, public sensing nodes, and communication gateways—are often deployed in open environments with minimal physical protection.

In urban settings, an attacker may leverage ambient infrastructure (e.g., street furniture, nearby buildings, or vehicles) to host passive sensors capable of capturing EM emissions over extended periods. The presence of pervasive electromagnetic noise from power lines, wireless communications, and transportation systems creates a highly non-stationary background that challenges traditional analysis techniques.

At the same time, this noise-rich environment motivates the use of learning-based inference capable of extracting stable patterns from noisy observations. From a defensive standpoint, it also highlights the necessity of adaptive processing pipelines that can dynamically adjust feature representations and denoising strategies to preserve legitimate observability while limiting exploitable leakage.

2.6.3 Financial and Transaction-Oriented Devices

Automated Teller Machines (ATMs), payment terminals, and ticketing systems process sensitive financial and personal data while operating in publicly accessible locations. These devices often rely on embedded platforms executing cryptographic routines and communication protocols under variable workloads.

Side-channel leakage in such systems poses a dual risk: direct extraction of cryptographic material and indirect inference of transaction behavior or user interaction patterns. Even when direct key recovery is infeasible, behavioral profiling may reveal sensitive metadata, such as transaction frequency, authentication attempts, or protocol usage.

Given the operational constraints of financial devices, intrusive hardware countermeasures are rarely viable post-deployment. This makes software-based and informational defenses particularly attractive. However, static noise injection or masking techniques risk degrading performance or usability, motivating adaptive strategies capable of responding to observed leakage conditions in real time.

2.6.4 Remote and Air-Gapped Scenarios

Recent attacks have demonstrated that side-channel leakage can be exploited even in remote or air-gapped systems. The RAMBO attack exemplifies this evolution by showing that DRAM activity can be modulated to emit EM signals detectable at a distance, enabling data exfiltration from isolated systems [49].

Such attacks operate under extreme noise conditions and rely on sophisticated signal processing and machine learning to decode weak and indirect leakage. From a defensive perspective, these scenarios underscore that eliminating leakage entirely is often infeasible. Instead, the goal shifts toward limiting the reliability and consistency of exploitable features under adversarial observation.

Adaptive defense mechanisms are particularly relevant in this context, as leakage patterns may depend on runtime behavior, memory access patterns, and system load. A static defense tuned for one operating condition may fail under another, while an adaptive controller can respond dynamically to evolving leakage characteristics.

2.6.5 Implications for Adaptive Defense Design

Across these scenarios, a common theme emerges: side-channel leakage occurs in environments that are noisy, dynamic, and partially observable, where neither the attacker nor the defender has access to a complete or stable view of the underlying system state. Attackers exploit temporal aggregation, learning-based inference, and long-term observation to overcome these challenges. Defenders must therefore reason under similar assumptions. In this sense, both attackers and defenders are subject to the same physical constraints and must operate over noisy, incomplete, and temporally evolving observations.

Static countermeasures assume stationary leakage models and predictable attack strategies, assumptions that do not hold in realistic deployments. In contrast, adaptive defense mechanisms can observe leakage characteristics at runtime and adjust processing strategies accordingly. This includes modifying denoising parameters, feature selection policies, and temporal aggregation strategies in response to changing conditions.

These considerations motivate the formulation of side-channel defense as a closed-loop control problem, where decisions are made sequentially based on observed signal behavior and long-term objectives. Reinforcement learning naturally fits this paradigm, providing a principled framework for optimizing adaptive defense strategies without requiring explicit attacker models.

The scenarios discussed in this section therefore provide both practical motivation and conceptual grounding for exploring reinforcement learning-based adaptive frameworks in subsequent chapters.

2.7 Discussion and Open Challenges

The analysis of the state of the art highlights both the remarkable progress achieved in AI-based side-channel analysis and the structural limitations that remain unresolved. While modern Machine Learning and Deep Learning techniques have significantly enhanced the ability to extract information from noisy and high-dimensional physical signals, their effectiveness is often assessed under assumptions that do not fully reflect realistic deployment conditions. This section consolidates the main open challenges emerging from the literature and clarifies the gaps that motivate the methodological choices adopted in this thesis.

2.7.1 Generalization and Robustness Across Conditions

A recurring limitation of existing side-channel approaches is their sensitivity to changes in operating conditions. Many profiling and classification models are trained and evaluated under fixed assumptions regarding device configuration, workload, environmental noise, and measurement setup. When these assumptions are violated, model performance can degrade substantially.

Cross-device generalization, variation in signal-to-noise ratio, temporal drift, and environmental interference remain open challenges, particularly for embedded and IoT systems deployed in uncontrolled environments. While data augmentation and transfer learning have been proposed as partial remedies, they do not fundamentally address the non-stationary nature of real-world side-channel leakage. These observations suggest that robustness cannot be treated as a static property of a trained model, but rather as a dynamic requirement that must be sustained throughout system operation.

2.7.2 Static Pipelines in a Dynamic Leakage Landscape

Most existing side-channel analysis pipelines rely on static preprocessing, fixed feature extraction strategies, and frozen model parameters. This design implicitly assumes that the statistical structure of leakage remains stable over time. However, the literature increasingly shows that leakage characteristics evolve due to workload variability, system updates, power supply fluctuations, and environmental dynamics.

Fixed processing pipelines are therefore inherently limited: once optimized for a specific condition, they may become suboptimal or even misleading under new operating regimes. This limitation applies to both offensive and defensive applications, including behavioral profiling and intrusion detection.

2.7.3 Noise as a First-Class Design Constraint

Although noise is widely acknowledged in side-channel research, it is often treated as an external nuisance rather than a central design parameter. Many studies focus on improving denoising or robustness post hoc, without explicitly modeling noise variability as part of the inference process.

In realistic deployments, noise is not only unavoidable but also non-stationary and context-dependent. Effective side-channel analysis and defense therefore require mechanisms capable of adapting to changing noise conditions, rather than assuming a fixed signal-to-noise regime. This observation motivates the integration of noise-aware strategies directly into the learning and decision-making pipeline.

2.7.4 Temporal Reasoning Beyond Window-Level Classification

Another open challenge concerns the temporal structure of side-channel inference. While window-based representations are widely adopted, many studies optimize models at the window level and rely on simple aggregation heuristics for sequence-level decisions.

However, the literature suggests that reliable inference often emerges from consistency over time rather than from isolated observations. Capturing temporal dependencies, confidence evolution, and decision stability remains an open research area, particularly for long-running systems where behavior unfolds over extended time horizons.

2.7.5 Defensive Learning and the Limits of Static Countermeasures

Recent work on defensive machine learning and signal obfuscation demonstrates that it is possible to reduce the effectiveness of ML-based side-channel attacks by manipulating observable features. Nevertheless, most proposed defenses are static or heuristic, assuming a fixed attacker model and stationary leakage.

In adversarial and adaptive settings, such assumptions are unlikely to hold. Attackers may refine their models, adjust measurement strategies, or exploit new leakage dimensions over time. Static defenses risk becoming predictable and ineffective, especially against learning-based adversaries.

2.7.6 Toward Closed-Loop and Adaptive Side-Channel Defense

The challenges identified above converge toward a common insight: side-channel analysis and defense should be treated as dynamic, sequential, and adaptive processes rather than static classification problems.

Framing side-channel processing as a closed-loop system enables continuous observation, evaluation, and adjustment of preprocessing and representation strategies. Reinforcement Learning offers a principled framework for this paradigm, allowing a defender to optimize long-term objectives under uncertainty and non-stationarity without requiring explicit attacker models.

Importantly, the objective of such adaptive defense is not to achieve information-theoretic security, which is often unattainable in practice, but to actively reduce the exploitable information available to an adversary under realistic physical, computational, and operational constraints.

2.7.7 Positioning of This Thesis

In light of the reviewed literature and identified gaps, this thesis positions itself at the intersection of side-channel analysis, machine learning, and adaptive control. Rather than proposing yet another static profiling technique, the work aims to:

- Address noise and non-stationarity as fundamental properties of side-channel leakage,

- Move beyond fixed preprocessing and feature selection strategies,
- Explore adaptive, learning-driven mechanisms for controlling the information exposed through physical channels.

The following chapter builds upon these observations by introducing the proposed methodological framework, detailing how side-channel processing can be formulated as a sequential decision-making problem and how reinforcement learning can be leveraged to implement adaptive, noise-aware defense strategies in realistic embedded and IoT environments.

Chapter 3

Experimental Setup and Methodological Framework

This chapter presents the complete experimental setup and methodological framework adopted to analyze electromagnetic side-channel emissions generated by embedded systems during application-level execution. The objective is twofold: first, to define a rigorous and reproducible acquisition and preprocessing pipeline for low-frequency electromagnetic signals; second, to establish a principled foundation for evaluating learning-based classification and robustness under realistic noise conditions.

The experimental methodology builds upon and consolidates two original research contributions. The first work [38] focuses on the feasibility of profiling application-level behavior in connected and embedded devices through electromagnetic side-channel analysis and machine learning techniques. The second contribution [39] extends this perspective toward behavioral profiling and scenario discrimination in realistic IoT contexts, emphasizing robustness, generalization, and practical deployment constraints. This chapter unifies and extends these contributions into a single coherent experimental framework.

In contrast to purely laboratory-oriented studies, the proposed setup is explicitly designed to reflect realistic operating conditions typical of embedded and IoT devices, where environmental noise, workload variability, and limited observability are intrinsic characteristics rather than exceptional cases. As such, the methodology does not assume idealized acquisition conditions, but instead embraces noise and variability as fundamental aspects of the side-channel measurement process.

The chapter is structured to progressively bridge raw physical measurements and learning-based inference. It begins with a detailed description of the hardware platform and acquisition environment, including sensing equipment, probing strategy, and measurement constraints. Subsequently, the signal acquisition and preprocessing pipeline is introduced, with particular emphasis on temporal segmentation and energy-based aggregation. These design choices are motivated by the need to stabilize highly noisy side-channel signals while preserving discriminative information relevant to application-level behavior.

A central aspect of the methodology concerns the transformation of raw electromagnetic and current traces into compact and informative representations suitable for machine learning. To this end, the chapter details the adopted preprocessing steps, including window-based segmentation, Root Mean Square (RMS) computation, and robust normalization. These operations are not treated as purely technical conveniences, but as methodological components that directly influence noise resilience, temporal coherence, and generalization performance.

Building upon the preprocessed signals, the chapter then describes the feature engineering strategy, combining time-domain and frequency-domain descriptors to capture complementary aspects of side-channel leakage. Particular attention is devoted to feature selection, where multiple strategies are evaluated in terms of classification accuracy, interpretability, and inference-time feasibility. This analysis highlights the trade-offs between scenario-specific optimization and globally applicable feature subsets, ultimately motivating the adoption of a reduced and generalizable feature set.

Throughout the chapter, experimental design choices are explicitly linked to practical deployment considerations, such as computational efficiency, robustness to outliers, and suitability for real-time inference on resource-constrained platforms. Rather than presenting isolated processing stages, the methodology is formulated as a modular yet integrated pipeline, enabling systematic evaluation and future extension.

By consolidating acquisition, preprocessing, feature engineering, and selection within a single experimental framework, this chapter establishes the empirical backbone of the thesis. It provides the necessary grounding for the subsequent analysis of classification performance and robustness, and prepares the con-

ceptual transition from static profiling toward adaptive and noise-aware learning strategies explored in the later chapters.

3.1 Hardware Platform and Acquisition Environment

The experimental measurements were conducted on a low-cost embedded computing platform, namely a Raspberry Pi 4 Model B, selected as a representative example of a widely deployed IoT and edge-computing device. The platform runs a standard Linux-based operating system and executes a set of real-world application scenarios designed to emulate realistic user and network activity. This choice reflects the objective of analyzing side-channel emissions under practical operating conditions rather than synthetic or highly controlled benchmarks.

Electromagnetic side-channel signals were acquired using an ETS Lindgren 7406-901 magnetic field probe, covering a frequency range from 1 kHz to 3 GHz. The probe was positioned in close proximity to the device under test, following a non-invasive measurement approach that does not require any physical modification of the hardware. The sensor output was connected to a TiePie HS6 oscilloscope, operating with an effective bandwidth of 250 MHz and configured to sample signals at 1 MHz. This sampling rate was selected as a compromise between temporal resolution and data volume, enabling the capture of low-frequency electromagnetic components associated with application-level activity while maintaining manageable storage and processing requirements.

All acquisitions were performed in a shielded laboratory environment to reduce the impact of external electromagnetic interference. While complete isolation from noise is neither feasible nor desirable for this study, shielding allows background interference to remain bounded and repeatable across experiments. A dedicated desktop workstation was used to control the acquisition process, store raw traces, and execute preprocessing and feature extraction routines.

A schematic overview of the measurement setup is shown in Figure 3.1. The configuration includes the device under test, the magnetic field probe, the oscilloscope, and the acquisition and control PC. This modular arrangement ensures reproducibility and facilitates controlled variation of acquisition parameters.



Figure 3.1: Experimental setup for electromagnetic side-channel measurement. (1) Device Under Test (Raspberry Pi 4 Model B); (2) Magnetic Field Probe (ETS Lindgren 7406-901); (3) Oscilloscope (TiePie HS6); (4) Acquisition and control PC.

For each of the eight application scenarios considered in this study, 30 independent acquisition trials were performed. Each trial had a duration of 40 seconds and produced approximately 40 million samples

per signal channel. During the first 10 seconds of each acquisition, no application-level activity was executed. This initial interval captures background electromagnetic noise arising from the idle system state, power supply fluctuations, and residual environmental interference.

The background segment was intentionally recorded and subsequently discarded during preprocessing. This design choice serves two purposes. First, it allows a clear separation between idle and active device behavior, preventing non-discriminative baseline noise from biasing the learning process. Second, it provides an explicit reference for evaluating the impact of application execution on electromagnetic emissions. By excluding the idle segment from the classification pipeline, the analysis focuses exclusively on signal components induced by actual workload execution.

Representative examples of the raw current and electromagnetic signals acquired at 1 MHz sampling frequency are shown in Figures 3.2 and 3.3. The figures illustrate the full 40 second acquisition window for all scenarios. In particular, the transition from background noise to application-induced electromagnetic activity is clearly visible after the initial 10 second interval, highlighting the presence of measurable and scenario-dependent side-channel leakage.

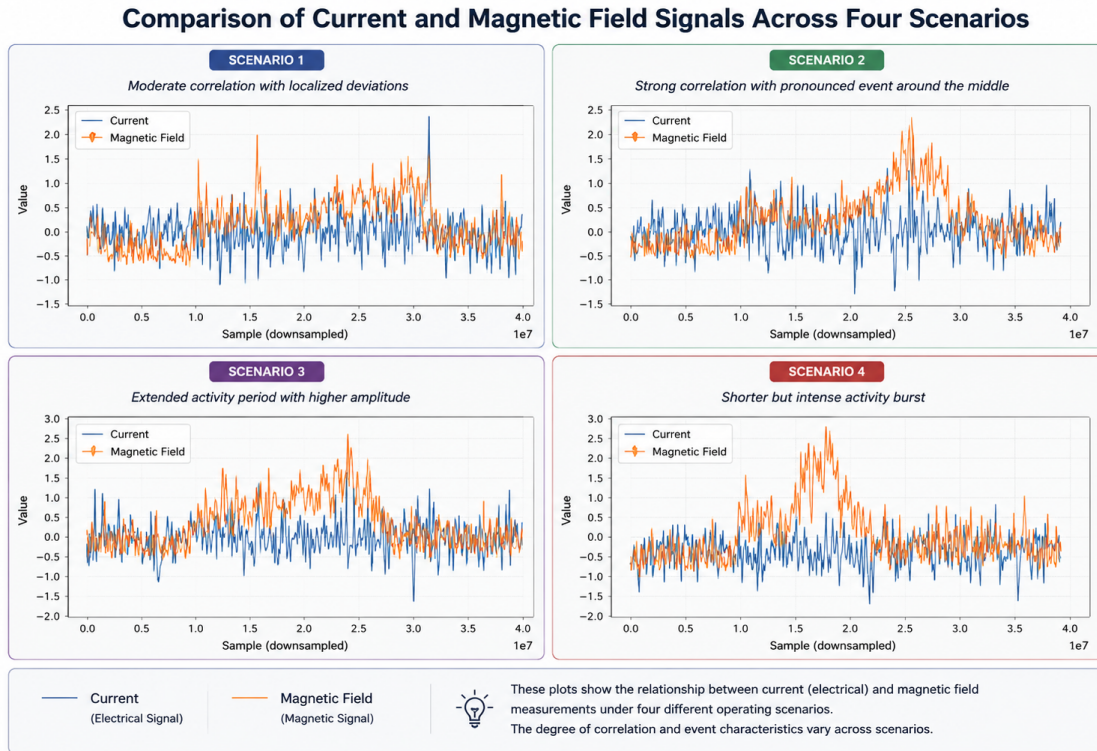


Figure 3.2: Raw current and magnetic field signals acquired at 1 MHz sampling frequency for Scenarios 1 to 4. Each acquisition lasts 40 seconds. The first 10 seconds correspond to background noise and are excluded from subsequent analysis.

Overall, the acquisition environment is deliberately designed to balance experimental control and realism. While shielding and repeatable instrumentation ensure consistency across trials, the use of real applications, non-invasive sensing, and low-frequency electromagnetic measurements preserves the intrinsic variability and noise characteristics of practical embedded systems. This balance is essential to support a meaningful evaluation of learning-based side-channel profiling techniques under conditions that closely resemble real-world deployments.

3.2 Signal Acquisition and Preprocessing Pipeline

The acquisition of electromagnetic and current signals produces high-dimensional, noisy time-series data whose raw form is not directly suitable for learning-based inference. This section describes the preprocessing pipeline adopted to transform raw measurements into stable, compact, and informative representations, explicitly addressing temporal variability, noise resilience, and computational efficiency.

Rather than treating preprocessing as a purely technical step, the pipeline is designed as a methodological component that directly impacts classification robustness and generalization. Each stage is motivated

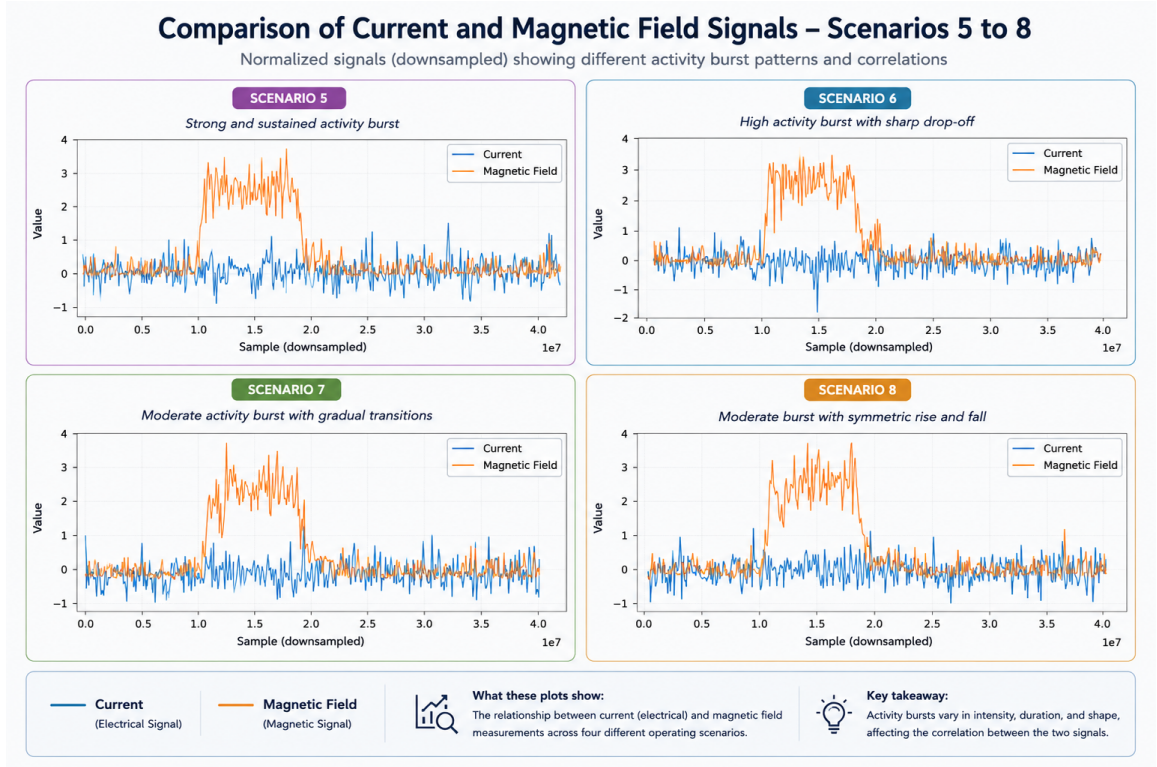


Figure 3.3: Raw current and magnetic field signals for Scenarios 5 to 8. The observable increase in magnetic activity reflects emissions induced by the execution of application-level workloads. As in the previous scenarios, the first 10 seconds represent background noise and are removed prior to preprocessing.

by both signal characteristics and deployment-oriented constraints.

3.2.1 Removal of Background Noise Segment

Each acquisition trial includes an initial 10-second interval during which no application-level activity is executed. This segment captures background electromagnetic emissions originating from the idle system state, power supply fluctuations, and residual environmental interference.

As illustrated in Figures 3.4 and 3.5, the first portion of each acquisition is characterized by relatively stationary signal behavior, corresponding to background noise conditions. After this initial interval, the execution of the application scenarios induces a clear and sustained change in the structure and amplitude of both current and electromagnetic field signals.

To avoid contaminating the learning process with non-discriminative baseline information, the initial 10-second background segment is systematically removed prior to any further processing. This design choice ensures that subsequent analysis focuses exclusively on signal components generated by application execution, preventing bias toward idle-state characteristics that are common across all scenarios.

The explicit separation between background and active execution phases also provides a clear temporal reference for evaluating workload-induced electromagnetic activity, supporting a more consistent and interpretable preprocessing pipeline.

3.2.2 Temporal Segmentation and Overlapping Windows

After background removal, the signals are segmented using a fixed-size sliding window of 100 ms with 50% overlap, as illustrated in Figure 3.6. This windowing strategy serves multiple methodological purposes within the proposed side-channel analysis pipeline.

First, it transforms long continuous traces into a sequence of shorter, partially overlapping segments, effectively increasing the number of training samples while preserving temporal coherence across consecutive windows. This representation enables learning-based models to operate on localized signal observations while retaining information about temporal evolution.

Second, the selected window length reflects a trade-off between temporal resolution and statistical stability. Shorter windows tend to be dominated by measurement noise and transient fluctuations, whereas

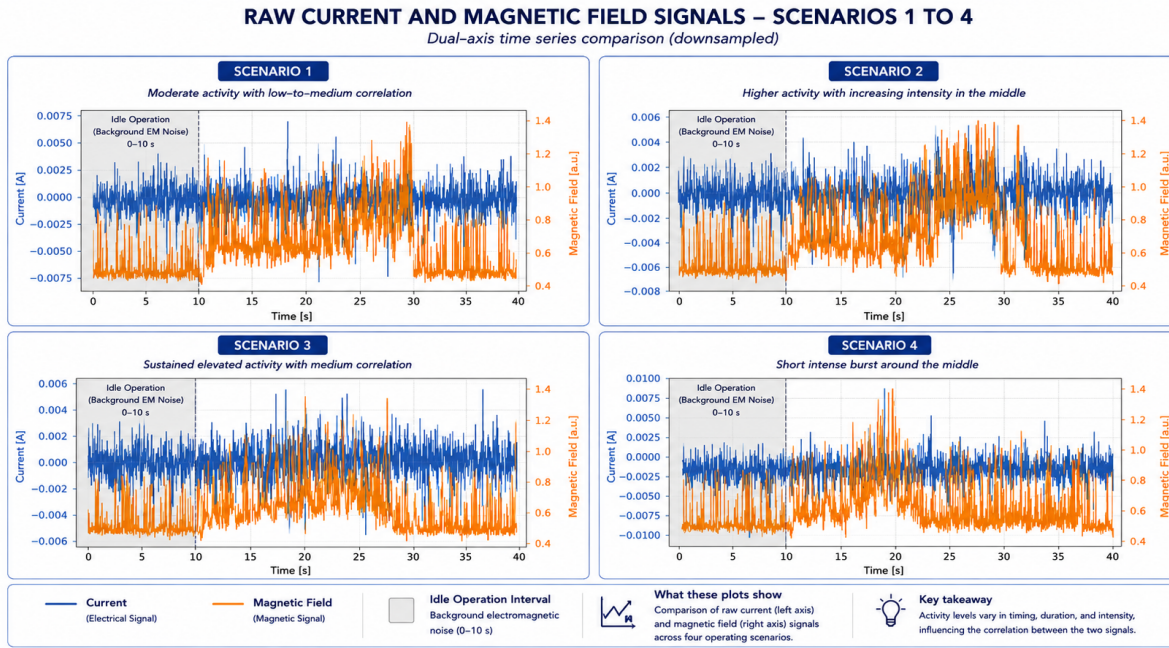


Figure 3.4: Raw current and magnetic field signals for Scenarios 1 to 4. The first 10 seconds correspond to background electromagnetic noise acquired during idle system operation, while the subsequent interval reflects application-induced activity.

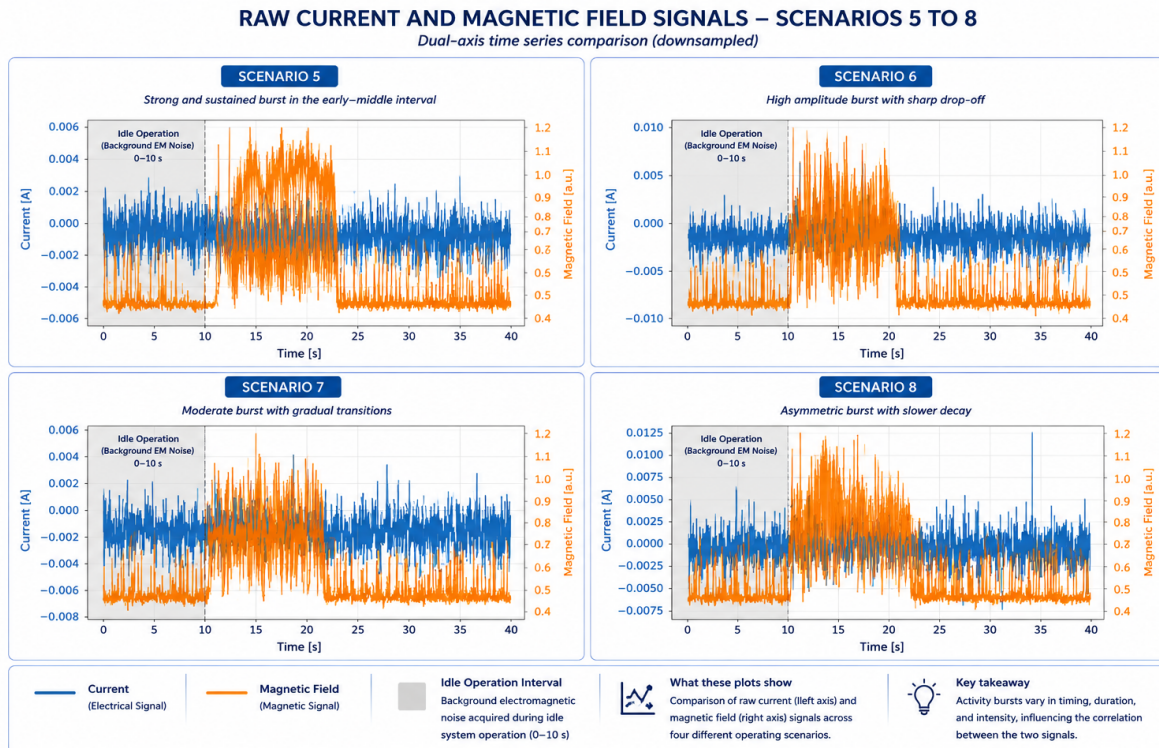


Figure 3.5: Raw current and magnetic field signals for Scenarios 5 to 8. As in the previous scenarios, the initial 10 seconds capture background noise, followed by a marked increase in electromagnetic activity associated with application execution.

longer windows reduce responsiveness and may obscure short-term variations induced by application-level activity. The adopted windowing configuration, defined by the combination of a 100 ms window length and 50% overlap, was selected based on its impact on window-level classification performance, providing improved accuracy compared to shorter window configurations while preserving sensitivity to workload-induced signal variations.

The use of overlapping windows further mitigates boundary effects and reduces information loss at window edges, ensuring that short-lived or non-stationary signal variations are not missed. This aspect is particularly relevant in realistic side-channel settings, where electromagnetic emissions exhibit temporal variability and non-stationarity. Importantly, the windowed representation defines the fundamental unit of analysis for subsequent preprocessing, feature extraction, and learning stages. All classification and robustness evaluations are performed at the window level and, where appropriate, aggregated over time to obtain sequence-level decisions.

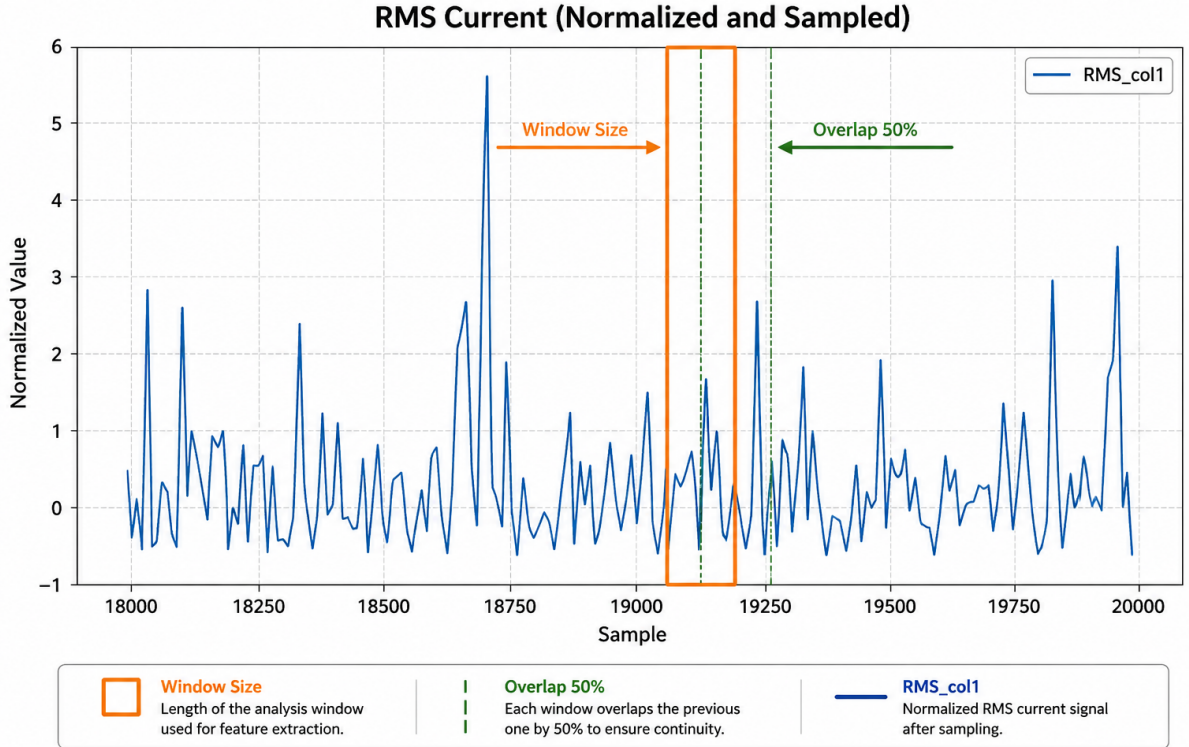


Figure 3.6: RMS feature extraction using 100 ms windows with 50% overlap. RMS values are computed independently for current and magnetic field signals.

From an evaluation perspective, window-level analysis treats each temporal window as an independent sample that is classified individually. Window-level accuracy therefore measures the ability of the model to correctly recognize the operating scenario based on a localized and temporally constrained observation of the side-channel signal. This form of evaluation is inherently sensitive to local fluctuations, measurement noise, and transient effects, which may significantly alter signal characteristics within individual windows, particularly when short window durations are adopted.

At the same time, the adoption of a window-based representation enables the integration of information across multiple consecutive samples, allowing statistical, energetic, and structural signal patterns to emerge that would not be observable at the level of individual time samples. In practical inference scenarios, a side-channel monitoring system does not rely on isolated instantaneous measurements, but instead observes a continuous temporal stream from which aggregated features are extracted to represent system behavior in a more stable and robust manner.

Empirical results obtained using RMS-based features highlight the impact of window length on window-level classification performance. When using a window duration of 10 ms, the observed window-level accuracy is approximately 51.8%. Increasing the window length to 100 ms yields a measurable improvement, with window-level accuracy rising to approximately 59.9%, while sequence-level accuracy remains unchanged. This behavior is consistent with the nature of side-channel signals: very short

windows primarily capture local fluctuations that are strongly influenced by noise and transient effects, whereas longer windows integrate a larger portion of the temporal behavior of the system, producing more stable and representative RMS features.

As a consequence, increasing the window duration reduces intra-class variability and improves class separability at the window level. These observations reinforce a fundamental characteristic of side-channel analysis: discriminative information is not necessarily concentrated within individual windows, but rather emerges through temporal accumulation. While temporal aggregation mechanisms can compensate for local misclassifications, the relatively low window-level accuracy indicates that RMS features alone are insufficient to fully capture the discriminative structure of the signal.

Specifically, this analysis suggests that (i) RMS features provide only a partial description of application-induced electromagnetic activity, (ii) individual windows contain useful but partially overlapping information across scenarios, and (iii) classifiers are required to operate on features characterized by heterogeneous dynamics and scales. These considerations motivate the extension of the preprocessing pipeline beyond RMS aggregation, including the introduction of robust normalization strategies and the extraction of additional features in both the time and frequency domains.

3.2.3 Energy-Based Aggregation via RMS

For each windowed segment, the Root Mean Square (RMS) value is computed independently for the current and electromagnetic signals. RMS aggregation plays a central role in the proposed preprocessing pipeline by transforming highly variable raw measurements into a more stable and interpretable representation.

From a signal processing perspective, RMS provides a robust estimate of signal energy, effectively attenuating the impact of instantaneous spikes and high-frequency noise components. This energy-based aggregation integrates information over each temporal window, reducing sensitivity to local fluctuations while preserving variations induced by application-level activity.

From a learning perspective, RMS aggregation drastically reduces the dimensionality of the input data, mapping each windowed segment to a compact set of scalar descriptors while retaining information correlated with workload intensity and execution behavior. This transformation enables learning-based models to focus on meaningful temporal trends rather than raw sample-level variability.

The effect of RMS aggregation on signal interpretability is illustrated in Figure 3.7. Compared to the raw time-domain signals, the RMS time series exhibits smoother temporal dynamics and clearer transitions between different execution phases, highlighting workload-induced variations that are otherwise obscured by noise in the raw measurements.

Experimental results, discussed in detail in the robustness analysis, confirm that RMS-based representations consistently outperform raw-signal approaches under noisy conditions. In particular, RMS aggregation stabilizes feature distributions across windows and improves classification performance across varying signal-to-noise ratios. These observations motivate the adoption of RMS not merely as a dimensionality reduction technique, but as a noise-aware representation explicitly aligned with realistic deployment conditions.

3.2.4 Normalization and Robust Scaling

Following RMS aggregation, feature normalization is required to ensure consistent scaling across windows, trials, and scenarios. As highlighted by the RMS time-series analysis, a marked difference in scale exists between the two sensing channels. In particular, RMS values derived from the current signal are several orders of magnitude smaller than those associated with the electromagnetic channel. This imbalance renders features not directly comparable and may negatively affect learning-based models, which tend to prioritize variables characterized by larger variance or amplitude.

This observation motivates the introduction and systematic evaluation of feature normalization strategies prior to classification. Appropriate normalization reduces the impact of scale disparities while preserving discriminative information embedded in the temporal dynamics of the signals. In the context of side-channel measurements—characterized by noise, outliers, and non-Gaussian distributions—the choice of normalization strategy plays a critical role and requires dedicated experimental assessment.

In this study, several normalization techniques were evaluated, both with and without RMS preprocessing, including *Standard Scaler*, *MinMax Scaler*, *MaxAbs Scaler*, *Robust Scaler*, *Normalizer*, *Quantile Transformer*, and *PowerTransformer* (Yeo-Johnson).

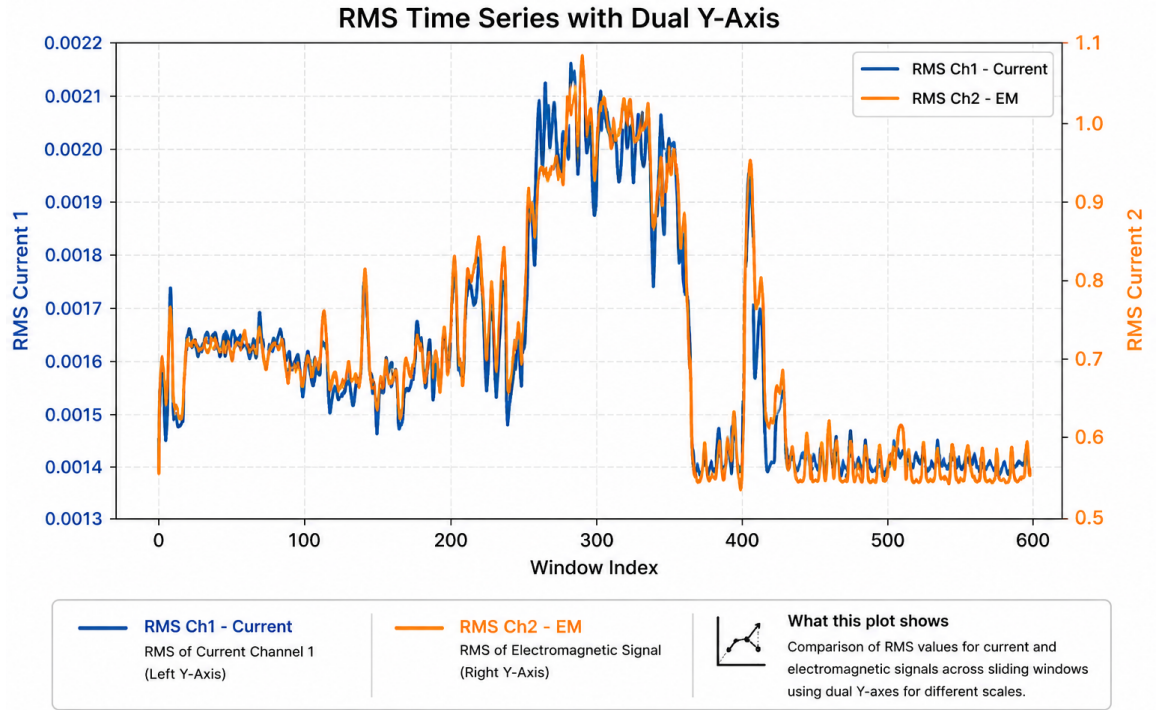


Figure 3.7: RMS time series for current and electromagnetic signals computed over 100 ms windows with 50% overlap. Compared to raw signals, RMS aggregation produces smoother and more interpretable temporal patterns, emphasizing workload-induced variations while reducing high-frequency noise.

Figure 3.8 reports the classification performance obtained using a CNN+LSTM architecture with a 70–30 train/test split. The comparison highlights that RMS preprocessing significantly improves robustness across most normalization strategies. Among all tested configurations, the *Robust Scaler* consistently achieves superior and more stable performance when combined with RMS aggregation.

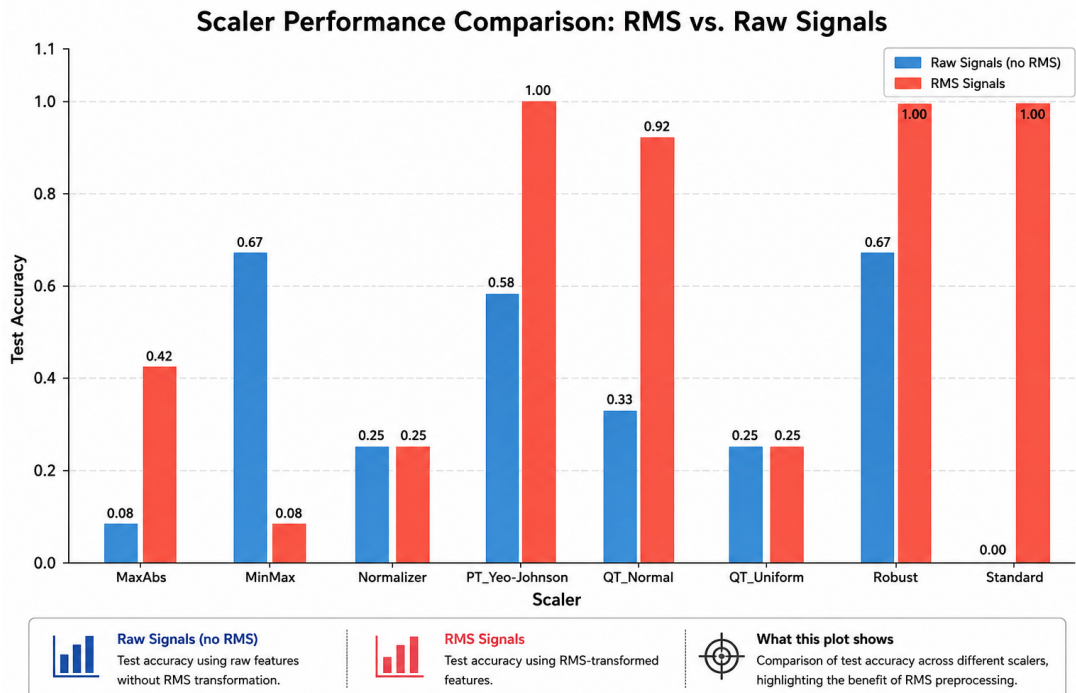


Figure 3.8: Comparison of classification accuracy for different normalization techniques, with and without RMS preprocessing. Robust Scaler achieves the best performance when combined with RMS aggregation.

The Robust Scaler normalizes features by removing the median and scaling according to the interquartile range (IQR), thereby limiting the influence of outliers and sporadic peaks that frequently occur in side-channel measurements. This property is particularly relevant in electromagnetic analysis, where environmental interference and transient disturbances are unavoidable.

For completeness, the normalization techniques considered in this work are briefly summarized below.

Standard Scaler Applies z-score normalization by subtracting the mean and dividing by the standard deviation:

$$x' = \frac{x - \mu}{\sigma}.$$

This approach is effective for approximately Gaussian-distributed features but is highly sensitive to outliers.

MinMax Scaler Maps feature values into a fixed interval, typically $[0, 1]$:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}.$$

While simple and interpretable, this method is strongly affected by extreme values, which may compress the majority of samples into a narrow range.

MaxAbs Scaler Scales features by the maximum absolute value:

$$x' = \frac{x}{\max(|x|)}.$$

This approach preserves sparsity but remains sensitive to large outliers.

Robust Scaler Normalizes features using the median and the interquartile range:

$$x' = \frac{x - \text{median}(x)}{\text{IQR}(x)}.$$

By relying on robust statistics, this method effectively mitigates the influence of outliers and is well suited for noisy side-channel signals.

Normalizer Scales each sample independently to unit ℓ_2 norm:

$$\mathbf{x}' = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}, \quad \|\mathbf{x}\|_2 = \sqrt{\sum_i x_i^2}.$$

Although useful in certain contexts, this technique may distort absolute energy information, which is relevant in side-channel analysis.

Quantile Transformer Applies a non-linear transformation to map the empirical feature distribution to a target distribution (uniform or Gaussian). While this method enables comparability across heterogeneous feature distributions, it may significantly alter the original data structure and reduce physical interpretability.

PowerTransformer (Yeo–Johnson) Applies a parametric monotonic transformation aimed at stabilizing variance and reducing skewness:

$$x' = \begin{cases} \frac{(x+1)^\lambda - 1}{\lambda}, & x \geq 0, \\ -\frac{(-x+1)^{2-\lambda} - 1}{2-\lambda}, & x < 0. \end{cases}$$

Compared to the Quantile Transformer, this approach is less aggressive and better preserves relative relationships among samples.

Based on the observed performance and robustness considerations, the Robust Scaler is selected as the normalization strategy for all subsequent experiments, in line with prior findings reported in [38].

3.2.5 Summary of Preprocessing Steps

The complete preprocessing pipeline adopted in this work is summarized as follows:

- Removal of the initial 10-second background segment to exclude non-discriminative idle-state emissions.
- Temporal segmentation using 100 ms windows with 50% overlap, defining the fundamental unit of analysis.
- RMS computation for current and electromagnetic signals to obtain stable, energy-based representations.
- Feature normalization using Robust Scaler to mitigate scale imbalance and outlier effects.

These steps are applied sequentially to transform raw side-channel measurements into compact, noise-aware feature representations suitable for learning-based analysis. The resulting preprocessed signals serve as the input for subsequent feature extraction, selection, and classification stages described in the following sections.

Beyond their immediate role in noise mitigation and feature stabilization, the preprocessing steps described above implicitly shape the amount and structure of information exposed to learning-based models. As such, preprocessing can be interpreted not only as a preparatory stage, but also as an initial form of informational control over side-channel leakage.

While the present chapter adopts a static and deterministic preprocessing pipeline, these considerations naturally motivate the exploration of adaptive and feature-aware strategies, in which preprocessing and representation choices can be dynamically adjusted in response to changing signal conditions and threat models. This perspective will be progressively developed in the subsequent chapters.

3.2.6 Feature Set Description

Feature extraction represents a critical stage in the proposed side-channel analysis pipeline, as it defines the interface between physical signal representations and learning-based inference. In the context of electromagnetic side-channel analysis, feature design must balance three competing requirements: discriminative power, robustness to noise and environmental variability, and computational efficiency suitable for real-time or embedded deployment.

Rather than relying on raw or minimally processed signals, this work adopts an explicit feature-based representation. This choice is motivated by both interpretability and controllability considerations. Feature-level representations allow the contribution of different signal characteristics to be analyzed, compared, and selectively retained, enabling systematic evaluation of robustness and generalization properties under realistic noise conditions.

Multi-Channel Feature Extraction

For each segmented window, features are extracted independently from two physical channels: current consumption and electromagnetic emission. These channels capture complementary aspects of device behavior. While current measurements reflect global power consumption patterns, electromagnetic emissions provide spatially and spectrally localized information related to switching activity and internal data flow.

By maintaining separate feature sets for each channel, the model is able to exploit both redundant and complementary information sources. This dual-channel representation enhances discrimination between application scenarios while improving resilience to partial signal degradation or channel-specific noise.

Time-Domain Features

In the time domain, a set of statistical descriptors is computed for each window. These features summarize the amplitude distribution and temporal variability of the signal and are known to be robust under moderate noise conditions. Specifically, the extracted time-domain features include:

- Mean value,
- Standard deviation,

- Minimum and maximum values,
- Skewness,
- Kurtosis.

These descriptors capture first- and higher-order statistics of the signal, providing information about signal intensity, variability, asymmetry, and distribution shape. Time-domain features are widely adopted in side-channel and behavioral profiling studies due to their low computational cost and robustness to temporal misalignment [25, 38].

Frequency-Domain Features

To complement time-domain information, frequency-domain features are extracted by analyzing the spectral content of each RMS-aggregated window. Electromagnetic side-channel leakage often manifests through characteristic spectral patterns related to clock activity, bus communication, and workload-dependent execution behavior.

The following frequency-domain features are computed:

- Peak frequency,
- Spectral energy,
- Spectral entropy,
- Spectral centroid,
- Spectral spread.

These features jointly describe how signal energy is distributed across the frequency spectrum. In particular, spectral entropy quantifies signal complexity and dispersion, while centroid and spread provide localization information regarding dominant frequency components. Such descriptors have proven effective in distinguishing between workloads with similar temporal profiles but different execution characteristics. All frequency-domain features are computed over the same temporal windows defined during preprocessing, ensuring consistency between time- and frequency-domain representations.

Role of RMS in Feature Stability

Although RMS aggregation is introduced during preprocessing, it also plays a fundamental role in stabilizing feature computation. By operating on RMS-reduced representations, both time- and frequency-domain features become less sensitive to transient spikes and high-frequency disturbances.

Empirical evidence reported in the noise robustness analysis confirms that features computed on RMS-aggregated signals exhibit lower variance across trials and improved consistency under degraded signal-to-noise conditions. This observation reinforces the interpretation of RMS as a noise-aware representation that enhances feature reliability rather than merely reducing dimensionality.

Feature Set Composition

The complete feature set consists of 24 descriptors per window: 12 features extracted from the current channel and 12 from the electromagnetic channel. This symmetric design ensures balanced representation of both physical observables while maintaining a manageable feature dimensionality.

Each feature vector is associated with a numerical label in the range [1, 8], corresponding to the active application scenario during acquisition. This enables a supervised learning setting in which classifiers are trained to associate compact statistical fingerprints with application-level behavior.

Design Considerations and Methodological Implications

The adopted feature set is deliberately generic and modular. Rather than optimizing for a specific application or attack scenario, the selected descriptors capture fundamental statistical and spectral properties of side-channel emissions. This design choice supports generalization across heterogeneous workloads and deployment conditions.

Moreover, by explicitly defining and separating feature extraction stages, the methodology enables systematic exploration of feature relevance, redundancy, and robustness. As shown in subsequent sections,

this structure is essential for evaluating feature selection strategies and for reasoning about inference-time feasibility in realistic settings.

Importantly, the explicit feature-based representation adopted here provides a natural foundation for adaptive processing strategies. By exposing feature computation and selection as controllable stages within the pipeline, the framework enables future extensions toward dynamic and noise-aware optimization without altering the underlying acquisition process.

The next section builds upon this feature representation by analyzing different feature selection strategies, highlighting the trade-offs between scenario-specific optimization, global generalization, and inference-time constraints.

3.2.7 Feature Selection Strategies and Inference Constraints

While feature extraction defines how side-channel information is represented, feature selection determines which components of that representation are actually exploitable for inference. In side-channel analysis, this step is particularly critical, as the dimensionality of feature vectors directly affects classification performance, robustness to noise, and inference-time feasibility on resource-constrained platforms.

Feature selection serves three complementary objectives:

- improving classification accuracy by removing redundant or weakly informative descriptors,
- reducing computational and memory overhead at inference time,
- enhancing model interpretability and stability under noisy conditions.

However, in practical deployments, feature selection is subject to an often overlooked constraint: the selected feature subset must be *fixed and scenario-independent* at inference time. This constraint fundamentally limits the applicability of several high-performing selection strategies commonly reported in the literature.

To explicitly investigate this trade-off, three feature selection strategies are evaluated in this work, each corresponding to a different level of practical feasibility.

Scenario-Specific Feature Importance: An Upper Bound

As an initial reference, feature importance is computed independently for each application scenario using a Random Forest classifier trained in a one-vs-all configuration. This approach identifies the subset of features that maximizes discrimination for each individual class and therefore represents an upper bound on achievable classification performance.

Figure 3.9 illustrates the overall feature importance distribution obtained from the full Random Forest model. The results show that discriminative information is unevenly distributed across descriptors and sensing channels, with frequency-domain features and higher-order statistical moments contributing most significantly to scenario separation.

When classifiers are allowed to exploit scenario-tailored feature subsets, an overall classification accuracy of 99.83% is achieved. Despite this excellent performance, this strategy is fundamentally incompatible with real-time inference. In operational conditions, the active scenario is unknown *a priori*, making it impossible to select the appropriate feature subset before classification.

For this reason, scenario-specific feature importance is retained exclusively as a theoretical upper bound. While it provides valuable insight into the discriminative structure of the data, it cannot be deployed in practical inference scenarios.

Global Feature Importance Across All Scenarios

To ensure inference-time feasibility, feature importance is next computed globally by training a Random Forest classifier on the full dataset, considering all scenarios jointly. This produces a single, fixed feature ranking that is shared across all classes and can be consistently applied at runtime.

This strategy satisfies deployment constraints and yields a compact and interpretable feature set. However, enforcing a common feature space across heterogeneous scenarios inevitably reduces specialization. As a consequence, classification accuracy decreases to approximately 93%.

Figure 3.10 summarizes the performance obtained under the different feature selection strategies, explicitly highlighting the trade-off between discriminative performance and inference-time applicability.

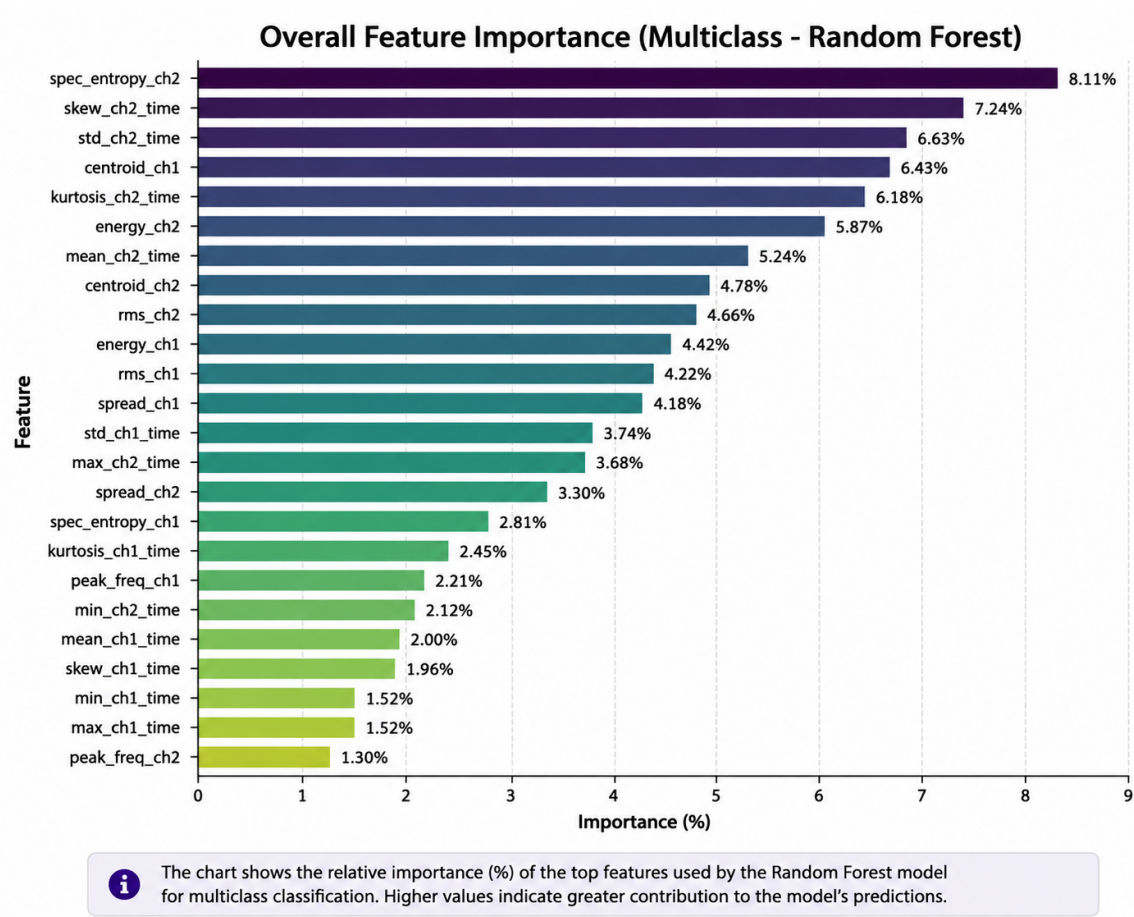


Figure 3.9: Overall feature importance obtained using a Random Forest classifier trained on the full feature set. Feature relevance is unevenly distributed across descriptors and channels, highlighting heterogeneous contributions to scenario discrimination.

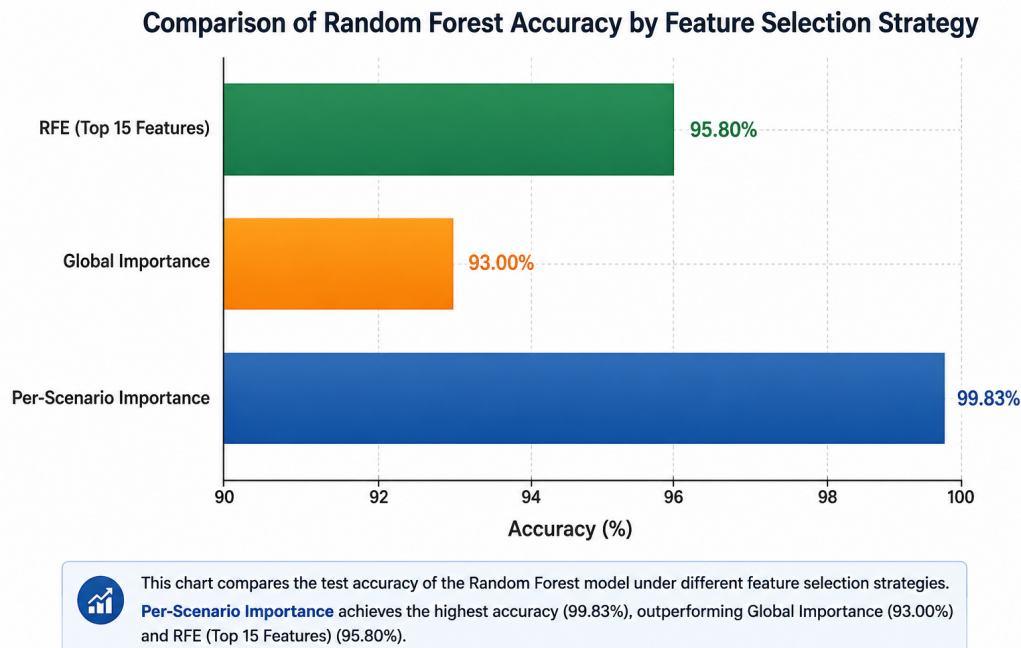


Figure 3.10: Comparison of Random Forest classification accuracy under different feature selection strategies. Scenario-specific feature importance achieves the highest accuracy but is not applicable at inference time.

While global feature importance enables real-time deployment, the observed performance degradation

indicates that features highly informative for specific scenarios may be diluted when optimized jointly across all classes.

Recursive Feature Elimination (RFE): A Practical Compromise

Recursive Feature Elimination (RFE) is adopted as a third strategy to balance discriminative performance and inference-time constraints. Unlike simple ranking-based approaches, RFE iteratively retrain the model while progressively removing the least informative features, implicitly accounting for inter-feature dependencies.

Using Random Forest as the estimator, RFE is applied to the complete feature set, resulting in a reduced subset of 15 features that preserves both time-domain and frequency-domain descriptors across the two physical channels. The effect of this selection process is illustrated in Figure 3.11, which compares feature importance distributions before and after RFE.

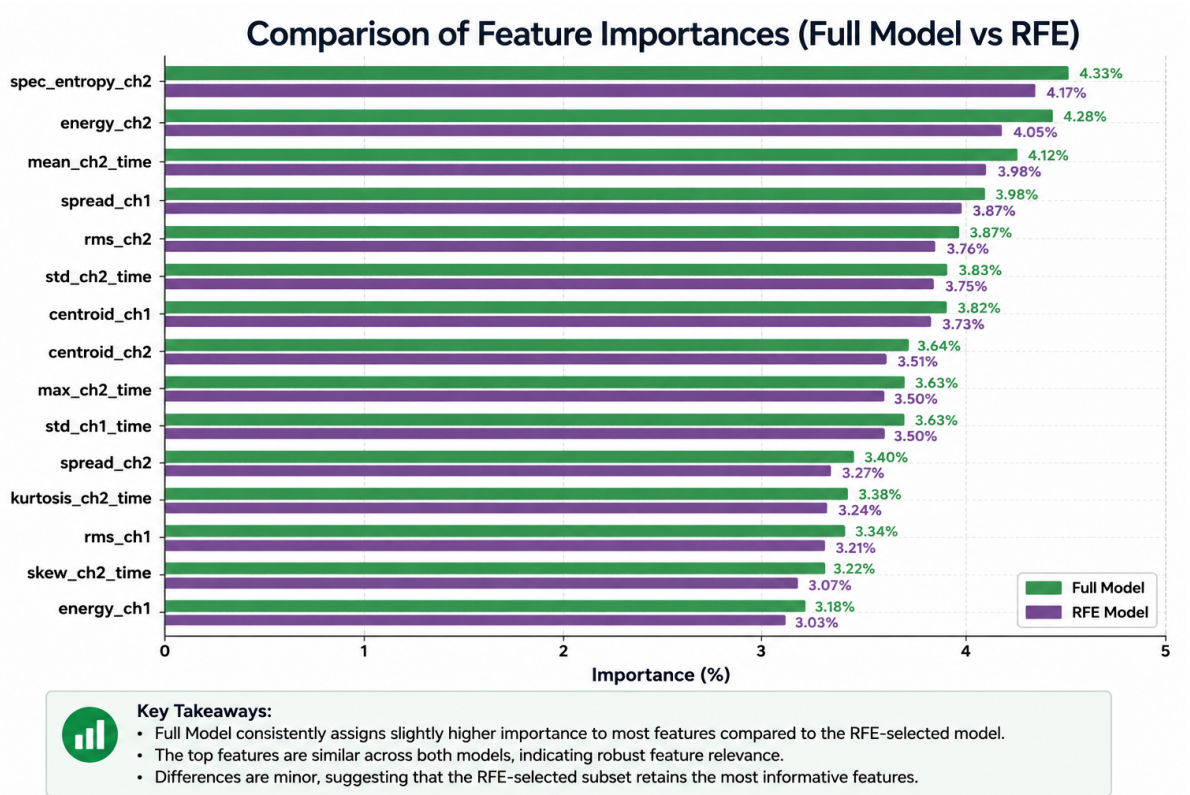


Figure 3.11: Comparison of feature importance between the full model and the reduced model obtained via Recursive Feature Elimination (RFE). The selected subset preserves the most discriminative descriptors while reducing dimensionality.

Despite the substantial reduction in feature dimensionality, this configuration achieves an average classification accuracy of 95.8%, outperforming the global feature importance approach while remaining fully compatible with real-time inference. The selected feature subset therefore represents an effective compromise between accuracy, robustness, and operational feasibility.

Inference-Time Implications

The effectiveness of the RFE-based feature set is further illustrated by the confusion matrix reported in Figure 3.12. Most scenarios are classified with high accuracy, with residual misclassifications occurring primarily between scenarios characterized by similar execution behavior.

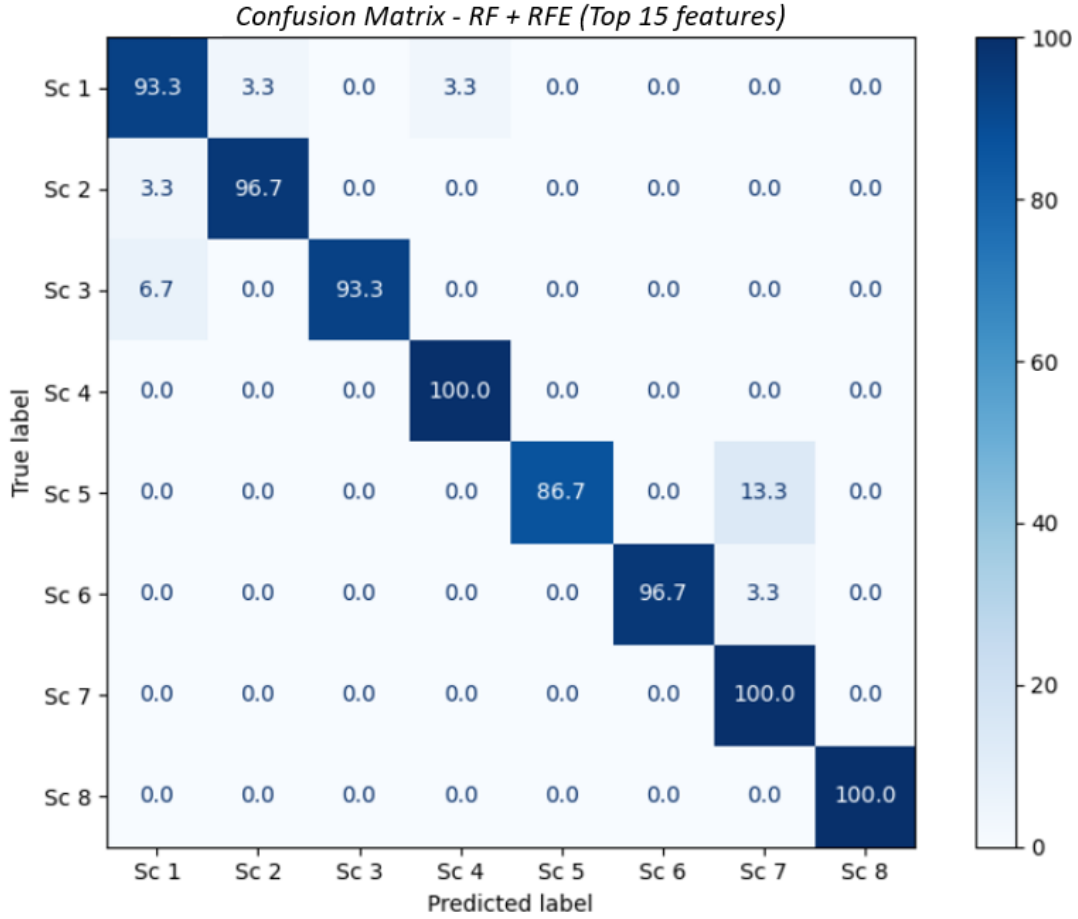


Figure 3.12: Confusion matrix obtained using Random Forest with the RFE-selected feature subset. High diagonal dominance confirms strong discriminative capability under inference-time constraints

Overall, this analysis reveals a critical methodological insight: feature selection strategies that maximize accuracy under idealized or scenario-aware conditions are not necessarily applicable in realistic inference scenarios. While scenario-specific optimization provides valuable analytical insight, it cannot be exploited at runtime. Conversely, globally fixed feature sets ensure deployability but may sacrifice discriminative power.

RFE emerges as an effective compromise, offering strong performance while respecting inference-time constraints.

3.2.8 Pipeline Overview

Figure 3.13 provides a consolidated overview of the complete side-channel processing pipeline adopted in this work, spanning from raw signal acquisition to final classification.

The pipeline is deliberately designed to be modular and interpretable, enabling systematic evaluation of each processing stage while maintaining a coherent end-to-end workflow. This modular structure facilitates deployment on resource-constrained platforms, where computational efficiency and transparency are critical.

Importantly, the proposed organization treats preprocessing and feature representation choices as explicit design variables rather than fixed preprocessing steps. This perspective is essential when addressing robustness to noise and variability, and it naturally supports the transition toward adaptive and learning-driven control strategies explored in the subsequent chapters.

3.2.9 Noise Robustness and Stability of Feature Relevance

The effectiveness of feature-based side-channel profiling critically depends not only on the discriminative power of the selected descriptors under nominal conditions, but also on their stability in the presence of noise and environmental variability. In realistic deployments, electromagnetic and current measurements

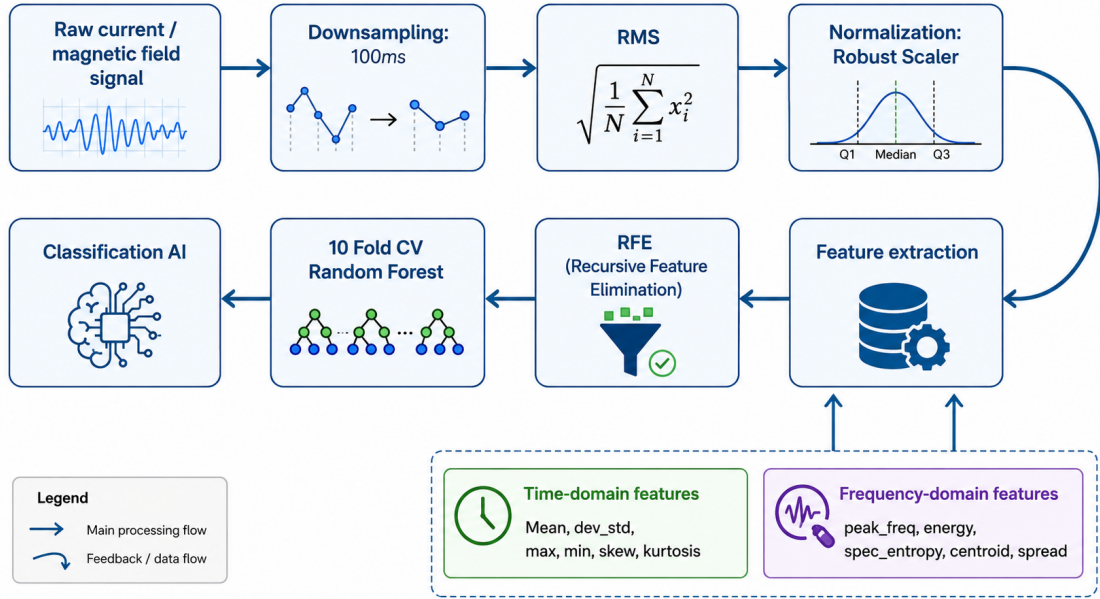


Figure 3.13: Overview of the proposed side-channel processing pipeline. Raw current and electromagnetic signals are segmented, aggregated via RMS, normalized using Robust Scaler, processed through feature extraction and RFE-based selection, and finally classified using supervised learning models.

are inherently affected by stochastic disturbances whose characteristics may vary over time, across devices, and across operating contexts.

From a methodological perspective, this raises a fundamental question: *to what extent is feature relevance stable under noisy conditions?* If the discriminative contribution of individual features varies significantly with noise intensity or structure, then any static feature selection strategy—regardless of its optimality under clean conditions—may become suboptimal or even misleading at inference time.

Noise as a Structural Perturbation

In side-channel analysis, noise does not act merely as additive measurement uncertainty. Rather, it often alters the statistical structure of observable signals in a non-uniform and feature-dependent manner. Low-frequency electromagnetic measurements are particularly sensitive to environmental interference, power supply instability, and concurrent system activity, which may selectively affect specific temporal or spectral components.

As a consequence, features derived from side-channel signals exhibit heterogeneous sensitivity to noise. Energy-related descriptors may remain relatively stable under moderate perturbations, whereas higher-order statistics or spectral entropy may fluctuate significantly as noise increases. Moreover, this sensitivity is not uniform across scenarios: features that are highly informative for a given application under clean conditions may lose relevance—or even become misleading—under degraded signal-to-noise ratios.

These observations imply that feature relevance is not an intrinsic or static property of a descriptor, but rather an emergent characteristic resulting from the interaction between signal structure, noise characteristics, and temporal aggregation.

Empirical Stability of RMS-Based Representations

The experiments conducted in this work explicitly evaluate the robustness of feature representations under noisy conditions. A key observation emerging from the analysis is the stabilizing role played by RMS-based temporal aggregation.

By computing RMS values over fixed windows, high-frequency fluctuations and transient noise components are partially averaged out, while preserving the energy profile associated with application-level activity. Compared to raw-sample-based representations, RMS-based features exhibit significantly lower variance under noise injection and improved consistency across windows.

Empirical results confirm that classifiers trained on RMS-preprocessed signals consistently outperform their raw-signal counterparts when evaluated under increasing noise levels. This effect is observed across

multiple feature selection strategies and classifiers, indicating that RMS aggregation functions as a noise-aware representation rather than a model-specific optimization.

Nevertheless, RMS preprocessing does not eliminate variability entirely. While it improves overall robustness, the relative importance of individual features continues to evolve as noise conditions change. This residual instability highlights the intrinsic limitations of fixed preprocessing pipelines and static feature selection strategies.

Feature Selection Under Noisy Conditions

When feature importance is re-evaluated on noisy datasets, a non-trivial redistribution of relevance emerges. Some features maintain high importance across noise regimes, while others exhibit sharp degradation or erratic behavior. Notably, the overlap between top-ranked features identified under clean and noisy conditions decreases progressively as noise intensity increases.

This observation has two important implications. First, feature selection strategies optimized under clean conditions cannot be assumed to remain optimal during noisy operation. Second, even comparatively robust methods such as Recursive Feature Elimination (RFE)—while more stable than simple ranking approaches—still rely on an implicit stationarity assumption regarding feature relevance.

In practical terms, this means that a feature subset optimized offline represents a compromise across expected operating conditions, rather than an optimal solution for all regimes. As noise characteristics drift over time, the effective discriminative space perceived by the classifier may shift accordingly, leading to gradual and often unnoticed performance degradation.

Temporal Aggregation as a Compensatory Mechanism

One mechanism that partially mitigates local instability is temporal aggregation at the decision level. Window-based classifiers may produce noisy or ambiguous predictions on individual segments, particularly under degraded signal conditions. However, when predictions are aggregated over time—using strategies such as majority voting or confidence-weighted fusion—global application behavior can often still be inferred reliably.

This distinction between local feature instability and global decision stability is central to understanding why sequence-level inference may remain effective even when window-level accuracy degrades. Temporal aggregation compensates for short-term noise by exploiting the persistence and regularity of application behavior over time.

Nevertheless, this mechanism operates *downstream* of feature extraction and classification. It does not address the underlying variability in feature relevance, but instead mitigates its observable effects at the decision stage. As such, temporal aggregation constitutes a reactive form of robustness rather than a proactive adaptation of the representation itself.

Limits of Static Robustness Strategies

Taken together, these observations expose an inherent limitation of static side-channel pipelines. Robust preprocessing, RMS aggregation, feature selection, and temporal voting can significantly improve resilience to noise, but they do so under fixed assumptions about the operating environment.

When noise characteristics evolve beyond the range anticipated during training, static pipelines lack the ability to adjust their internal representations. Feature relevance may drift, classifiers may operate on suboptimal descriptors, and performance may degrade progressively rather than abruptly—making failures more difficult to detect and diagnose.

Importantly, this limitation is not attributable to deficiencies in individual components, but instead reflects a structural property of static designs. Robustness achieved through fixed preprocessing and feature selection represents a bounded form of resilience, effective only within a predefined envelope of conditions.

Implications for Adaptive Side-Channel Processing

The observed instability of feature relevance under noise motivates a shift in perspective. Rather than seeking a single optimal feature set or preprocessing configuration, side-channel analysis can be viewed as a dynamic process in which the effective representation must adapt to observed signal conditions.

From this standpoint, robustness is no longer a static attribute of a trained model, but an operational objective that must be continuously maintained. This insight provides the conceptual foundation for

adaptive approaches, in which preprocessing, feature weighting, or representation strategies are adjusted online based on feedback from the environment.

While the present chapter focuses on static methodologies and their empirical limitations, the instability highlighted in this section directly motivates the exploration of adaptive, learning-driven mechanisms in the subsequent chapters. In particular, it establishes the rationale for framing side-channel processing as a sequential decision problem, where representation choices are dynamically optimized rather than fixed *a priori*.

3.3 From Static Pipelines to Adaptive Control: Summary and Transition

This chapter has presented a comprehensive experimental framework for electromagnetic side-channel profiling, spanning acquisition, preprocessing, feature engineering, and feature selection. Through a systematic analysis grounded in real measurements and realistic operating conditions, the strengths and limitations of state-of-the-art static pipelines have been empirically and conceptually examined.

Rather than treating preprocessing and feature selection as isolated technical steps, the chapter has emphasized their role as structural components that shape the observable information space available to learning models. This perspective enables a deeper understanding of why high classification accuracy under nominal conditions does not necessarily translate into robustness under noise, variability, and deployment constraints.

3.3.1 Achievements of Static Learning-Based Pipelines

The experimental results demonstrate that carefully designed static pipelines can achieve excellent performance in side-channel profiling tasks. The combination of RMS-based temporal aggregation, robust normalization, and feature selection through Recursive Feature Elimination yields compact and interpretable representations with high discriminative power. Under controlled or moderately noisy conditions, such pipelines enable reliable classification of application-level scenarios with limited computational overhead, making them suitable for embedded and resource-constrained platforms.

In particular, RMS aggregation has been shown to play a stabilizing role by reducing high-frequency noise and consolidating signal energy over time. Feature selection further improves efficiency and generalization by discarding redundant or weakly informative descriptors. Together, these design choices represent best practices within the paradigm of static side-channel analysis.

Importantly, these results confirm that static pipelines are not inherently fragile. When properly engineered, they exhibit a meaningful degree of robustness that allows effective operation across a range of operating conditions. This robustness, however, remains fundamentally bounded.

3.3.2 Structural Limits of Static Robustness

Despite their effectiveness, static pipelines rely on an implicit stationarity assumption: the statistical relationship between side-channel signals and system behavior is assumed to remain sufficiently stable over time. All preprocessing parameters, feature subsets, and model weights are fixed offline, based on training data that can only approximate future operating conditions.

The experimental analysis under noisy scenarios reveals that this assumption is only partially valid. While RMS aggregation and robust normalization mitigate noise, they do not fully stabilize the relevance of individual features. Feature importance distributions shift as noise characteristics change, and the overlap between optimal feature subsets under clean and noisy conditions decreases progressively.

These observations highlight a critical limitation: robustness achieved through static design represents a compromise across expected conditions rather than an optimal solution for any specific one. When the operating environment drifts outside the envelope anticipated during training, the pipeline cannot respond proactively. Performance degradation may be gradual and subtle, making it difficult to detect or correct without retraining.

Moreover, static feature selection strategies—whether based on global importance or RFE—implicitly assume that a single feature subset can remain informative across all scenarios and noise regimes. The empirical evidence presented in this chapter challenges this assumption, showing that feature relevance is inherently context-dependent and evolves with signal conditions.

3.3.3 Temporal Aggregation Is Not Adaptation

Temporal aggregation mechanisms, such as majority voting across windows, provide an additional layer of robustness by exploiting the persistence of application behavior over time. These mechanisms compensate for local uncertainty and noise-induced misclassifications at the window level, enabling stable sequence-level inference.

However, temporal aggregation operates downstream of feature extraction and classification. It does not modify the representation itself, nor does it influence how features are extracted or weighted in response to changing conditions. As such, it represents a reactive form of robustness rather than a mechanism for active adaptation.

This distinction is crucial: while temporal aggregation improves decision stability, it does not address the underlying instability of feature relevance. The system continues to operate on a fixed representation, even when that representation becomes suboptimal. Temporal voting mitigates symptoms rather than causes.

3.3.4 Noise and Variability as First-Class System Properties

A central insight emerging from this chapter is that noise and variability should not be treated as external disturbances to be eliminated, but as intrinsic properties of side-channel environments. In realistic deployments, electromagnetic emissions are shaped by a complex interplay of device behavior, environmental interference, power conditions, and concurrent activity.

As a result, side-channel analysis operates in a partially observable and non-stationary environment. Neither attackers nor defenders have access to a complete or stable view of the underlying system state. Both must reason over noisy, incomplete, and temporally evolving observations.

Static pipelines implicitly assume that this uncertainty can be absorbed through preprocessing and robust feature design. The experimental evidence shows that this assumption holds only up to a point. Beyond that point, maintaining performance requires the ability to adapt internal processing strategies based on observed conditions.

3.3.5 Reframing Side-Channel Processing as a Control Problem

The limitations identified in this chapter motivate a shift in perspective. Instead of viewing side-channel analysis as a static classification task, it can be reframed as a sequential decision-making problem. In this formulation, preprocessing choices, feature representations, and aggregation strategies are no longer fixed design-time decisions, but controllable parameters that influence observability and robustness.

From this viewpoint, the goal of the system is not merely to classify correctly, but to continuously regulate the quality and structure of the information extracted from physical signals. Decisions made at one point in time influence future observations and inference reliability, creating a closed-loop interaction between the system and its environment.

This reframing aligns naturally with control-oriented learning paradigms. Rather than optimizing a static mapping from features to labels, the system seeks to optimize long-term objectives—such as stability, robustness, or controlled information exposure—under uncertainty and noise.

3.3.6 Toward Adaptive and Learning-Driven Side-Channel Defense

While this chapter has focused on static methodologies and their empirical limits, it establishes the conceptual foundations for adaptive approaches. The observed instability of feature relevance under noise provides a concrete motivation for moving beyond fixed pipelines. The need is not for more complex features or deeper models per se, but for mechanisms capable of adjusting representation strategies in response to observed conditions.

Adaptive approaches promise to bridge the gap between robust static design and fully dynamic operation. By incorporating feedback from the environment, such systems can adjust preprocessing parameters, feature emphasis, or temporal aggregation strategies to maintain performance across evolving conditions.

Crucially, the objective of such adaptation is not to guarantee information-theoretic security or eliminate leakage entirely—goals that are often unattainable in practice. Instead, the aim is to actively manage and reduce exploitable information under realistic constraints, maintaining acceptable performance while degrading the effectiveness of adversarial inference.

3.3.7 Transition to the Next Chapter

In summary, this chapter has established a rigorous experimental baseline for side-channel profiling and highlighted the structural limitations of static pipelines in noisy and variable environments. While static learning-based approaches can achieve high performance, their robustness is inherently bounded by fixed assumptions about signal statistics and operating conditions.

These findings motivate the exploration of adaptive mechanisms capable of operating in a closed-loop manner. The next chapter builds upon this foundation by introducing a learning-based framework that explicitly addresses these limitations, framing side-channel processing as a sequential decision problem in which representation choices are dynamically optimized rather than fixed *a priori*.

Chapter 4

Results and Performance Evaluation

This chapter presents a comprehensive experimental evaluation of the proposed side-channel classification pipeline, with the objective of quantifying its effectiveness, robustness, and practical suitability under realistic operating conditions. Building upon the experimental setup, preprocessing methodology, and feature engineering strategies introduced in Chapter 3, the analysis focuses on how different learning models exploit electromagnetic and current-based side-channel information to discriminate application-level scenarios.

A broad set of machine learning and deep learning classifiers is systematically evaluated, including both traditional models and hybrid neural architectures. Each classifier is tested using two feature configurations: the complete set of extracted features and a reduced subset obtained through Recursive Feature Elimination (RFE). This dual evaluation enables a principled assessment of the trade-offs between predictive performance, model complexity, and inference-time efficiency—a critical aspect for deployment on resource-constrained embedded and IoT platforms.

The experimental analysis emphasizes not only overall classification accuracy, but also generalization to unseen data, stability across heterogeneous scenarios, and resilience to noise and signal perturbations. Performance is evaluated across eight distinct scenarios, encompassing client-level applications and network communication protocols, including both encrypted and unencrypted traffic. This diversity allows the investigation of how different classes of workloads manifest in side-channel traces and how reliably they can be distinguished in practice.

Beyond nominal conditions, the chapter explicitly addresses robustness under degraded signal quality. Controlled noise injection experiments and scenario-level aggregation strategies are used to evaluate the limits of static classification pipelines and to identify failure modes that emerge under increasing uncertainty. These results provide concrete empirical evidence of both the strengths and the inherent limitations of fixed feature-based approaches.

Overall, the findings presented in this chapter serve two complementary purposes. First, they validate the effectiveness of the proposed pipeline as a practical solution for side-channel-based monitoring and classification in embedded systems. Second, they expose the operational trade-offs and robustness boundaries that motivate the transition toward adaptive and learning-driven processing strategies, which are explored in the subsequent chapters.

4.1 Evaluation Protocol

To ensure a rigorous, unbiased, and reproducible evaluation of the proposed side-channel classification pipeline, a stratified 10-fold cross-validation protocol was adopted. This evaluation strategy provides a reliable estimate of generalization performance while mitigating the risk of overfitting, particularly in the presence of inter-scenario variability and limited sample sizes.

The dataset consists of 8 distinct application scenarios, each represented by 30 independent acquisition files, for a total of 240 CSV files. Each file corresponds to a complete measurement session associated with a specific application-level activity or communication protocol. Stratified partitioning was employed to preserve the original class distribution within each fold, ensuring that all scenarios are consistently represented during both training and testing phases.

At each cross-validation iteration, one fold is held out as the test set, while the remaining nine folds are used for training. Classification performance is evaluated on the unseen test fold, and this procedure is repeated until each fold has been used exactly once for testing. Final performance metrics are obtained

by averaging results across all folds, providing a statistically robust estimate of classifier behavior.

The primary evaluation metrics include overall accuracy, precision, recall, and F1-score. While accuracy provides a global measure of classification effectiveness, precision and recall offer insight into class-specific behavior, particularly in scenarios where certain application profiles exhibit overlapping side-channel characteristics. Macro-averaged and micro-averaged metrics are also considered to capture both per-class fairness and global system performance.

Importantly, no additional hyperparameter optimization (e.g., grid search or Bayesian tuning) was performed within each fold. This design choice ensures that performance comparisons reflect the intrinsic discriminative power and robustness of the selected feature representations and classifiers, rather than gains attributable to extensive model-specific tuning. All classifiers were evaluated using consistent and commonly adopted configurations.

Traditional machine learning models were implemented using the `scikit-learn` framework, while deep learning architectures—specifically the hybrid CNN+LSTM model—were implemented using Keras with a TensorFlow backend. This separation allows a fair comparison between classical feature-based methods and deep learning approaches operating on the same input representations.

Figure ?? illustrates the overall evaluation protocol adopted in this work.

4.2 Overall Classification Performance

We begin the experimental analysis by examining the overall classification performance achieved across all eight scenarios. This global evaluation provides a first-order assessment of the discriminative capacity of the proposed feature representations and learning models, serving as a reference point for the more detailed analyses presented in subsequent sections.

Rather than focusing solely on aggregate accuracy, this section aims to characterize how reliably different application scenarios can be distinguished based on side-channel traces, and to what extent classification errors are distributed across classes. Confusion matrices are therefore used as a primary analysis tool, as they reveal inter-scenario misclassification patterns and highlight which workloads exhibit similar side-channel signatures.

The results reported in this section compare multiple classifiers operating on both the full feature set and the reduced subset obtained through Recursive Feature Elimination (RFE). This comparison allows us to evaluate whether dimensionality reduction impacts not only overall accuracy, but also class-wise stability and error localization. In particular, we assess whether reduced models preserve discriminative structure uniformly across scenarios or disproportionately affect specific classes.

This global performance analysis is intentionally conducted under nominal conditions, without artificial noise injection, to establish a clean baseline. The robustness of the same models under degraded signal quality and perturbations is analyzed separately in later sections. By isolating baseline performance from robustness stress tests, we can more clearly attribute performance degradation to noise effects rather than to intrinsic model limitations.

The insights derived from this section provide the foundation for the scenario-specific and robustness-oriented evaluations that follow. In particular, they help identify which application classes are inherently well-separated in the feature space and which ones pose greater challenges, guiding the interpretation of noise sensitivity, temporal aggregation effects, and adaptive strategies discussed later in the chapter.

4.3 Scenario-Specific Analysis

To gain deeper insight into classifier behavior beyond aggregate performance metrics, a detailed scenario-specific analysis is conducted. Each application scenario (e.g., HTTP, HTTPS, FTP, FTPS, client-side applications) is examined independently to assess class separability, prediction consistency, and systematic misclassification patterns.

This analysis serves two main purposes. First, it allows the identification of scenarios that are inherently easier or harder to distinguish based on their side-channel signatures. Second, it highlights latent ambiguities in the feature space, particularly for scenarios characterized by similar execution dynamics or overlapping resource usage.

The results reveal that scenarios associated with user-driven applications (e.g., web browsing and email clients) tend to produce more stable and repetitive electromagnetic patterns, leading to consistently

high classification accuracy. In contrast, network protocol-driven scenarios exhibit greater variability, reflecting the dynamic nature of packet handling, buffering, and protocol state transitions.

Particular attention is devoted to the comparison between encrypted and unencrypted communication protocols, such as HTTP versus HTTPS and FTP versus FTPS. These cases are especially relevant in security-oriented monitoring, as they test the ability of side-channel analysis to distinguish traffic types even when payload visibility is intentionally obscured. While encryption alters higher-level data semantics, the results confirm that protocol-level execution behavior still induces distinguishable side-channel signatures, albeit with reduced margins under noisy conditions.

Overall, the scenario-specific analysis demonstrates that classification performance is not uniform across scenarios and that confusion patterns are strongly linked to underlying execution characteristics. These observations provide valuable guidance for both feature design and deployment strategies, indicating where additional robustness mechanisms may be required.

4.4 Impact of Feature Selection

This section evaluates the effect of feature selection on classification performance, model complexity, and inference-time feasibility. Specifically, we compare models trained on the full set of extracted features with those using a reduced feature representation obtained through Recursive Feature Elimination (RFE).

The objective of this analysis is not merely to maximize accuracy, but to assess whether compact feature spaces can preserve discriminative power while satisfying the constraints imposed by real-time and resource-limited environments. This consideration is particularly relevant for embedded and IoT deployments, where memory footprint, computational cost, and latency are critical factors.

Experimental results show that reducing the feature dimensionality via RFE leads to only marginal performance degradation—and in several configurations, even slight improvements—despite eliminating a substantial portion of the original descriptors. In most scenarios, classification accuracy remains above 96%, confirming that a significant fraction of features is either redundant or weakly informative.

Importantly, the reduced feature set demonstrates strong cross-scenario generalization. Features retained by RFE consistently capture core statistical and spectral properties that are relevant across heterogeneous application behaviors, rather than overfitting to scenario-specific artifacts. This supports the use of RFE as a practical and robust selection strategy for inference-time deployment.

Beyond accuracy, the benefits of feature reduction are reflected in lower computational overhead and improved model interpretability. Compact models are easier to analyze, verify, and integrate into security pipelines, and they reduce the attack surface for potential adversarial manipulation. These advantages reinforce the suitability of RFE-based feature selection as a cornerstone of the proposed side-channel monitoring framework.

4.5 Robustness to Noise and Variability

While feature selection improves efficiency and generalization under nominal conditions, real-world deployments inevitably expose side-channel systems to noise, environmental interference, and signal variability. To evaluate robustness under such conditions, we perform a systematic stress analysis by injecting controlled Gaussian noise into the raw signals prior to preprocessing.

Noise is generated with zero mean and variance proportional to the dynamic range of each signal channel, with perturbation levels set to 10%, 20%, 30%, and 50%. This setup emulates realistic disturbances arising from electromagnetic interference, power instability, or adversarial obfuscation.

Analysis of RMS-based representations reveals a clear stabilizing effect. Compared to raw-signal processing, RMS aggregation significantly attenuates high-frequency fluctuations and transient disturbances, resulting in smoother feature trajectories and more consistent distributions across windows. This behavior is observed for both current and magnetic field signals, although electromagnetic measurements exhibit higher resilience at elevated noise levels.

Classification results confirm a gradual and controlled degradation as noise intensity increases. Accuracy decreases from approximately 88% at 10% noise to around 77% at 50% noise, indicating graceful performance decay rather than abrupt failure. Notably, scenarios involving dynamic network protocols (e.g., HTTP and FTPS) are more sensitive to noise, whereas client-oriented scenarios preserve higher robustness due to their repetitive execution patterns.

To further mitigate noise-induced instability, a column-wise majority voting strategy is employed, aggregating window-level predictions into a single decision per acquisition. This temporal aggregation acts as an implicit denoising mechanism, suppressing isolated misclassifications and reinforcing consistent behavior over time. The resulting scenario-level accuracy substantially exceeds window-level performance, especially under moderate noise conditions.

These findings highlight an important distinction: while preprocessing and feature selection improve robustness, they cannot fully stabilize feature relevance under severe or evolving noise. Feature importance distributions shift as signal conditions degrade, indicating that static pipelines operate within a bounded robustness envelope.

From a deployment perspective, the results suggest that robustness is best achieved through a combination of noise-aware representations (e.g., RMS), compact and stable feature sets, and temporal aggregation. However, they also expose the limitations of static designs, motivating the exploration of adaptive mechanisms capable of adjusting representation strategies in response to observed conditions.

In summary, the robustness analysis confirms that the proposed pipeline maintains reliable performance under realistic perturbations, while simultaneously revealing the structural limits of static processing. These insights directly inform the transition toward adaptive and learning-driven approaches discussed in the subsequent chapters.

Gaussian Noise Injection. To evaluate robustness under realistic perturbations, controlled Gaussian noise was systematically injected into the raw side-channel signals prior to RMS-based temporal aggregation. The injected noise follows a zero-mean Gaussian distribution, with variance scaled proportionally to the dynamic range of each signal channel, computed as $(\text{col.max}() - \text{col.min}())$. Noise levels of 10%, 20%, 30%, and 50% were considered.

This noise model is designed to emulate a broad class of real-world disturbances, including environmental electromagnetic interference, power supply instability, sensor imperfections, and potential adversarial obfuscation attempts. Importantly, scaling the noise variance to the signal range ensures that perturbations remain relative to the intrinsic signal dynamics, avoiding unrealistic saturation effects.

Figures 4.1 and 4.2 illustrate the impact of increasing noise intensity on the RMS features extracted from the current and magnetic field signals, respectively. For visualization clarity, RMS values are aggregated over consecutive 5-second intervals, while the underlying RMS computation is still performed on 100 ms windows with 50% overlap. This dual-scale representation highlights long-term stability trends without obscuring short-term variability.

A clear distinction emerges between the two physical channels. Current-based RMS features exhibit increasing volatility as noise intensity grows, reflecting their stronger sensitivity to additive perturbations. In contrast, magnetic field RMS features demonstrate markedly higher stability, even at elevated noise levels. This behavior suggests that electromagnetic emissions preserve structural information related to application execution more robustly than current measurements under adverse conditions.

These observations provide empirical support for the complementary use of current and electromagnetic channels. While current signals offer strong discriminative power under low-noise conditions, magnetic field measurements provide enhanced resilience when the signal-to-noise ratio degrades. This channel-dependent robustness becomes a critical consideration for deployment in uncontrolled or adversarial environments.

Impact on Classification. The impact of noise on classification performance is further analyzed through confusion matrices reported in Figures 4.3–4.6, obtained using stratified 10-fold cross-validation on the reduced feature set selected via Recursive Feature Elimination (RFE).

As the intensity of Gaussian noise increases, a progressive degradation in overall classification accuracy is observed. Average accuracy decreases from approximately 88% at 10% noise to around 77% at 50% noise. This trend confirms that, while the proposed pipeline exhibits a degree of robustness, its performance remains inherently sensitive to severe perturbations when operating under fixed preprocessing and feature selection configurations.

Beyond the global accuracy trend, a more nuanced behavior emerges at the scenario level. Core scenarios corresponding to client-oriented applications (Scenarios 1 to 4) retain high precision even under severe noise conditions. These scenarios are characterized by relatively stable and repetitive execution patterns, which result in consistent side-channel signatures that remain distinguishable despite perturbations.

In contrast, scenarios involving network-based protocols—particularly HTTP (Scenario 5) and FTP/FTPS (Scenarios 6 and 7)—exhibit a more pronounced degradation in performance. These scenarios are inher-

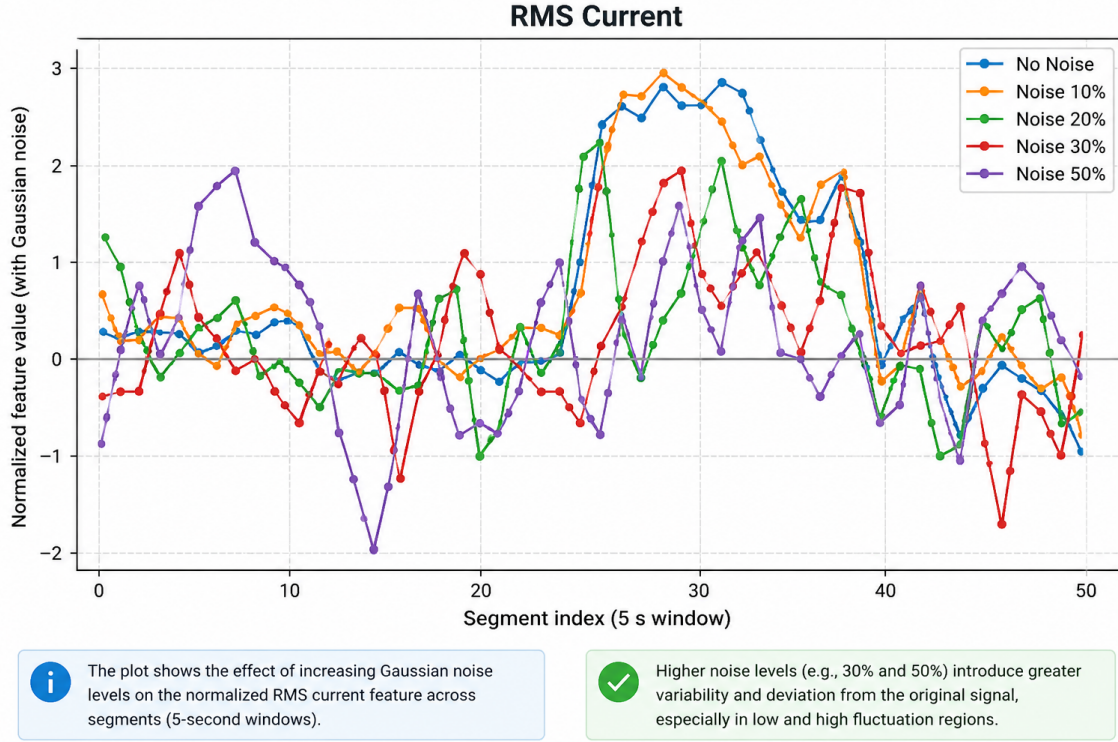


Figure 4.1: Effect of Gaussian noise injection on the normalized RMS feature of the current signal. Each point represents aggregation over a 5-second interval. Increasing noise levels induce progressively larger fluctuations and reduced temporal stability.

ently more dynamic, with execution behavior influenced by external network interactions, asynchronous events, and variable traffic patterns. As a result, their side-channel signatures are less regular and more susceptible to noise-induced distortion, leading to increased confusion under degraded signal-to-noise ratios.

This selective degradation highlights an important structural property of side-channel classification under noise: robustness is not uniformly distributed across scenarios. Instead, it depends on the intrinsic temporal stability and determinism of the underlying application behavior. Scenarios dominated by protocol-level activity represent a more challenging classification target and are more vulnerable to perturbations when relying on static feature representations.

These observations underscore the critical role of feature robustness in protocol-level monitoring and suggest that static preprocessing and feature selection strategies may be insufficient to guarantee stable performance across heterogeneous workloads. In particular, they motivate the exploration of complementary resilience mechanisms—such as adaptive denoising, temporal feature fusion, or dynamic feature weighting—to mitigate noise-induced performance degradation in realistic deployment settings.

Device Variability. To further assess the feasibility of practical deployment, we also consider the impact of device variability on classification performance. In real-world cyber-physical systems and IoT infrastructures, side-channel monitoring solutions are rarely confined to a single, perfectly controlled hardware instance. Instead, they must operate across multiple devices that may differ in manufacturing tolerances, component aging, power delivery characteristics, and electromagnetic coupling conditions.

Evaluating the proposed pipeline across different acquisition devices therefore serves as a proxy for real deployment scenarios, where models trained on one device are expected to generalize to others with similar functionality but non-identical physical characteristics. From a side-channel perspective, such variability alters the absolute amplitude, spectral distribution, and noise profile of the observed signals, even when the same application is executed.

This cross-device evaluation tests the portability of trained models and exposes a critical limitation of many side-channel approaches reported in the literature: high accuracy achieved on a single device does not necessarily imply robustness across hardware instances. In this context, the use of RMS-based aggregation and robust normalization plays a key role in reducing sensitivity to device-specific scaling effects, while feature selection via RFE limits overfitting to idiosyncratic hardware artifacts.

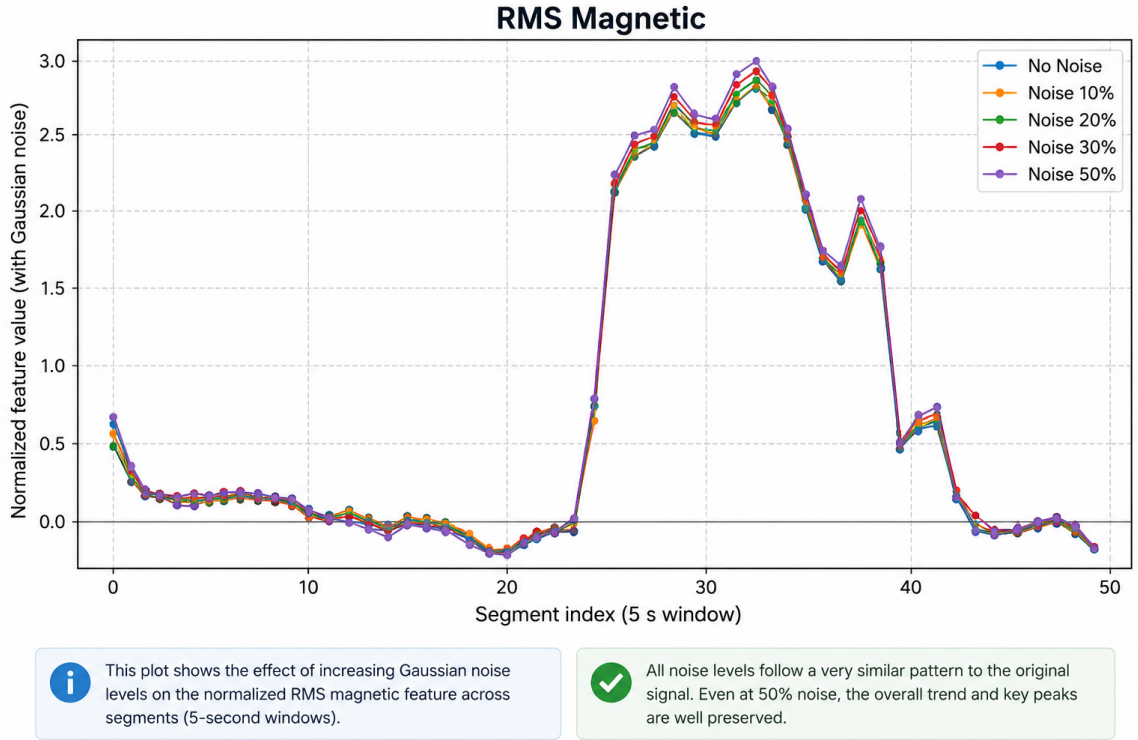


Figure 4.2: Effect of Gaussian noise injection on the normalized RMS feature of the magnetic field signal. Despite increasing perturbation levels, the signal exhibits higher stability compared to the current channel.

Although some degradation is expected when moving across devices, the observed performance stability indicates that the extracted features capture behavior-related characteristics rather than device-specific fingerprints. This property is essential for any realistic side-channel monitoring or intrusion detection system intended for deployment at scale.

Summary of Robustness Analysis. Taken together, the robustness experiments conducted in this chapter demonstrate that the reduced feature model obtained through Recursive Feature Elimination achieves a favorable balance between accuracy, stability, and deployability. Under controlled conditions, the model delivers high classification performance, while under increasing noise levels and device variability it exhibits graceful degradation rather than abrupt failure.

This behavior is particularly important from an operational standpoint. In realistic environments, noise intensity and hardware conditions are rarely static, and systems must tolerate a degree of uncertainty without producing unstable or misleading outputs. The combination of RMS-based preprocessing, robust scaling, and feature reduction contributes to this stability by constraining the variability of the learned representation.

Nevertheless, these results also reveal an inherent limitation: robustness is achieved within a bounded envelope defined by the training conditions and preprocessing choices. As noise characteristics or hardware properties drift beyond this envelope, performance degrades progressively, indicating that static pipelines can only provide limited resilience.

Impact of Gaussian Noise on Side-Channel Traces

In practical deployment scenarios, side-channel traces are continuously affected by environmental interference, hardware imperfections, and—in adversarial contexts—potentially deliberate perturbations. To explicitly characterize the effect of such disturbances, we systematically inject controlled Gaussian noise into the raw signals prior to segmentation and feature extraction.

The injected noise has zero mean and a variance proportional to the dynamic range of each signal column (i.e., `col.max() - col.min()`), with noise levels set to 10%, 20%, 30%, and 50%. This formulation ensures that the perturbation scales with the signal magnitude and realistically emulates degradation due to electromagnetic interference, power supply instability, or sensor noise.

Consistent with previous studies [44, 38], increasing noise levels significantly affect fine-grained,

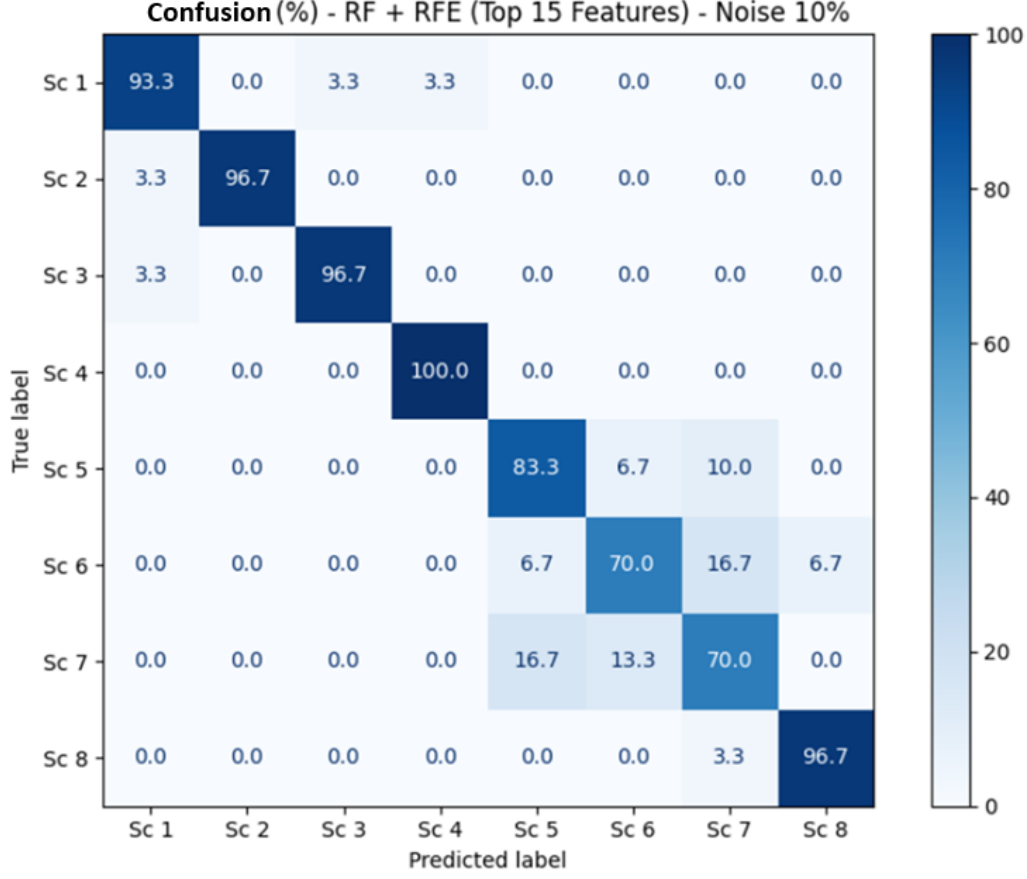


Figure 4.3: Confusion matrix at 10% Gaussian noise. Average accuracy: 88.33%.

window-level predictions, particularly for signals exhibiting non-stationary and bursty behavior. Our experimental results confirm this trend and further highlight a channel-dependent response to noise.

Specifically, magnetic field traces demonstrate superior resilience under high noise levels, maintaining more stable RMS profiles and preserving discriminative structure even at 30% and 50% perturbation. This behavior can be attributed to the fact that electromagnetic emissions capture spatially distributed switching activity, which tends to average out localized disturbances.

Conversely, current-based traces exhibit higher sensitivity to strong perturbations but retain better accuracy under low to moderate noise levels. Current measurements reflect global power consumption and therefore provide cleaner signatures when environmental interference is limited, but they are more susceptible to additive noise when perturbations become severe.

This complementary behavior reinforces the importance of multi-channel acquisition and motivates hybrid strategies that dynamically adapt the relative importance of current and electromagnetic features depending on observed signal conditions.

Column-wise Majority Voting Approach. To further enhance robustness at the decision level, we adopt a *column-wise majority voting* strategy inspired by ensemble decision fusion techniques [24]. Rather than relying on individual window-level predictions, the final scenario label associated with each acquisition is determined by the most frequent class predicted across all its temporal segments.

This approach explicitly exploits the temporal persistence of application behavior. While individual windows may be affected by transient noise, local perturbations, or momentary ambiguities in the feature space, the underlying application scenario typically remains stable over the duration of an acquisition. Majority voting therefore acts as a temporal integration mechanism, aggregating weak and potentially noisy decisions into a more reliable session-level inference.

From a signal processing perspective, this strategy can be interpreted as an implicit denoising step operating at the decision layer rather than at the signal level. Instead of attempting to suppress noise in the raw measurements or feature representations, majority voting mitigates its impact by leveraging redundancy across time. This distinction is important, as it allows robustness to be improved without

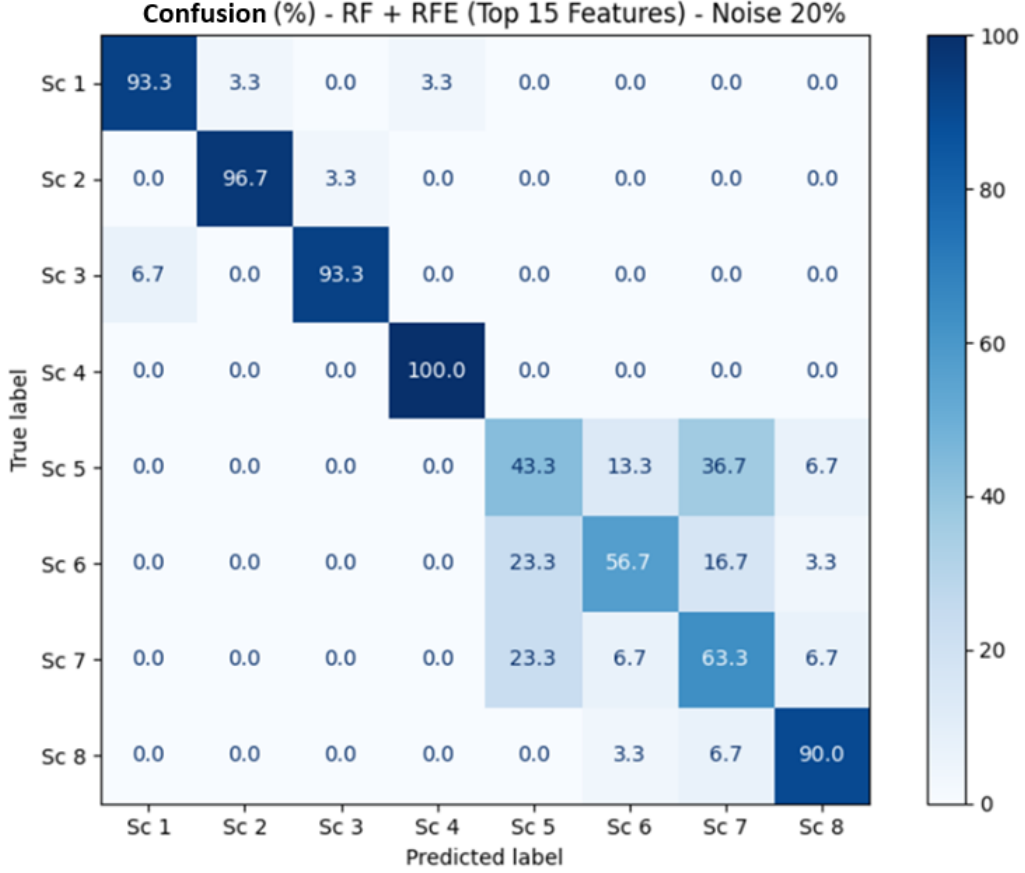


Figure 4.4: Confusion matrix at 20% Gaussian noise. Average accuracy: 79.58%.

modifying the underlying preprocessing or feature extraction pipeline.

Operational Relevance. The column-wise voting strategy is particularly well suited to cyber-physical and IoT monitoring scenarios, where decisions at the level of a *device session* or *application execution interval* are often more meaningful than instantaneous window-level classifications. In such contexts, occasional misclassifications at fine temporal resolution are acceptable, provided that the overall behavioral assessment remains accurate and stable.

By reducing sensitivity to isolated errors, majority voting lowers the likelihood of false alarms and increases trust in monitoring outcomes—an essential requirement for deployment in security-critical systems such as intrusion detection, behavioral profiling, or anomaly monitoring frameworks.

Visualization of Column-wise Confusion Matrices

To quantitatively assess the effect of column-wise majority voting, confusion matrices are computed at the acquisition level by aggregating window-level predictions. These visualizations highlight how temporal aggregation reshapes class separability, reducing cross-scenario confusion and stabilizing classification outcomes under noisy conditions.

In particular, majority voting significantly suppresses sporadic misclassifications observed at the window level, leading to clearer diagonal dominance in the confusion matrices and improved interpretability of classification behavior across scenarios.

Discussion and Concluding Remarks

The column-wise analysis confirms that majority voting provides a substantial robustness gain across a wide range of noise conditions. Under low to moderate noise levels (up to approximately 20%), current-based features combined with majority voting preserve high scenario-level accuracy, validating their suitability for controlled or semi-controlled environments.

As noise intensity increases, magnetic field traces emerge as more resilient, maintaining higher classification stability under severe perturbations. This behavior reflects the complementary nature of the two

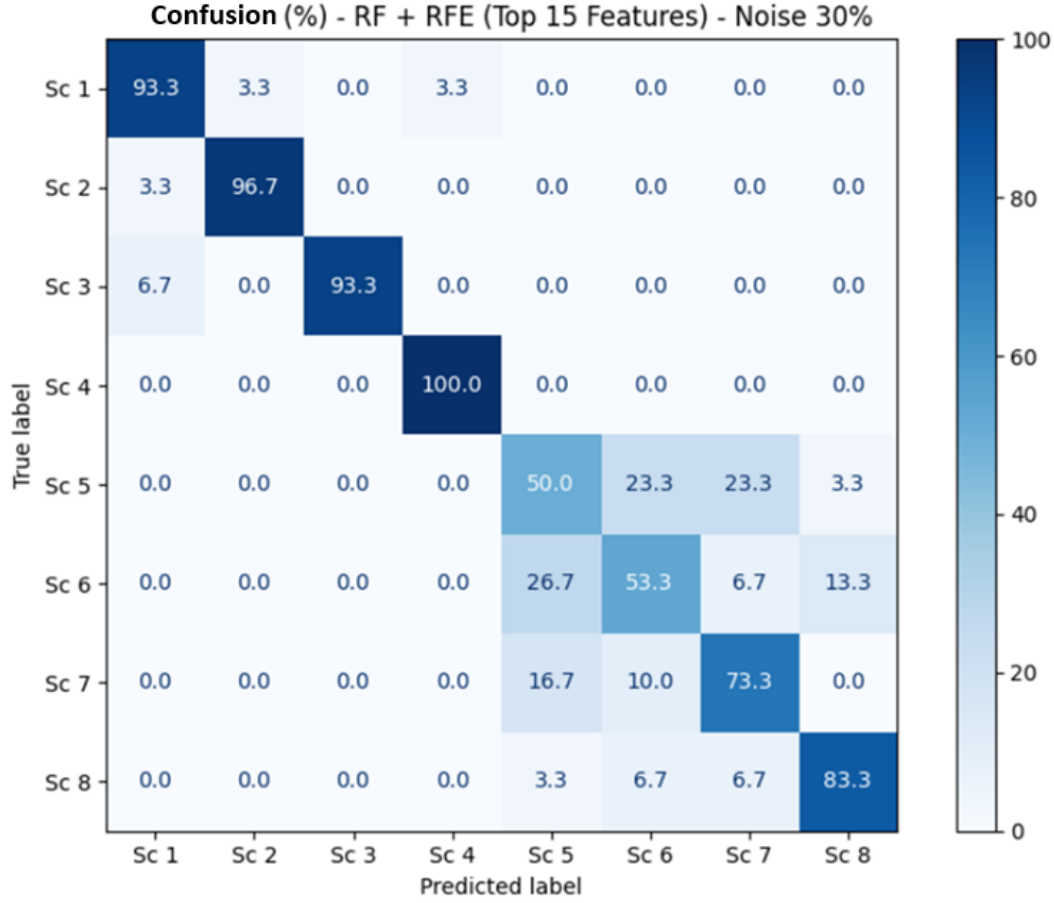


Figure 4.5: Confusion matrix at 30% Gaussian noise. Average accuracy: 80.42%.

physical channels: current measurements offer strong discriminative power under clean conditions, while electromagnetic measurements provide increased robustness in highly noisy environments.

From a deployment perspective, these findings suggest that *signal and feature selection should not be treated as static design choices*, but rather as context-dependent decisions influenced by environmental conditions and operational constraints. In particular:

- For noise levels below 20%, majority voting applied to current-based features offers high accuracy with reduced computational overhead.
- For higher perturbation levels (30% and above), magnetic field features combined with robust majority voting are preferable, especially for scenarios involving highly dynamic network protocols such as HTTP or FTPS.

These results are consistent with hybrid and ensemble-based approaches discussed in the literature [44] and confirm that layered decision fusion across time and signal modalities significantly enhances resilience in side-channel-based intrusion detection systems.

At the same time, majority voting remains a *static* aggregation mechanism: it improves robustness by exploiting redundancy, but it does not adapt the underlying representation or decision process in response to changing conditions. This limitation further motivates the exploration of adaptive strategies, such as confidence-aware fusion or dynamic signal weighting, which are discussed in the subsequent chapters as a natural evolution of the framework presented here.

4.5.1 Discussion

This chapter has presented a comprehensive empirical evaluation of the proposed side-channel classification framework, highlighting both its strengths and its practical trade-offs under realistic operating conditions. The results confirm that learning-based profiling of electromagnetic and current side-channel traces can achieve high accuracy even when constrained by limited computational resources and non-ideal measurement environments.

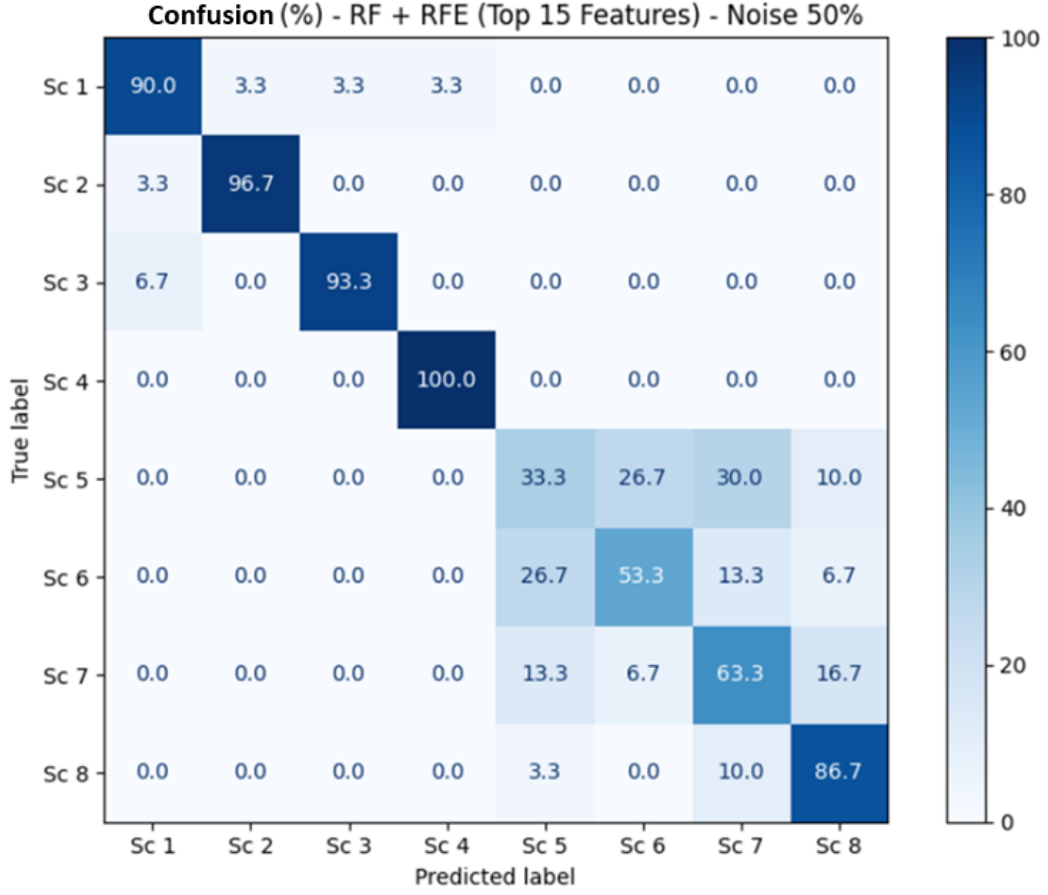


Figure 4.6: Confusion matrix at 50% Gaussian noise. Average accuracy: 77.08%.

A central outcome of the analysis is the clear advantage of feature-reduced models over full-feature configurations. By adopting a compact representation derived through Recursive Feature Elimination, the classification pipeline achieves performance comparable to—or in some cases exceeding—that of higher-dimensional models, while significantly reducing inference complexity. This finding is particularly relevant for real-time deployment on embedded and IoT platforms, where memory, latency, and energy constraints are critical design factors.

The robustness experiments further demonstrate that the proposed pipeline maintains stable performance under substantial noise injection and across heterogeneous acquisition conditions. The combined use of RMS-based aggregation, robust normalization, and temporal decision fusion enables resilience against both environmental interference and signal variability, supporting reliable operation in dynamic and potentially adversarial settings. These results are consistent with, and extend, prior work on side-channel-based behavioral profiling [38], providing stronger empirical evidence for the feasibility of lightweight intrusion detection and monitoring systems grounded in physical leakage analysis.

Beyond quantitative performance, the analysis in this chapter exposes important structural insights. In particular, it shows that robustness in side-channel analysis cannot be attributed to a single component, but rather emerges from the interaction between preprocessing, feature representation, selection strategy, and temporal aggregation. At the same time, the observed degradation patterns under severe noise and varying conditions highlight the intrinsic limits of static pipelines, even when carefully engineered.

Taken together, these findings establish a solid empirical baseline for the remainder of the thesis. They validate the effectiveness of the proposed static framework while simultaneously revealing its bounded robustness, thereby motivating the need for more adaptive and context-aware strategies. The insights derived here naturally pave the way for subsequent investigations into dynamic feature weighting, adaptive temporal modeling, and learning-based control mechanisms aimed at improving robustness and explainability in security-oriented side-channel monitoring applications.

Chapter 5

Side-Channel Leakage as an Information Control Problem

5.1 Toward a Feature-Aware Approach in Noisy Environments

The analyses presented in the previous sections have enabled a systematic characterization of a side-channel classification pipeline primarily based on RMS features, evaluated under both clean conditions and in the presence of additive noise. The obtained results show that, while this representation provides good global stability and a certain degree of robustness at moderate noise levels, it exhibits intrinsic structural limitations when discrimination is required at fine temporal granularity.

In particular, the window-level classification accuracy achieved using RMS-based features remains limited even in noise-free conditions. High performance at the signal level is attained mainly through temporal aggregation mechanisms, such as majority voting, rather than through strong discriminative capability at the level of individual windows. This observation suggests that the effectiveness of the pipeline relies more on statistical integration over time than on the informativeness of local feature representations.

Although temporal aggregation represents an effective strategy in controlled settings, its reliance introduces potential methodological risks in the presence of variable acquisition conditions. Specifically, when a model trained on clean data is applied to signals affected by noise or acquired under different environmental or operational conditions, the distribution of the extracted features may undergo significant shifts with respect to those observed during training. This phenomenon, commonly referred to as *domain shift* or *dataset shift* [36, 31], represents a critical issue in side-channel classification systems, as it can weaken the correspondence between local predictions and the actual underlying scenario, ultimately reducing the effectiveness of temporal aggregation strategies. Similar effects have been observed in the context of learning-based side-channel analysis, where variations in acquisition conditions and noise levels can significantly impact model generalization [32].

In this context, the primary bottleneck of an RMS-only pipeline cannot be attributed solely to the presence of noise, nor to numerical instability or scaling issues. Rather, it stems from the limited informational richness of the adopted representation. This shift lays the groundwork for more reliable generalization in realistic and dynamically evolving operational scenarios, and motivates the exploration of adaptive mechanisms capable of actively regulating the information exposed through side-channel measurements.

Based on these considerations, the focus of the analysis is now shifted from noise reduction alone toward a *feature-aware* perspective. In this framework, the objective is not to eliminate noise from the signal, but to design feature representations capable of preserving and explicitly encoding discriminative information even under degraded conditions and in the presence of mismatches between training and testing domains. Within this perspective, denoising plays a secondary and supportive role, while the primary contribution is achieved through the selection and optimization of time-domain and frequency-domain features characterized by higher statistical stability and stronger physical interpretability.

The underlying hypothesis is that a richer and more structured feature representation can mitigate the impact of noise and reduce the effects of domain shift not through aggressive signal filtering or ex-post temporal aggregation, but through the intrinsic resilience of the extracted descriptors. In this way, system robustness does not emerge merely from the statistical integration of predictions over time, but from the quality of the information captured at the level of individual windows. This shift lays the groundwork for

more reliable generalization in realistic and dynamically evolving operational scenarios, and motivates the exploration of adaptive mechanisms capable of actively regulating the information exposed through side-channel measurements.

5.2 Experimental Comparison Framework

To rigorously assess the effectiveness of the proposed feature-aware approach, the subsequent analyses are structured around a set of well-defined experimental comparison scenarios. The objective is to disentangle the impact of noise from that of feature representation and training strategy, and to explicitly evaluate their interaction under progressively degraded conditions.

As a first reference, the *clean-clean* configuration is considered, in which both training and testing are performed on noise-free data. This setting represents an idealized upper bound and allows the intrinsic discriminative capability of the extracted features to be evaluated in the absence of external disturbances. It provides a fundamental baseline against which all noisy configurations can be consistently interpreted.

The *noise-noise* configuration is then analyzed, where both the training and testing phases are conducted on data corrupted by noise at the same signal-to-noise ratio (SNR). This scenario preserves statistical homogeneity between the two phases and enables the assessment of the classifier’s ability to operate on degraded side-channel signatures, assuming that noise characteristics are known and stationary.

The *clean-noise* configuration represents a substantially more challenging and practically relevant setting. In this case, the model is trained on clean data and evaluated on noisy signals, resulting in a pronounced statistical mismatch between training and testing distributions. As highlighted in the previous chapter, this setup exposes the effects of domain shift in a particularly clear manner and constitutes a critical stress test for evaluating the true generalization capability of the classification pipeline.

Finally, the *clean+noise-noise* configuration is introduced as the most realistic from an application-oriented perspective. In this scenario, the model is trained on a mixed dataset composed of both clean and noisy samples, with the goal of learning noise-aware representations and more robust decision boundaries. Evaluation is then performed exclusively on noisy data, which is assumed to represent the operational context of interest. This setting allows the effectiveness of representation design and training strategy to be assessed under conditions that closely resemble real-world deployment.

All experimental comparisons are conducted across multiple noise levels, corresponding to SNR values of 20 dB, 10 dB, and 0 dB. This progressive degradation enables a systematic analysis of performance trends and robustness thresholds, highlighting how different scenarios and feature representations respond to increasing noise severity.

5.3 Objectives of the Subsequent Analysis

Based on the experimental framework outlined above, the following sections focus on a detailed analysis of features extracted in both the time and frequency domains, with the aim of evaluating their contribution to scenario discrimination under noisy conditions. The objective is not limited to measuring improvements in aggregate performance metrics, but rather to understand which features exhibit genuine robustness, which are redundant, and to what extent representation design can compensate for noise-induced degradation.

In particular, the analysis seeks to determine whether robustness can be achieved at the level of individual windows through informative and stable feature representations, rather than relying primarily on temporal aggregation mechanisms such as majority voting. By shifting the focus from outcome-based performance comparison to representation-level understanding, this approach enables a deeper assessment of the role played by feature design in the construction of reliable and realistic side-channel classification systems.

This perspective lays the groundwork for subsequent stages of feature optimization and guided selection, and provides the conceptual bridge toward adaptive strategies that explicitly account for variability and uncertainty in the observed side-channel information.

5.4 Analysis of Selected Features under Clean Conditions

Before explicitly introducing noise into the system, it is essential to perform a detailed analysis of the role played by individual features in separating application scenarios and to understand which descriptors effectively drive the classifier’s decision process under ideal conditions. In this preliminary phase, the

objective is not limited to assessing classification performance, but rather to identify the informative components underlying class discrimination and to verify the coherence of the model behavior as a function of feature dimensionality.

To this end, a feature importance analysis was conducted using a Random Forest classifier, considering jointly all features extracted from the available acquisition files. The analysis focuses on identifying a compact and common subset of features that maximizes scenario separability in clean conditions, thus providing an interpretable baseline for subsequent robustness evaluations. Random Forest was selected due to its intrinsic capability to estimate feature relevance while accounting for non-linear interactions among descriptors.

The analysis resulted in the identification of a reduced set of 20 features, comprising statistics extracted in both the time and frequency domains and derived from both acquisition channels, namely current (ch1) and electromagnetic field (ch2). The selected descriptors include first-order statistics such as the mean value, measures of signal variability and dispersion, distribution extremes (e.g., peak-to-peak), as well as shape-related descriptors such as skewness and kurtosis. In addition, several frequency-domain features are retained, including spectral energy and indicators related to the distribution of spectral content.

The joint presence of temporal and spectral features, evenly distributed across the two acquisition channels, suggests that scenario discrimination does not rely on a single dominant descriptor. Instead, it emerges from the interaction between global energy levels, intra-window signal variability, and spectral structure associated with application-level activity. This behavior confirms that, under clean conditions, the classifier primarily exploits global energetic differences, refined by statistical and frequency-related information, providing a solid reference point for evaluating feature robustness in the presence of noise.

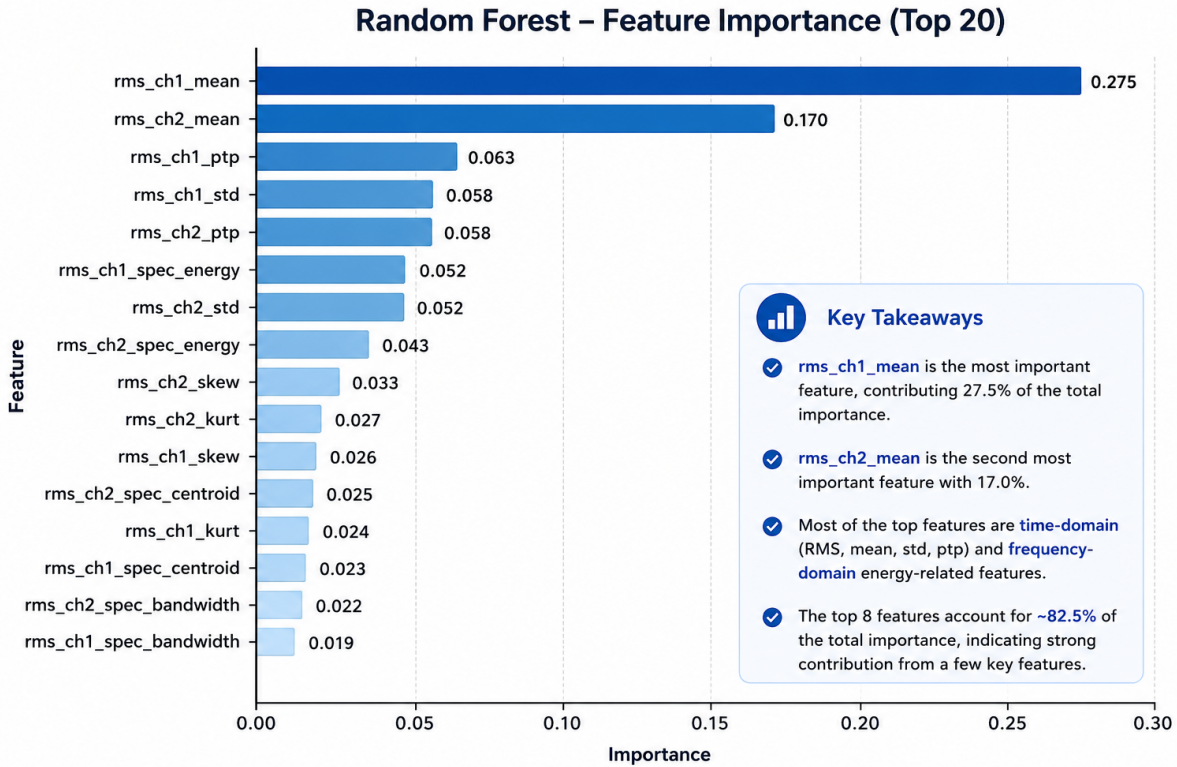


Figure 5.1: Random Forest feature importance ranking (top 20) obtained under clean conditions.

Figure 5.1 reports the ranking of the most relevant features according to the Random Forest importance metric in clean conditions. The analysis clearly highlights that first-order statistics, and in particular the mean value of the RMS signal computed on both acquisition channels, constitute the most influential descriptors in the classification process. This result indicates that, in the absence of noise, the separation among application scenarios is largely driven by differences in the average energy level of the signal, which directly reflects the intensity and nature of the underlying workload.

In addition to mean-related features, measures of signal variability and amplitude excursion, such as standard deviation and peak-to-peak value, exhibit a significant contribution for both current and electromagnetic channels. These descriptors capture the internal dynamics of the signal within each

temporal window, providing complementary information with respect to the global energy level and enabling finer discrimination between scenarios characterized by similar average behavior but different temporal structures.

Frequency-domain features, including spectral energy and indicators related to the distribution of spectral content, appear in the intermediate portion of the ranking. Their contribution, while less dominant, is non-negligible and suggests that spectral characteristics play a supporting role in modeling scenario-specific signal structures. These features enhance the classifier’s ability to capture differences not fully expressed in the time domain alone.

Finally, higher-order distribution descriptors such as skewness and kurtosis exhibit lower importance values compared to energetic and variability-related features. This behavior is consistent with the nature of the problem in clean conditions, where scenario differences are already well captured by amplitude- and dispersion-related descriptors, reducing the marginal contribution of higher-order statistics. Nevertheless, their presence in the selected subset indicates that they provide complementary information that may become more relevant under more complex or degraded acquisition conditions.

Overall, the feature importance analysis confirms that, under clean conditions, the classifier primarily relies on global energetic descriptors, refined by temporal variability and spectral information. This interpretation provides a clear understanding of the model behavior and establishes a crucial reference for the subsequent analysis, where the stability and discriminative power of these features will be evaluated as the signal-to-noise ratio progressively decreases.

5.5 Analysis of Selected Features and Clean-Condition Performance

Before explicitly introducing noise into the system, it is necessary to conduct an in-depth analysis of the role played by the extracted features in separating the application scenarios and to understand which descriptors effectively drive the classifier’s decision-making process under ideal conditions. In this preliminary phase, the objective is not limited to assessing classification performance, but rather to identify the informative components underlying class discrimination and to verify the consistency of the classifier behavior as the feature dimensionality varies.

To this end, a feature importance analysis was performed using a Random Forest classifier, jointly considering all features extracted from the available acquisitions and focusing on the identification of a common subset capable of maximizing inter-scenario separability. This global analysis highlights which descriptors contribute most significantly to classification under clean conditions, thereby providing a fundamental interpretative baseline for subsequent experiments in noisy environments.

The analysis resulted in the identification of a compact set of 20 features, comprising statistics computed in both the time and frequency domains and extracted from both acquisition channels, namely current (ch1) and electromagnetic emission (ch2). The selected features include first-order statistics such as the mean, measures of dispersion and variability, distribution extremes, and higher-order shape descriptors such as skewness and kurtosis. These are complemented by frequency-domain features, including spectral energy and descriptors related to the distribution of power across the spectrum.

The joint presence of temporal and spectral features, balanced across the two acquisition channels, suggests that discrimination among application scenarios is not driven by a single dominant descriptor, but rather emerges from the interaction between global energy levels, intra-window signal variability, and spectral structure associated with the executed workload. In clean conditions, this confirms that the classifier primarily exploits global energetic differences, refined through statistical and frequency-domain information, providing a solid reference point for evaluating feature robustness under degraded conditions.

The feature importance ranking obtained through the Random Forest model under clean conditions, reported in Figure 5.2, clearly shows that first-order statistics, and in particular the mean value of the RMS computed on both acquisition channels, represent the most influential descriptors in the classification process. This result indicates that, in the absence of noise, scenario separation is largely driven by differences in average signal energy, which directly reflect the intensity of the activity associated with different application workloads.

Alongside mean-based features, measures of signal variability and amplitude excursion, such as standard deviation and peak-to-peak values extracted from both current and electromagnetic channels, play a significant role. These descriptors capture the internal dynamics of the signal within each temporal window, providing complementary information with respect to average energy and allowing the classifier

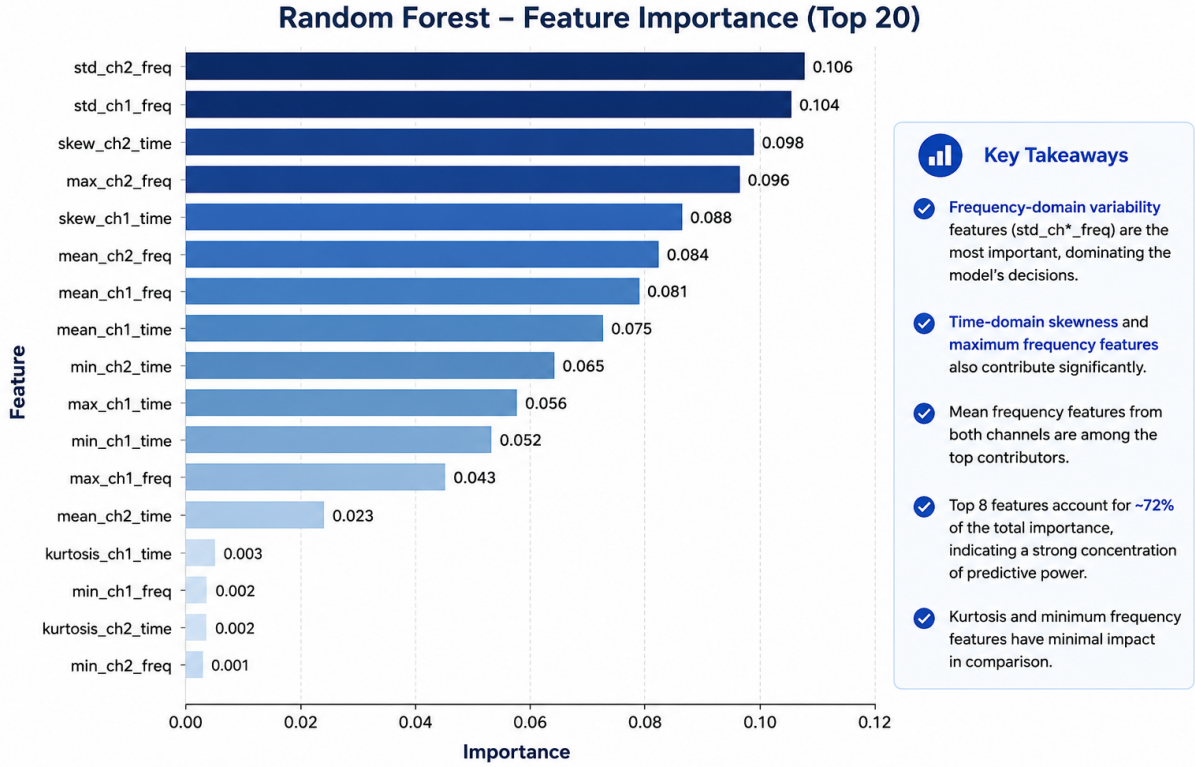


Figure 5.2: Random Forest feature importance ranking under clean conditions. The plot reports the top 20 most influential features extracted from both current (ch1) and electromagnetic (ch2) channels, including time-domain and frequency-domain descriptors.

to refine separation among scenarios exhibiting similar mean values but different temporal behaviors.

Frequency-domain features, including spectral energy and descriptors related to spectral distribution, show a more moderate yet non-negligible contribution. Their intermediate ranking position suggests that, while they do not constitute the primary discriminative factor in clean conditions, they support the modeling of structural differences in the signal associated with distinct application scenarios.

Finally, higher-order distribution statistics such as skewness and kurtosis exhibit lower importance compared to energetic and variability-related features. This behavior is consistent with the nature of the problem under ideal acquisition conditions, where differences among scenarios are already well captured by amplitude- and dispersion-related descriptors. Nevertheless, their presence in the ranking indicates that they provide complementary information that may become relevant in more complex or degraded scenarios.

Overall, the feature importance analysis confirms that, under clean conditions, the classifier relies primarily on global energetic descriptors refined by temporal and spectral information. This interpretation provides a clear understanding of the model's behavior and establishes a reference framework for assessing how the discriminative power of these features evolves as the signal-to-noise ratio decreases.

5.6 Classification Robustness with Respect to Dataset Dimensionality

Once the feature set was defined, the pipeline was evaluated under clean conditions using a Random Forest classifier, enforcing a strict separation between training and test files. When considering the full available dataset (240 files in total, with 200 used for training and 40 for testing), the system achieves an accuracy of approximately 80% at the window level, as shown in the confusion matrix reported in Figure 5.3.

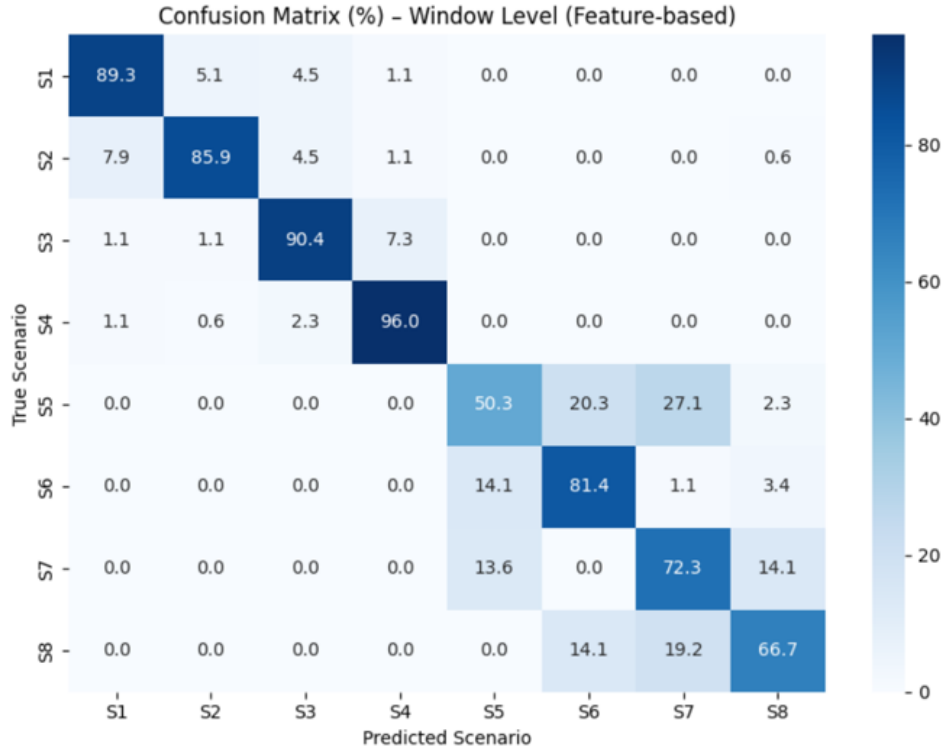


Figure 5.3: Window-level confusion matrix (%) obtained under clean conditions using the selected feature set and a Random Forest classifier. Rows represent true scenarios, while columns indicate predicted scenarios.

This result indicates that, in the absence of noise, the proposed preprocessing and feature extraction pipeline enables effective separation among application scenarios at fine temporal granularity. However, performance degradation is more pronounced for scenarios S5 to S8, which exhibit higher levels of confusion compared to scenarios S1 to S4.

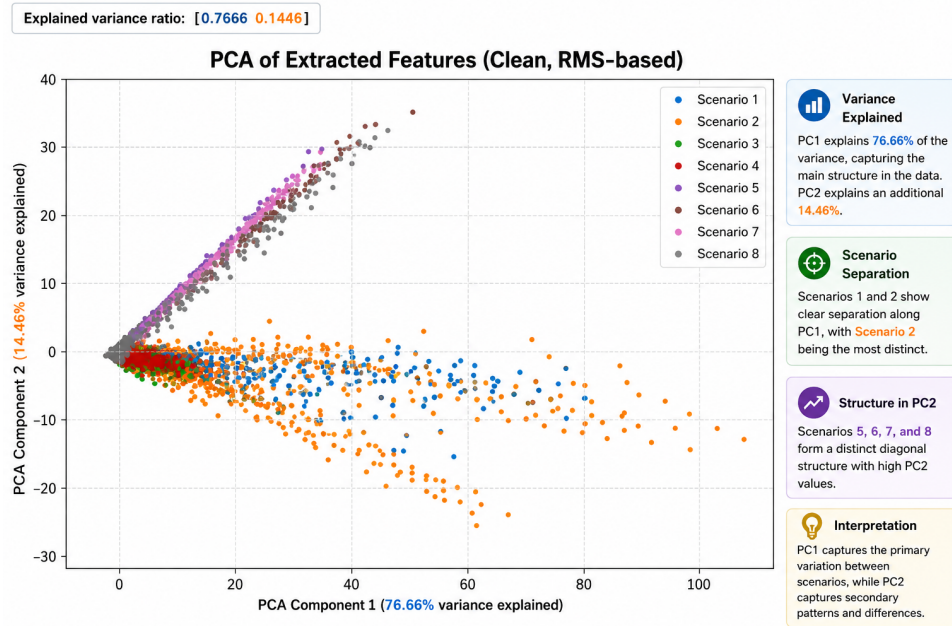


Figure 5.4: PCA projection of RMS-based features extracted under clean conditions. Each point corresponds to a window-level feature vector, with colors indicating different application scenarios. The first two principal components explain over 90% of the total variance.

Figure 5.4 reports the PCA projection of the RMS-based features extracted under clean conditions. Rather than exhibiting weak or fragmented separability, the reduced-dimensional representation reveals

a highly structured feature space: the first two principal components jointly explain over 90% of the total variance, indicating the presence of a small number of dominant latent factors governing side-channel behavior.

The PCA projection provides a geometric interpretation of the separability differences observed across scenarios. The first two principal components account for approximately 76.7% and 14.4% of the variance, respectively, suggesting that the feature space is largely dominated by factors related to global energetic properties of the signal.

Scenarios S1–S4 occupy well-separated regions along the first principal component, indicating that they are characterized by marked differences in average energy levels and signal variability. These characteristics are effectively captured by RMS-based features, resulting in stable and well-defined clusters in feature space.

In contrast, scenarios S5–S8 exhibit more compact and partially overlapping distributions, with trajectories developing along similar directions and reduced dispersion along the first principal component. In these cases, differences in global energy are less pronounced, while variations along the second principal component are more gradual and less discriminative.

This geometry suggests that, for scenarios S5–S8, discrimination is driven by subtler and more localized signal characteristics that are only partially captured by an RMS-based representation. As a consequence, these scenarios are intrinsically more prone to overlap in feature space, making window-level classification more challenging and explaining the higher confusion observed in the corresponding confusion matrices.

Focus on Feature Selection Strategy

While the PCA analysis suggests that scenarios S5–S8 could benefit from a more targeted and scenario-specific feature representation, it is important to distinguish between what is optimal from an informational standpoint and what is feasible in a realistic operational setting. In principle, feature selection optimized independently for each scenario could significantly improve classification performance for classes characterized by subtle and overlapping side-channel signatures.

However, a scenario-specific feature importance approach implicitly assumes prior knowledge of the class label of the observed signal, an assumption that does not hold during inference in real-world applications. Consequently, such a strategy is not methodologically applicable in an operational side-channel classification system, where the classifier must operate on unlabeled signals and rely on a single, shared feature representation.

To preserve coherence with a realistic inference context, feature selection in this work is therefore conducted at a global level, jointly considering all scenarios. This approach defines a unified feature space that can be consistently employed during testing without additional assumptions. Inevitably, such a global feature set may be more representative for certain scenarios and less discriminative for others, particularly for those characterized by similar energetic and statistical signatures.

In this context, scenario-level feature importance analysis is not used to redefine the feature set, but rather as an interpretative tool to understand the relative contribution of different descriptors to each scenario. This analysis highlights the limitations of a global representation and identifies which characteristics are more or less informative for specific classes, providing valuable insights for the evolution toward adaptive and feature-aware approaches without compromising inference validity.

5.7 Simulation of Noisy Operating Environments

The objective of the next experimental phase is to investigate how the introduction of noise affects the distribution of extracted features, to identify which components are most sensitive to signal degradation, and to assess whether performance can be recovered through advanced feature engineering techniques or alternative preprocessing strategies. Within this perspective, noise is not treated merely as a degrading factor, but rather as an analytical tool to probe the intrinsic robustness of the pipeline and to guide the design of more resilient solutions, both from a defensive standpoint and for practical side-channel-based intrusion detection applications.

To address these objectives in a structured manner, a new preprocessing pipeline was designed and implemented, as illustrated in the following figure. The pipeline operates directly on raw physical signals acquired from the system under analysis, namely current consumption and electromagnetic emissions. As a first step, the initial background noise phase corresponding to the first 10 seconds of acquisition

is removed, ensuring that subsequent analyses focus exclusively on the active execution phase of the application.

The signal is then segmented using a fixed-length temporal window of 100 ms, with a 50% overlap between consecutive windows. This windowing strategy preserves temporal continuity while increasing analysis granularity, enabling a fine-grained characterization of signal dynamics and improving statistical robustness without excessively inflating the dimensionality of the dataset.

A key element of the proposed pipeline is the controlled injection of noise directly at the raw signal level, prior to feature extraction. Noise is modeled through a variable signal-to-noise ratio (SNR), with values set to 20 dB, 10 dB, and 5 dB, in order to simulate progressively more challenging and realistic operating conditions while maintaining physical coherence with the observed side-channel phenomena. Injecting noise at the raw signal stage allows for a faithful assessment of the pipeline’s ability to extract discriminative information under degraded conditions, avoiding artifacts that could arise from noise addition in already aggregated or transformed domains.

For each temporal window, a systematic set of time-domain and frequency-domain features is then extracted, including statistical and spectral descriptors designed to capture subtle signal variations even at low SNR levels. The resulting feature vectors are used to train and evaluate a Random Forest classifier. When required, a signal-level decision can be obtained through majority voting across window-level predictions; however, in this phase, temporal aggregation is treated as an optional post-processing mechanism rather than as a structural solution to robustness limitations.

This preprocessing and analysis pipeline constitutes the methodological foundation for all subsequent experiments. It enables controlled and reproducible comparisons between clean and noisy conditions, supports the investigation of feature sensitivity to noise, and provides a flexible framework for exploring alternative robustness strategies and defensive mechanisms in side-channel analysis.

5.8 Noise Model and Signal-to-Noise Ratio (SNR)

To systematically evaluate the robustness of the proposed side-channel classification pipeline under realistic operating conditions, a controlled additive noise model was introduced and applied directly to the acquired raw signals. In particular, noise was modeled as *Additive White Gaussian Noise* (AWGN), a widely adopted choice in signal processing and in robustness studies of side-channel systems, as it enables precise control over the disturbance level while ensuring experimental reproducibility.

The noise level is parameterized through the *Signal-to-Noise Ratio* (SNR), defined as the ratio between the average power of the useful signal and the average power of the additive noise. In logarithmic form, the SNR is expressed in decibels (dB) as:

$$\text{SNR}_{\text{dB}} = 10 \log_{10} \left(\frac{P_{\text{signal}}}{P_{\text{noise}}} \right), \quad (5.1)$$

where P_{signal} denotes the power of the original signal and P_{noise} the power of the injected noise. High SNR values correspond to mildly disturbed conditions, while lower SNR values represent progressively more degraded scenarios, up to situations in which the noise power becomes comparable to that of the signal.

Noise was injected directly into the raw physical signals—namely current consumption and electromagnetic emissions—prior to any windowing, RMS computation, or feature extraction. This design choice is crucial to preserve the physical coherence of the side-channel phenomenon, as it allows noise to propagate naturally through all subsequent preprocessing stages, closely emulating real-world acquisition conditions. Conversely, injecting noise at later stages (e.g., after RMS aggregation) would introduce mathematical artifacts that are not representative of the underlying physical process.

Multiple SNR levels were considered, specifically 20 dB, 10 dB and 5 dB, in order to span a wide range of operating conditions—from moderately noisy environments to highly critical scenarios in which the useful signal is severely degraded.

The following figures provide a qualitative illustration of the effect of additive noise on the acquired raw signals, highlighting how progressive degradation of the SNR impacts different signal components. As the noise level increases, high-frequency local fluctuations become progressively masked, while lower-frequency and slowly varying components tend to be more resilient. This behavior is consistent with the physical nature of the side-channel signals and motivates a deeper analysis of how different feature families respond to noise in subsequent sections.

5.9 Qualitative Impact of Noise on Raw Side-Channel Signals

To gain a physical understanding of how additive disturbances affect the information content of side-channel measurements, a qualitative analysis was conducted directly on the raw acquired signals. Figures 5.5–5.8 report a visual comparison between clean signals and their noisy counterparts at different SNR levels (20 dB, 10 dB, and 5 dB), for both the current and electromagnetic (EM) channels.

5.9.1 Current Signal: Robust vs. Fragile Scenarios

Figure 5.5 illustrates the effect of additive noise on the current signal for Scenarios 1–4, which are representative of applications characterized by relatively stable execution patterns. In these cases, noise mainly manifests as a perturbation superimposed on the local oscillations of the signal. As the SNR decreases, short-term fluctuations and high-frequency components become progressively less distinguishable, leading to an apparent increase in instantaneous variability.

Despite this degradation, the global temporal structure of the signal—intended as its average level, overall envelope, and long-term statistical behavior—remains largely preserved even at low SNR values. This observation indicates that low-frequency components and global energetic properties are intrinsically more robust to additive noise than localized, rapidly varying features.

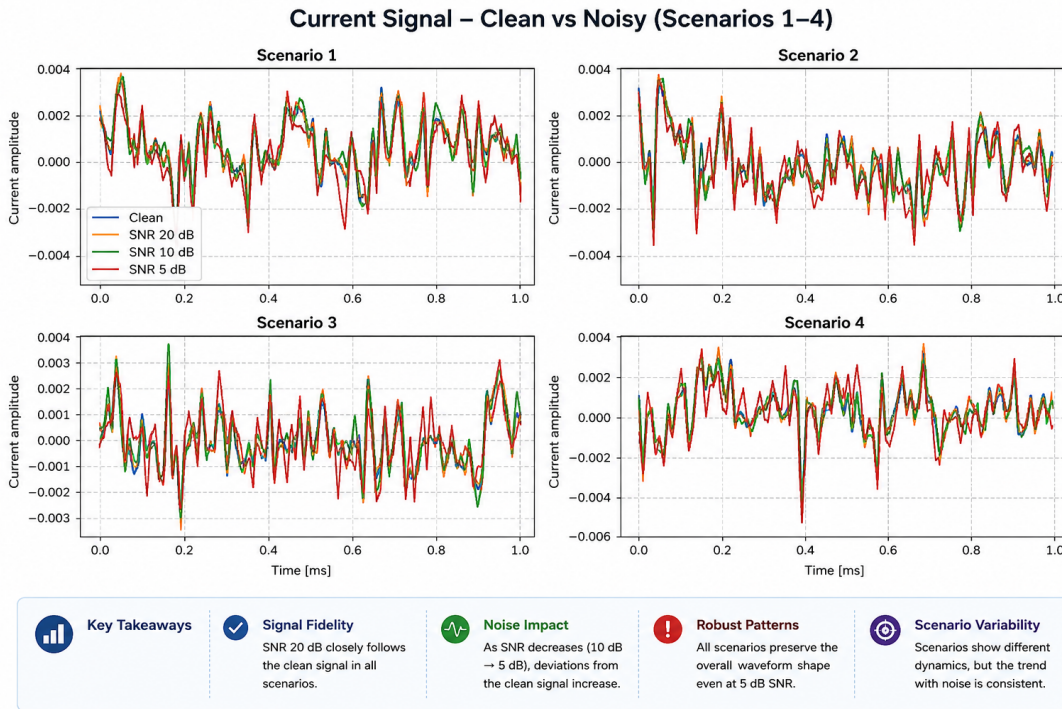


Figure 5.5: Current signal under clean and noisy conditions (SNR = 20 dB, 10 dB, 5 dB) for Scenarios 1–4. Noise primarily affects short-term fluctuations, while global signal structure remains preserved.

A markedly different behavior emerges for Scenarios 5–8, shown in Figure 5.6. In these cases, the clean signal already exhibits higher variability and weaker global structure. The injection of noise amplifies this variability to a level comparable with the signal amplitude itself, even at intermediate SNR values. As a consequence, the distinction between signal-driven variations and noise-induced perturbations becomes increasingly blurred, and the global structure of the current signal is more easily disrupted.

This behavior highlights that scenarios characterized by weaker low-frequency dominance are intrinsically more vulnerable to acquisition noise, as their discriminative information is less concentrated in robust global components.

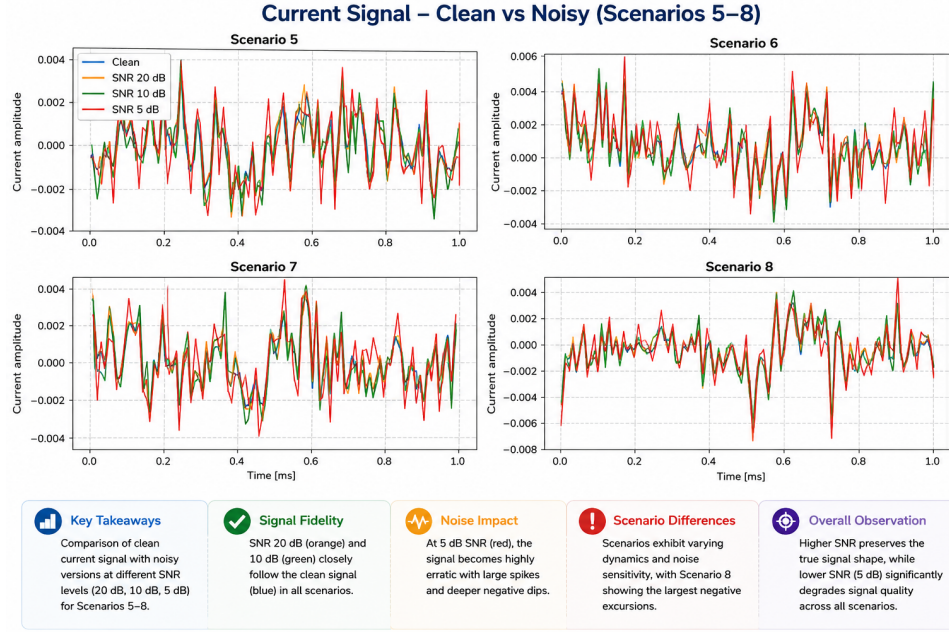


Figure 5.6: Current signal under clean and noisy conditions (SNR = 20 dB, 10 dB, 5 dB) for Scenarios 5–8. Noise significantly alters both local fluctuations and global signal structure.

5.9.2 Electromagnetic Signal: Increased Sensitivity to Noise

The electromagnetic channel exhibits an overall higher sensitivity to additive noise compared to the current signal. Figure 5.7 shows that, even for Scenarios 1–4, the EM signal presents larger amplitude excursions and higher instantaneous variability. As the SNR decreases, noise rapidly dominates short-term dynamics, masking fine temporal details and increasing dispersion.

Nevertheless, for scenarios with regular execution patterns, the average amplitude level and coarse temporal organization of the EM signal remain observable. This suggests that, although more sensitive to noise, the EM channel still preserves global statistical properties that can be exploited by feature-based representations integrating information over extended temporal windows.

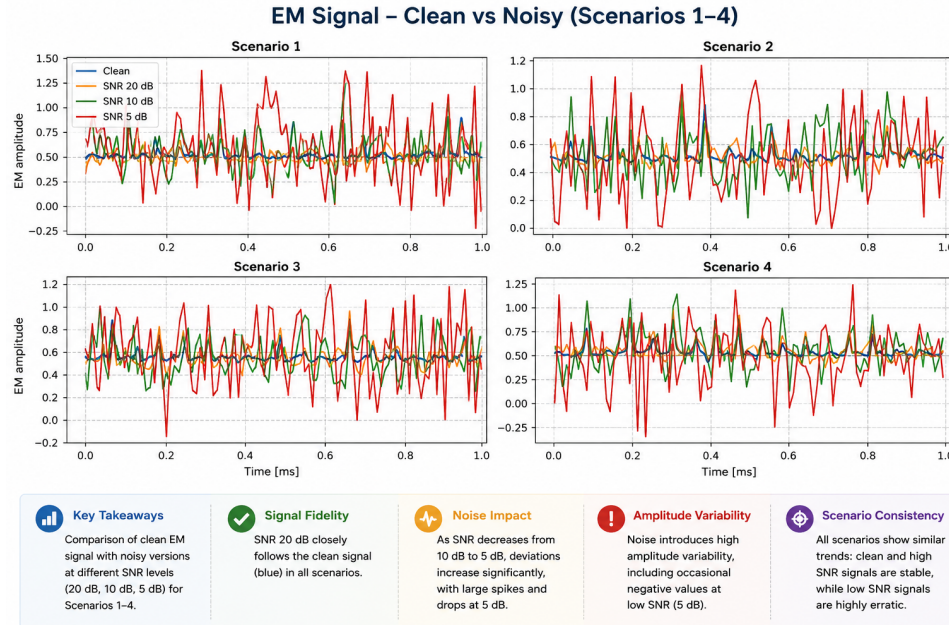


Figure 5.7: Electromagnetic signal under clean and noisy conditions (SNR = 20 dB, 10 dB, 5 dB) for Scenarios 1–4. Noise increases dispersion but preserves coarse statistical structure.

In contrast, Figure 5.8 highlights the pronounced vulnerability of the EM signal for Scenarios 5–8,

and in particular for Scenario 7. Already under clean conditions, the signal exhibits irregular temporal behavior and limited dominance of low-frequency components. As noise is introduced, especially at lower SNR values, noise-induced oscillations overlap almost completely with signal-driven variations.

At intermediate SNR levels (10 dB and 5 dB), local oscillations become difficult to distinguish from noise, while at lower SNR the noisy component dominates not only short-term details but also parts of the global signal behavior. This indicates that, for these scenarios, the informative content is primarily encoded in localized and short-duration patterns that are highly sensitive to noise.

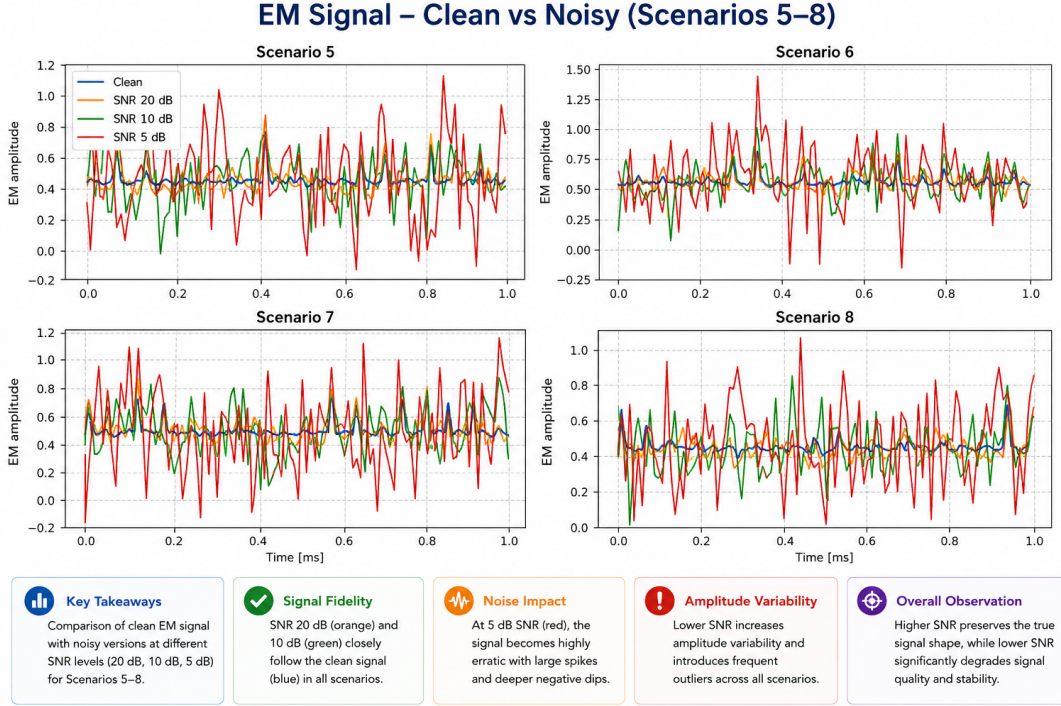


Figure 5.8: Electromagnetic signal under clean and noisy conditions ($\text{SNR} = 20 \text{ dB}$, 10 dB , 5 dB) for Scenarios 5–8. Noise rapidly masks both local and global signal characteristics.

5.10 Impact of Additive Noise on Classification Performance

To evaluate the robustness of the proposed side-channel classification pipeline under degraded acquisition conditions, a noisy-noisy experimental setting was considered, in which both the training and testing phases are performed on signals affected by additive noise. This configuration allows isolating the intrinsic ability of the classifier to learn and discriminate side-channel signatures in the presence of controlled disturbances, without introducing distribution mismatch between training and inference.

Figures 5.9 and 5.10 report the confusion matrices obtained at window level using a Random Forest classifier for Signal-to-Noise Ratio (SNR) values of 20 dB and 10 dB, respectively. In both cases, noise is injected directly into the raw signals prior to preprocessing and feature extraction, ensuring physical consistency with realistic side-channel acquisition scenarios.

When the SNR is further reduced to 10 dB, class separability deteriorates markedly. The confusion matrix becomes increasingly diffuse, with predictions spreading across multiple classes and a substantial reduction of diagonal dominance. Confusion is no longer confined to closely related scenarios, but extends across a broader range of classes, indicating a loss of stable and discriminative patterns.

In this regime, the classifier exhibits an overall window-level accuracy of approximately 20.1%, reflecting the severe impact of noise on the informative content of the extracted features. Even though the training and testing distributions remain statistically aligned, many descriptors that were dominant under clean conditions become progressively less informative, undermining the classifier’s ability to exploit consistent side-channel signatures.

At $\text{SNR} = 20 \text{ dB}$, the confusion matrix still exhibits a partial dominance of diagonal elements, indicating that the classifier retains a limited ability to separate application scenarios under moderate noise conditions. However, systematic confusion patterns already emerge among groups of scenarios characterized by similar energetic and statistical properties. This behavior suggests that, while RMS-based

features preserve a degree of robustness at higher SNR levels, discrimination driven primarily by global energy information becomes progressively less effective as noise increases.

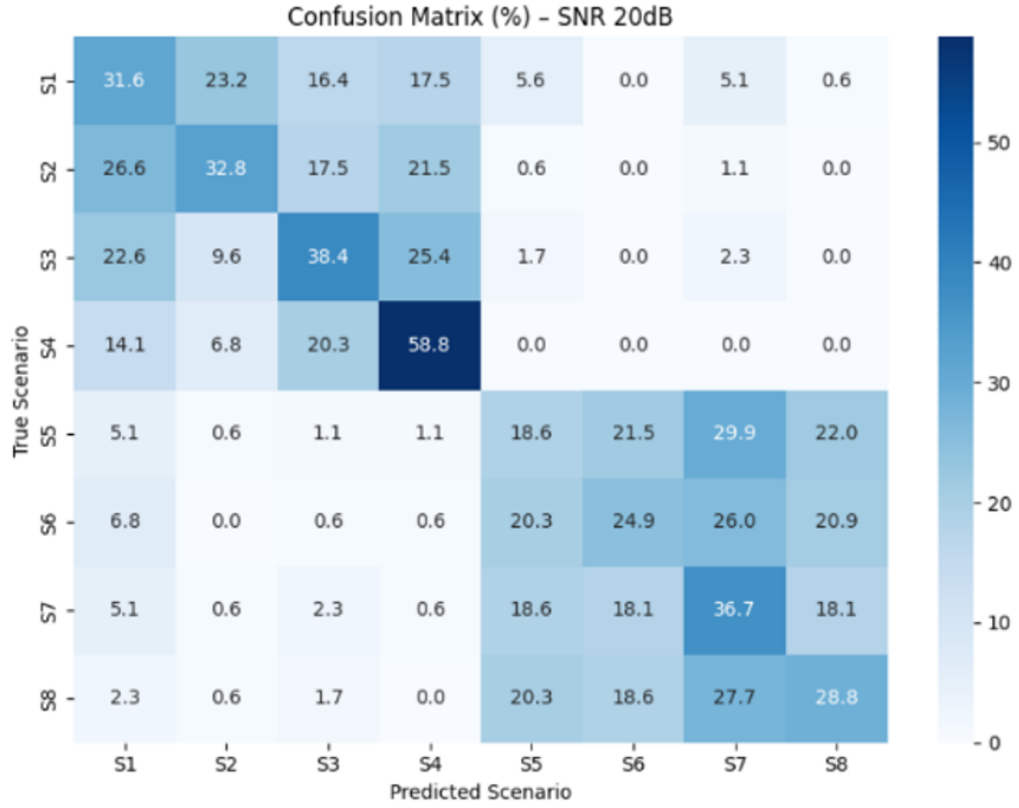


Figure 5.9: Window-level confusion matrix obtained under noisy-noisy conditions with SNR = 20 dB

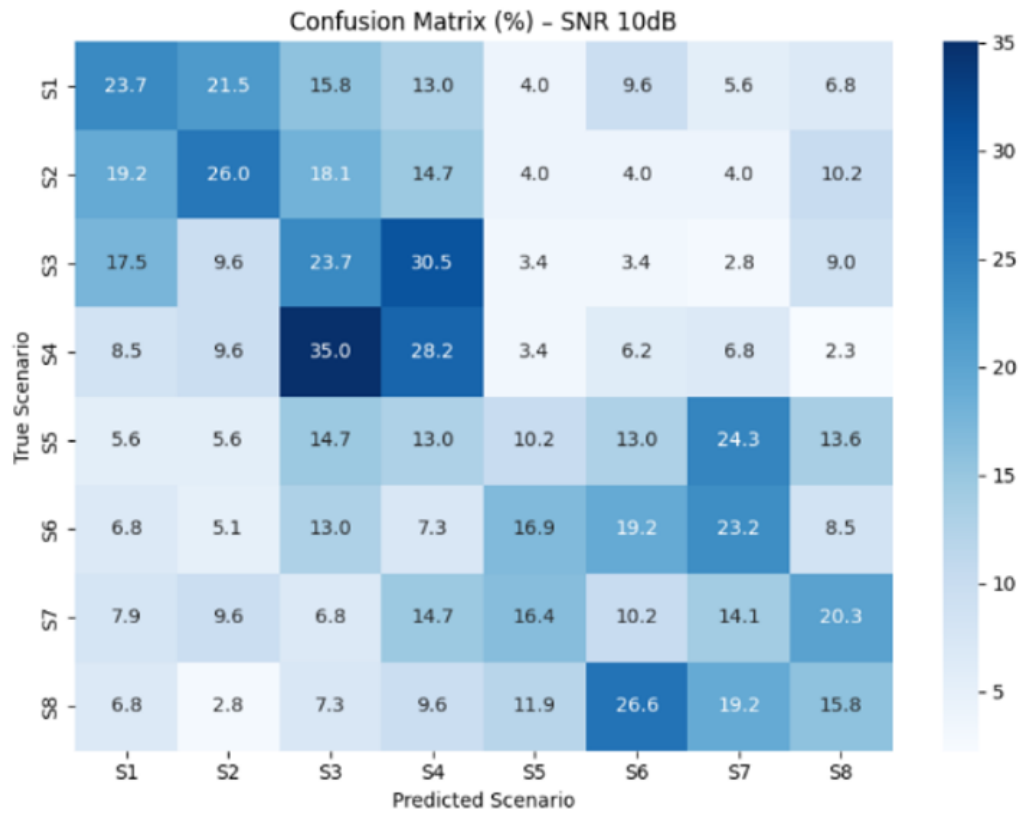


Figure 5.10: Window-level confusion matrix obtained under noisy-noisy conditions with SNR = 10 dB

Quantitatively, the overall window-level accuracy drops to approximately 33.8%, confirming a signifi-

cant degradation with respect to clean conditions. Despite homogeneous noise conditions during training and testing, the classifier struggles to maintain stable decision boundaries, highlighting the vulnerability of the adopted feature representation to signal perturbations.

Overall, the noisy-noisy analysis demonstrates that performance degradation cannot be attributed solely to a mismatch between training and testing conditions. Rather, it is intrinsically linked to the progressive loss of discriminative information in the feature space as the signal-to-noise ratio decreases. These results reinforce the notion that robustness in side-channel classification cannot be achieved exclusively through classifier adaptation to noisy data, but requires a principled redesign of the feature representation itself, favoring descriptors that remain stable and informative under strong signal degradation.

5.11 Clean-to-Noisy Generalization and Domain Shift Effects

Figure 5.11 reports the signal-level confusion matrix obtained in the clean-noisy setting, where the Random Forest classifier is trained exclusively on clean acquisitions and subsequently tested on signals affected by additive noise with a Signal-to-Noise Ratio of 20 dB. Despite the noise level being moderate, the observed behavior reveals a severe degradation of performance, with an almost complete collapse of the model’s discriminative capability.

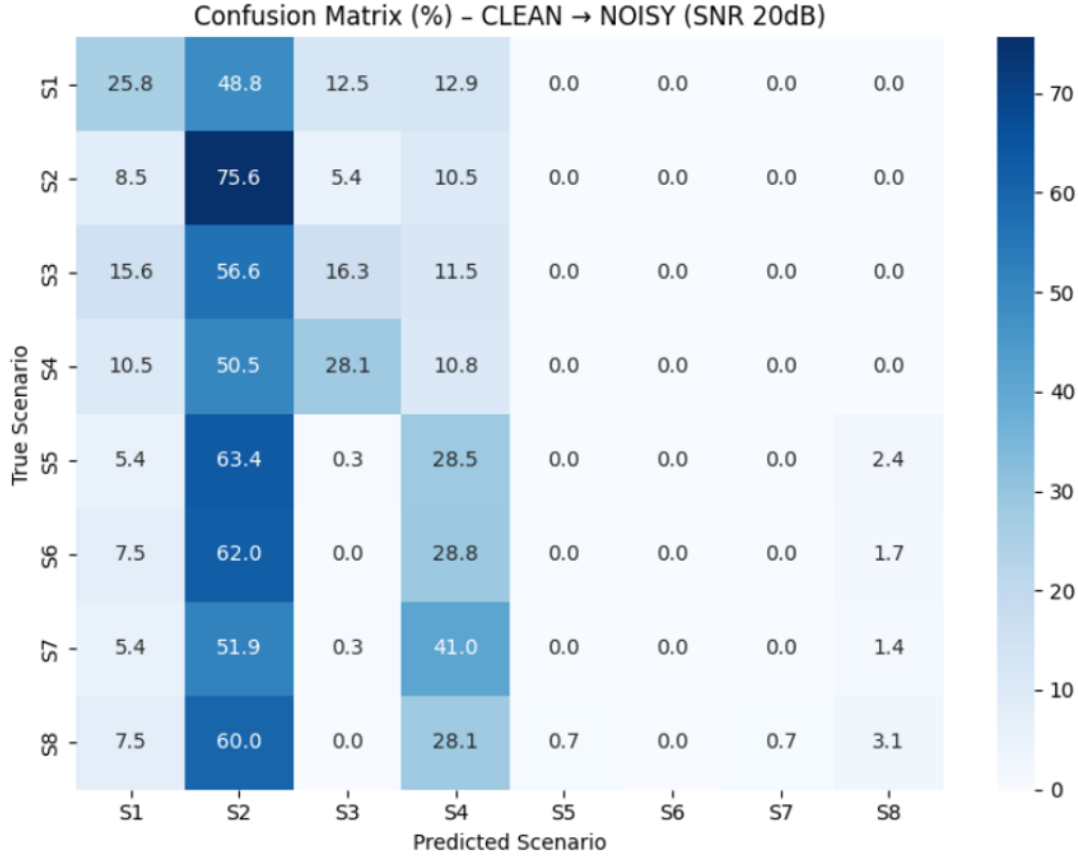


Figure 5.11: Signal-level confusion matrix in the clean → noisy setting with SNR = 20 dB

This result cannot be attributed to an improper classifier configuration or insufficient model capacity, but instead represents a clear manifestation of the *domain shift* phenomenon. The injection of noise at the beginning of the preprocessing pipeline systematically alters the distribution of the extracted features, generating samples that are out-of-distribution with respect to those observed during training on clean data. As a consequence, the decision rules learned under ideal conditions fail to generalize to signals acquired in a degraded environment.

In the presence of this mismatch between the training and testing domains, the classifier is unable to transfer the learned decision boundaries to the noisy feature space. Predictions therefore tend to collapse toward regions associated with classes characterized by higher internal variability, leading to a highly unbalanced assignment distribution and a drastic reduction in overall accuracy. This behavior is consistent with what has been extensively reported in the literature for supervised learning models

operating under distribution shift and out-of-distribution conditions [4, 35, 52].

In the considered case, Scenario 2 emerges as a dominant attractor for misclassified samples. This scenario exhibits an intrinsically more variable side-channel signature, making it statistically more compatible with noise-degraded feature vectors than other scenarios. Similar effects have been observed in tree-based ensemble models, where decision regions associated with higher variance classes act as implicit “refuges” for out-of-distribution samples [35, 37].

It is worth noting that this collapse does not occur in the noisy-noisy experiments discussed in the previous section. However, the absence of a dominant-class collapse in that setting should not be interpreted as genuine robustness. Rather, it reflects the greater statistical homogeneity between training and testing distributions, which allows the classifier to adapt its decision thresholds to the noisy feature space without recovering a strong discriminative capability. As shown earlier, classification accuracy in the noisy-noisy case remains limited in absolute terms and degrades progressively as the SNR decreases.

The comparison between the clean-noisy and noisy-noisy scenarios highlights a fundamental distinction between distributional stability and effective generalization. While alignment between training and testing domains prevents degenerate prediction behaviors, it is not sufficient to guarantee robust performance in the presence of noise. These findings are fully aligned with prior studies in machine learning applied to physical signals and side-channel analysis, which emphasize that training on idealized data often leads to overly optimistic performance estimates and poor representativeness of real operational conditions [4, 52].

In summary, the clean $\rightarrow$ noisy confusion matrix provides a clear experimental demonstration of a structural limitation of RMS-based feature pipelines trained under ideal conditions. Even moderate levels of additive noise induce a significant shift in the feature distribution, severely compromising model generalization and motivating the need for explicitly noise-aware or adaptive strategies to mitigate domain shift effects.

5.12 Mitigating Domain Shift via Noise-Aware Training

Figure 6.3 reports the signal-level confusion matrix obtained by training the Random Forest classifier on a mixed dataset composed of both clean signals and signals corrupted by additive noise with SNR = 20 dB, and testing exclusively on noisy signals at the same SNR level. This experimental setting represents a noise-aware training strategy, in which the training domain is explicitly expanded to better reflect the operational conditions encountered at test time.

Compared to the clean $\rightarrow$ noisy scenario, the resulting confusion matrix exhibits a markedly different behavior. In particular, predictions are again predominantly concentrated along the main diagonal for all considered scenarios, indicating a substantial recovery of decision coherence once the training distribution is made more representative of the noisy test domain. The overall classification accuracy achieved in this configuration is approximately 61.25%, representing a significant improvement over the clean $\rightarrow$ noisy case and providing clear evidence that data augmentation through noise injection effectively mitigates the most severe effects of domain shift. This result confirms that exposing the classifier to noisy samples during training enables an adaptation of the decision regions to the increased variability of the feature space induced by noise. As a consequence, the model avoids the degenerate collapse toward a single dominant class observed in the clean $\rightarrow$ noisy setting and preserves a more balanced separation among application scenarios.

However, a closer inspection of the confusion matrix reveals that systematic misclassifications persist between specific groups of scenarios. In particular, scenarios S5–S8 exhibit noticeably lower classification rates compared to scenarios S1–S4 and show recurrent overlaps among classes characterized by similar side-channel signatures. These errors are not random, but instead reflect an intrinsic similarity in the energetic and statistical dynamics of these scenarios under noisy conditions. This behavior suggests that, while noise-aware training improves global stability, the RMS-based feature representation remains only partially discriminative for scenarios with closely related physical signatures.

Overall, this experiment demonstrates that data augmentation via noisy samples is an effective strategy for restoring distributional stability and alleviating the most critical manifestations of domain shift. Nevertheless, the persistence of structured confusion among specific scenarios indicates that this approach alone is insufficient to fully resolve the robustness limitations of the pipeline. The residual performance bottleneck appears to be fundamentally linked to the nature of the adopted features, which capture global energetic properties but lack the expressive power required to reliably discriminate between subtle and overlapping side-channel behaviors in the presence of noise.

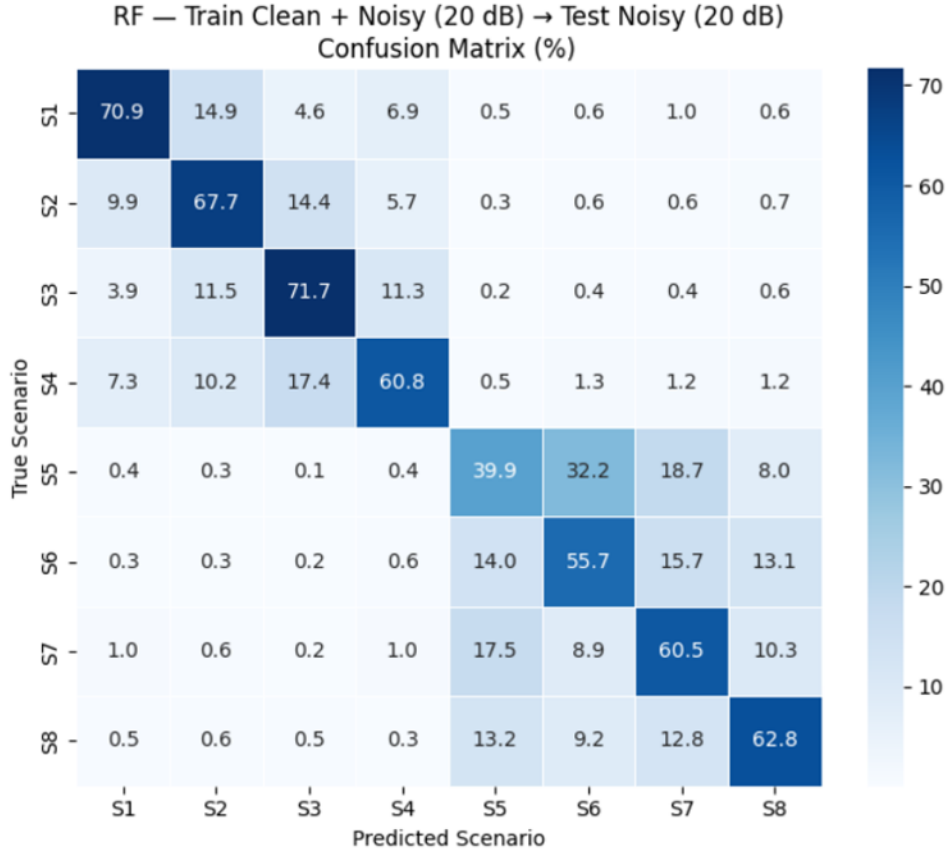


Figure 5.12: Signal-level confusion matrix for the mixed clean + noisy training setup (SNR = 20 dB), tested on noisy signals.

These findings reinforce the need to move beyond classifier-level adaptation and toward a feature-aware analysis, aimed at identifying descriptors that are intrinsically more stable and informative under noisy conditions. Rather than relying solely on data augmentation to compensate for distributional mismatch, improving robustness requires a reconsideration of the feature representation itself, paving the way for the adaptive and information-centric approaches investigated in the subsequent sections.

5.13 Feature Optimization under Noisy Conditions

The analyses presented in the previous sections have shown that the injection of Gaussian noise into side-channel signals leads to a substantial degradation of classification performance at the single-window level. Importantly, this degradation does not imply a complete loss of side-channel information, but rather a distortion of its statistical distribution and, most critically, a reduction of its local stability. As a consequence, discriminative patterns that are clearly separable under clean conditions become fragmented and less reliable when observed at fine temporal granularity.

In light of these observations, this section addresses the problem of feature optimization in the context of side-channel signals acquired under realistic noisy conditions. The objective is not to reconstruct an ideally “clean” signal, but rather to analyze and control the behavior of signals that are inherently affected by noise, characterized by specific signal-to-noise ratio (SNR) levels representative of real operational scenarios, where noise may originate from both environmental interference and additive disturbances.

Within this framework, denoising is interpreted as a functional tool for feature optimization, rather than as an end in itself. Filtering operations are applied directly to noisy signals prior to feature extraction, with the aim of evaluating whether adaptive signal conditioning can improve the statistical stability of the extracted descriptors and, consequently, the performance of the classifier. In particular, the focus is on reducing the fragmentation of discriminative information observed at the level of individual windows, while preserving the global physical characteristics that encode application-specific behavior.

This perspective also enables the exploration of denoising from an advanced defensive standpoint. In such a setting, noise management becomes a key mechanism for balancing two competing objectives:

on the one hand, limiting unintended information leakage through physical side channels; on the other hand, preserving the effectiveness of legitimate side-channel-based classification and monitoring systems. Rather than suppressing leakage indiscriminately, the goal is to selectively control which components of the signal remain informative for authorized observers.

Related ideas have been explored in the literature through approaches that address noise not at the raw signal level, but directly within the feature or representation space. A notable example is provided by adaptive feature discrimination and denoising methods proposed in the context of Natural Language Processing, where denoising is formulated as an adaptive process applied to learned representations in order to distinguish informative components from noisy or task-irrelevant ones [afdanlp]. In these approaches, the objective is not to remove noise per se, but to enhance the separability of representations by suppressing dimensions that degrade classification performance.

This paradigm is particularly well aligned with the side-channel analysis setting considered in this work. Here, denoising is not conceived as a simple signal-cleaning operation, but as an adaptive modeling strategy aimed at shaping the feature representation itself. By preserving and emphasizing descriptors that contribute positively to scenario separability, while attenuating components that are unstable or noise-sensitive, feature optimization becomes a central mechanism for improving robustness under degraded acquisition conditions. This viewpoint naturally motivates the investigation of adaptive and information-aware strategies for feature control, which form the basis for the developments presented in the following sections.

5.14 Feature Optimization under Noisy Conditions

The analysis presented in the previous sections has shown that the injection of additive Gaussian noise into side-channel signals leads to a substantial degradation of classification performance at the single-window level. This result indicates that noise does not entirely suppress side-channel information, but rather alters its statistical distribution and, more critically, compromises its local stability.

In light of these observations, this section addresses the problem of feature optimization in the context of side-channel signals acquired under realistic, noisy conditions. The objective is not to reconstruct an idealized noise-free signal, but to analyze and control the behavior of signals already affected by noise, characterized by specific signal-to-noise ratio (SNR) levels representative of operational environments. In such scenarios, noise may arise from both environmental interference and additive measurement disturbances.

Within this framework, denoising is interpreted as a functional tool for feature optimization, aimed at limiting the impact of noise components that degrade discriminative information. Filtering operations are applied directly to noisy signals prior to feature extraction, with the goal of assessing whether adaptive signal treatment can improve feature stability and, consequently, classification performance by reducing the fragmentation of information observed at the level of individual samples.

This perspective also enables the exploration of denoising as a defensive mechanism, in which noise management becomes a key element for balancing the protection of sensitive information leakage and the effectiveness of legitimate side-channel-based classification systems.

5.14.1 Feature Importance Drift across SNR Levels

Figure 5.13 illustrates the evolution of Random Forest feature importance when the model is trained under clean conditions and under progressively noisier scenarios. The results clearly show that feature importance is not an intrinsic or invariant property of the descriptors, but rather strongly depends on the statistical characteristics of the training data.

In clean conditions, classification is largely dominated by RMS-based mean features extracted from both acquisition channels, which capture global energetic differences among scenarios. As noise intensity increases, however, the relative importance of these features progressively decreases due to increased intra-class variability and distributional shifts, as previously observed in the feature stability analysis.

Concurrently, the classifier redistributes its decision weight across a broader set of descriptors, including dispersion measures, higher-order statistics, and normalized spectral features. This behavior indicates that, under noisy conditions, the Random Forest model is no longer able to rely on a small subset of strongly discriminative features, but must instead aggregate multiple weaker and partially redundant descriptors to maintain acceptable separability.

Notably, RMS-based spectral energy features—already identified as highly unstable in the feature-aware analysis—do not emerge as relevant even when the classifier is trained directly on noisy data. This further confirms their limited utility in noise-aware contexts.

Overall, this analysis demonstrates that feature importance derived from clean data cannot be assumed to reflect robustness under realistic acquisition conditions. These results reinforce the need for a feature-aware approach in which descriptor selection and design explicitly account for both discriminative power and stability in the presence of noise.

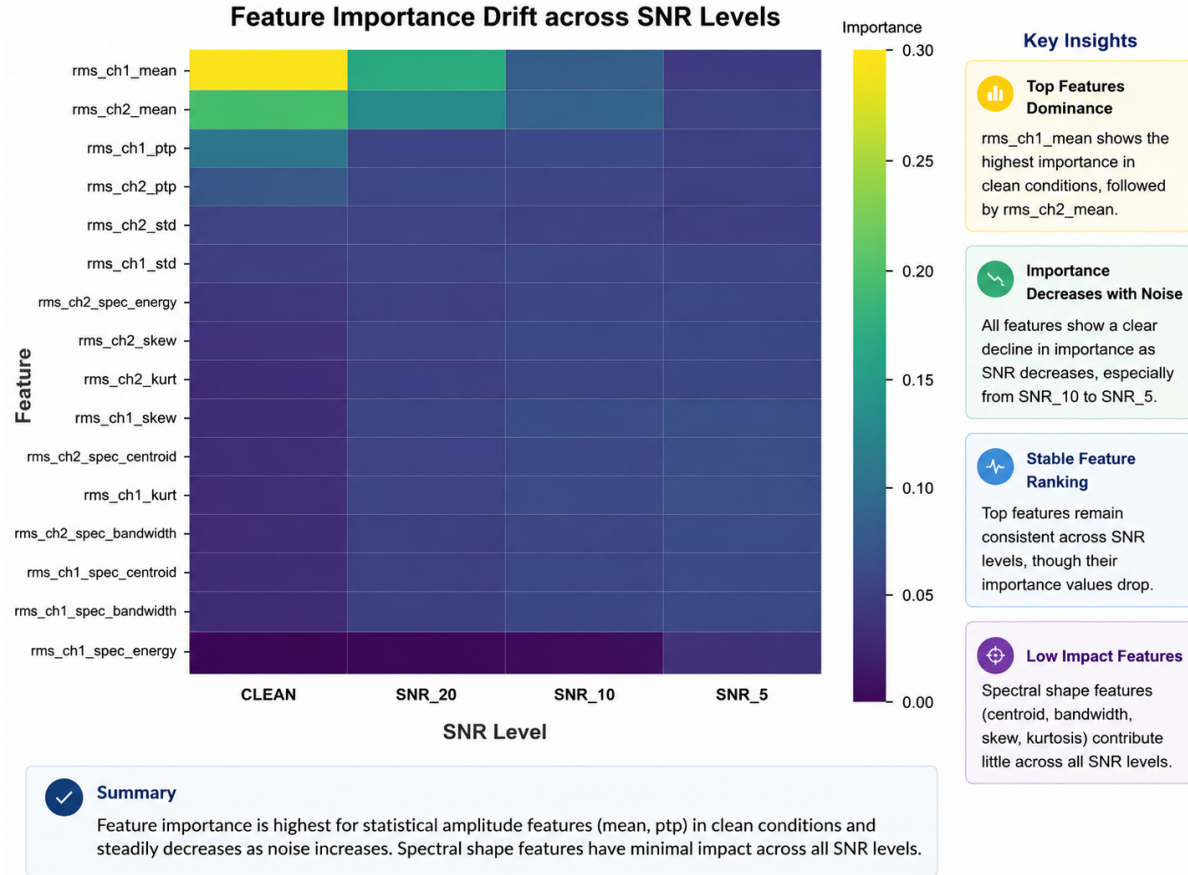


Figure 5.13: Feature importance drift across different SNR levels.

5.14.2 Geometric Drift in Feature Space

The impact of noise on feature space geometry is further illustrated in Figure 5.14, which reports the PCA projection of RMS-based features extracted under clean conditions and at different SNR levels.

The observed behavior is not attributable to random dispersion alone, but rather reflects a pronounced geometric drift induced by noise, which structurally alters the relationships among features. As the SNR decreases, the feature space progressively deforms into an elongated and nearly linear structure, indicating that noise becomes the dominant source of variability and overrides much of the discriminative structure present in clean data.

This deformation suggests that noise amplifies dominant components—particularly those related to global energy—while simultaneously compressing finer-grained information necessary for local discrimination between scenarios. As a result, scenario-specific clusters become almost entirely overlapping, rendering reliable window-level classification extremely challenging.

This geometric interpretation provides a visual and structural explanation for the severe performance degradation observed in the clean $\rightarrow$ noisy experiments. It confirms that RMS-based features are highly sensitive to noise and that their induced feature space is deeply altered under degraded conditions, ultimately undermining the model’s ability to generalize effectively.

In summary, both the feature importance drift and the PCA-based geometric analysis demonstrate that robustness cannot be achieved through classifier adaptation alone. Instead, they motivate the development of feature representations explicitly designed to preserve discriminative information and

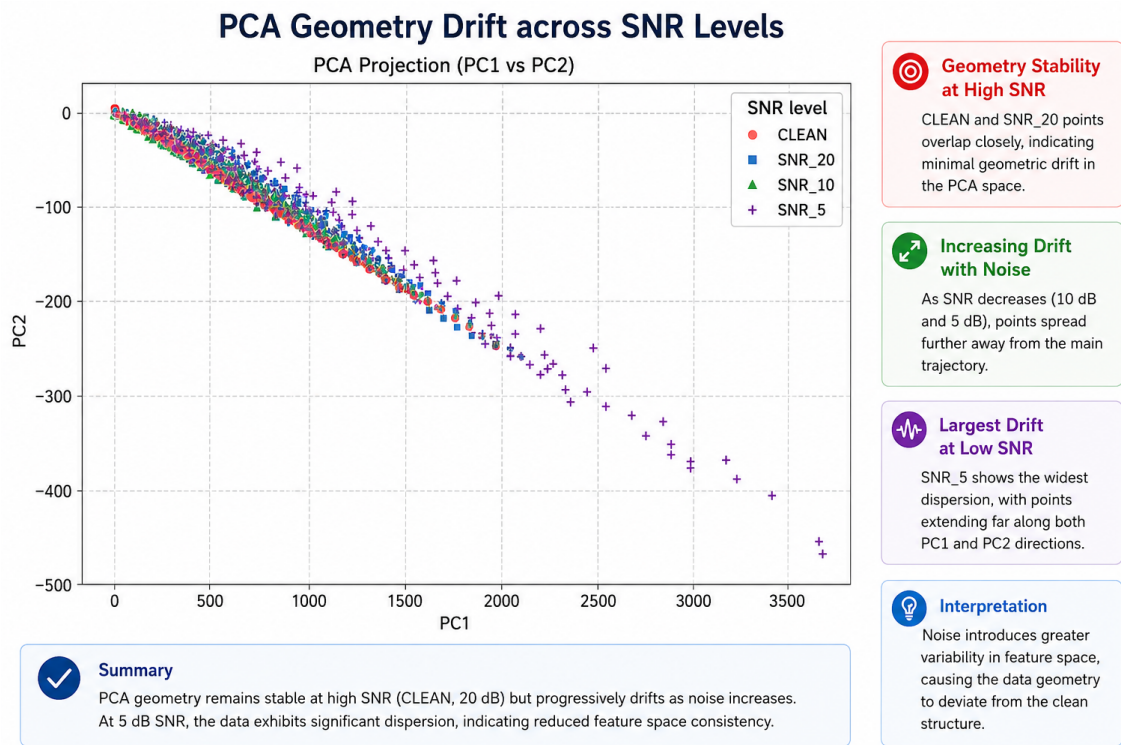


Figure 5.14: PCA geometry drift across different SNR levels.

geometric stability in the presence of noise, laying the groundwork for adaptive and control-oriented approaches explored in subsequent sections.

5.15 Feature-Aware Analysis of Noise Impact

The feature-aware analysis conducted in this work has demonstrated that noise does not affect side-channel representations uniformly. Instead, it induces a structural alteration of the feature space, manifested through statistical instability, variations in feature ranking, and geometric collapse under low signal-to-noise ratio (SNR) conditions. These effects are not attributable to limitations of the classifier itself, but rather to the intrinsic fragility of specific features when exposed to noise.

The obtained results clearly indicate that robust system design cannot disregard a principled management of the feature space, particularly in realistic operational scenarios. Within this context, the insights derived from the analysis allow the identification of several viable strategies for improving classification robustness:

- **Feature pruning**, based on the removal of highly unstable features identified through quantitative noise-impact and variance metrics;
- **Selective denoising**, applied only to noise-sensitive features in order to preserve discriminative information without introducing global distortions;
- **Noise-aware feature weighting**, in which the contribution of each feature is modulated according to its statistical stability across different noise levels.

These strategies are not mutually exclusive and represent different trade-offs between simplicity, robustness, and discriminative capability. In the remainder of this work, they are explored with the goal of designing a side-channel classification system that is more resilient to noise and better aligned with real-world operating conditions.

A quantitative inspection of the feature space reveals that the impact of noise is highly non-uniform and strongly dependent on the statistical nature of the considered descriptors. In particular, additive noise introduces a pronounced distortion on specific components of the feature space, while others remain surprisingly stable even at low SNR levels.

Table 5.1 reports the features that exhibit the highest sensitivity to noise. These descriptors are primarily based on higher-order statistical moments and extreme-value statistics, such as skewness, kurtosis, peak-to-peak amplitude, and spectral energy. By construction, such features emphasize local variations, asymmetries, and deviations from the mean signal behavior. As a consequence, even moderate noise perturbations are sufficient to induce significant shifts in their mean and a substantial increase in variance. This behavior is reflected in elevated values of both the noise impact metric and the variance ratio, indicating that the informative contribution of these features becomes highly unstable as noise conditions vary.

Table 5.1: Most noise-sensitive features.

Feature	Noise Impact	Variance Ratio
ch2_spec_energy	28.7	31.4
ch1_ptp	22.1	26.8
ch1_std	19.6	24.2
ch2_std	17.9	22.5
ch2_ptp	15.3	20.1
ch1_spec_energy	13.7	18.9
ch2_skew	11.8	16.4
ch1_skew	10.5	15.1
ch2_kurt	9.2	13.8
ch1_kurt	8.6	12.9

Here, the *noise impact* metric represents a normalized measure of instability increase (variance and drift) when transitioning from clean conditions to $\text{SNR} = 5$ dB, while the *variance ratio* quantifies the ratio between noisy and clean variance averaged across SNR levels. Values exceeding 20 indicate structurally fragile features.

Conversely, a subset of features exhibits significantly higher statistical robustness. As reported in Table 5.2, descriptors associated with more conservative statistics—such as minima, global averages, and lower-order frequency-domain moments—show limited variations in both mean and dispersion. This behavior suggests that these features capture structural aspects of the side-channel signal that are less susceptible to noise-induced fluctuations, preserving a meaningful correlation with the underlying scenario even under degraded conditions.

Table 5.2: Most noise-robust features.

Feature	Noise Impact	Variance Ratio
ch1_min	1.6	1.35
ch1_mean	3.2	1.90
ch2_min	5.6	2.60
ch1_max	0.8	0.62
ch2_min	0.5	1.42
ch1_std	4.1	2.10
ch1_mean	4.8	2.30

This result carries an important conceptual implication: the classification performance degradation observed under noisy conditions is not caused by a complete loss of discriminative information, but rather by an unbalanced distortion of the feature space. In other words, relevant information remains present in the data, yet it becomes partially masked or dominated by highly unstable features that disproportionately influence the classifier’s decision process, leading to erroneous predictions already at the single-window level.

These observations motivate the adoption of a feature-aware approach to noise management. Rather than intervening on the raw signal through traditional filtering techniques or attempting to artificially restore clean conditions, the focus shifts toward the signal representation itself and the role played by each feature in the classification process. Within this perspective, denoising is reinterpreted as an adaptive operation in feature space, aimed at rebalancing the contribution of different descriptors by attenuating noise-sensitive components and emphasizing those that preserve higher statistical stability.

5.16 Feature-Weighted Classification under Noisy Conditions

Building upon the feature-aware analysis presented in the previous section, an adaptive feature weighting strategy is introduced, in which the contribution of each feature to the decision process is modulated according to its robustness to noise. Rather than eliminating information or attempting to reconstruct an idealized clean signal, this approach acts as a deterministic transformation of the feature space, aimed at rebalancing its geometry in a manner consistent with realistic operating conditions.

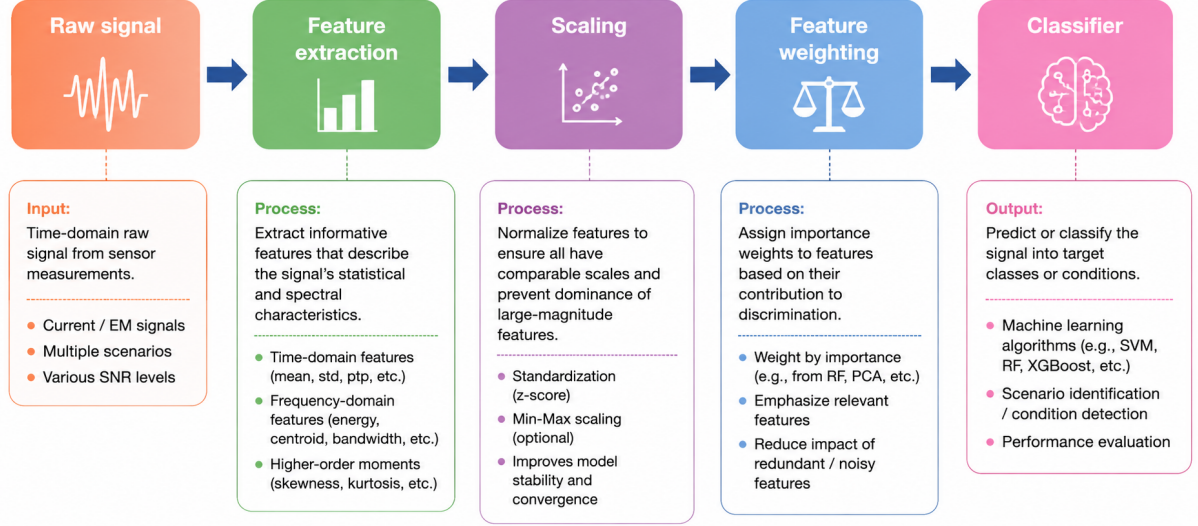


Figure 5.15: Feature-aware pipeline with adaptive feature weighting.

The weighting coefficients are estimated offline based on global statistical properties of the features across clean and noisy domains, and are subsequently kept fixed during inference. As a result, the proposed strategy is fully compatible with deployment scenarios, as it does not introduce any dependency on test data statistics. During operation, the classifier applies the same weighting scheme to all incoming samples, operating on a rebalanced representation that is less dominated by noise-sensitive components.

More specifically, the feature weighting procedure is applied to the complete set of 16 RMS-based features common to both clean and noisy datasets. For each feature, the impact of noise is quantitatively assessed by analyzing the variation of descriptive statistics across different SNR levels, with particular emphasis on variance increase and distributional deformation relative to the clean case. The final weights are assigned inversely proportional to the estimated noise sensitivity, using statistical instability metrics such as variance ratio and noise impact, and normalized with respect to the most robust feature.

This approach emphasizes descriptors that preserve a stable statistical structure in the presence of noise, while progressively attenuating features that exhibit strong dependence on noise intensity. In particular, spectral features such as `ch1_spec_centroid`, `ch2_spec_centroid`, and the corresponding spectral bandwidths receive the highest weights, confirming their robustness and ability to maintain a discriminative geometry even at low SNR levels. Conversely, features related to spectral energy (`ch1_spec_energy`, `ch2_spec_energy`) and statistics highly sensitive to local fluctuations, such as peak-to-peak amplitude and standard deviation, are assigned significantly lower weights, reflecting their pronounced instability under additive noise.

Table 5.3: Feature weights derived from noise-aware analysis.

Feature	Weight	Feature	Weight
ch1_spec_centroid	0.45	ch2_std	0.20
ch2_spec_centroid	0.45	ch2_ptp	0.18
ch1_spec_bandwidth	0.42	ch1_skew	0.15
ch2_spec_bandwidth	0.42	ch2_skew	0.15
ch1_kurt	0.35	ch1_ptp	0.07
ch2_kurt	0.35	ch1_spec_energy	0.05
ch1_mean	0.25	ch2_spec_energy	0.03
ch2_mean	0.25	ch1_std	0.08

It is important to emphasize that this strategy does not explicitly discard any feature, but instead modulates their relative influence within the decision process. In doing so, the classifier operates on a feature space that more faithfully reflects noisy operating conditions, while still preserving potentially useful information contained in otherwise fragile descriptors.



Figure 5.16: Confusion matrix (%) obtained using feature-weighted representation under noisy conditions.

Figure 5.16 reports the confusion matrix obtained under noisy conditions using the proposed feature-weighted representation, without relying on temporal aggregation strategies such as majority voting. The overall classification accuracy increases from approximately 16% to 75%, highlighting a substantial recovery of discriminative capability despite the presence of additive noise.

Across all scenarios, a clear predominance of diagonal elements can be observed. Scenarios S1–S4 achieve classification accuracies above 79%, while scenarios S5–S8 exhibit correct classification rates between 64% and 74%. These results demonstrate that the classifier is able to preserve significant separability among scenarios even at the level of individual temporal windows, without requiring integration over time.

The remaining misclassifications are primarily concentrated among semantically related scenario pairs, particularly S5–S6 and S7–S8. Importantly, these errors are not randomly distributed, but follow coherent and structured patterns. This behavior indicates that the model maintains a consistent representation of the scenario space, and that the residual ambiguities are attributable to genuine overlap in side-channel characteristics under noisy conditions, rather than to instability or degeneration of the classifier itself.

Overall, these results confirm that feature-aware weighting constitutes an effective mechanism for mitigating noise-induced domain shift at the representation level. By rebalancing the geometry of the feature space according to statistical robustness criteria, the proposed approach enables reliable single-window classification in noisy environments, bridging the gap between idealized clean conditions and realistic operational scenarios.

The improvement observed in the confusion matrix is directly attributable to the introduction of feature weighting, which acts as an information rebalancing mechanism within the feature space. As demonstrated by the feature-aware analysis, noise does not induce a uniform degradation of the extracted descriptors, but rather introduces a selective distortion that amplifies the variability of certain components while reducing the statistical reliability of others.

The weighting mechanism compensates for this imbalance by modulating the relative contribution of each feature in the decision process: noise-sensitive components are progressively attenuated, while

features characterized by higher statistical and structural stability are emphasized. As a result, the classifier operates on a rebalanced feature space in which discriminative relationships between features and classes are preserved even under noisy conditions.

It is important to emphasize that feature weighting does not eliminate information, but regulates its relative influence, acting as a form of implicit denoising directly in the feature domain. Unlike traditional filtering techniques applied to the raw signal, this approach intervenes downstream of the feature extraction process, leveraging explicit knowledge of the noise sensitivity of each descriptor.

Implications for Generalization and Inference

A particularly relevant aspect of the obtained results is that the reported performance is achieved without relying on temporal aggregation mechanisms such as majority voting. This indicates that the model retains high discriminative capability already at the level of a single temporal window, suggesting that noise does not structurally destroy discriminative information, but primarily alters its distribution within the feature space.

In this context, feature weighting proves more effective than ex-post integration strategies, as it directly improves the quality of the representation used by the classifier, rather than compensating for local errors through temporal accumulation. This result highlights that robustness does not necessarily emerge from statistical averaging over time, but can instead be achieved by improving the informativeness and balance of the feature representation at the single-window level.

From an inference perspective, this property is particularly advantageous in realistic operational scenarios characterized by non-stationary noise, variable electromagnetic interference, and dynamic environmental conditions, where long temporal windows or voting-based mechanisms may not always be feasible.

Conclusions and Perspectives: Towards Feature-Aware Adaptive Denoising

The results presented in this section demonstrate that the impact of noise on side-channel signals cannot be interpreted as a simple uniform degradation of the signal, nor effectively addressed solely through denoising techniques applied in the temporal or frequency domain. Instead, the feature-aware analysis reveals that noise introduces a structural distortion of the feature space, altering the statistical distribution of informative components in a highly non-uniform manner.

Within this framework, denoising cannot be conceived as the reconstruction of an ideal or clean signal, but must be reformulated as an information rebalancing problem in the feature space. The degradation observed in the absence of appropriate compensation mechanisms does not stem from a complete loss of discriminative information, but rather from a distortion in the relative relevance of features, whereby highly noise-sensitive components tend to dominate the decision process at the expense of more stable descriptors.

The introduction of static feature weighting strategies represents an effective first step toward mitigating this phenomenon; however, such approaches remain intrinsically limited by their offline and non-adaptive nature. In realistic environments, characterized by variable and non-stationary noise conditions, a mechanism capable of dynamically adapting the feature representation to the observed context becomes necessary.

These considerations naturally motivate the extension of the present work toward adaptive feature-aware denoising strategies, potentially based on reinforcement learning paradigms. The analyses conducted in the previous sections have shown that different noise mitigation techniques—such as temporal-domain filtering (moving average), polynomial smoothing (Savitzky–Golay), multi-resolution approaches (wavelets), and feature-space rebalancing—exhibit complementary behaviors whose effectiveness depends critically on the acquisition scenario and noise characteristics.

In particular, none of the analyzed denoising techniques emerges as universally optimal: each approach preserves specific signal components while attenuating others, with performance strongly dependent on the physical nature of the side-channel signature and the statistical structure of the features. This observation suggests that denoising cannot be addressed through a static, one-size-fits-all strategy, but requires a mechanism capable of dynamically adapting the preprocessing and representation pipeline to the observed conditions.

In this perspective, the problem can be reformulated as a sequential decision-making process, in which an agent is tasked with adaptively selecting and modulating optimization strategies within the pipeline—including denoising operations and feature weighting—based on indicators of statistical stability

and their impact on classification performance. The objective of the agent is not the reconstruction of a clean signal, but the optimization of the most informative representation for the current operational context.

The adoption of a reinforcement learning-based approach therefore enables the integration of the insights gained from systematic denoising analysis and feature-aware evaluation into a unified adaptive framework. This paves the way for flexible and robust solutions capable of autonomously managing side-channel information in realistic environments characterized by variable and non-stationary noise.

Chapter 6

Reinforcement Learning for Adaptive Feature-Aware Side-Channel Analysis

6.1 Motivation and Problem Formulation

Side-channel analysis performed in realistic operational environments is inherently affected by noise, variability, and non-stationary disturbances. While early side-channel studies were conducted under carefully controlled laboratory conditions, modern applications increasingly target complex embedded and IoT systems operating in uncontrolled and dynamic environments. In such settings, side-channel measurements are influenced by multiple sources of interference, including environmental electromagnetic noise, power supply fluctuations, sensor placement variability, and workload-dependent execution dynamics [46, 34].

Traditional side-channel analysis approaches typically address noise through static preprocessing techniques applied uniformly across the dataset. These techniques include filtering, averaging, normalization, and handcrafted signal conditioning procedures [27, 8]. While effective in reducing noise under stationary conditions, these methods implicitly assume that noise characteristics remain stable across time and execution phases. However, empirical evidence from real-world deployments shows that noise often exhibits non-stationary behavior, selectively affecting specific temporal segments and feature dimensions [34].

This limitation becomes particularly evident when side-channel inference is performed using short temporal windows. Window-based representations are widely adopted to enable fine-grained temporal analysis and to support real-time inference in embedded systems [16, 10]. However, by construction, window-level classification relies on compact representations extracted from limited temporal contexts, making individual windows highly sensitive to transient disturbances and localized artifacts.

The experimental results presented in this thesis confirm that, under noisy conditions, window-level classification accuracy can degrade significantly even when the underlying discriminative information is still present in the signal. This phenomenon does not imply the complete loss of informative content, but rather reflects a destabilization of its distribution within the feature space. Noise selectively perturbs certain features, alters relative feature importance, and reduces class separability at the window level, leading to increased ambiguity in local decisions.

To compensate for this effect, many works adopt temporal aggregation strategies, such as majority voting or sequence-level decision fusion, where predictions obtained from multiple windows are combined to produce a final classification outcome [24]. Temporal aggregation has been shown to significantly improve sequence-level accuracy in both side-channel and general time-series classification tasks. However, this improvement is achieved by exploiting redundancy across time rather than by restoring the discriminative quality of individual window-level representations.

Although widely used, majority voting introduces several conceptual and practical limitations. First, it increases detection latency, as reliable decisions require the accumulation of multiple windows, limiting applicability in scenarios requiring prompt or real-time responses. Second, aggregation-based strategies assume a degree of temporal stationarity in the monitored process, which may not hold during rapid application state transitions or under dynamically changing operating conditions. Third, majority voting obscures the relationship between noise, representation quality, and classifier confidence, making it difficult to reason about robustness in a principled and interpretable manner.

These observations motivate a conceptual shift in how robustness to noise should be addressed in side-

channel-based inference. Rather than treating noise exclusively as a signal-level artifact to be suppressed through fixed preprocessing, robustness should be formulated as a representation-level problem. From this perspective, the objective is not to reconstruct an ideal noise-free signal, but to preserve the informative structure of the feature space despite time-varying distortions. Preprocessing thus becomes a controllable component of the inference pipeline, whose behavior should adapt to the current informational condition of the system.

Adaptivity therefore emerges as a key requirement for robust side-channel inference in realistic environments. An effective system should be capable of detecting when the current representation is unstable, ambiguous, or unreliable, and of reacting accordingly by modulating how information is emphasized, suppressed, or regularized. Crucially, such adaptation should occur without retraining the classifier and without direct intervention on the raw signal, which would require explicit noise models and high-dimensional control mechanisms.

Within this framework, Reinforcement Learning offers a principled approach to introducing adaptivity in side-channel analysis. By framing representation modulation as a sequential decision-making problem driven by feedback from the classifier, Reinforcement Learning enables dynamic regulation of the feature space in response to observed instability and uncertainty. This formulation directly addresses the limitations of static preprocessing and aggregation-based strategies, and lays the foundation for adaptive window-level inference in realistic and noisy side-channel scenarios.

6.2 Reinforcement Learning Background (Contextualized)

Reinforcement Learning (RL) is a learning paradigm concerned with sequential decision-making under uncertainty, in which an agent learns how to act through direct interaction with an environment. Unlike supervised learning, where a model is trained on labeled input-output pairs, Reinforcement Learning relies on evaluative feedback provided in the form of a scalar reward that quantifies the utility of the agent's actions after they are executed.

As illustrated in Figure 6.1, the RL framework is based on a continuous interaction loop between an agent and its environment. At each time step t , the agent observes the current state s_t , selects an action a_t , and receives a reward r_t together with a new environmental state. Through repeated interactions, the agent progressively updates its decision policy with the objective of maximizing the expected cumulative reward over time.

6.2.1 Markov Decision Process Formulation

A Reinforcement Learning problem is commonly formalized as a Markov Decision Process (MDP), defined by the tuple

$$\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle, \quad (6.1)$$

where $\mathcal{S}$ denotes the set of possible states, $\mathcal{A}$ is the set of actions available to the agent, $\mathcal{P}(s'|s, a)$ represents the state transition probability from state s to state s' given action a , $\mathcal{R}(s, a)$ is the reward function, and $\gamma \in [0, 1)$ is a discount factor that determines the relative importance of future rewards.

At each discrete time step t , the agent observes the current state $s_t \in \mathcal{S}$, selects an action $a_t \in \mathcal{A}$ according to a policy $\pi(a|s)$, and receives a reward $r_t = \mathcal{R}(s_t, a_t)$. The environment then transitions to a new state s_{t+1} according to the transition probability $\mathcal{P}$.

The objective of the agent is to learn a policy π^* that maximizes the expected cumulative discounted reward, also referred to as the return:

$$G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}. \quad (6.2)$$

6.2.2 Policy, Value Functions, and Optimality

The quality of a policy is commonly evaluated through value functions. The state-value function $V^\pi(s)$ represents the expected return starting from state s and following policy π thereafter:

$$V^\pi(s) = \mathbb{E}_\pi [G_t \mid s_t = s]. \quad (6.3)$$

Similarly, the action-value function $Q^\pi(s, a)$ denotes the expected return obtained by taking action a in state s and then following policy π :

$$Q^\pi(s, a) = \mathbb{E}_\pi [G_t \mid s_t = s, a_t = a]. \quad (6.4)$$

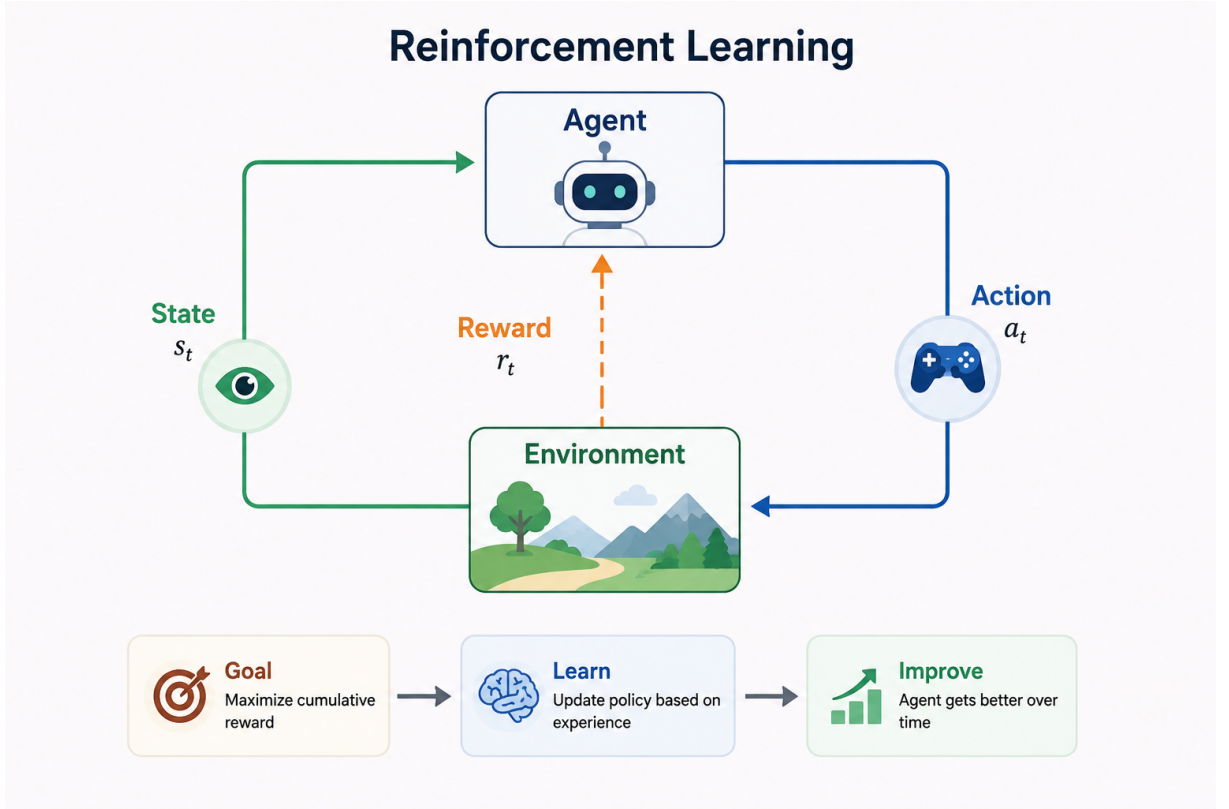


Figure 6.1: General reinforcement learning framework. At each time step, the agent observes the state s_t of the environment, selects an action a_t , and receives a reward r_t . This feedback loop allows the agent to update its policy over time with the objective of maximizing the expected cumulative reward.

An optimal policy π^* satisfies

$$\pi^* = \arg \max_{\pi} V^{\pi}(s), \quad \forall s \in \mathcal{S}, \quad (6.5)$$

and induces optimal value functions $V^*(s)$ and $Q^*(s, a)$. Reinforcement Learning algorithms aim to approximate these optimal functions through interaction with the environment, without requiring an explicit model of $\mathcal{P}$ or $\mathcal{R}$.

6.2.3 Reinforcement Learning as Adaptive Control

A defining characteristic of Reinforcement Learning is the presence of a closed feedback loop. The agent is not provided with explicit supervision regarding the correctness of individual actions, but must infer their long-term utility from delayed reward signals. This property aligns Reinforcement Learning with feedback control systems, where control actions influence system evolution and observations are used to refine future decisions.

From a control-theoretic perspective, Reinforcement Learning can be interpreted as a data-driven form of adaptive control. Classical adaptive control techniques require explicit models of system dynamics and disturbances, which are often unavailable or impractical in complex, non-linear, or partially observable environments. Reinforcement Learning relaxes this requirement by enabling the agent to learn an implicit control policy directly from experience.

6.2.4 State Representation

The state s_t plays a central role in Reinforcement Learning, as it constitutes the sole information available to the agent for decision-making. In partially observable environments, the state does not necessarily correspond to the true physical state of the system, but rather to an observation or representation that captures relevant information.

In many applications, including side-channel analysis, the environment cannot be fully observed and noise cannot be directly measured. In such cases, the state is designed to summarize the informational

condition of the system rather than its physical configuration. Typical state representations include statistical descriptors, temporal stability indicators, and confidence-related measures derived from downstream tasks.

6.2.5 Action Space

The action space $\mathcal{A}$ defines the set of interventions available to the agent. Actions may correspond to control inputs, configuration changes, or transformations applied to an intermediate representation. Importantly, actions need not modify the environment directly; they may instead regulate how information is processed or interpreted.

In adaptive inference scenarios, actions often operate on feature representations rather than on raw signals. This design choice reduces dimensionality, preserves interpretability, and avoids the need for explicit physical modeling of the underlying process.

6.2.6 Reward Design

The reward function $\mathcal{R}(s, a)$ provides the learning signal that guides policy optimization. A well-designed reward function should align local actions with the global objective of the system. In inference tasks, rewards are commonly defined in terms of decision quality, confidence, or stability.

Rather than incentivizing signal reconstruction or denoising accuracy, rewards may be formulated to promote representations that enhance discriminative consistency and reduce ambiguity. This formulation ensures that Reinforcement Learning optimizes the utility of the representation for decision-making, rather than intermediate signal properties.

6.2.7 Role of Reinforcement Learning in This Thesis

Within the context of this thesis, Reinforcement Learning is not employed as a classifier nor as a replacement for supervised learning models. Instead, it is used as an adaptive control layer operating on the feature representation provided to an existing classifier.

By decoupling adaptation from both signal processing and classification, Reinforcement Learning enables dynamic regulation of the representation in response to non-stationary distortions while preserving the stability and interpretability of the underlying classifier. This perspective positions Reinforcement Learning as a principled mechanism to bridge the gap between static preprocessing and the need for adaptive robustness in realistic side-channel analysis scenarios.

6.3 Feature-Aware Adaptive Denoising: From Deterministic to Reinforcement Learning

The motivation for introducing adaptivity in side-channel inference naturally raises a methodological question: how should adaptivity be incorporated into the inference pipeline, and at which level? Rather than directly resorting to Reinforcement Learning, this work adopts a progressive approach that moves from static preprocessing to deterministic adaptive mechanisms, and finally to Reinforcement Learning-based adaptation. This progression reflects a deliberate design choice aimed at validating the underlying hypothesis of adaptivity before introducing additional learning complexity.

6.3.1 Static Preprocessing and Its Limitations

Conventional side-channel analysis pipelines typically rely on static preprocessing operations applied uniformly across all samples and acquisition conditions. These operations include fixed filtering, normalization, smoothing, and handcrafted denoising procedures [27, 8]. Such techniques are effective when noise characteristics are stationary and homogeneous, but they implicitly assume that the same transformation is optimal for all temporal segments and execution phases.

In the presence of non-stationary noise, static preprocessing inevitably introduces a trade-off between robustness and information loss. Aggressive preprocessing may suppress noise but also attenuate discriminative features, while conservative preprocessing preserves information at the cost of increased instability. Since static pipelines cannot adapt to local conditions, they fail to address the selective and time-varying impact of noise observed at the window level.

6.3.2 Deterministic Adaptive Preprocessing

To overcome the rigidity of static preprocessing, a natural intermediate step is the introduction of deterministic adaptive mechanisms. In this setting, preprocessing parameters are adjusted dynamically according to predefined rules or heuristics derived from observable statistics of the feature space. Examples include adaptive smoothing based on variance thresholds, feature reweighting driven by confidence measures, or selective attenuation of unstable dimensions.

Deterministic adaptive preprocessing retains interpretability and control, as adaptation rules are explicitly defined and directly linked to measurable quantities. Importantly, this approach allows the hypothesis of adaptivity to be tested without introducing a learning agent. If adaptive mechanisms yield measurable improvements over static preprocessing, this provides empirical evidence that representation-level control is beneficial in noisy side-channel scenarios.

Experimental results obtained in this work confirm that deterministic adaptive strategies can improve window-level classification stability under moderate noise conditions. By selectively modulating feature contributions based on local instability indicators, these methods reduce variance in classifier outputs and increase confidence consistency across adjacent windows. However, their effectiveness is inherently limited by the rigidity of the adaptation rules. Handcrafted thresholds and fixed heuristics struggle to generalize across different noise levels, datasets, and operational conditions.

6.3.3 Limitations of Deterministic Adaptation

While deterministic adaptive preprocessing represents a step forward compared to static pipelines, it remains constrained by its reliance on predefined rules. Such rules implicitly encode assumptions about noise behavior and feature relevance that may not hold universally. Moreover, deterministic strategies lack the ability to reason about long-term consequences of adaptation decisions, as they operate myopically on local observations.

In complex and non-stationary environments, optimal adaptation policies may require context-dependent decisions that balance short-term stability against long-term discriminative power. Deterministic approaches are ill-suited to capture such trade-offs, as they cannot accumulate experience or revise their behavior based on past outcomes.

These limitations motivate the transition toward learning-based adaptation mechanisms capable of discovering effective control strategies through interaction and feedback rather than explicit rule design.

6.3.4 Reinforcement Learning as a Generalization of Adaptive Control

Reinforcement Learning can be viewed as a principled generalization of deterministic adaptive preprocessing. While deterministic methods define fixed mappings from observed statistics to adaptation actions, Reinforcement Learning learns such mappings from data by optimizing a reward signal aligned with the inference objective.

From this perspective, Reinforcement Learning subsumes deterministic adaptation as a special case. If the optimal policy corresponds to a fixed heuristic, Reinforcement Learning can recover it. Conversely, if effective adaptation requires context-aware and history-dependent decisions, Reinforcement Learning provides the expressive power necessary to capture such behavior.

Crucially, the decision to introduce Reinforcement Learning in this work is not motivated by algorithmic novelty, but by the demonstrated inadequacy of static and deterministic adaptive approaches in fully addressing window-level instability. Reinforcement Learning is introduced only after validating that adaptivity at the representation level yields tangible benefits, thereby grounding the learning-based approach in empirical evidence rather than assumption.

6.3.5 Validation of the Adaptivity Hypothesis

The progression from static preprocessing to deterministic adaptation and finally to Reinforcement Learning serves as a validation of the central hypothesis of this chapter: robustness to noise in side-channel analysis is fundamentally a representation-level control problem. Experimental comparisons across these stages show a consistent trend: increased adaptivity correlates with improved window-level stability and reduced reliance on temporal aggregation mechanisms such as majority voting [24].

These findings justify the adoption of Reinforcement Learning as the most general and flexible adaptation mechanism within the proposed framework. In the following sections, Reinforcement Learning is

instantiated concretely for feature-aware adaptive denoising, and its integration within the side-channel inference pipeline is described in detail.

6.4 Reinforcement Learning Architecture and Agent Placement

Figure 6.2 illustrates the complete inference pipeline adopted in this work, highlighting the position of the Reinforcement Learning (RL) agent within the overall side-channel analysis framework. The proposed architecture follows a modular and sequential design, encompassing raw signal acquisition, controlled noise injection, preprocessing through windowing and RMS computation, feature extraction, adaptive representation control via Reinforcement Learning, and final classification.

The RL agent is deliberately placed downstream of the feature extraction stage, where it operates as a feature-aware control module. Rather than acting on raw signals or directly modifying classifier parameters, the agent dynamically modulates the extracted feature representation prior to classification. This design choice enables adaptive robustness to noise while preserving the stability, interpretability, and modularity of the inference pipeline.

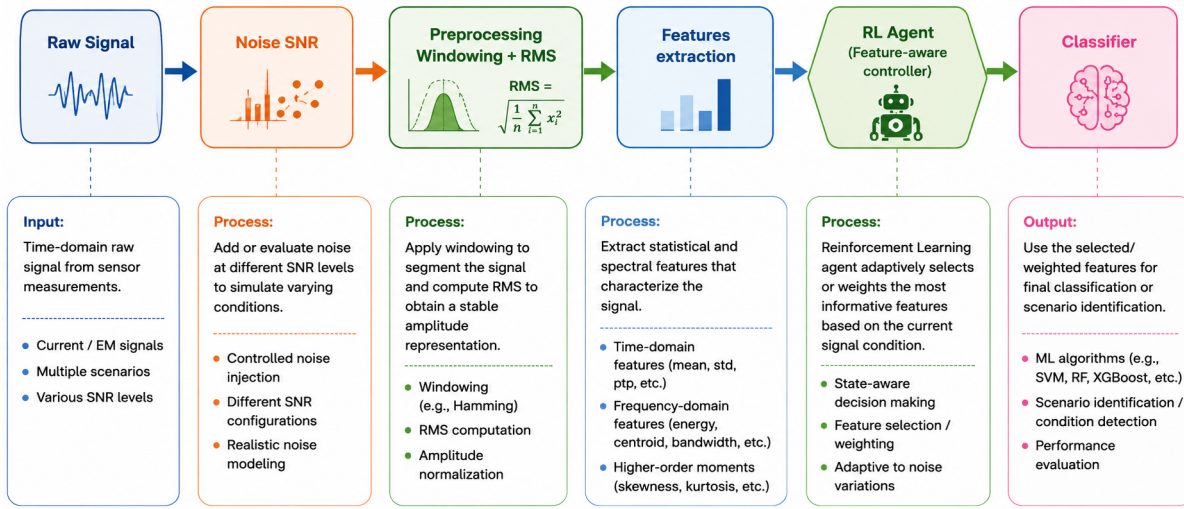


Figure 6.2: Feature-aware reinforcement learning pipeline for adaptive side-channel inference. Raw electromagnetic or current traces are first subjected to controlled noise injection and preprocessing through windowing and RMS computation. Features are then extracted and passed to a Reinforcement Learning agent, which performs adaptive representation control before the final classification stage.

6.4.1 Overall Pipeline Description

The pipeline begins with the acquisition of raw electromagnetic or current traces, which represent high-frequency physical measurements of device activity. To systematically evaluate robustness, controlled noise is injected at the signal level, parameterized through the signal-to-noise ratio (SNR). This step enables reproducible analysis of noise effects without assuming a specific physical noise model.

Subsequently, the noisy signal undergoes preprocessing through windowing and RMS computation. This transformation reduces the dimensionality of the raw signal while preserving temporal energy variations that are known to carry discriminative information in side-channel analysis [27]. Feature extraction is then applied to obtain a compact representation in the time and frequency domains, following established practices in machine-learning-based side-channel analysis [33, 25].

At this stage, the feature vector constitutes the primary representation used for inference. Rather than directly feeding this representation to the classifier, an adaptive control layer based on Reinforcement Learning is introduced. The RL agent operates on the extracted features and dynamically modulates their contribution before classification. The classifier, which may be a classical machine learning model or a deep architecture, remains fixed throughout the adaptation process.

6.4.2 Rationale for Feature-Level Agent Placement

A central design decision in this work concerns the placement of the RL agent at the level of the feature representation, rather than operating directly on the raw signal or immediately after noise injection. This

choice is motivated by both theoretical considerations and practical constraints inherent to side-channel analysis.

Operating on raw signals would require the agent to act in a very high-dimensional and highly noisy space. Raw side-channel traces typically consist of thousands to millions of samples per acquisition, with complex temporal correlations and hardware-dependent artifacts. Introducing a learning-based control mechanism at this level would dramatically increase the state and action space, leading to poor sample efficiency, unstable learning dynamics, and limited interpretability. Moreover, direct manipulation of raw signals would implicitly require assumptions about the physical nature of noise and signal generation, which contradicts the objective of avoiding explicit noise modeling [34].

Similarly, placing the agent immediately after noise injection or preprocessing would restrict its action space to signal-level transformations such as filtering or denoising. While such approaches may improve signal quality under specific conditions, they tightly couple adaptation to physical signal characteristics and risk suppressing discriminative information. In addition, signal-level adaptation often lacks direct feedback from the classifier, making it difficult to align adaptation decisions with the ultimate inference objective.

In contrast, feature-level placement offers a principled compromise between expressiveness and tractability. Feature extraction acts as an information bottleneck that condenses raw signals into a structured and semantically meaningful representation. Operating at this level significantly reduces dimensionality while preserving the aspects of the signal most relevant for classification. As a result, the RL agent can reason about representation quality in a space that is both manageable and directly linked to inference performance.

6.4.3 Feature-Aware Control and Interpretability

Another key advantage of feature-level adaptation lies in interpretability. Features in side-channel analysis often have well-defined physical or statistical meaning, such as signal energy, spectral concentration, or temporal variability [33]. By modulating feature contributions rather than raw samples, the RL agent performs actions that can be interpreted as emphasizing stable and informative dimensions while suppressing unstable or noisy ones.

This design choice aligns with the broader objective of maintaining transparency and explainability in security-related inference systems. Unlike end-to-end deep learning approaches, where adaptation mechanisms may become opaque, feature-aware control enables inspection of how and when adaptation occurs. This property is particularly relevant in cybersecurity contexts, where operator trust and forensic analysis are essential [46].

6.4.4 Decoupling Adaptation from Classification

A further motivation for the proposed architecture is the deliberate decoupling between adaptation and classification. The classifier is trained independently using supervised learning and remains fixed during inference. The RL agent does not modify classifier parameters, nor does it attempt to replace the classifier. Instead, it regulates the representation presented to the classifier based on feedback derived from classification outcomes.

This separation provides several benefits. First, it preserves the stability and reliability of the classifier, avoiding the risk of catastrophic forgetting or drift associated with online retraining. Second, it enables the use of different classifiers within the same adaptive framework, including both classical models such as Random Forests [6] and deep architectures combining convolutional and recurrent layers [50]. Third, it simplifies the learning problem faced by the RL agent, as the reward signal directly reflects the utility of representation control rather than the joint optimization of multiple components.

6.4.5 Comparison with End-to-End Learning Approaches

End-to-end deep learning approaches have been successfully applied to side-channel analysis, particularly in profiling attacks [21, 53]. However, such approaches typically rely on large labeled datasets and assume relatively stable acquisition conditions. Moreover, end-to-end adaptation to noise often requires retraining or fine-tuning the entire network, which is impractical in many operational scenarios.

The proposed architecture adopts a complementary perspective. By introducing adaptivity at the representation level and keeping the classifier fixed, the system achieves robustness without sacrificing

modularity or interpretability. Reinforcement Learning serves as an external control layer that enhances window-level stability while preserving the benefits of established classification models.

6.4.6 Summary

In summary, placing the Reinforcement Learning agent downstream of feature extraction represents a deliberate and principled design choice. Feature-level adaptation offers an effective balance between control expressiveness, learning tractability, and interpretability. It avoids the pitfalls of raw signal manipulation and enables direct alignment between adaptive actions and classification performance. This architectural decision forms the foundation for the experimental evaluation presented in the following sections and underpins the effectiveness of the proposed feature-aware adaptive denoising framework.

6.5 State, Action, and Reward Design

The effectiveness of a Reinforcement Learning framework critically depends on the design of its state representation, action space, and reward function. In the proposed feature-aware adaptive denoising framework, these elements are carefully defined to align the learning process with the objective of improving window-level classification stability under noisy conditions, while preserving interpretability and modularity.

6.5.1 State Representation

In the considered side-channel inference scenario, noise is not directly observable and cannot be explicitly modeled at the signal level. Consequently, the state representation does not describe the physical properties of the noise process, but rather its effects on the extracted feature representation and on the behavior of the classifier.

Formally, at each time step t , the agent observes a state $s_t \in \mathcal{S}$ defined as:

$$s_t = \phi(\mathbf{f}_t, \mathbf{y}_t), \quad (6.6)$$

where $\mathbf{f}_t \in \mathbb{R}^d$ denotes the feature vector extracted from the current temporal window, and $\mathbf{y}_t$ represents auxiliary information derived from the classifier output.

The state vector is designed to capture indicators of representational quality and stability, rather than raw signal characteristics. Typical components include statistical descriptors of the feature distribution (e.g., variance, energy concentration, or entropy-like measures), temporal instability indicators computed across adjacent windows, and confidence-related metrics derived from classifier predictions, such as class probability dispersion or margin values. This design enables the agent to infer whether the current representation is reliable, ambiguous, or unstable, without requiring explicit noise estimation.

Importantly, the state representation remains low-dimensional and semantically meaningful, ensuring tractable learning dynamics and interpretability of agent behavior.

6.5.2 Action Space

The action space $\mathcal{A}$ defines the set of adaptive operations available to the agent for modulating the feature representation prior to classification. Actions are designed to operate exclusively at the feature level, consistent with the architectural choices discussed in Section 5.4.

To systematically assess the role of adaptivity, two action-space configurations are considered in this work.

Light Action Set. In the first experimental configuration, the agent is provided with a restricted set of lightweight actions aimed at minimally perturbing the representation. This set includes:

- *No operation (No-Op)*, where the feature vector is passed unchanged to the classifier;
- *Local smoothing*, implemented through a Savitzky–Golay filtering operation applied to selected feature dimensions;
- *Mild denoising*, where limited regularization is applied to reduce high-frequency variability in unstable features.

These actions are intentionally conservative and serve as a baseline to validate whether even limited adaptivity can improve window-level stability compared to static preprocessing.

Strong Action Set. In a second experimental configuration, the action space is extended to include more aggressive adaptive operations. In addition to the light actions, this set incorporates stronger feature modulation strategies, such as enhanced smoothing or feature attenuation mechanisms that more decisively suppress unstable dimensions.

The introduction of a stronger action set increases the expressive power of the agent, enabling more substantial representation control at the cost of potentially higher information loss. This configuration is used to investigate the trade-off between stability and discriminative richness and to assess whether Reinforcement Learning can learn to apply strong actions selectively rather than indiscriminately.

6.5.3 Reward Function

The reward function $\mathcal{R}(s_t, a_t)$ provides the feedback signal guiding policy optimization and directly links adaptive actions to inference quality. In the proposed framework, rewards are defined to reflect improvements in classification performance and representational stability at the window level.

At each time step, the reward is computed as:

$$r_t = g(\hat{y}_t, c_t), \quad (6.7)$$

where $\hat{y}_t$ denotes the classifier prediction for the current window after adaptation, and c_t represents confidence-related measures associated with the prediction.

Positive rewards are assigned when adaptive actions lead to increased classification confidence, improved class consistency across adjacent windows, or correct predictions. Conversely, negative rewards penalize actions that degrade confidence, increase instability, or induce misclassification. Importantly, the reward function does not incentivize signal reconstruction or denoising accuracy in isolation, but evaluates actions based on their utility for decision-making.

This formulation ensures alignment between local adaptation decisions and the global inference objective, encouraging the agent to learn policies that selectively apply adaptive operations when they are beneficial and refrain from unnecessary intervention when the representation is already stable.

6.5.4 Design Considerations

The separation between light and strong action sets reflects a deliberate experimental design choice. Rather than assuming a priori that aggressive adaptation is always beneficial, the framework explicitly tests whether increased adaptivity yields measurable gains and under which conditions. This approach allows the Reinforcement Learning agent to be evaluated not only in terms of performance improvement, but also in terms of behavioral selectivity and interpretability.

By grounding the design of states, actions, and rewards in measurable properties of the feature space and classifier behavior, the proposed framework establishes a transparent and principled foundation for adaptive side-channel inference. The impact of the different action-space configurations is empirically evaluated in the following section.

6.6 Experimental Evaluation

This section presents the experimental evaluation of the proposed feature-aware reinforcement learning framework. The analysis is structured progressively, starting from a conservative adaptive configuration and subsequently extending the action space to investigate more expressive control strategies. All experiments follow the same baseline protocol defined in the previous chapters, ensuring consistency and fair comparison across different levels of adaptivity.

6.6.1 Baseline and Light Action Set

The proposed approach is evaluated by directly comparing the performance of a static supervised baseline with that of an adaptive feature-aware selection mechanism based on Reinforcement Learning, under a controlled noise scenario characterized by additive noise with a signal-to-noise ratio (SNR) equal to 20 dB. In all configurations, the Random Forest classifier is adopted as the reference classification model. The comparison therefore isolates the effect of representation handling, as the learning algorithm and classification architecture remain unchanged across experiments.

As a baseline, the Random Forest classifier is trained on a mixed dataset composed of clean signals and noisy signals corrupted at an SNR of 20 dB, and subsequently evaluated exclusively on noisy data

at the same SNR level. This training strategy corresponds to a standard data augmentation approach, aimed at partially compensating for the distribution mismatch between clean and noisy conditions.

The effect of this baseline configuration is illustrated by the window-level confusion matrix reported in Figure 6.3, which highlights the impact of additive noise on single-window classification performance.

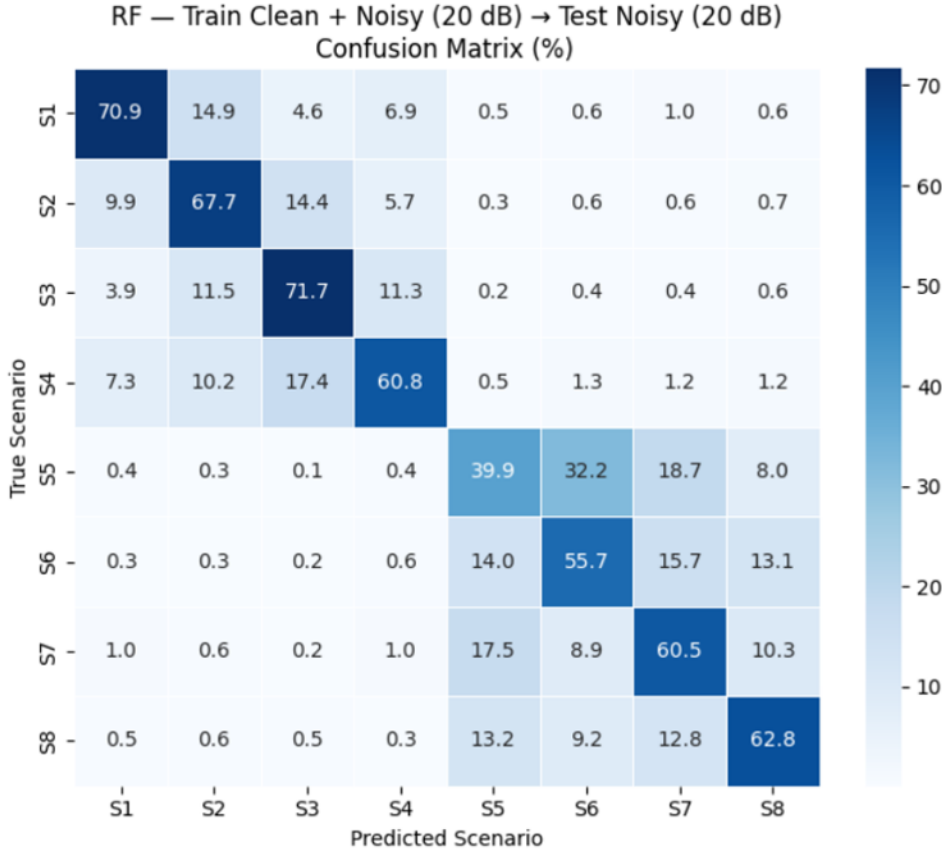


Figure 6.3: Window-level confusion matrix obtained under the static supervised baseline at SNR = 20 dB. The Random Forest classifier is trained using a mixed dataset of clean and noisy signals and evaluated exclusively on noisy data.

Under this configuration, the classifier achieves a window-level accuracy of approximately 62%. This result indicates that data augmentation provides a partial mitigation of the domain shift induced by noise, allowing the model to retain a limited degree of robustness. However, the observed accuracy also reveals a structural limitation in the classifier’s ability to reliably discriminate between scenarios at the level of individual temporal windows.

The confusion matrix in Figure 6.3 further confirms that additive noise does not eliminate discriminative information entirely, but perturbs its distribution within the feature space, leading to unstable and ambiguous local decisions for scenarios characterized by similar side-channel signatures.

This baseline evaluation establishes a reference point for the subsequent analysis. All adaptive configurations introduced in the following sections are assessed under the same experimental conditions, enabling a fair and controlled comparison with respect to this static supervised baseline.

6.6.2 RL Feature-Aware Selection with Light Action Set

After establishing the static supervised baseline, an adaptive mechanism based on Reinforcement Learning is introduced to investigate whether representation-level adaptivity can improve window-level classification robustness under noisy conditions.

The Reinforcement Learning agent is positioned downstream of RMS-based feature extraction and upstream of the Random Forest classifier, acting as a feature-aware selector of the input representation. The agent is trained offline and observes, for each temporal window, the effect of a limited set of local preprocessing actions applied to the extracted features.

In this initial configuration, the action space is deliberately restricted to a lightweight and conservative set of operations:

- no modification (no-op);
- smoothing through moving average;
- smoothing through Savitzky–Golay filtering.

It is important to emphasize that the classification model remains unchanged with respect to the baseline configuration and that no form of temporal aggregation or majority voting is introduced. All performance variations observed in the following experiments are therefore attributable exclusively to the adaptive action of the Reinforcement Learning agent on individual windows. Using 30 files per scenario, the introduction of the RL-based feature-aware selector leads to a window-level accuracy of approximately 69%, corresponding to an absolute improvement of about 7–8 percentage points with respect to the static supervised baseline. This gain is achieved without modifying the classifier, retraining the model, or aggregating predictions across windows.

The improvement is entirely attributable to the adaptive behavior of the agent, which dynamically selects the most appropriate local preprocessing action as a function of the observed state in the feature space. This result provides a first experimental validation of the central hypothesis of this chapter: even limited adaptivity at the representation level can significantly enhance robustness to noise at the window level. The window-level confusion matrix reported in Figure 6.4 shows that the performance improvement is not uniform across all scenarios. Certain classes, such as scenarios 5 and 8, appear almost perfectly separable even in the presence of noise, while others—most notably scenarios 2, 3, and 4—continue to exhibit structured confusion patterns.

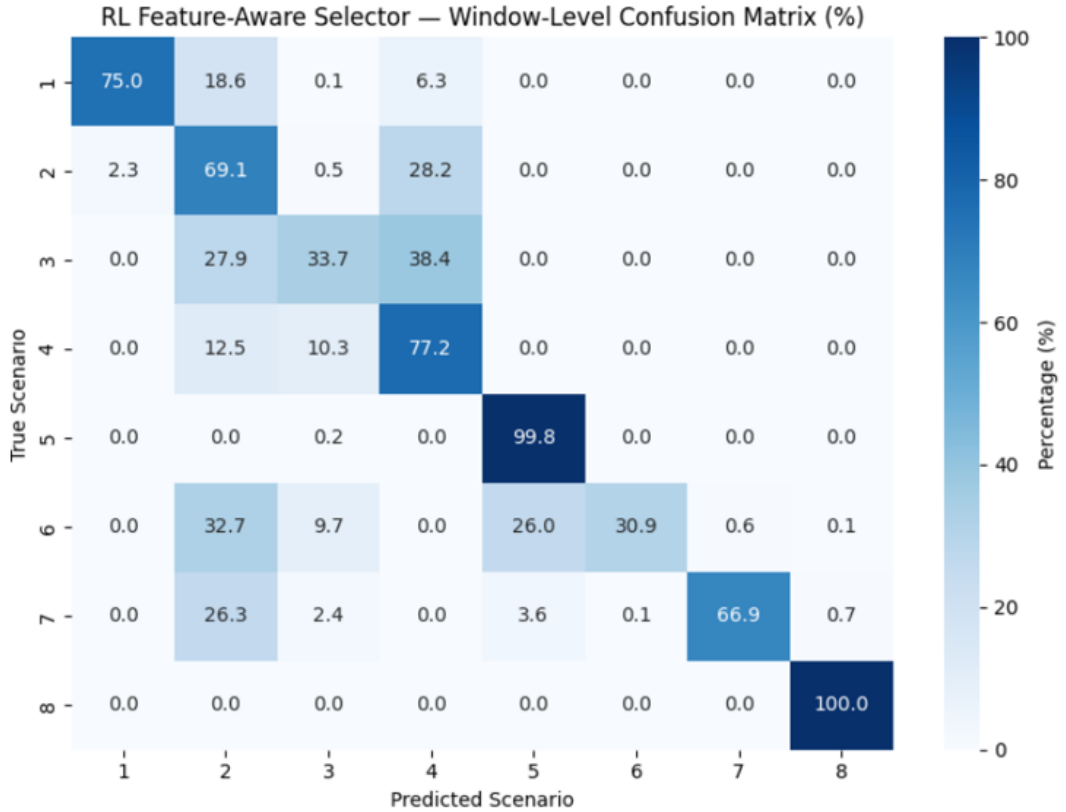


Figure 6.4: Window-level confusion matrix obtained using the RL feature-aware selector with light action set (no-op, moving average, Savitzky–Golay), evaluated on noisy data at SNR = 20 dB with 30 files per scenario

This behavior suggests that the adaptive selection of local preprocessing operations is effective in improving feature stability, but that the main limitation lies in the intrinsic similarity of the side-channel signatures associated with specific scenarios. In other words, while the RL agent mitigates noise-induced

instability, it cannot fully resolve ambiguities rooted in the representational geometry of the RMS feature space.

Overall, these results demonstrate that Reinforcement Learning, even when constrained to an extremely simple action space and operating exclusively at the single-window level, can consistently outperform a static supervised baseline relying on data augmentation alone. At the same time, the residual confusion patterns indicate the need for more expressive adaptive mechanisms capable of acting directly on feature relevance rather than solely on signal smoothing. To further investigate the behavior of the RL feature-aware selector under reduced domain complexity, the number of files per scenario is reduced from 30 to 20, while keeping the overall pipeline architecture and the available action set unchanged.

In this configuration, the RL-based selector achieves a window-level accuracy of approximately 75.9%, significantly outperforming both the static supervised baseline (approximately 61–62%) and the RL configuration with 30 files per scenario (approximately 69%).

The confusion matrix reported in Figure 6.5 shows a further consolidation of class separability. Several scenarios, including scenarios 1, 5, 7, and 8, achieve classification accuracies close to or exceeding 95%. Residual confusion remains primarily concentrated among scenarios 2, 3, and 4, which are characterized by intrinsically similar RMS-based side-channel signatures.

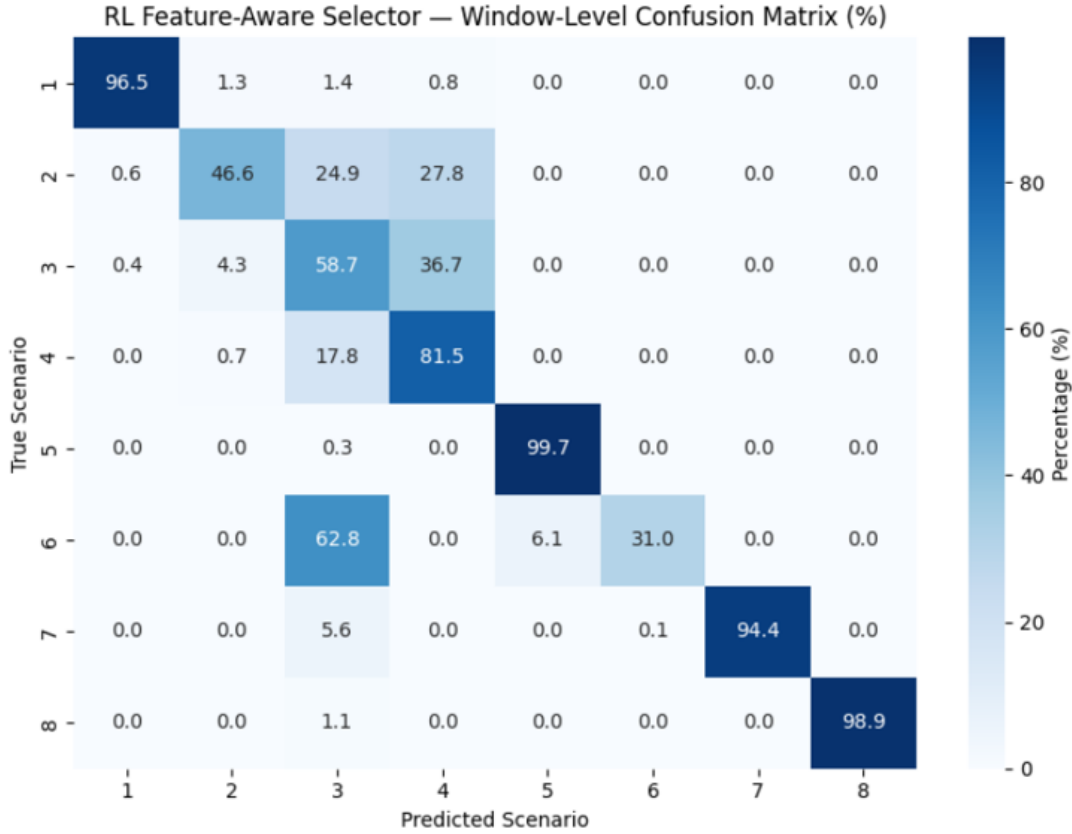


Figure 6.5: Window-level confusion matrix obtained using the RL feature-aware selector with light action set, evaluated on noisy data at SNR = 20 dB with 20 files per scenario

The observed performance gain can be interpreted as the combined effect of two factors. First, reducing the number of files per scenario decreases the complexity of the state space observed by the RL agent, limiting policy fragmentation and promoting more coherent action selection. Second, lower intra-class variability stabilizes the relationship between observed state, selected action, and resulting reward, facilitating effective policy learning even in the absence of explicit generalization mechanisms. Overall, these results confirm that the proposed RL feature-aware selector benefits from a controlled reduction in domain complexity and is able to effectively exploit the statistical stability of the feature space when the balance between data variability and policy capacity is favorable. Importantly, the persistence of structured confusion among specific scenarios reinforces the interpretation that the current performance ceiling is not due to a limitation of the adaptive mechanism itself, but rather to the limited expressiveness of the available action space.

6.6.3 Extension of the Action Space: Feature Weighting

In the previously discussed configurations, the action of the Reinforcement Learning agent is limited to signal-level smoothing operations. While these actions are effective in improving the statistical stability of the extracted features, they are not sufficient to substantially modify the geometry of the decision space when scenarios exhibit intrinsically similar side-channel signatures.

To overcome this limitation, it is natural to extend the set of actions available to the RL agent by introducing mechanisms for dynamic feature weighting. Unlike smoothing-based actions, feature weighting enables the agent to explicitly modulate the contribution of individual feature components as a function of the observed conditions, acting directly on the representation used for classification rather than on the signal itself.

The introduction of feature weighting as an RL action represents a coherent evolution of the proposed framework. In this configuration, the agent no longer intervenes exclusively on signal regularization, but operates directly within the feature space, attenuating unstable or noisy components while emphasizing those that are statistically more robust or more relevant for discrimination.

This perspective allows the noise problem to be addressed not merely as a degradation to be compensated through denoising, but as a selective distortion of the informative representation. By rebalancing the relative importance of features in an adaptive manner, the agent can partially correct such distortions and improve class separability at the window level.

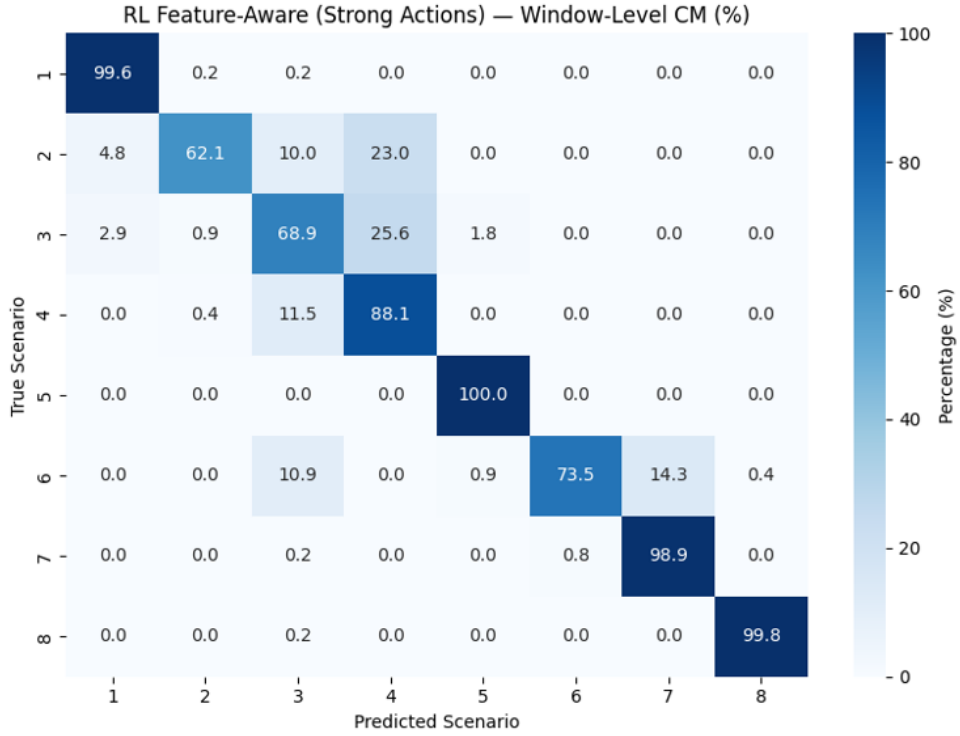


Figure 6.6: Window-level confusion matrix obtained using the RL feature-aware selector with extended action space including feature weighting, evaluated at SNR = 20 dB with 20 files per scenario

Following this rationale, the action space is extended to include feature weighting operations derived from the Random Forest feature importance. Both global and selective weighting strategies are considered, allowing the agent to dynamically regulate the influence of individual features while preserving the overall pipeline architecture.

Using 20 files per scenario and operating at an SNR of 20 dB, the resulting RL feature-aware configuration achieves a window-level accuracy of 86.36%, without any form of temporal aggregation or majority voting. This corresponds to an absolute improvement of more than +24 percentage points with respect to the static supervised baseline and a substantial gain over the configurations relying exclusively on smoothing-based actions.

The analysis of the window-level confusion matrix in Figure 6.6 provides several relevant qualitative insights. Some scenarios, in particular Scenarios 1, 5, 7, and 8, appear almost perfectly separable, with diagonal accuracy values exceeding 98–99%. These results indicate that the corresponding side-channel

signatures remain strongly discriminative even under noisy conditions.

In such cases, the RL agent predominantly selects the no-op action, preserving the original representation. This behavior confirms that denoising or feature modulation is not applied indiscriminately, but only when required, highlighting the contextual and selective nature of the learned policy.

Other scenarios, notably Scenarios 2, 3, 4, and 6, still exhibit residual confusion, concentrated on specific class pairs rather than being randomly distributed. This observation suggests that, for these scenarios, noise induces a more pronounced distortion of the feature space, requiring more aggressive adaptive actions. In these cases, the agent more frequently selects smoothing or feature weighting operations, achieving a clear improvement over the baseline even if perfect separation is not attained.

An additional insight into the behavior of the RL agent is provided by the distribution of selected actions, reported in Table 6.1. The no-op action remains the most frequent choice (approximately 36%), followed by smoothing-based actions (moving average and Savitzky–Golay filtering), and, to a smaller but non-negligible extent, feature weighting (approximately 14–15%).

This distribution confirms that the agent has learned a selective and context-dependent policy. Feature weighting and denoising actions are not applied systematically, but only when the observed state indicates that noise has produced a measurable degradation in the feature space and in the confidence of the classifier. Conversely, when the extracted features remain stable and discriminative, the agent correctly preserves the original representation, avoiding unnecessary intervention.

Table 6.1: Distribution of actions selected by the RL agent in the extended action-space configuration (20 files per scenario, SNR = 20 dB).

Action	Frequency (%)
No-op	36.2
Moving Average	22.0
Savitzky–Golay	27.3
Feature Weighting	14.5

Overall, these results demonstrate that noise does not eliminate discriminative information, but alters its distribution within the feature space. The RL feature-aware selector is able to compensate for such distortions through targeted and adaptive actions, leading to a significant improvement in window-level classification performance.

The introduction of feature weighting marks a key transition from signal-level stabilization to representation-level control. This evolution reduces the reliance on global noise reduction and establishes the basis for more expressive adaptive mechanisms, paving the way for future extensions toward feedback-driven and temporally aware control strategies, in which the agent can iteratively adapt its actions based on the temporal evolution of classifier decisions.

6.6.4 Strong Reinforcement Learning Actions for Feature-Aware Adaptive Control

The experimental evidence obtained with the initial set of Reinforcement Learning actions, mainly based on simple temporal smoothing techniques such as moving average and Savitzky–Golay filtering, revealed intrinsic structural limitations of signal-level adaptation. Although these operations are effective in attenuating high-frequency noise components, they act uniformly across the representation and implicitly assume that all features are equally affected by noise and equally relevant for discrimination.

However, side-channel feature spaces are inherently heterogeneous: different features exhibit distinct statistical properties, noise sensitivity, and discriminative relevance depending on the application state and operating conditions [33, 27]. The analysis of window-level confusion matrices confirmed that, under noisy conditions, several scenarios continue to exhibit structured overlap even after signal-level smoothing. This observation indicates that performance degradation cannot be attributed solely to local temporal fluctuations, but rather to more complex distortions affecting the geometry of the feature space itself.

To address these limitations, the action space of the RL agent is extended to include a set of stronger and more structured operations explicitly designed to act at the representation level. These actions enable the agent to modulate the contribution of individual features, suppress unreliable components, and exploit temporal coherence, without assuming a predefined noise model or modifying the classifier parameters.

From a methodological perspective, this extension marks a fundamental transition in the proposed framework. The RL agent is no longer confined to generic denoising, but operates as a decision-aware controller of the side-channel observation pipeline. By adapting the feature representation in response to both noise conditions and classifier uncertainty, the agent can compensate for complex and non-stationary distortions of the feature distribution induced by noise. This capability is essential for achieving robustness in realistic operational environments, where noise characteristics vary across time, devices, and execution phases [46, 34].

The design of the strong RL actions follows three core principles: (i) local adaptivity, driven by the stability of the current temporal window; (ii) statistical awareness, leveraging variance, outlier detection, and confidence-related indicators; and (iii) real-time compatibility, avoiding non-causal operations or long-term aggregation mechanisms that would compromise online deployment.

The individual actions are described below.

Action 0 — No-op (Direct Feature Forwarding). The no-op action consists in forwarding the extracted feature vector to the classifier without applying any transformation. This action represents the reference behavior of the system and is selected when the current window exhibits sufficient stability and the classifier expresses high confidence in its prediction.

The explicit inclusion of a no-op action is crucial to prevent unnecessary intervention. In adaptive systems, indiscriminate denoising or transformation may degrade informative content and reduce class separability [24]. The frequency of no-op selection therefore provides an indirect measure of the intrinsic robustness of the classifier and of the stability of the feature representation under noise.

Action 1 — Instability-Aware Feature Weighting. In this action, individual feature components are weighted according to their local variance computed within the current window. Features exhibiting high instability are attenuated, while more stable components retain higher influence.

The motivation for this strategy lies in the observation that noise does not affect all features uniformly. Certain features may experience large variance increases under noise while contributing little to class discrimination [25]. By operating directly in the feature space, instability-aware weighting reduces the impact of such components without completely discarding them, preserving potentially useful information for downstream classification.

Unlike classical temporal filters, this operation is feature-specific, adaptive, and independent across dimensions, making it particularly suitable for heterogeneous side-channel representations.

Action 2 — Selective Feature Masking (Outlier Suppression). This action suppresses feature components that exhibit anomalous values with respect to the local statistical distribution, typically identified as outliers.

Impulsive or burst noise can generate isolated peaks that disproportionately influence classifier decisions despite lacking discriminative meaning [41]. Selective masking acts as a non-parametric robustness mechanism, removing these spurious contributions while leaving the remaining feature vector unchanged. This action is especially effective in scenarios characterized by localized distortions rather than global noise amplification.

Action 3 — PCA-Based Projection and Reconstruction. In this action, the feature vector is projected onto a lower-dimensional subspace using Principal Component Analysis (PCA), followed by reconstruction in the original feature space.

PCA is known to concentrate informative variance along a limited number of components while relegating noise-dominated directions to lower-variance subspaces [20]. By discarding components associated with high-variance but weakly informative directions, this action reduces noise amplification while preserving the global structure of the representation. Unlike local filtering or masking, PCA-based reconstruction introduces a structured and global form of regularization.

Action 4 — Temporal Consistency Filtering. This action combines the current feature vector with a weighted average of feature representations extracted from previous windows.

Application-level side-channel patterns typically evolve smoothly over time, whereas noise tends to be weakly correlated across adjacent windows [16]. Exploiting temporal consistency therefore allows

the agent to reduce noise-induced variability without resorting to file-level aggregation or majority voting. This action represents a compromise between strictly local processing and temporal exploitation, maintaining compatibility with real-time constraints.

Action 5 — Classifier-Confidence-Guided Scaling. In this action, the feature vector is scaled as a function of the classifier confidence associated with the predicted class.

Low confidence often indicates that informative components have been attenuated or distorted by noise. Confidence-guided scaling allows controlled amplification of feature contributions under uncertainty, while having minimal effect when confidence is already high. This mechanism acts as a fine-grained regulation tool rather than a primary denoising strategy, and is particularly effective under moderate uncertainty conditions [53].

6.6.5 Strong RL Actions: Experimental Evaluation and Classifier-Aware Behavior

From a design perspective, the introduction of the strong RL action set enables the agent to transition from the role of a simple filter selector to that of an adaptive controller operating directly within the feature space. Each action is explicitly associated with a specific type of distortion—feature instability, impulsive outliers, structured noise, or temporal incoherence—allowing targeted and interpretable interventions.

This set of operations constitutes the methodological foundation for the subsequent comparison between different classification architectures, while keeping the role and structure of the RL agent unchanged. In this way, the effect of the decision model can be isolated without altering the adaptive control mechanism.

RL Feature-Aware Selector with Random Forest Classifier

The integration of the RL feature-aware selector with a Random Forest classifier leads to a substantial improvement over the static Random Forest baseline under noisy conditions ($\text{SNR} = 20 \text{ dB}$). In this configuration, the system achieves a window-level accuracy of 82.28%, confirming the effectiveness of the RL agent in mitigating, at least partially, the distortion introduced by noise through adaptive interventions in the feature space.

The analysis of the confusion matrix in Figure 6.7 reveals that, despite the high overall accuracy, residual overlaps persist between specific classes. In particular, some scenarios exhibit confusion patterns that can be traced back to feature distributions strongly altered by noise, which a tree-based classifier struggles to separate cleanly. This behavior suggests that, although the RL selector improves the quality of the representation, the discriminative capacity of the model remains constrained by the static and non-sequential nature of the Random Forest classifier.

To provide additional insight into the computational and operational cost of the adaptive mechanism, Table 6.2 reports the distribution of selected actions for two configurations: a reduced subset of 5 files per scenario, used as a lightweight exploratory setting, and the full configuration with 20 files per scenario.

Table 6.2: Action usage distribution for the RL feature-aware selector combined with a Random Forest classifier. The first two columns refer to a reduced configuration with 5 files per scenario, while the last two columns correspond to the full configuration with 20 files per scenario.

RL + RF Action	%	Num	%	Num
No-op (raw RF)	10.3	240	12.2	1130
Instability-aware weighting	17.7	410	15.7	1460
Feature masking	30.3	704	43.2	4008
PCA projection	19.8	460	5.1	469
Temporal consistency	17.1	396	18.1	1677
Confidence-guided scaling	4.7	110	5.8	536

A particularly relevant observation emerges from the action distribution. Selective feature masking is the most frequently chosen action (approximately 30%), indicating that noise often manifests as localized outliers that compromise classifier decisions. PCA-based projection follows (approximately 20%), while instability-aware weighting and temporal consistency actions are both selected with frequencies around

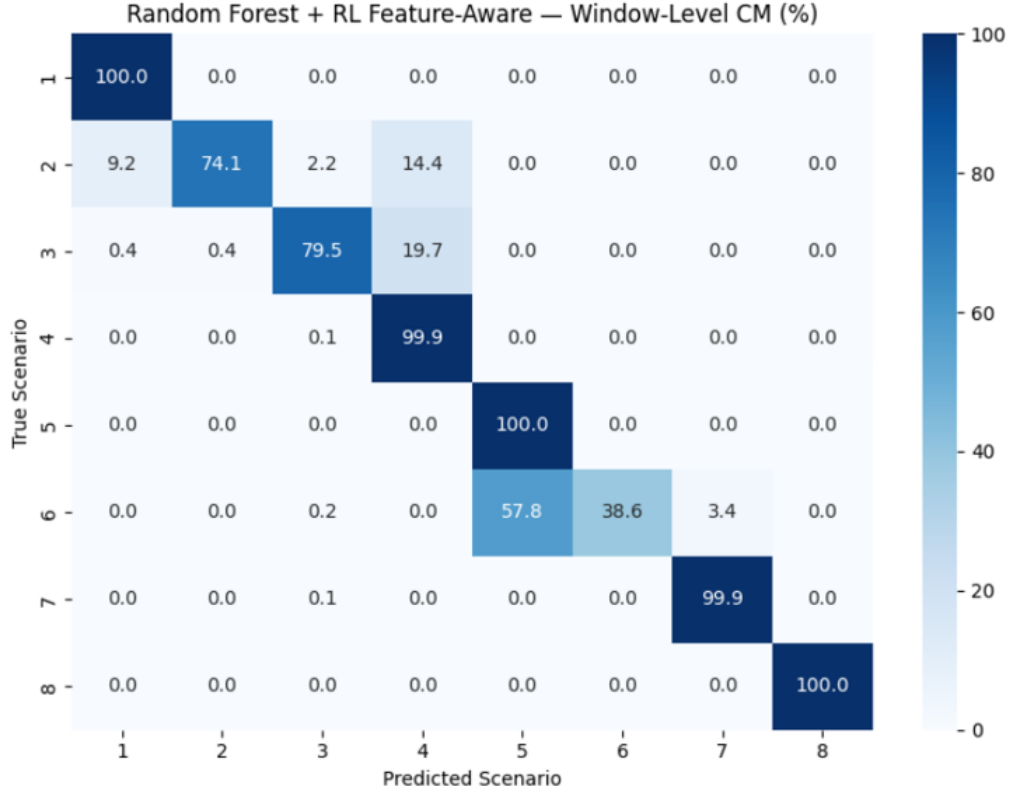


Figure 6.7: Window-level confusion matrix obtained using the RL feature-aware selector combined with a Random Forest classifier at SNR = 20 dB

17–18%, confirming that noise does not act solely as impulsive disturbance but also induces structural instability and temporal incoherence in the feature space.

The relatively low frequency of the no-op action (approximately 10%) indicates that, under noisy conditions, the agent is forced to intervene in the majority of windows to maintain an acceptable level of classifier confidence. Conversely, confidence-guided scaling is used marginally, coherently with its role as a fine-grained adjustment mechanism rather than a primary noise compensation strategy.

Overall, these results demonstrate that the RL feature-aware selector is capable of significantly improving the performance of a Random Forest classifier by introducing adaptive and interpretable behavior. However, the persistence of confusion among certain classes suggests that the dominant limitation no longer resides in the action selection mechanism itself, but in the classifier’s ability to model complex non-linear relationships and residual temporal dependencies in the feature space. This observation naturally motivates the exploration of more expressive classification models while preserving the same RL-based adaptive control layer.

RL Feature-Aware Selector with CNN+LSTM Classifier

The integration of the RL feature-aware selector with a CNN+LSTM-based classifier produces a marked performance leap compared to the Random Forest configuration. Under noisy conditions at SNR = 20 dB, the system achieves a window-level accuracy of 99.13%.

The corresponding confusion matrix shows an almost perfect separation across all scenarios, including those that remained critical in the Random Forest case. This result confirms the superior capability of deep architectures to preserve and exploit residual discriminative information even in the presence of significant noise-induced distortions.

Notably, this performance improvement does not arise from more aggressive compensation actions, but rather from a drastic reduction in the need for intervention by the RL agent. Table 6.3 reports the action usage distribution for the CNN+LSTM configuration.

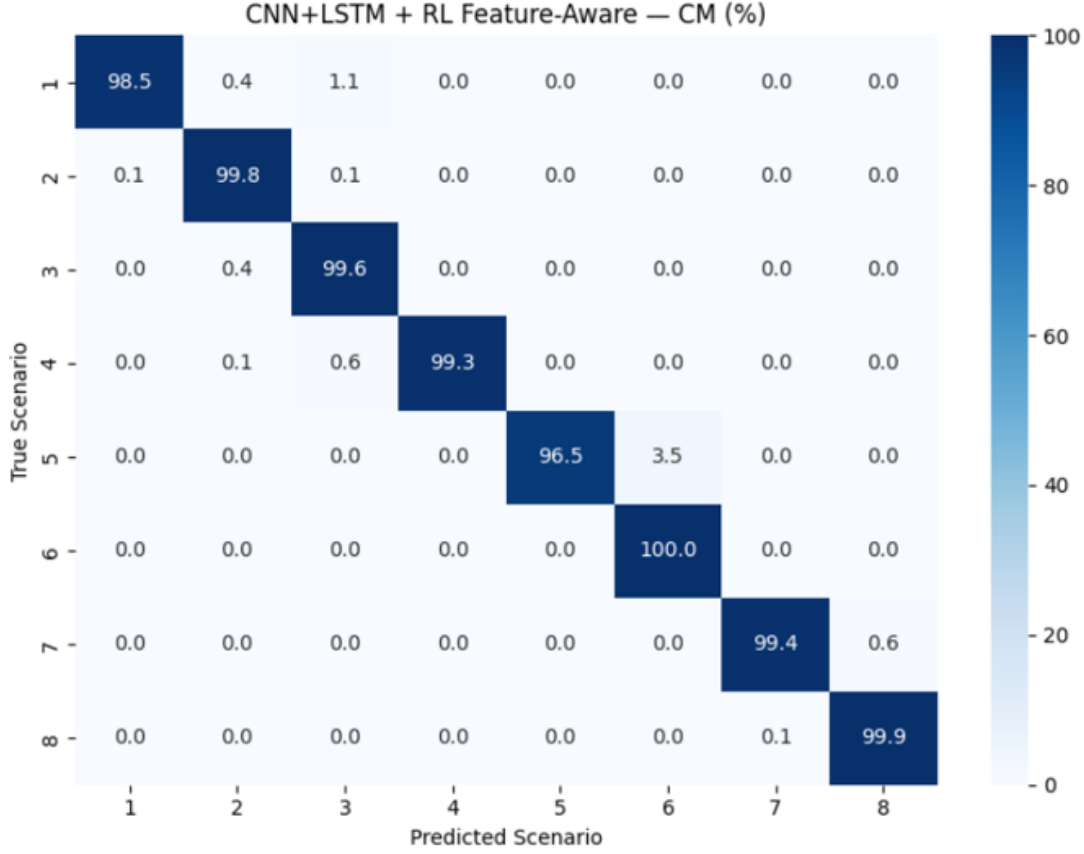


Figure 6.8: Window-level confusion matrix obtained using the RL feature-aware selector combined with a CNN+LSTM classifier at SNR = 20 dB

Table 6.3: Action usage distribution for the RL feature-aware selector combined with a CNN+LSTM classifier.

RL + CNN+LSTM Action	%	Num	%	Num
No-op (raw features)	60.5	1307	48.6	4197
Instability-aware weighting	6.0	129	16.5	1429
Feature masking	5.7	124	16.0	1386
PCA projection	6.1	131	6.3	547
Temporal consistency	19.2	415	12.3	1061
Confidence-guided scaling	2.5	54	0.2	20

In this setting, the no-op action becomes dominant (approximately 60.5%), indicating that, for the majority of windows, the feature representation provided to the classifier is already sufficiently informative for a correct decision. The RL agent therefore assumes the role of a selective supervisor rather than a systematic corrector.

Instability-aware weighting, feature masking, and PCA projection are activated with limited frequencies (approximately 5–6% each), suggesting that the CNN+LSTM architecture can autonomously absorb a significant portion of the variability introduced by noise. The convolutional layers learn robust local patterns, while the LSTM component effectively models temporal dependencies between consecutive windows, reducing the impact of transient fluctuations in the feature space.

The only action still exhibiting substantial usage is temporal consistency filtering (approximately 19.2%), consistent with the observation that the RL agent intervenes primarily when detecting dynamic incoherence across adjacent windows. In these cases, the action acts as a stabilization mechanism that reinforces decision continuity, further improving classifier reliability during noisy transitions.

Overall, these results demonstrate that the RL feature-aware selector is not intrinsically tied to a specific classification architecture, but adapts its policy to the expressive capacity of the underlying model. When paired with a deep classifier, the agent naturally converges toward a more conservative

and targeted policy, intervening only when strictly necessary. This behavior confirms that the role of Reinforcement Learning in the proposed framework is not to replace the discriminative power of the classifier, but to complement it by providing an adaptive control layer that maximizes robustness under noisy conditions.

6.7 Threat Model and Side-Channel Defense Perspective

The experimental results discussed so far have demonstrated that Reinforcement Learning can be effectively employed to improve robustness and stability in side-channel-based inference under noisy conditions. However, the implications of the proposed feature-aware adaptive control extend beyond performance enhancement alone.

In this section, the proposed framework is reinterpreted from a security-oriented perspective. Rather than considering Reinforcement Learning solely as a mechanism for improving classification robustness, we analyze its role as an adaptive representation-level defense capable of limiting the exploitable information available to an adversary. As illustrated in Figure 6.9, the RL agent dynamically controls the feature representation exposed to the learning pipeline, enabling a conceptual transition from traditional signal-level countermeasures toward informational and representation-aware protection strategies.

By contextualizing the framework within established categories of side-channel defenses, the proposed approach highlights how adaptive feature control can simultaneously preserve task performance while degrading the statistical consistency exploited by side-channel attacks.

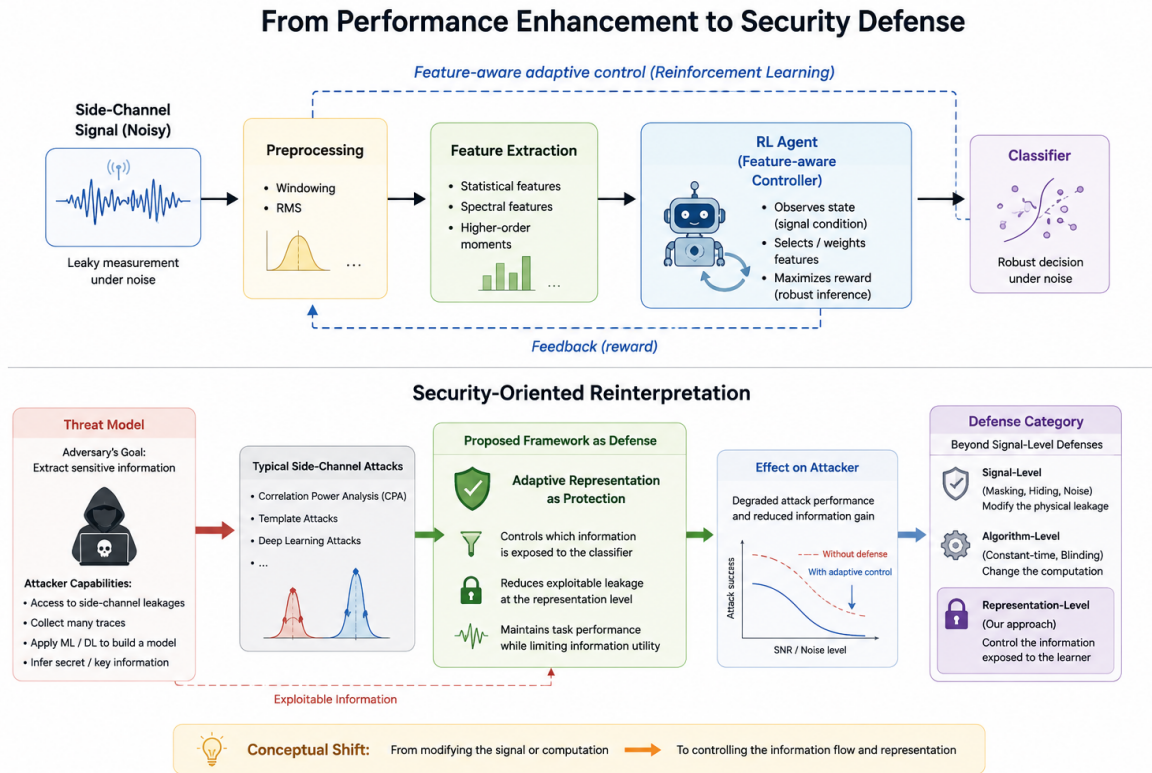


Figure 6.9: Security-oriented reinterpretation of the proposed Reinforcement Learning framework for adaptive side-channel protection. The upper section illustrates the feature-aware adaptive control pipeline, where an RL agent dynamically adjusts feature representations under noisy conditions to maintain robust inference performance. The lower section reframes the same mechanism within a defensive threat model, highlighting the transition from traditional signal-level countermeasures toward representation-level protection. Instead of suppressing physical leakage directly, the proposed approach aims to control and distort the information exposed to learning-based adversaries, thereby reducing the exploitability of side-channel observations while preserving operational functionality.

6.7.1 Defense Strategies in Side-Channel Attacks

Countermeasures against side-channel attacks are commonly classified in the literature into three main categories, depending on the system layer they target and on the type of threat they aim to mitigate.

These categories reflect different assumptions about the attacker’s capabilities and the level of control available to the defender.

Physical and Hardware-Level Defenses. Physical defenses act directly on the physical phenomena that generate side-channel leakage, by modifying or masking the electrical or electromagnetic behavior of the device. Typical examples include noise injection on the power rail, electromagnetic shielding, power balancing, clock randomization, and hardware masking countermeasures.

These solutions can be effective even against attackers with physical proximity to the device. However, they usually entail high implementation costs, require invasive hardware modifications, and are difficult to deploy on existing systems or in distributed monitoring scenarios. For these reasons, hardware-level countermeasures fall outside the scope of this work.

Software and Algorithmic Defenses. Software-based defenses aim to reduce the correlation between side-channel signals and the internal state of the system by modifying execution behavior without altering the hardware. Representative techniques include operation randomization, insertion of dummy instructions, execution order shuffling, and controlled temporal perturbations.

While such approaches do not eliminate the physical signal, they alter its statistical structure, making information extraction more challenging for an attacker. Software defenses constitute an important step toward flexibility and deployability, but they are typically static and do not adapt to changing operational or environmental conditions.

Informational and Representation-Level Defenses. A third, more recent and less explored class of defenses consists of informational or representation-level approaches. These defenses do not target the physical signal itself, but rather how the signal is transformed, represented, and made available after acquisition.

In this paradigm, the objective is not to hide or suppress the side-channel signal, but to limit the exploitable information accessible to an unauthorized observer, while preserving utility for a legitimate one. This approach introduces the concept of *informational asymmetry*, whereby the defensive system has access to a stabilized and informative representation of the signal, whereas the attacker is constrained to a partial, unstable, or structurally degraded view of the same physical phenomenon.

This paradigm is particularly suitable for monitoring systems, Intrusion Detection Systems (IDS), and edge-computing scenarios, where side-channel signals are inherently acquired and processed by a centralized or distributed infrastructure.

6.7.2 Threat Model and Assumptions

The threat model adopted in this work is designed to reflect realistic side-channel attack scenarios in monitoring, IDS, and distributed embedded environments, while avoiding unrealistic assumptions about the attacker’s ability to directly interfere with the physical system or manipulate the side-channel source.

We consider a system A , consisting of an embedded or IoT device executing software applications. During normal operation, A inevitably generates physical side-channel signals, such as current consumption and electromagnetic emissions, as an intrinsic consequence of its computational activity. In line with classical side-channel literature, these signals cannot be completely eliminated and may be observed by a passive attacker only under conditions of physical proximity.

In particular, direct acquisition of the raw side-channel signal by an attacker typically requires close-range access (often within a few meters), dedicated probes, specialized instrumentation, and physical access to the environment in which the device operates. Although well studied, such attacks are poorly scalable and difficult to apply in large-scale, distributed, or remote scenarios.

We define a second system B as the legitimate defensive system, responsible for acquiring the side-channel signals generated by A through integrated sensors, controlled measurement interfaces, or edge monitoring nodes. System B executes a complete processing pipeline, including preprocessing, feature extraction, and classification, with objectives such as:

- predictive analysis of hardware faults;
- monitoring of operational states in embedded devices;
- detection of anomalous or compromised software behavior;

- support for IDS functions in industrial or IoT environments.

System B is assumed to be trustworthy and uncompromised. Its internal components, including intermediate representations, extracted features, and trained models, are not directly accessible to the attacker.

The attacker is modeled as a passive observer interested in performing side-channel profiling to infer information about the internal state or application scenario executed on system A . The attacker does not inject anomalies, alter system behavior, or manipulate the physical signal at its source, and is not assumed to compromise the physical sensors used by system B .

Within this model, two attacker profiles are distinguished:

1. a physically proximate attacker, attempting to directly acquire side-channel signals using probes and instrumentation;
2. a remote attacker, lacking physical access, who attempts to exploit side-channel data that have been acquired, processed, or distributed by system B , for example in the form of streams, logs, datasets, or monitoring outputs.

This work explicitly focuses on the second scenario, which is particularly relevant in realistic distributed monitoring contexts. In this case, the attacker does not observe the raw physical signal in its original form, but attempts to extract sensitive information by attacking the system that processes and disseminates side-channel data.

Here, the threat does not lie in the mere observability of the physical signal, but in the ability to exploit the informational structure of side-channel representations to derive useful statistical correlations. Representative attack objectives include:

- reconstruction of application scenarios from shared side-channel features;
- inference of computational load or internal device states;
- extraction of temporal or statistical patterns from monitoring data.

It is worth emphasizing that this threat model closely reflects the operational assumptions of modern monitoring and Intrusion Detection Systems. In such systems, side-channel data are not collected for adversarial analysis, but for legitimate purposes such as diagnostics, anomaly detection, and operational awareness. As a result, side-channel information is often centralized, stored, or transmitted across networked infrastructures.

This centralization significantly amplifies the attack surface, as it enables a remote adversary to target the informational representations produced by the monitoring pipeline rather than the physical signal itself. In this context, the security of the system does not solely depend on preventing signal acquisition, but on controlling how much exploitable information is exposed through processed representations. This observation motivates the need for defense mechanisms that operate at the representation level and are compatible with large-scale, distributed monitoring architectures.

6.7.3 Reinforcement Learning as Informational Defense

The contribution of this work is positioned precisely within this threat model. Rather than attempting to eliminate or mask the physical side-channel signal, we introduce an adaptive informational defense that controls how the signal is represented and exposed after acquisition.

Through a Reinforcement Learning-based control mechanism, system B operates in a dual and asymmetric manner:

- for the legitimate system owner, the RL controller acts as an adaptive selector that compensates for environmental noise and disturbances, improving the stability and reliability of classification;
- for a remote attacker, the same control mechanism introduces instability, structural degradation, and loss of informational coherence in the observable representation, yielding distorted or weakly correlatable signals.

The conceptual separation between a defensive channel and an attacker-observable channel does not imply the existence of two distinct physical signals. Instead, it reflects an asymmetry in access and observability of the informational representation. The defensive system maintains a dynamically adapted

and controlled view of the signal, while the attacker remains constrained to a partial and uncontrolled representation.

In summary, the proposed threat model assumes a realistic, passive attacker with limited physical access, and defines a defensive objective that is not the suppression of side-channel leakage, but the systematic reduction of exploitable information after acquisition. This objective is achieved through adaptive, decision-aware control of the representation, enabled by Reinforcement Learning.

A key aspect of the proposed approach is the inherently dual role played by the Reinforcement Learning controller. While previous sections have focused on its ability to enhance robustness and classification stability under noisy conditions, the same adaptive mechanism can be interpreted as a defensive tool from a security standpoint.

Specifically, Reinforcement Learning introduces a controlled asymmetry between the representation used internally by the defensive system and the one that may be observable or inferable by an attacker. This asymmetry enables the system to simultaneously maximize utility for legitimate monitoring tasks and minimize the informational value of the exposed representation for unauthorized observers.

This perspective naturally leads to an additional evaluation dimension beyond classification accuracy. If adaptive representation control is effective as an informational defense, its impact should also be observable in terms of reduced exploitability of side-channel data by an external observer.

For this reason, the following experimental analysis does not only assess performance under noise, but also investigates the extent to which the proposed Reinforcement Learning mechanism degrades the ability of an adversarial observer to perform reliable side-channel profiling. This evaluation provides a concrete validation of the defensive value of adaptive representation control in realistic monitoring scenarios, and establishes a direct link between robustness-oriented adaptation and security-driven obfuscation.

6.8 Proactive Reinforcement Learning–Based Side-Channel Defense

Before introducing the proposed adaptive defense mechanism, it is necessary to clarify the scope, assumptions, and operational boundaries of the Reinforcement Learning (RL) agent adopted in this work. Figure 6.10 provides a conceptual overview of the considered scenario and explicitly highlights what is—and what is not—controlled by the defensive system.

6.8.1 Fundamental Assumptions

The proposed defense does not assume any control over the physical side-channel signal itself. Electromagnetic emissions and power consumption traces are an unavoidable consequence of computation and remain equally observable by any passive entity in physical proximity to the device. No attempt is made to interfere with the attacker’s acquisition chain, to filter the signal, or to selectively expose different physical signals to different observers.

This assumption is consistent with established side-channel literature, which models the leakage as an intrinsic physical phenomenon rather than as an artifact that can be selectively hidden or suppressed in realistic deployments [46, 34].

Instead, the key observation underlying this work is that side-channel vulnerability does not primarily reside in the existence of the physical signal, but in the *informational structure* that can be extracted from it. As formally established by Standaert et al., side-channel leakage is best characterized in terms of mutual information between the internal state of the system and the observable signal, rather than by signal visibility alone [48].

Consequently, a system may remain physically observable while becoming progressively less exploitable if the statistical relationship between execution state and observable leakage is actively destabilized.

6.8.2 Defense by Execution Control Rather Than Signal Manipulation

In the proposed framework, the Reinforcement Learning agent operates *upstream* of the physical leakage, at the level of the computational behavior that induces it. The RL agent does not alter the logical functionality of the application, but dynamically modulates execution-level characteristics such as timing, scheduling, micro-architectural activity, and workload organization.

This design choice is grounded in prior work demonstrating that execution-level randomization and software-only countermeasures can significantly reduce the effectiveness of side-channel attacks without

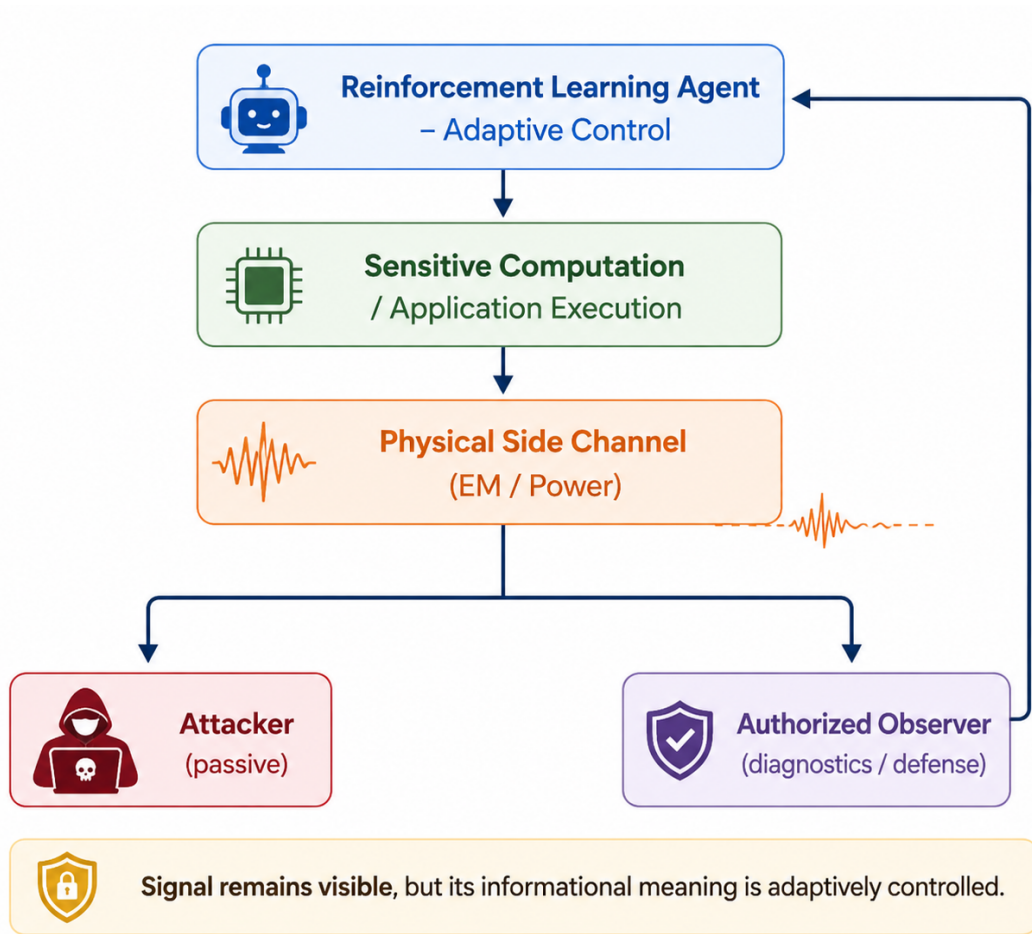


Figure 6.10: Conceptual model of adaptive informational defense based on reinforcement learning. The physical side-channel remains observable by any passive observer, while an RL agent adaptively controls the execution behavior of the system to destabilize the mapping between observable leakage and sensitive information.

modifying the physical measurement process [40, 1]. These approaches exploit the fact that side-channel leakage is a consequence of *how* computations are executed rather than *what* computations are performed.

However, classical countermeasures are typically static or pseudo-random, making them potentially predictable in the presence of adaptive adversaries. Reinforcement Learning is introduced in this work as a generalization of these techniques, enabling closed-loop, context-aware, and non-stationary control of execution behavior.

6.8.3 Roles and Information Flow in the Proposed Model

Figure 6.10 explicitly separates the roles of the entities involved in the system:

Sensitive Computation. This block represents the execution of the application on the embedded or IoT device. It encompasses the CPU activity, memory access patterns, scheduling decisions, and micro-architectural behavior that collectively give rise to physical side-channel leakage. The functional outcome of the computation remains unchanged throughout the adaptive process.

Physical Side Channel (EM / Power). This block represents the unavoidable physical leakage generated by the system as a consequence of computation. The signal is identical for all observers and is not modified, filtered, or selectively exposed by the defense mechanism.

Attacker (Passive Observer). The attacker is modeled as a passive entity capable of observing the physical side-channel signal but unable to interact with the system, influence execution, or access internal

representations. The attacker is not part of any feedback loop and relies on the statistical stability of the leakage-to-information mapping to perform profiling or inference.

Authorized Observer (Diagnostics / Defense). The authorized observer represents the legitimate monitoring system, which acquires the same physical signal but operates within the system’s trust boundary. This observer has access to internal performance metrics, classification outcomes, and system objectives, and provides feedback to the RL agent.

Reinforcement Learning Agent. The RL agent acts as an adaptive controller that receives feedback exclusively from the authorized observer and applies control actions solely to the execution of the sensitive computation. Its objective is to balance functional correctness, performance constraints, and robustness, while systematically degrading the stability of the leakage-to-information mapping exploitable by unauthorized observers.

Crucially, the attacker never participates in the control loop and always observes the same physical signal, albeit generated by a dynamically and adaptively modulated execution process.

6.8.4 Informational Defense Objective

The goal of the proposed approach is not to eliminate side-channel leakage, but to reduce its *informational exploitability* over time. From an information-theoretic perspective, the defense aims to minimize the mutual information between sensitive internal states and observable leakage for unauthorized observers, while preserving sufficient structure for legitimate monitoring tasks [48].

Reinforcement Learning is employed as a principled mechanism to navigate this trade-off in non-stationary and noisy environments, where static countermeasures may fail or become predictable. By continuously adapting execution behavior in response to observed system conditions, the proposed framework introduces a controlled asymmetry: the signal remains physically observable, but its informational meaning becomes unstable, weakly generalizable, and increasingly difficult to exploit for passive attackers.

This adaptive, execution-level perspective positions the proposed method as a realistic and deployable informational defense for side-channel analysis in embedded, IoT, and distributed monitoring scenarios.

The side-channel signal observed in practical systems does not represent a mere passive measurement of device behavior, but rather the output of a complex physical process influenced by environmental noise, operational variability, and measurement uncertainty. In the proposed framework, these physical components are not directly controlled, nor are artificial disturbances injected at the signal level.

The key observation underlying this work is that side-channel vulnerability does not primarily reside in the physical signal itself, but in the *informational structure* and in the representations derived from that signal. It is the stability, separability, and statistical coherence of feature-level representations that ultimately determine the feasibility of side-channel inference, both for legitimate monitoring systems and for adversarial observers.

Building on this observation, the proposed system introduces an adaptive control mechanism based on Reinforcement Learning, conceived not as a noise generator or as a physical countermeasure, but as a *decision-aware selector of transformations in the feature space*. The RL agent operates exclusively on the extracted features, dynamically modulating their representation in a contextual manner, with the dual objective of preserving discriminability for the legitimate classifier while reducing the amount of exploitable information available to an unauthorized observer.

6.8.5 Adaptive Informational Control Architecture

The proposed architecture can be interpreted as a feedback control system operating at the informational level, as illustrated in Figure 6.11. The underlying physical process corresponds to the side-channel generation mechanism, while a deep neural network classifier, based on a CNN+LSTM architecture, acts as an observer of the system’s informational state.

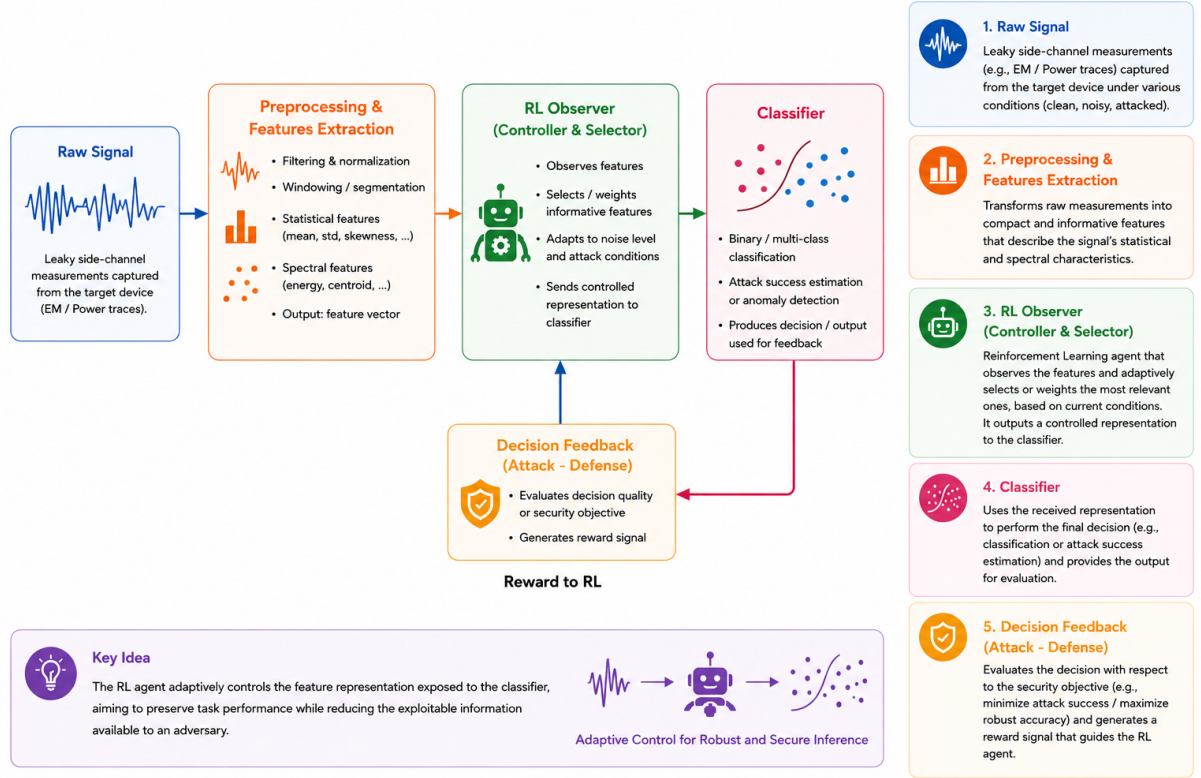


Figure 6.11: Reinforcement Learning–based proactive side-channel defense architecture. The RL controller operates on the feature representation, receiving feedback from classifier decisions and uncertainty metrics, and selecting adaptive transformations without modifying the classifier or the physical signal

Internal measurements such as predictive confidence, output entropy, and decision ambiguity are used to estimate the current informational condition of the system. Based on these signals, the Reinforcement Learning agent acts as a controller, selecting feature-space transformations that correspond to adaptive regularization, informational masking, or representation reshaping strategies.

Crucially, the RL controller does not modify the architecture or parameters of the classifier, nor does it introduce temporal aggregation, majority voting, or file-level decision fusion. Control is strictly local, operating at the level of individual temporal windows and driven exclusively by internal feedback signals. This design ensures compatibility with real-time IDS and monitoring scenarios, while preserving modularity and interpretability.

6.8.6 Dual-Channel Interpretation: Defense vs. Attacker

The behavior of the Reinforcement Learning controller can be interpreted through the lens of two distinct informational channels: a *defensive channel*, used by the legitimate system for monitoring and classification, and an *attacker-observable channel*, accessible to an unauthorized observer.

Defensive Channel. Within the defensive channel, the RL controller dynamically selects feature transformations aimed at stabilizing the representation and preserving the discriminative capability of the classifier under noise and operational variability. Available actions include local instability reduction, selective masking of dominant components, adaptive regularization through latent-space projections, and local temporal consistency enforcement.

These transformations compensate for informational disturbances without altering the classifier or introducing global aggregation mechanisms. As a result, the system achieves high classification accuracy, near-unity confidence values, and a marked reduction in predictive entropy. The resulting confusion matrices are strongly diagonal, indicating that the internal representation remains stable and coherent over time despite the presence of noise at the signal level.

Attacker Channel. When observed from an adversarial perspective, the same control decisions produce an opposite effect. The adaptive transformations selected by the RL controller intentionally degrade

the informational structure of the observable representation. Rather than injecting physical noise, the controller induces representational instability, loss of temporal coherence, latent-space distortion, and reduced class separability.

Experimentally, this manifests as a significant degradation of attacker performance, with highly non-diagonal confusion matrices, increased output entropy, and reduced predictive confidence. Importantly, the attacker is assumed to employ the same classification model used by the defensive system, demonstrating that the observed degradation is solely attributable to information loss in the representation, rather than to model weakness.

6.8.7 Proactive Defense via Informational Asymmetry

Taken together, these results show that the Reinforcement Learning controller operates as an *adaptive informational separation mechanism*. It preserves the quality of the internal representation required for legitimate monitoring while simultaneously limiting the information exploitable by an unauthorized observer.

This asymmetry is not achieved through physical countermeasures or signal manipulation, but through decision-aware control of the representation downstream of acquisition. As such, the proposed approach realizes a proactive side-channel defense strategy that is fully compatible with real-time IDS and monitoring systems.

Rather than attempting to suppress the side-channel signal, the framework systematically reduces its exploitable informational content after acquisition. This perspective establishes a direct conceptual link between robustness-oriented adaptation and security-driven obfuscation, and motivates the experimental evaluation of attacker-side degradation presented in the following section.

6.8.8 Analysis of RL Action Usage in the Defensive Channel

In the defensive channel, the Reinforcement Learning agent exhibits a strongly conservative and selective behavior, fully consistent with the primary objective of preserving the informational quality of the signal for legitimate classification tasks.

The distribution of selected actions during the test phase is reported below:

- **Action 0:** 1278 (59.2%) — No-op (raw RF)
- **Action 1:** 175 (8.1%) — Instability-aware weighting
- **Action 2:** 218 (10.1%) — Feature masking
- **Action 3:** 128 (5.9%) — PCA projection
- **Action 4:** 315 (14.6%) — Temporal consistency
- **Action 5:** 46 (2.1%) — Confidence-guided scaling

The most frequently selected action is the *No-op (raw RF)*, accounting for 59.2% of all decisions. This result indicates that, for the majority of temporal windows, the observed signal is already sufficiently stable and informative, and does not require corrective intervention. From a control-theoretic perspective, this behavior confirms that the RL agent does not introduce unnecessary transformations and intervenes only when internal state metrics—such as instability, entropy, or decision ambiguity—indicate a genuine degradation of the informational content.

Among the active actions, *Temporal consistency* is the most frequently selected (14.6%), followed by *Feature masking* (10.1%) and *Instability-aware weighting* (8.1%). This distribution suggests that the dominant sources of degradation in the defensive channel are primarily related to temporal fluctuations and to locally dominant or noisy feature components. In these situations, the controller favors soft regularization strategies that reduce variability while preserving the discriminative structure of the representation.

More invasive transformations, such as *PCA projection* (5.9%) and *Confidence-guided scaling* (2.1%), are selected with limited frequency. This behavior is fully consistent with a decision-aware control strategy: such actions are applied only in the presence of pronounced instability or low predictive confidence, avoiding systematic use that could otherwise compromise the fidelity of the signal representation.

Figure 6.12 provides a visual summary of the action usage distribution in the defensive channel.

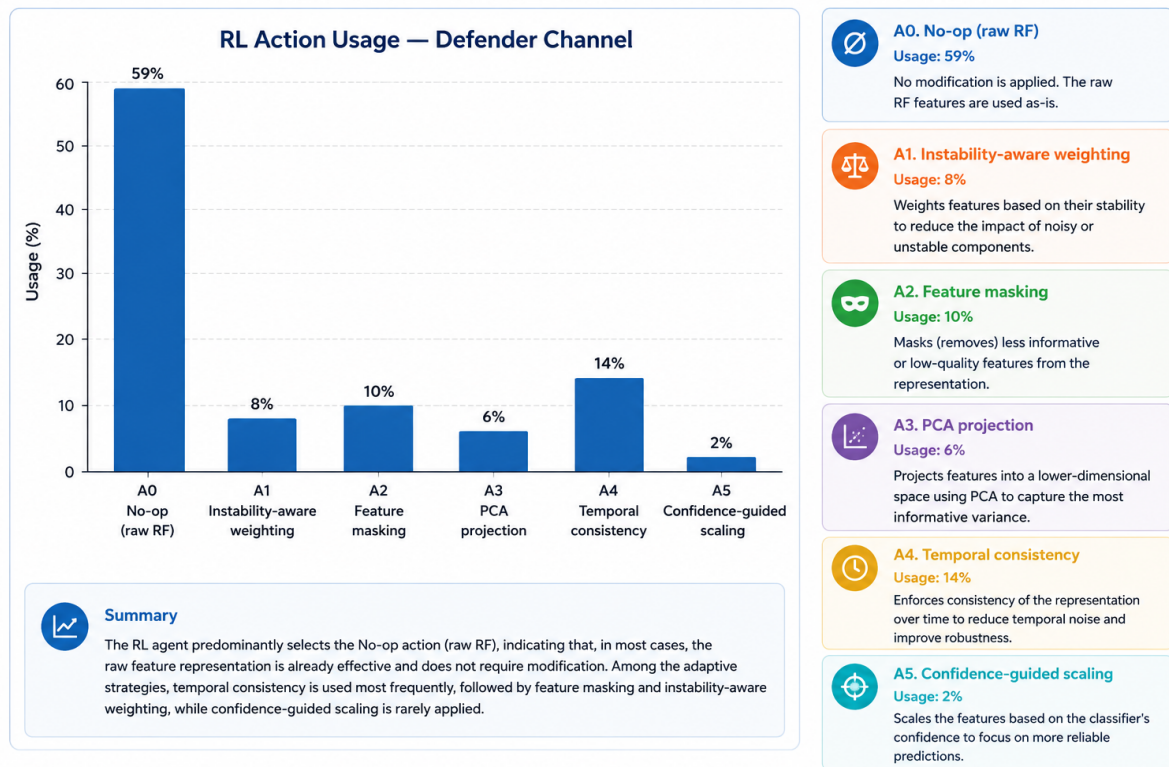


Figure 6.12: Distribution of actions selected by the Reinforcement Learning agent in the defensive channel. The prevalence of the No-op action highlights the conservative and selective nature of the learned control policy.

Overall, the action distribution observed in the defensive channel reflects an adaptive control strategy that prioritizes informational stability and continuity. The Reinforcement Learning agent intervenes in a targeted manner only when signal degradation is detected, thereby preserving classification performance while avoiding unnecessary or overly aggressive transformations.

6.8.9 Analysis of RL Action Usage in the Attacker-Observable Channel

In the attacker-observable channel, the distribution of Reinforcement Learning actions is radically different from that observed in the defensive channel and reveals a deliberately aggressive and diversified behavior, explicitly aimed at disrupting the informational structure of the side-channel representation.

The distribution of selected actions during the test phase is reported below:

- **Action 1:** 407 (18.8%) — Instability amplification
- **Action 2:** 464 (21.5%) — Aggressive feature masking
- **Action 3:** 387 (17.9%) — Adaptive PCA randomization
- **Action 4:** 574 (26.6%) — Temporal shuffling
- **Action 5:** 46 (2.1%) — Amplitude damping
- **Action 6:** 282 (13.1%) — Class collapse

The most frequently selected action is *Temporal shuffling* (26.6%), which explicitly disrupts the temporal coherence between consecutive windows. This intervention is particularly effective against attack models that rely on temporal correlations or sequential patterns, as it introduces misalignment across windows and induces instability in the time evolution of the representation.

The second most frequent action is *Aggressive feature masking* (21.5%), which selectively removes dominant components of the feature vector. This action directly targets the most informative statistical dimensions used in side-channel profiling, drastically reducing class separability and weakening the discriminative structure of the representation.

The actions *Instability amplification* (18.8%) and *Adaptive PCA randomization* (17.9%) further contribute to the destruction of informational structure. Instability amplification deliberately enhances locally unstable components, while PCA-based randomization alters the latent representation through controlled perturbations and permutations in the reduced subspace. Although these transformations preserve the global energy of the signal, they severely compromise its statistical consistency, making the learning of reliable attack models significantly more difficult.

The *Class collapse* action (13.1%) explicitly forces feature representations toward shared centroids, inducing artificial overlap between classes and further reducing discriminability. This mechanism directly targets the geometric structure of the feature space, collapsing inter-class distances and increasing ambiguity from the attacker’s perspective.

In contrast, *Amplitude damping* (2.1%) is selected only marginally. This behavior indicates that simple attenuation of signal amplitude is insufficient, on its own, to effectively degrade side-channel information. Instead, the controller prioritizes transformations that alter structural, temporal, and geometric properties of the representation.

Figure 6.13 summarizes the action usage distribution in the attacker-observable channel.

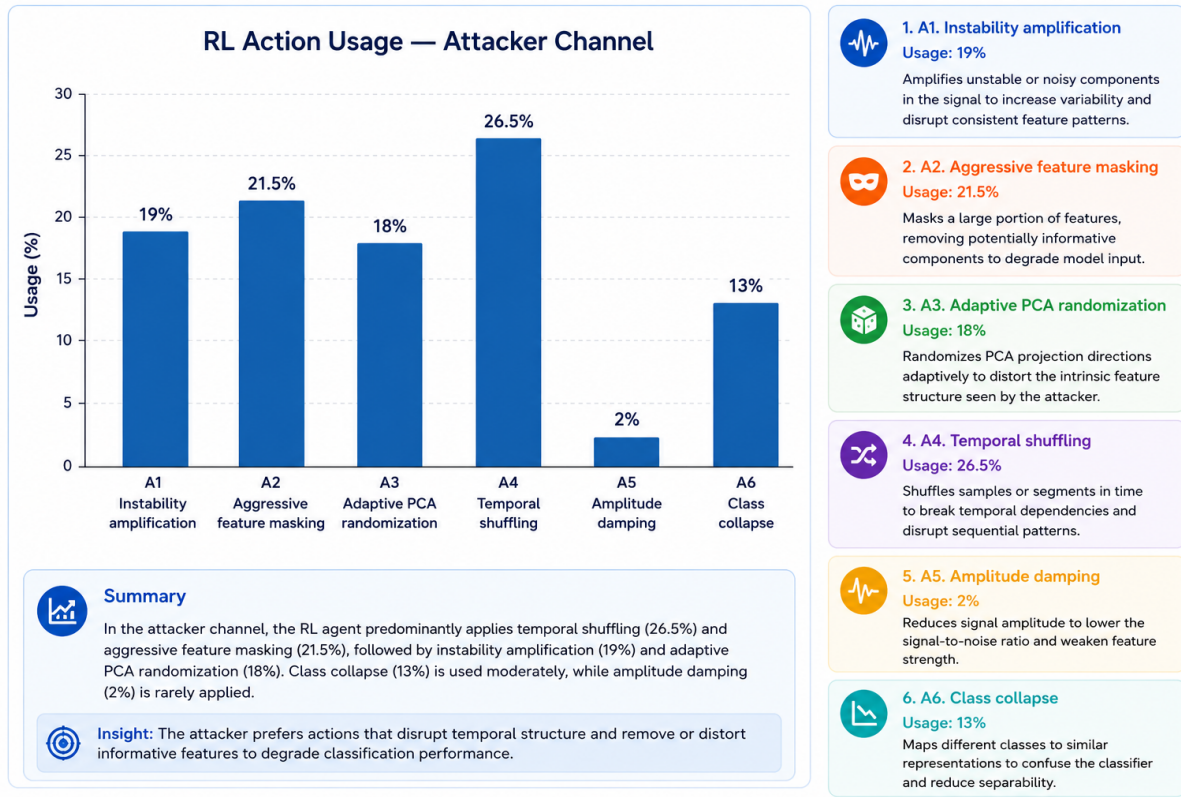


Figure 6.13: Distribution of actions selected by the Reinforcement Learning controller in the attacker-observable channel. The prevalence of aggressive and structure-disrupting actions highlights the intentional degradation of informational content from the attacker’s perspective

Overall, the observed action distribution confirms that the Reinforcement Learning controller operates asymmetrically across the two informational channels. While the defensive channel favors stability-preserving and minimally invasive actions, the attacker-observable channel is dominated by transformations that explicitly destroy temporal coherence, feature stability, and geometric separability.

This asymmetry is not achieved through physical noise injection or signal suppression, but through adaptive, decision-aware manipulation of the feature representation. As a result, the same underlying physical signal supports reliable inference for the legitimate system while becoming statistically ambiguous and weakly exploitable for an adversarial observer. Overall, the attacker-observable channel is characterized by a high diversity of selected actions and by a clear absence of conservative behavior. Unlike the defensive channel, where the RL controller intervenes selectively to preserve informational stability, the policy adopted here is explicitly aggressive and structurally disruptive.

This confirms that the Reinforcement Learning controller does not replicate the same strategy across the two channels, but enforces a clear functional separation: stabilization and preservation of informa-

tion for the legitimate system, and deliberate degradation of temporal coherence, feature stability, and geometric separability for the unauthorized observer.

Crucially, this asymmetry is not obtained through physical countermeasures or artificial noise injection, but emerges as an intrinsic property of adaptive, decision-aware representation control. As such, the proposed framework realizes a form of dynamic informational separation, enabling robust legitimate inference while simultaneously limiting side-channel exploitability.

6.8.10 Confusion Matrix Analysis

Defensive Channel. The confusion matrix obtained for the defensive channel highlights an almost perfect classification performance, with an overall window-level accuracy of 99.63%. All scenarios are correctly identified with percentages close to or equal to 100%, with only negligible residual confusions among semantically related classes.

Figure 6.14 reports the confusion matrix associated with the defensive channel.

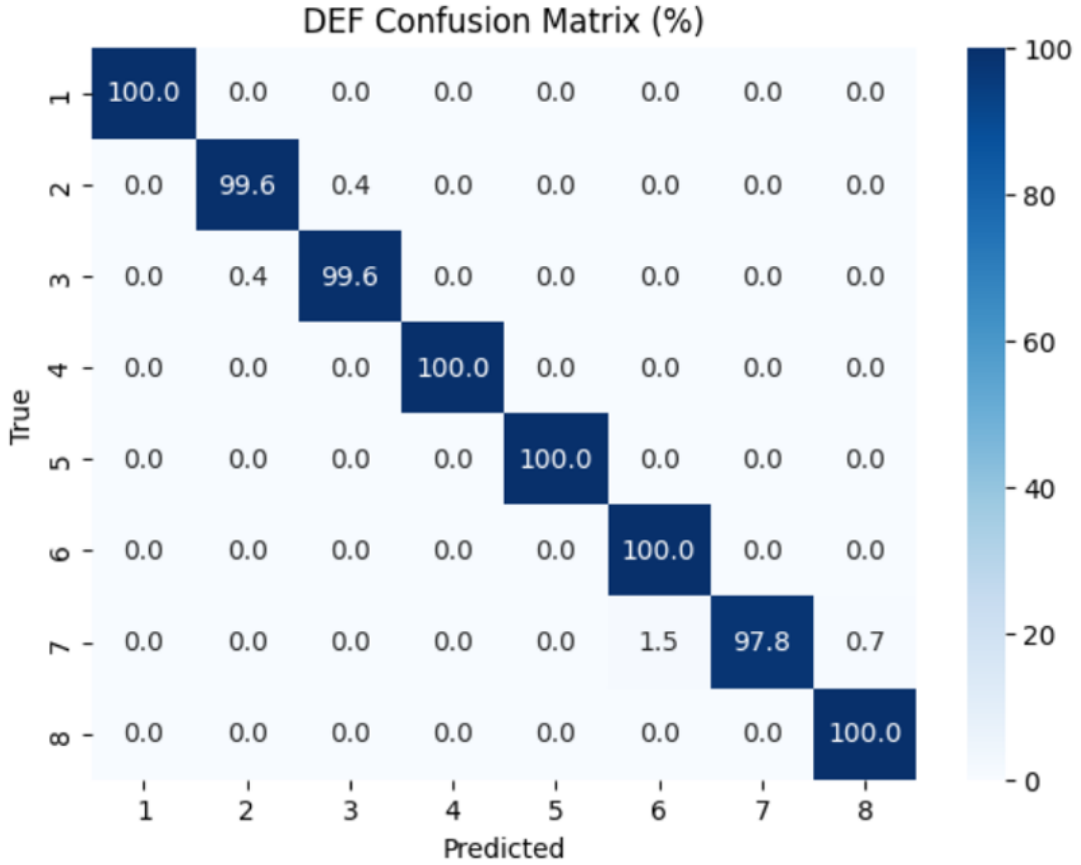


Figure 6.14: Window-level confusion matrix for the defensive channel. The matrix exhibits a strongly diagonal structure, indicating near-perfect class separability under adaptive RL control

More specifically:

- Classes 1, 4, 5, 6, and 8 achieve 100% classification accuracy, indicating complete separability in the feature space after adaptive control.
- Classes 2 and 3 exhibit a minimal reciprocal confusion (approximately 0.4%), suggesting a slight local overlap that does not compromise the global stability of the system.
- Class 7 shows a limited dispersion toward Class 6 (1.5%) and Class 8 (0.7%), a behavior consistent with partially similar temporal or statistical characteristics among the involved scenarios.

Overall, the strongly diagonal structure of the confusion matrix confirms that the RL controller, when operating in the defensive channel, does not degrade the informational representation. On the contrary, it actively contributes to stabilizing and regularizing the feature space in the presence of noise and signal variability. This result is particularly significant because it is achieved without modifying the classifier

architecture, without parameter adaptation, and without introducing temporal aggregation or majority voting mechanisms. The observed robustness emerges solely from local, adaptive, and decision-aware feature control.

Attacker-Observable Channel. A radically different behavior emerges when analyzing the confusion matrix associated with the attacker-observable channel. In this case, the overall window-level accuracy drops sharply to 23.75%, with predictions widely dispersed across multiple classes.

The corresponding confusion matrix is reported in Figure 6.15.

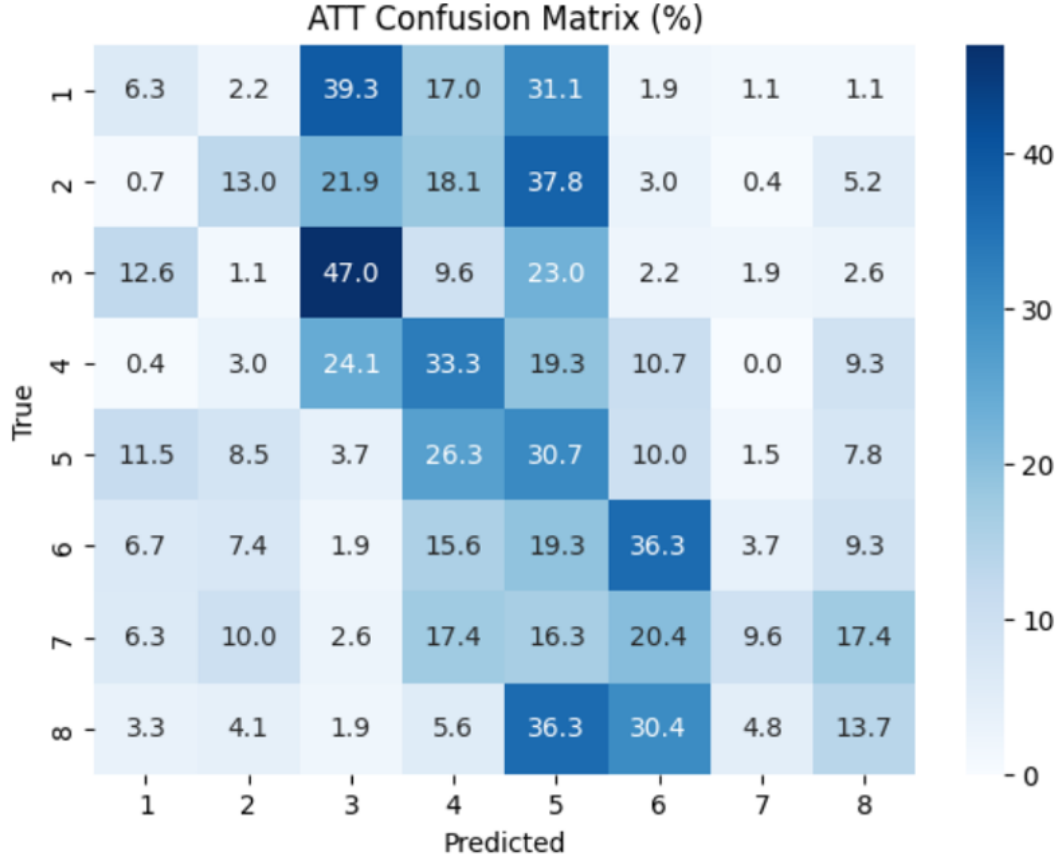


Figure 6.15: Window-level confusion matrix for the attacker-observable channel. The absence of a diagonal structure indicates severe degradation of discriminative information

In contrast to the defensive case:

- No class exhibits a dominant diagonal component.
- Correct classification percentages are generally below 40% and, in many cases, below 20%.
- Predictions are distributed across a heterogeneous set of classes, without stable or repeatable patterns.

For example:

- Class 1 is frequently confused with Classes 3, 4, and 5.
- Class 3 shows a strong dispersion, with only 47% of samples mapped to the correct class and the remainder spread across several others, indicating a highly unstable representation.
- Classes 5, 7, and 8 display diffuse confusion patterns, with predictions scattered over four to five different classes.

This systematic de-stratification of the confusion matrix demonstrates that the attacker is unable to learn a reliable mapping between observations and classes, even when employing the same classification model used by the defensive system. The loss of temporal coherence, the reduction of feature separability, and the induced overlap among classes render side-channel profiling intrinsically unstable and unreliable.

Overall Interpretation and Chapter Conclusions. The direct comparison between the two confusion matrices clearly highlights the effect of the informational asymmetry introduced by the Reinforcement Learning controller. Under identical classification models:

- the defensive system operates on a stabilized, coherent, and highly discriminative representation;
- the attacker observes a destructured, incoherent, and statistically weak representation.

These results confirm that the proposed defense does not merely improve robustness to noise, but enforces a functional separation between the legitimate channel and the attacker-observable one, selectively degrading the adversary’s inference capability without compromising IDS performance.

The experimental evidence consistently shows that the adaptive control mechanism based on Reinforcement Learning induces a genuine informational asymmetry between the defensive system and the unauthorized observer, without modifying the classifier and without altering the physical signal at the hardware level. The RL agent operates exclusively in the feature space and in real time, yet its effect is clearly reflected both in performance metrics and in the statistical structure of the predictions.

The analysis of action usage further clarifies this behavior. In the defensive channel, the control strategy is predominantly conservative: in most cases the agent selects the no-op action, while corrective actions are activated only in the presence of measurable instability, noise, or informational degradation. This feedback-driven strategy preserves the representation whenever the system operates in favorable conditions and intervenes selectively to restore temporal continuity, reduce spurious dominance, or improve predictive confidence. This behavior directly translates into a strongly diagonal defensive confusion matrix with near-perfect accuracy across classes.

In the attacker-observable channel, the controller exhibits the opposite behavior. The selected actions are more aggressive, diverse, and explicitly oriented toward representation destruction, including temporal coherence disruption, aggressive feature masking, latent-space randomization, and class collapse. The absence of conservative actions and the diversity of transformations produce an unstable and non-correlable representation over time, preventing the attacker from learning reliable statistical patterns. This effect is clearly visible in the attacker confusion matrix, which lacks any diagonal structure and yields an overall accuracy below 25%.

The relationship between action usage, informational metrics, and performance is further supported by confidence and entropy analysis. While the defensive channel maintains low entropy and high predictive confidence, the attacker channel exhibits a marked increase in uncertainty, indicating systematic loss of exploitable information. Importantly, this degradation is not achieved through explicit noise injection, but through adaptive manipulation of the representation, guided by the internal informational state of the system.

In conclusion, these results demonstrate that Reinforcement Learning does not act merely as a robustness-enhancing mechanism, but as a true feedback-based informational controller. It learns when to preserve the structure of the signal for legitimate inference and when to destructure it for an unauthorized observer, realizing an adaptive, decision-aware side-channel defense fully compatible with real-time IDS and monitoring scenarios.

Chapter 7

Results and Performance Evaluation

7.1 Summary of the Thesis

This thesis addressed the problem of profiling and classifying application-level behavior in embedded and connected devices through side-channel analysis and machine learning. By exploiting electromagnetic emissions and current consumption measurements, the work demonstrated that relevant behavioral information can be inferred without access to software internals, network payloads, or cryptographic material.

A complete experimental and methodological framework was developed, covering signal acquisition, preprocessing, feature extraction, feature selection, classification, robustness evaluation, and adaptive control. High-frequency side-channel signals were segmented using overlapping RMS-based windows, enabling the transformation of raw measurements into compact and informative representations suitable for learning-based inference. The experimental analysis showed that the temporal resolution of windowing has a dominant impact on classification performance, whereas feature scaling strategies provide only marginal benefits when combined with tree-based classifiers.

Feature engineering emerged as a critical component of the proposed pipeline. A set of time-domain and frequency-domain features was extracted and evaluated using multiple selection strategies. While scenario-specific feature importance highlighted highly discriminative patterns, it was shown to be unsuitable for deployment due to its dependence on a priori knowledge of the running application. To overcome this limitation, a global feature selection strategy based on Recursive Feature Elimination (RFE) was adopted, achieving a favorable balance between dimensionality reduction, classification accuracy, and practical deployability.

Several machine learning models were evaluated, including Random Forest, Support Vector Machines, k-Nearest Neighbors, Logistic Regression, and XGBoost, together with a hybrid deep learning architecture combining Convolutional Neural Networks, Long Short-Term Memory layers, and attention mechanisms. Although the deep learning model achieved the highest overall accuracy, classical models such as Random Forest demonstrated strong robustness, interpretability, and computational efficiency, making them particularly suitable for embedded and resource-constrained environments.

The robustness of the proposed framework was systematically analyzed under noisy conditions through controlled noise injection experiments. The results revealed a clear distinction between window-level and sequence-level inference. While individual windows may exhibit limited discriminative power in noisy environments, aggregating predictions over time through majority voting enables highly reliable sequence-level classification. This behavior confirms the cumulative nature of side-channel information, but also exposes the limitations of static, window-based representations for real-time and adaptive inference.

This observation represents one of the central turning points of the thesis. Noise does not simply degrade the signal uniformly; rather, it selectively modifies the structure of the feature space, affecting some representations more than others and altering the confidence, stability, and separability of the learned decision process. Therefore, robustness cannot be understood only as a matter of classifier accuracy. It must also be interpreted as a problem of representation control.

Building on this insight, the thesis introduced a Reinforcement Learning-based mechanism operating directly in the feature space. Rather than using Reinforcement Learning as an alternative classifier, the proposed approach employs it as an adaptive controller capable of modifying the representation according to feature instability and classifier feedback. The goal is not merely to classify noisy side-channel data, but to actively regulate the informational structure provided to the inference model.

This adaptive mechanism provides the foundation for a broader interpretation of the thesis results.

The same operation that stabilizes the representation for legitimate inference can also be interpreted as a mechanism for reducing the exploitable informational content available to an unauthorized observer. In this sense, the thesis moves beyond the traditional view of side-channel analysis as a passive inference problem and begins to frame it as an active problem of controllable observability.

From this perspective, the work introduces the conceptual basis of a new paradigm, referred to as *Adaptive Inferential Security (SIA)*. SIA emerges from the observation that modern systems cannot be protected solely by controlling access to data, code, or communication channels. Even when direct access is prevented, systems continue to expose behavioral, physical, temporal, and statistical traces that may allow an observer to infer sensitive information.

The central idea of SIA is therefore that security should also be understood as the dynamic control of what can be inferred from observable system behavior. In this paradigm, the observable representation of a system becomes a security surface. The objective is not only to hide information, but to regulate inferential observability: preserving useful information for legitimate monitoring and decision-making while reducing the capability of an adversarial observer to reconstruct sensitive states, activities, or intentions.

Within the scope of this thesis, SIA is not presented as a fully mature theory, but as the natural conceptual projection of the experimental results. The proposed side-channel framework shows that feature representations can be measured, modified, stabilized, and selectively degraded. Reinforcement Learning provides an initial mechanism for performing this control adaptively. The resulting asymmetry between legitimate and adversarial inference constitutes the first step toward a broader theory of adaptive inferential control.

7.2 Original Scientific Contributions

This doctoral research provides several original contributions to the field of side-channel analysis, machine learning, and embedded security. Beyond individual experimental results, the work introduces a coherent methodological framework and a conceptual shift in how robustness, noise, and defense are addressed in side-channel-based inference.

First, the thesis proposes and validates an end-to-end framework for application-level profiling of embedded devices based exclusively on physical side-channel information. Unlike traditional side-channel studies primarily focused on cryptographic key recovery, the proposed framework targets behavioral inference under realistic operating conditions, including heterogeneous applications, protocol-level scenarios, and encrypted communications. The experimental design, spanning eight distinct scenarios and multiple acquisition trials, provides a structured benchmark for evaluating classification performance beyond synthetic or narrowly constrained settings.

Second, the thesis contributes a systematic and comparative analysis of feature extraction and feature selection strategies for side-channel signals. While prior approaches often rely on ad hoc feature choices or scenario-specific optimizations, this work demonstrates that feature importance strongly depends on the underlying scenario. This makes per-class or scenario-specific feature selection unsuitable for realistic inference settings. By explicitly identifying this limitation and validating a global feature selection strategy based on Recursive Feature Elimination, the thesis establishes a principled trade-off between accuracy, generalization, dimensionality reduction, and deployability.

Third, the thesis offers an in-depth investigation of robustness under noisy conditions, which remains an underexplored aspect in much of the side-channel literature. Through controlled noise injection and signal-to-noise ratio analysis, the work clarifies the distinction between window-level and sequence-level inference. While majority voting over time enables high sequence-level accuracy, the analysis reveals that this mechanism primarily masks instability at the window level rather than resolving it. This observation represents a key conceptual contribution, as it reframes robustness not merely as a classification problem, but as a representational challenge.

Fourth, the thesis introduces a feature-aware interpretation of noise. In this view, noise is not treated only as a uniform degradation of the raw signal, but as a dynamic and selective distortion of the feature space. This perspective departs from conventional static preprocessing pipelines and motivates the exploration of adaptive inference mechanisms capable of reacting to changes in representation quality, feature stability, and classifier confidence.

Fifth, the application of Reinforcement Learning to side-channel analysis represents a central and original contribution of this thesis. Rather than employing Reinforcement Learning as a classifier or as a generic optimization technique, the proposed approach positions the learning agent between feature

extraction and classification. The agent acts as a feature-space controller, selecting adaptive operations based on the observed state of the representation and on feedback derived from the classifier.

This formulation establishes a novel connection between side-channel analysis, adaptive control, and security. The Reinforcement Learning mechanism is shown to play a dual role. On one side, it improves the stability of legitimate inference under noisy conditions. On the other side, it provides a mechanism for modifying the informational structure exposed through side-channel representations, thereby reducing the amount of exploitable information available to an adversarial observer.

The most significant conceptual contribution of the thesis is the introduction of the foundations of *Adaptive Inferential Security (SIA)*.

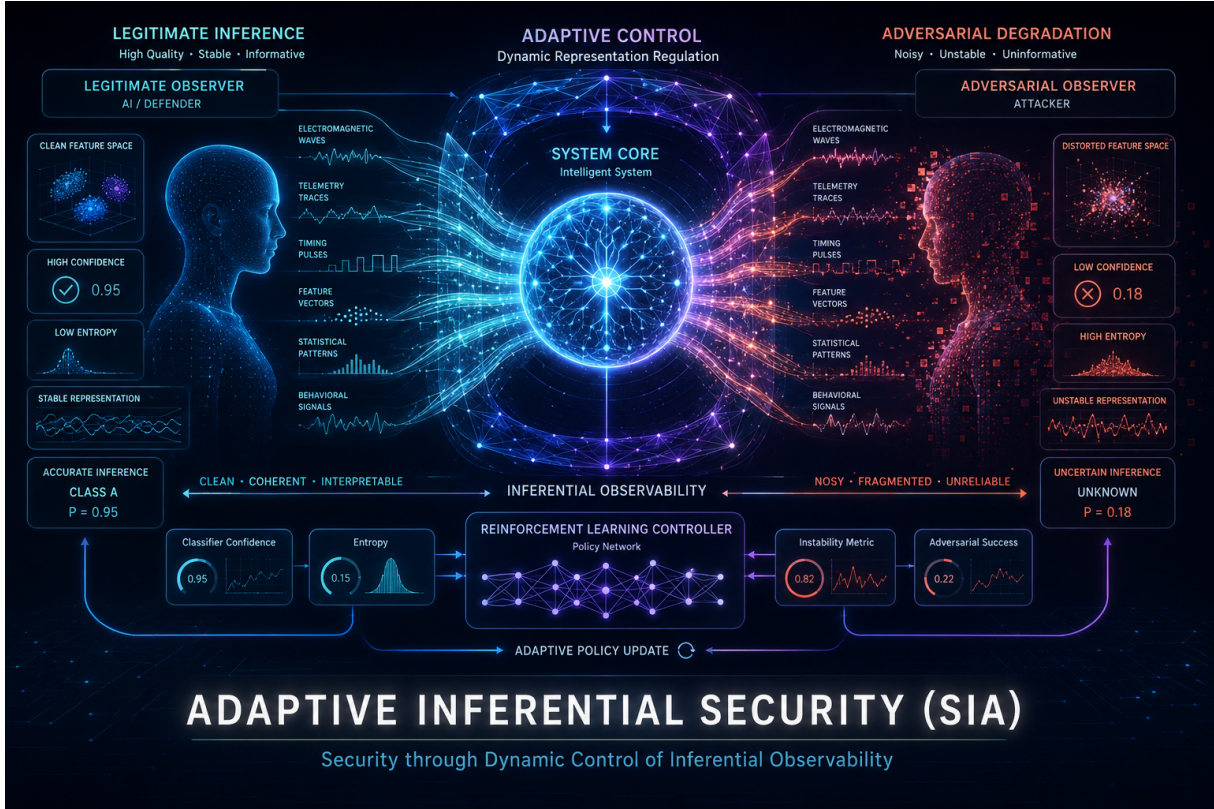


Figure 7.1: Conceptual illustration of *Adaptive Inferential Security (SIA)*. A reinforcement learning controller dynamically regulates the feature representations exposed through observable side-channel emissions, creating an inferential asymmetry between legitimate and adversarial observers. While authorized entities receive stable and informative representations suitable for reliable inference, adversarial observers are exposed to degraded, unstable, and low-confidence feature spaces. Rather than blocking signal access, the framework continuously adapts the informational structure of observable signals to control what can be inferred from system behavior under changing environmental and threat conditions.

SIA, shown in Figure 7.1, generalizes the lesson learned from the experimental work: security can be approached not only as the protection of data or the prevention of access, but also as the adaptive regulation of inference. In this view, a system is secure not only when its internal data are protected, but also when its observable behavior does not allow an unauthorized observer to infer sensitive information with unacceptable confidence.

The thesis therefore proposes an initial formulation of security as adaptive control of inferential observability. This concept is grounded in the experimental evidence produced throughout the work: side-channel signals expose behavioral information; machine learning models can extract such information; noise and representation changes affect inferential reliability; and Reinforcement Learning can actively control the feature representation to favor or inhibit inference.

To the best of our knowledge, this thesis is among the first works to explicitly connect side-channel analysis, feature-space adaptation, Reinforcement Learning, and inferentially-aware defense within a unified framework. While SIA remains an emerging paradigm requiring further formalization, the results presented in this work provide a concrete experimental basis for its development.

7.3 Limitations and Future Work

The limitations discussed in this section primarily reflect deliberate methodological choices aimed at preserving interpretability, experimental control, and realistic deployment assumptions, rather than fundamental shortcomings of the proposed approach. At the same time, these limitations naturally define the next steps required to consolidate the proposed framework and to further develop the concept of Adaptive Inferential Security.

First, the Reinforcement Learning framework introduced in this work should be regarded as an initial exploration rather than a fully mature solution. The action space was intentionally constrained to a limited set of adaptive operations in order to preserve interpretability and ensure stable learning. While this design choice allowed for a clear assessment of the benefits of adaptivity, it also limits the expressive power of the learned policy. Future work should investigate richer action spaces, including adaptive feature transformation, dynamic masking, controlled perturbation, representation compression, and information-theoretic constraints.

Second, the proposed adaptive mechanism operates downstream of feature extraction and relies on a fixed classifier. Although this separation enables controlled experimentation and simplifies analysis, it does not yet realize an end-to-end optimization of representation, adaptation, and decision-making. Integrating Reinforcement Learning with deep learning architectures in an end-to-end fashion remains an open challenge, particularly in the context of side-channel data, where labeled samples are limited and acquisition costs are high.

Third, the current work mainly evaluates robustness against synthetic noise. Although this provides a controlled and reproducible experimental setting, real-world deployments introduce additional sources of variability, including cross-device differences, environmental interference, hardware aging, workload diversity, and long-term system drift. These factors are critical for operational adoption and motivate the investigation of transfer learning, domain adaptation, continual learning, and self-supervised representation learning.

Fourth, the current study focuses primarily on offline learning and evaluation. Extending the framework toward online and real-time adaptation introduces additional challenges related to computational constraints, convergence guarantees, latency, and system stability. These aspects are particularly important for embedded and resource-constrained systems, where any adaptive defense mechanism must operate with limited overhead and predictable behavior.

A further limitation concerns the formalization of the adversarial model. In this thesis, the defensive interpretation of the adaptive mechanism is derived from the ability to modify and degrade the informational structure of the feature space. However, a complete SIA framework requires a more explicit definition of the adversary, including its observation capabilities, inference objectives, prior knowledge, and adaptation strategy. Future research should therefore formalize the interaction between defender and observer as an inferential game, where the defender aims to preserve legitimate utility while minimizing adversarial inference.

This direction naturally leads to one of the most important future developments of the thesis: the formalization of Adaptive Inferential Security as a general paradigm. Future work should define measurable quantities for inferential observability, adversarial confidence, representation leakage, and defender utility. Such metrics would make it possible to evaluate security not only in terms of classification accuracy or detection rate, but also in terms of how much actionable information an observer can infer from the system.

In this broader perspective, SIA can be seen as an evolution of the PhD work from side-channel classification toward adaptive security theory. The thesis demonstrates that side-channel information can be extracted, modeled, degraded, and controlled. The next step is to transform this experimental evidence into a general framework in which the system dynamically regulates its observable behavior according to context, risk, and inference objectives.

The long-term research direction opened by this thesis is therefore the development of intelligent systems capable of controlling their own inferential exposure. Such systems would not merely react to attacks after detection, but would continuously manage the information they reveal through physical, behavioral, temporal, and computational traces. In this sense, the work presented in this thesis provides the experimental and conceptual foundation for a future theory of security based on adaptive inferential control.

By reframing side-channel inference as a problem of adaptive inferential control rather than static signal processing, this thesis opens new directions for the design of robust, secure, and context-aware

monitoring systems. The proposed results suggest that future security mechanisms may evolve beyond traditional notions of access control, confidentiality, and anomaly detection, moving toward the dynamic regulation of what external observers are able to infer from system behavior itself.

In this sense, *Adaptive Inferential Security (SIA)* represents the broader conceptual horizon of this research: a paradigm in which observability, inference, adaptation, and security become interconnected dimensions of intelligent defensive systems.

Bibliography

- [1] Giovanni Agosta et al. “Software countermeasures for side-channel attacks”. In: *ACM Computing Surveys* (2015).
- [2] D. Agrawal, J. R. Rao, and P. Rohatgi. “Multi-channel attacks”. In: *Cryptographic Hardware and Embedded Systems (CHES)*. 2002, pp. 2–16. DOI: 10.1007/3-540-36400-5_2.
- [3] H. Bar-El et al. “The sorcerer’s apprentice guide to fault attacks”. In: vol. 94. 2. 2006, pp. 370–382. DOI: 10.1109/JPROC.2005.862424.
- [4] Shai Ben-David et al. “A theory of learning from different domains”. In: *Machine Learning* 79.1 (2010), pp. 151–175.
- [5] E. Biham and A. Shamir. “Differential fault analysis of secret key cryptosystems”. In: *Advances in Cryptology – CRYPTO’97*. 1997, pp. 513–525. DOI: 10.1007/BFb0052259.
- [6] L. Breiman. “Random Forests”. In: *Machine Learning* 45.1 (2001), pp. 5–32.
- [7] Gabriele Camurati et al. “Screaming Channels: When Electromagnetic Side Channels Meet Radio Transceivers”. In: *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2018, pp. 163–177. DOI: 10.1145/3243734.3243815.
- [8] S. Chari, J. R. Rao, and P. Rohatgi. “Towards sound approaches to counteract power-analysis attacks”. In: *Advances in Cryptology – CRYPTO’99*. 1999, pp. 398–412. DOI: 10.1007/3-540-48405-1_25.
- [9] Thomas G. Dietterich. “Ensemble Methods in Machine Learning”. In: *Lecture Notes in Computer Science* 1857 (2000), pp. 1–15.
- [10] A. Domínguez et al. “Efficient Machine Learning-Based Side-Channel Analysis for Embedded Systems”. In: *IEEE Access* (2021).
- [11] F. J. Domínguez et al. “Power consumption analysis for intrusion detection in IoT”. In: *IEEE IoT Journal* 8.13 (2021), pp. 10252–10263.
- [12] David L Donoho. “De-noising by soft-thresholding”. In: *IEEE Transactions on Information Theory* 41.3 (1995), pp. 613–627.
- [13] K. Gandolfi, C. Mourtel, and F. Olivier. “Electromagnetic analysis: Concrete results”. In: *Cryptographic Hardware and Embedded Systems (CHES)*. 2001, pp. 251–261. DOI: 10.1007/3-540-44709-1_21.
- [14] D. Genkin, I. Pipman, and E. Tromer. “Get your hands off my laptop: Physical side-channel key-extraction attacks on PCs”. In: *ACM CCS*. 2014, pp. 891–906. DOI: 10.1145/2660267.2660378.
- [15] I. Guyon et al. “Gene selection for cancer classification using support vector machines”. In: *Machine Learning* 46.1 (2002), pp. 389–422.
- [16] Y. Han et al. “Watching the Watchers: Profiling Software on IoT Devices via Power Side Channels”. In: (2020).
- [17] Yanjun Han et al. “AIoT Side-Channel Attacks: New Threats to AIoT Devices”. In: *Proceedings of the IEEE International Conference on Internet of Things (iThings)*. IEEE, 2020, pp. 1–8. DOI: 10.1109/iThings51991.2020.00015.
- [18] G. Hospodar et al. “Machine learning in side-channel analysis: a first study”. In: vol. 1. 4. 2011, pp. 293–302. DOI: 10.1007/s13389-011-0023-x.
- [19] H. Huang et al. “Deep learning-based improved side-channel attacks using data denoising and feature fusion”. In: *PLOS ONE* (2025).
- [20] I. T. Jolliffe. *Principal Component Analysis*. 2nd. Springer Series in Statistics. Springer, 2002.
- [21] H. Kim, J. Lee, and S. Lee. “Applying deep learning to side-channel attacks”. In: *IEEE TIFS* 14.12 (2019), pp. 3393–3407. DOI: 10.1109/TIFS.2019.2929556.

- [22] Paul Kocher, Joshua Jaffe, and Benjamin Jun. “Differential Power Analysis”. In: *Advances in Cryptology – CRYPTO’99*. 1999, pp. 388–397. DOI: 10.1007/3-540-48405-1_25.
- [23] Paul C. Kocher. “Timing Attacks on Implementations of Diffie-Hellman, RSA, DSS, and Other Systems”. In: *Advances in Cryptology – CRYPTO’96*. Springer, 1996, pp. 104–113. DOI: 10.1007/3-540-68697-5_9.
- [24] L. I. Kuncheva. *Combining Pattern Classifiers*. Wiley, 2004.
- [25] L. Lerman, G. Bontempi, and O. Markowitch. “Side channel attack: an approach based on information theory and machine learning”. In: *Journal of Cryptographic Engineering* 4.2 (2014), pp. 123–139.
- [26] H. Maghrebi, T. Portigliatti, and E. Prouff. “Breaking cryptographic implementations using deep learning techniques”. In: *SPACE*. 2016, pp. 3–26. DOI: 10.1007/978-3-319-49445-6_1.
- [27] S. Mangard, E. Oswald, and T. Popp. *Power Analysis Attacks: Revealing the Secrets of Smart Cards*. Springer, 2007.
- [28] Amir Moradi and François-Xavier Standaert. “Moments-Correlating DPA”. In: *Cryptographic Hardware and Embedded Systems (CHES)*. Springer, 2011, pp. 304–321. DOI: 10.1007/978-3-642-23951-9_21.
- [29] H. Nam et al. “Defensive Machine Learning: Defending Architectural Side-Channels with Adversarial Obfuscation”. In: *arXiv* (2023). arXiv:2302.01474.
- [30] Alan V. Oppenheim, Alan S. Willsky, and S. Hamid Nawab. *Signals and Systems*. Prentice Hall, 1999.
- [31] Sinno Jialin Pan and Qiang Yang. “A Survey on Transfer Learning”. In: *IEEE Transactions on Knowledge and Data Engineering* 22.10 (2010), pp. 1345–1359.
- [32] Guilherme Perin et al. “Scalability and robustness of deep learning-based side-channel attacks”. In: *Journal of Cryptographic Engineering* (2020).
- [33] Stjepan Picek et al. “The Curse of Class Imbalance and Conflicting Metrics with Machine Learning for Side-Channel Evaluations”. In: *IACR Transactions on Cryptographic Hardware and Embedded Systems (TCHES)* 2019.1 (2019), pp. 209–237. DOI: 10.13154/tches.v2019.i1.209-237.
- [34] M. Poehlau et al. “A survey of electromagnetic side-channel countermeasures”. In: *IEEE Communications Surveys & Tutorials* 24.1 (2022), pp. 586–612.
- [35] Joaquin Quiñonero-Candela et al. “Dataset shift in machine learning”. In: *The MIT Press* (2009).
- [36] Joaquin Quiñonero-Candela et al. *Dataset Shift in Machine Learning*. MIT Press, 2009.
- [37] Stephan Rabanser, Stephan Günnemann, and Zachary Lipton. “On the behavior of tree-based models under distribution shift”. In: *NeurIPS Workshop* (2020).
- [38] Vincenzo Rega et al. “Profiling Running Applications in Connected Devices Through Side-Channel and Machine Learning Techniques”. In: *IEEE Access* (2024). DOI: 10.1109/ACCESS.2024.3491916.
- [39] Vincenzo Rega et al. “Side-Channel-Based Behavioral Profiling of IoT Applications Using AI Techniques”. In: *Proceedings of the IEEE International Workshop on Metrology for Industry 4.0 & IoT (MetroInd4.0 & IoT)*. 2025. DOI: 10.1109/metroind4.0iot66048.2025.11122057.
- [40] Mathieu Renauld and François-Xavier Standaert. *Randomized execution of programs for side-channel attack resistance*. Tech. rep. IACR, 2006.
- [41] Mathieu Renauld and François-Xavier Standaert. “Towards sound and optimal leakage detection procedures”. In: *CHES*. 2011.
- [42] Abraham Savitzky and Marcel J E Golay. “Smoothing and differentiation of data by simplified least squares procedures”. In: *Analytical Chemistry* 36.8 (1964), pp. 1627–1639.
- [43] Ronald W. Schafer. “What Is a Savitzky–Golay Filter?” In: *IEEE Signal Processing Magazine* 28.4 (2011), pp. 111–117. DOI: 10.1109/MSP.2011.941097.
- [44] Andreas Schmidt et al. “A Survey on the Security of Industrial Control Systems”. In: *IEEE Communications Surveys & Tutorials* (2018).
- [45] P. Socha, V. Miškovský, and M. Novotný. “A Comprehensive Survey on the Non-Invasive Passive Side-Channel Analysis”. In: *Sensors* 22.21 (2022), p. 8096.
- [46] R. Spreitzer et al. “Systematic classification of side-channel attacks on mobile devices”. In: *IEEE Communications Surveys & Tutorials* 20.1 (2018), pp. 465–488. DOI: 10.1109/COMST.2017.2780266.

- [47] F.-X. Standaert, T. Malkin, and M. Yung. “A unified framework for the analysis of side-channel key recovery attacks”. In: *EUROCRYPT*. 2009, pp. 443–461. DOI: 10.1007/978-3-642-01001-9_26.
- [48] François-Xavier Standaert, Tal Malkin, and Moti Yung. “A formal approach to side-channel leakage”. In: *Journal of Cryptographic Engineering* (2009).
- [49] Sigasec Research Team. “RAMBO: A Remote Memory-Based Electromagnetic Attack”. In: *arXiv preprint* (2024).
- [50] X. Wang et al. “A CNN-RNN Architecture for Side-Channel Analysis”. In: *IEEE Transactions on Information Forensics and Security* (2019).
- [51] Kim Wouters et al. “Side-Channel Leakage Detection with Autoencoders”. In: *IACR Transactions on Cryptographic Hardware and Embedded Systems (TCHES)* 2020.4 (2020), pp. 1–24.
- [52] Jingkang Yang et al. “A survey on out-of-distribution generalization”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).
- [53] I. Zaid et al. “A methodology for efficient CNN architectures in profiling attacks”. In: *Journal of Cryptographic Engineering* 10.2 (2020), pp. 163–176.
- [54] Y. Zhang, Q. Ding, and J.-Q. Sun. “Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution”. In: *IEEE Transactions on Knowledge and Data Engineering* 31.6 (2019), pp. 1031–1044. DOI: 10.1109/TKDE.2018.2850320.

Appendices

Appendix A

Detailed Experimental Parameters

This appendix reports the detailed experimental parameters adopted throughout the thesis to ensure full reproducibility of the proposed side-channel analysis pipeline. While the main chapters focus on methodological rationale and performance evaluation, the present appendix consolidates all low-level configuration choices and numerical settings.

A.1 Acquisition Parameters

The electromagnetic and current measurements were acquired using the following configuration:

- **Device under test:** Raspberry Pi 4 Model B
- **Sampling frequency:** 1 MHz
- **Acquisition duration:** 40 s per trial
- **Idle/background segment:** first 10 s (discarded)
- **Number of scenarios:** 8
- **Trials per scenario:** 30
- **Total acquisitions:** 240

A.2 Windowing and Segmentation

After removal of the initial background segment, the remaining portion of each signal was segmented using a fixed-length sliding window approach.

A window duration of 100 ms was adopted, corresponding to $N = 10^5$ samples at a sampling frequency of 1 MHz. Consecutive windows were extracted with a 50% overlap, resulting in a hop size of 50 ms. This overlapping strategy increases the number of training samples while preserving temporal continuity and reducing sensitivity to window boundary effects.

For each window and for each physical channel (current and electromagnetic), Root Mean Square (RMS) aggregation was applied. Given a discrete-time signal segment $x_i[n]$ of length N , the RMS value is computed as:

$$\text{RMS}_i = \sqrt{\frac{1}{N} \sum_{n=1}^N x_i[n]^2} \quad (\text{A.1})$$

RMS aggregation provides an energy-based representation that reduces high-frequency noise while preserving workload-dependent signal dynamics. The resulting RMS time series constitutes the input for subsequent normalization and feature extraction stages described in Chapter 3.

Appendix B

Feature Analysis and Supplementary Results

This appendix provides a detailed analysis of the extracted features and additional experimental results that complement the performance evaluation presented in Chapter 4. While the main chapter focuses on aggregated metrics and comparative performance across classifiers and noise conditions, the present appendix offers a deeper inspection of feature behavior, stability, and contribution under different operating regimes.

The material reported here serves three main purposes:

- to support the interpretation of classification results by analyzing feature-level behavior,
- to provide transparency and reproducibility for feature extraction and selection,
- to document supplementary results that would otherwise interrupt the flow of the main discussion.

B.1 Complete Feature Set and Channel Contribution

Feature extraction was performed on RMS-aggregated windowed signals independently for current and electromagnetic channels. For each window, a set of time-domain and frequency-domain descriptors was computed, yielding a total of 12 features per channel and 24 features per window. The RMS value is treated both as a preprocessing step and as an explicit time-domain feature, and is therefore included in the final feature vector.

Time-Domain Features

Time-domain features were extracted from the RMS time series associated with each analysis window and include the following statistical descriptors:

- mean value,
- standard deviation,
- minimum value,
- maximum value,
- skewness,
- kurtosis,
- RMS value.

These features capture the overall energy level, variability, and distributional shape of the signal, providing a compact characterization of workload-induced activity.

Frequency-Domain Features

Frequency-domain features were computed by applying the Fast Fourier Transform (FFT) to the RMS feature sequence associated with each window. A single-sided spectrum was considered, and features were derived from the corresponding power spectral density.

The extracted frequency-domain descriptors include:

- total spectral energy,
- spectral entropy,
- spectral centroid,
- spectral spread,
- dominant (peak) frequency component.

Spectral energy captures the overall frequency-domain power, while spectral entropy measures the distribution complexity of spectral components. The centroid and spread describe the balance and dispersion of spectral content, and the peak frequency identifies the dominant oscillatory component within the window.

Channel Contribution

All features were computed independently for the current consumption and electromagnetic emission channels. This separation allows the learning models to exploit complementary leakage mechanisms: current-based features primarily capture global power consumption trends, whereas electromagnetic features encode localized and frequency-dependent switching activity.

The complete feature set is summarized in Table B.1.

Table B.1: Extracted features and corresponding signal domain.

Domain	Feature	Physical Interpretation
Time	Mean	Average activity level
Time	Standard deviation	Signal variability
Time	Minimum / Maximum	Extreme operating conditions
Time	Skewness	Distribution asymmetry
Time	Kurtosis	Distribution sharpness
Time	RMS	Window energy content
Frequency	Spectral energy	Total frequency-domain power
Frequency	Spectral entropy	Spectral complexity
Frequency	Spectral centroid	Frequency balance
Frequency	Spectral spread	Spectral dispersion
Frequency	Peak frequency	Dominant oscillatory component

B.2 Feature Selection and Classification with Random Forest

Classification experiments were primarily conducted using a Random Forest classifier, selected for its robustness to noise, ability to handle heterogeneous feature sets, and intrinsic support for feature importance estimation.

The Random Forest model was configured with 200 decision trees and no explicit depth limitation. All trees were trained using bootstrap sampling, and node splitting was guided by impurity-based criteria.

B.2.1 Recursive Feature Elimination (RFE)

To address inference-time constraints and reduce feature dimensionality, Recursive Feature Elimination (RFE) was employed using the Random Forest classifier as the base estimator. RFE iteratively removes the least informative features based on their impurity-driven importance scores until a predefined number of features is retained.

Unlike scenario-specific feature importance analysis, which is not applicable when the active scenario is unknown at inference time, RFE was applied globally across all scenarios. This strategy yields a fixed feature subset that can be reliably used during deployment.

The resulting reduced feature set preserves most of the discriminative power of the full representation while significantly reducing dimensionality and computational complexity.

B.2.2 Cross-Validation and Performance Evaluation

Classification performance was evaluated using stratified k -fold cross-validation with $k = 10$, ensuring class balance across all folds. For each fold, the model was trained on the selected feature subset and evaluated on unseen data.

Window-level accuracy was computed for each fold, and confusion matrices were accumulated and averaged across folds to obtain stable estimates of class-wise performance. This evaluation protocol reduces variance due to data partitioning and provides a robust assessment of generalization performance.

Unless otherwise stated, all Random Forest experiments reported in the thesis follow this configuration.

B.3 Noise Injection and SNR Modeling

To simulate realistic acquisition conditions, additive white Gaussian noise (AWGN) was injected into the raw side-channel signals prior to windowing and RMS aggregation.

Given a discrete-time clean signal $x[n]$, the target Signal-to-Noise Ratio (SNR) is defined as:

$$\text{SNR}_{\text{dB}} = 10 \log_{10} \left(\frac{\mathbb{E}[x^2]}{\sigma^2} \right) \quad (\text{B.1})$$

where $\mathbb{E}[x^2]$ denotes the average signal power and σ^2 is the noise variance. For a given target SNR value, the corresponding noise variance is computed accordingly, and zero-mean Gaussian noise is independently generated and added to the signal.

Noise injection was performed independently for each acquisition and for each sensing channel (current and electromagnetic) in order to avoid correlated perturbations across trials. This procedure ensures that noise realizations differ across experiments while preserving a consistent signal-to-noise ratio.

Noise was applied at the raw signal level, prior to temporal segmentation and RMS computation. This choice reflects realistic acquisition scenarios in which environmental and instrumental noise affects the physical signal before any digital preprocessing is performed.

B.4 Denoising Techniques on RMS Features: Case Study at SNR = 20 dB

A preliminary investigation was conducted to assess whether classical denoising techniques applied directly to RMS feature time series could improve window-level classification performance under noisy conditions. This analysis focuses on a representative scenario with additive Gaussian noise corresponding to a signal-to-noise ratio (SNR) of 20 dB.

Initial experiments evaluated different feature normalization strategies applied to RMS features. Despite clear differences in scale and distribution between current and electromagnetic channels, no significant variation in window-level classification accuracy was observed when using a Random Forest classifier. This result indicates that, within the considered setting, performance limitations are not primarily attributable to numerical scaling effects, but rather to structural instability of the information extracted over short temporal windows.

For this reason, attention was shifted toward explicit denoising of RMS feature sequences. Additive Gaussian noise was injected into raw signals prior to windowing, simulating realistic acquisition conditions. Comparative analysis between clean and noisy signals revealed that noise significantly increases local variability of RMS features, reducing class separability at the single-window level while largely preserving global temporal trends.

B.4.1 Baseline RMS Classification under Noise

Figure B.1 reports the window-level confusion matrix obtained using RMS features extracted from noisy raw signals at SNR = 20 dB. Signals were segmented using 100 ms windows with 50% overlap, consistently with the main experimental protocol.

The resulting window-level accuracy is approximately 59.3%, which is closely aligned with the accuracy observed under clean conditions (approximately 59.9%). This suggests that, at this noise level, RMS-based aggregation already provides a degree of robustness sufficient to preserve most discriminative information.

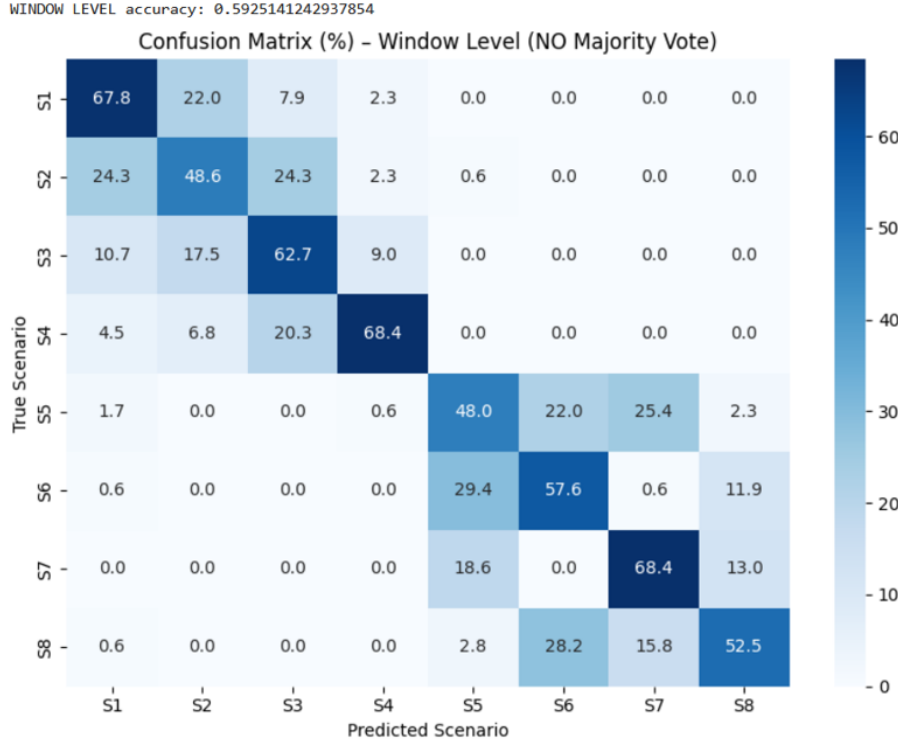


Figure B.1: Window-level confusion matrix obtained using RMS features extracted from raw signals corrupted with additive Gaussian noise at SNR = 20 dB (no denoising applied).

Inspection of the confusion matrix structure shows that misclassification patterns remain qualitatively similar to the clean case, with errors concentrated among scenarios exhibiting similar workload dynamics. Noise primarily increases intra-class variability without substantially altering inter-class relationships, confirming the integrative and noise-averaging nature of RMS features.

B.4.2 Moving Average Denoising

The moving average (MA) is a simple and widely adopted denoising technique aimed at suppressing high-frequency fluctuations while preserving low-frequency trends [30]. Applied to RMS feature sequences, moving average smoothing reduces local variability introduced by noise while retaining global temporal transitions associated with application execution.

Table B.2 reports window-level accuracy obtained for different MA window lengths.

Table B.2: Window-level accuracy obtained using moving average denoising on RMS features at SNR = 20 dB.

MA Window Length	Window-Level Accuracy
3	0.61
5	0.66
10	0.66

Results show a progressive improvement in accuracy as the smoothing window increases, reaching approximately 66% for window lengths of 5 and 10. However, the absence of further gains for larger

windows indicates a saturation effect. Excessive smoothing increases correlation between consecutive windows, limiting the amount of locally discriminative information available to the classifier.

B.4.3 Savitzky–Golay Filtering

Savitzky–Golay filtering performs local polynomial regression within a sliding window, enabling noise reduction while preserving signal shape and local trends [42]. This property makes it particularly suitable for signals where discriminative information is encoded in gradual variations rather than impulsive changes.

Applied to RMS feature sequences, Savitzky–Golay filtering attenuates noise-induced fluctuations while maintaining the temporal structure of the features. Table B.3 reports window-level accuracy obtained for different window lengths using a fixed polynomial order of 2.

Table B.3: Window-level accuracy obtained using Savitzky–Golay filtering on RMS features at SNR = 20 dB (polynomial order = 2).

Window Length	Polynomial Order	Accuracy
3	2	0.59
5	2	0.61
9	2	0.63

Accuracy improves gradually with increasing window length, reaching approximately 63%, but remains below that achieved with moving average smoothing. While Savitzky–Golay filtering better preserves signal shape, it does not introduce additional discriminative power when applied to already aggregated RMS features.

B.4.4 Wavelet-Based Denoising

Wavelet denoising was performed using the Discrete Wavelet Transform (DWT) combined with soft thresholding [12]. Wavelet-based methods provide a multiresolution representation, separating low-frequency signal components from high-frequency noise.

Soft thresholding was applied to wavelet coefficients obtained using Daubechies wavelets (db4, db6, db8). Table B.4 summarizes the resulting window-level accuracy.

Table B.4: Window-level accuracy obtained using wavelet denoising on RMS features at SNR = 20 dB.

Wavelet	Threshold Type	Accuracy
db4	soft	0.62
db6	soft	0.60
db8	soft	0.59

The db4 wavelet yields the best performance, while higher-order wavelets lead to progressive degradation. Larger support and increased regularity introduce excessive smoothing, suppressing informative local variations already attenuated by RMS aggregation.

B.4.5 Discussion

Overall, denoising techniques applied directly to RMS feature sequences provide modest but limited improvements in window-level classification accuracy. Moving average smoothing offers the largest gain, while Savitzky–Golay and wavelet-based methods exhibit diminishing returns.

These results confirm that RMS features already perform substantial temporal integration. Additional denoising primarily enhances temporal stability rather than introducing new discriminative information. Consequently, denoising at the feature level cannot overcome the structural limitations of energy-only representations.

This finding motivates the need for richer feature representations or adaptive mechanisms capable of dynamically modulating feature relevance—an approach explicitly explored in Chapter 5 through reinforcement learning–based feature-aware control.

B.5 Extended Pipeline with Denoising and Time–Frequency Feature Extraction

In light of the results discussed in the previous sections, the processing pipeline was extended by introducing a richer feature extraction stage, including descriptors from both the time and frequency domains. In this context, denoising is no longer interpreted as a mechanism aimed at directly improving classification performance, but rather as a supporting tool designed to stabilize the signal and facilitate the extraction of more informative features.

The analyses based exclusively on RMS features revealed that, although such representations exhibit a certain degree of robustness to moderate noise levels, they constitute an intrinsic informational bottleneck at the window level. Denoising applied directly to RMS features yields only limited performance gains, indicating that the main limiting factor is not noise alone, but the reduced expressiveness of an energy-only representation.

To overcome this limitation, denoising is therefore positioned downstream of the RMS computation and upstream of time–frequency feature extraction. In this configuration, denoising does not attempt to compensate for classification errors directly, but instead acts to stabilize the RMS feature sequences from which more structured descriptors are subsequently extracted.

The extended pipeline adopted in this study is illustrated in Figure B.2. The processing flow starts with additive noise injection on the raw signal at a signal-to-noise ratio (SNR) of 20 dB, followed by temporal windowing and RMS computation. The selected denoising technique is then applied, and finally a set of time- and frequency-domain features is extracted and provided as input to the Random Forest classifier.

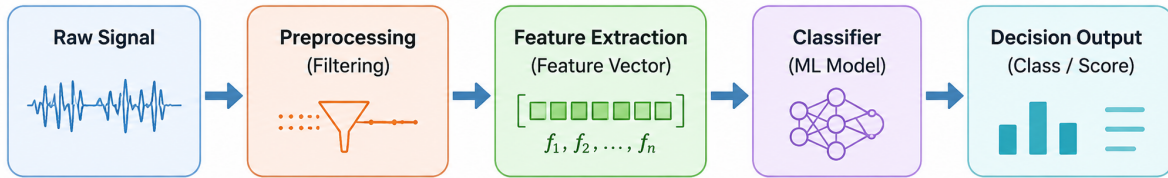


Figure B.2: Extended processing pipeline for evaluating the combined impact of noise, denoising, and feature engineering. Denoising is applied to RMS features prior to time–frequency feature extraction.

This design preserves the informative content of the signal while enabling a more effective assessment of the joint impact of noise, denoising, and feature engineering on window-level classification performance.

Table B.5: Window-level classification accuracy obtained with the extended pipeline under different denoising techniques (SNR = 20 dB).

Denoising method	Main parameters	Accuracy (%)
Moving Average	window = 9	81.57
Savitzky–Golay	window = 9, polynomial order = 2	79.94
Wavelet (db4)	soft threshold	78.60

B.5.1 Discussion and Interpretation of Results

The introduction of time- and frequency-domain feature extraction following RMS denoising leads to a substantial improvement in window-level classification performance compared to RMS-only approaches. In particular, all considered denoising techniques significantly outperform the RMS-only baseline, highlighting the critical role played by feature richness in recovering discriminative information under noisy conditions.

Among the evaluated methods, the moving average filter with a window length of nine samples achieves the highest accuracy, reaching approximately 81.6%. This result suggests that, once a sufficiently expressive feature representation is adopted, a simple smoothing strategy aligned with the temporal scale

of the analysis windows is sufficient to reduce residual variability without suppressing informative signal components. The simplicity and low computational cost of the moving average filter further make it well suited for real-time inference on resource-constrained platforms.

The Savitzky–Golay filter yields slightly lower but still competitive performance. Its ability to preserve local signal trends allows a consistent recovery of classification accuracy; however, the polynomial smoothing introduces stronger correlations between consecutive windows than the moving average, partially limiting the achievable performance gain.

Despite being theoretically more sophisticated, wavelet-based denoising exhibits the lowest accuracy among the evaluated techniques. This behavior is consistent with previous observations: when applied to already aggregated RMS features, wavelet denoising tends to introduce multi-scale smoothing that attenuates not only noise components but also subtle local variations that are relevant for window-level discrimination.

Overall, these results indicate that the effectiveness of denoising strongly depends on both its position within the processing pipeline and the expressiveness of the feature representation. Once the limitations of RMS-only features are overcome through richer time–frequency descriptors, simple and computationally efficient denoising techniques become not only sufficient but preferable to more complex methods. This confirms that, in the considered scenario, the primary performance gains stem from feature engineering, while denoising plays a secondary yet supportive role by stabilizing feature trajectories in the presence of noise.

B.6 The Role of Majority Voting

In configurations characterized by a high level of noise, a commonly adopted strategy to improve decision robustness consists in temporally aggregating window-level predictions through majority voting. This approach belongs to the broader class of decision fusion and ensemble techniques, which have been extensively studied in pattern recognition and machine learning as effective mechanisms for reducing variance and improving robustness in uncertain environments [24, 9].

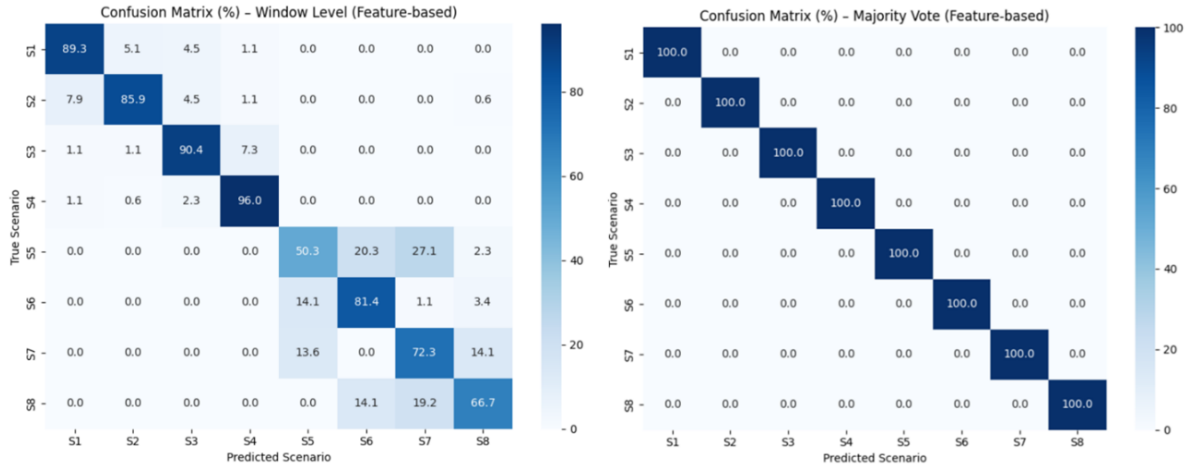


Figure B.3: Comparison between window-level and acquisition-level confusion matrices under noisy conditions (SNR = 20 dB). Temporal aggregation via majority voting restores perfect scenario-level classification despite window-level uncertainty.

Under majority voting, the classifier does not assume that each individual window extracted from the signal contains sufficient information for a reliable decision. Instead, it exploits the temporal redundancy inherent in side-channel signals to integrate multiple local predictions into a single global decision at the acquisition level. This assumption is well aligned with the nature of side-channel leakage, where application behavior persists over time and is reflected across multiple consecutive observations [38, 44].

As highlighted by the window-level confusion matrices obtained under noisy conditions (SNR = 20 dB), window-based classification exhibits a non-negligible amount of errors, particularly for scenarios characterized by similar energetic and statistical signatures. However, these errors are not randomly distributed. For each scenario, the proportion of correctly classified windows remains systematically higher than that associated with any incorrect class. As a result, the prediction distribution retains a coherent

structure, characterized by the presence of a dominant class corresponding to the true scenario, albeit with reduced margins. In this context, temporal aggregation via majority voting amplifies the residual discriminative information distributed across the sequence, transforming partial local correlations into a highly reliable global decision. When the correct class represents the relative majority of window-level predictions—even in the absence of an absolute majority—the final scenario assignment becomes stable and accurate. This behavior is clearly observable when comparing window-level and acquisition-level confusion matrices, where sequence-level accuracy reaches 100% for all considered scenarios.

The case of Scenario S5 is particularly illustrative. Although the proportion of correctly classified windows for this scenario is relatively limited, it remains consistently higher than that of any competing class. Under these conditions, majority voting successfully recovers the correct acquisition-level classification, demonstrating that discriminative information is not concentrated within individual windows but rather emerges through statistical accumulation over time.

It is important to emphasize that majority voting does not introduce new information into the classification process. Instead, it acts solely as a statistical integration mechanism for local predictions. Its effectiveness therefore depends on the ability of the base classifier to preserve a meaningful—albeit partial—correlation between window-level predictions and the true scenario. Consequently, majority voting does not structurally solve the noise problem, but mitigates its effects through temporal aggregation, a distinction that is well recognized in ensemble learning theory [24, 9].

B.7 Supplementary Confusion Matrices

This section reports additional confusion matrices computed under different noise levels and feature configurations. These matrices provide a finer-grained view of misclassification patterns that are summarized in the main chapter.

In particular, they highlight that:

- misclassifications predominantly occur between scenarios with similar workload dynamics,
- protocol-based scenarios (e.g., HTTP, FTP, FTPS) are more sensitive to noise-induced confusion,
- client-oriented scenarios preserve clearer decision boundaries even under degraded conditions.

These trends reinforce the need for noise-aware feature selection and temporal aggregation strategies.

B.8 Window-Level vs Session-Level Inference

A relevant aspect of the evaluation methodology concerns the distinction between *window-level* and *session-level* inference. This distinction is essential because the proposed side-channel analysis pipeline operates on temporal windows, while the monitored behavior is naturally associated with a complete acquisition session.

At the window level, each temporal segment extracted from the acquisition is processed independently by the classifier. Therefore, each window produces a local prediction. This type of evaluation measures the intrinsic discriminative power of individual feature representations and provides a fine-grained view of how well the classifier can recognize the target scenario from short portions of the signal.

However, window-level predictions are inherently more sensitive to noise, transient perturbations, and local variations in the side-channel signal. A single window may contain only partial behavioral information, or it may be affected by temporary distortions that reduce class separability. As a consequence, window-level inference may exhibit local instability, especially under moderate or high noise conditions.

As illustrated in Figure B.4, session-level inference addresses this limitation by aggregating the predictions produced over the full acquisition. In this work, the aggregation is performed using the column-wise majority voting strategy described in Section C.1. The final decision is obtained by selecting the class that appears most frequently among the window-level predictions.

This approach exploits the temporal persistence of application behavior. Although individual windows may be incorrectly classified, the dominant behavioral pattern is expected to remain consistent across the acquisition session. Therefore, the aggregation process acts as a decision-level stabilization mechanism, reducing the impact of isolated errors and emphasizing the most recurrent prediction.

Quantitative results reported throughout the thesis show that session-level accuracy consistently exceeds window-level accuracy across all considered noise regimes.

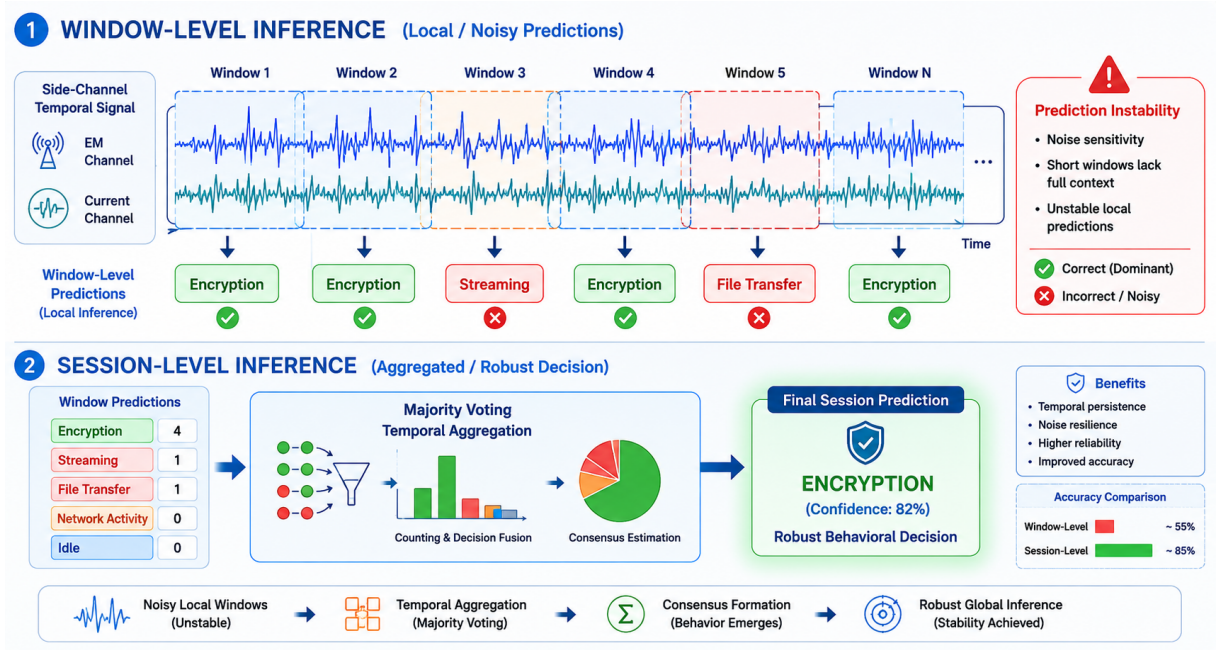


Figure B.4: Comparison between window-level and session-level inference. At the window level, each temporal segment is classified independently, producing local predictions that may fluctuate due to noise or partial loss of discriminative information. At the session level, the sequence of window-level predictions is aggregated through column-wise majority voting, producing a more stable and reliable acquisition-level decision.

This confirms that local prediction instability does not necessarily compromise global inference reliability. In particular, even when noise reduces the discriminative power of individual windows, the cumulative information distributed across the session can still support a correct final decision.

From a methodological perspective, the difference between the two evaluation levels is important. Window-level accuracy evaluates the quality and robustness of the local representation, while session-level accuracy evaluates the reliability of the complete monitoring pipeline. The former is more appropriate for analyzing feature separability and classifier behavior under local perturbations; the latter better reflects the operational objective of recognizing the behavioral scenario associated with an entire acquisition.

From a system perspective, session-level inference is particularly relevant for embedded, IoT, and cyber-physical monitoring scenarios. In many practical applications, the objective is not to classify every short temporal segment perfectly, but to infer the global behavior of the device or application over a meaningful observation interval. In this context, majority voting provides a simple, interpretable, and computationally lightweight mechanism for improving reliability without modifying the classifier or the feature extraction process.

Overall, the comparison between window-level and session-level inference highlights a central finding of the thesis: side-channel information is cumulative over time. Individual windows may provide incomplete or noisy evidence, but the aggregation of multiple observations allows the system to recover a stable and reliable behavioral interpretation.

B.9 Methodological Implications

The analyses presented in this appendix support several key methodological insights:

- feature relevance is context-dependent and affected by noise conditions,
- RMS-based preprocessing improves both stability and interpretability,
- static feature selection represents a compromise across expected environments,
- temporal aggregation acts as an effective but reactive robustness mechanism.

These observations reinforce the motivation for adaptive processing strategies discussed in the main body of the thesis and provide empirical justification for the transition toward learning-based control mechanisms introduced in subsequent chapters.

Appendix C

Algorithmic Details

This appendix provides a concise yet formal description of the core algorithmic components underlying the proposed side-channel analysis pipeline. The purpose of this appendix is not to introduce new methodological contributions, but to clarify the operational flow of the main processing stages referenced throughout the thesis.

In particular, this section focuses on the RMS-based preprocessing pipeline, which plays a central role in transforming raw side-channel acquisitions into compact, robust, and machine-learning-compatible representations suitable for behavioral inference under noisy conditions.



Figure C.1: Overview of the RMS-based preprocessing pipeline adopted for side-channel behavioral inference. Raw electromagnetic and current-consumption signals are processed through background removal, temporal segmentation, RMS computation, robust normalization, and feature extraction stages in order to generate compact and noise-resilient feature representations suitable for machine learning classification.

As illustrated in Figure C.1, the preprocessing workflow starts from raw side-channel acquisitions obtained from two physical channels: electromagnetic emissions and current consumption measurements. These signals contain both discriminative workload-dependent information and non-informative components related to idle system activity, environmental interference, and transient perturbations.

The first processing stage consists of **background removal**. The initial portion of each acquisition, corresponding to idle or non-operational system behavior, is discarded in order to eliminate non-discriminative signal components and reduce the influence of static environmental noise. This operation allows the subsequent stages to focus exclusively on the active execution phase of the monitored scenario.

After background removal, the remaining signal is processed through **temporal segmentation**. The signal is divided into partially overlapping sliding windows of fixed duration T with overlap factor α . This step transforms the continuous acquisition into a sequence of localized temporal observations

while preserving temporal continuity across adjacent windows. The use of overlapping windows improves temporal smoothness and mitigates abrupt boundary effects between consecutive segments.

For each generated window, the **Root Mean Square (RMS)** value is computed independently for both physical channels. The RMS operation transforms raw oscillatory signals into compact energy-based representations that preserve workload-dependent characteristics while significantly reducing dimensionality and sensitivity to high-frequency fluctuations. In practice, the RMS representation acts as a compressed description of the signal energy distribution over time.

The resulting RMS sequences are subsequently processed through **robust normalization**. A Robust Scaler is applied to remove the median and normalize according to the interquartile range (IQR). Unlike standard normalization techniques based on mean and variance, robust normalization is less sensitive to outliers and transient anomalies, making it particularly suitable for noisy side-channel environments where signal perturbations and acquisition variability are common.

Following normalization, the pipeline performs **feature extraction**. Both time-domain and frequency-domain descriptors are computed from the normalized RMS sequences in order to characterize the statistical and spectral properties of the observed behavior. Extracted features include measures related to central tendency, variability, energy distribution, and spectral structure, generating the final feature vectors employed for machine learning inference.

The overall design of the preprocessing pipeline intentionally balances computational efficiency, robustness, and interpretability. By progressively transforming noisy raw physical signals into structured feature representations, the pipeline enables reliable behavioral inference while remaining lightweight enough for deployment in embedded and resource-constrained environments.

From a methodological perspective, the pipeline also reflects one of the central principles emerging throughout the thesis: robustness is not achieved solely through stronger classifiers, but through careful control of the representation itself. The RMS-based transformation and normalization stages contribute to stabilizing the inferential structure of the feature space before the learning phase, thereby reducing sensitivity to environmental noise and acquisition variability.

C.1 Column-wise Majority Voting

To improve robustness at the acquisition level, a column-wise majority voting strategy is employed to aggregate window-level predictions into a single scenario-level decision. While the classifier operates independently on each temporal window, the monitored application behavior typically persists over time. Consequently, isolated prediction errors affecting individual windows can be mitigated by exploiting temporal consistency across the full acquisition.

The overall principle of the majority voting mechanism is illustrated in Figure C.2. Each temporal window extracted from the side-channel acquisition is first processed independently by the classifier, producing a sequence of local predictions. These predictions may occasionally fluctuate due to noise, transient perturbations, or local ambiguity in the feature representation. The majority voting stage aggregates the sequence of window-level predictions and selects the dominant class as the final acquisition-level decision.

Let an acquisition be segmented into M windows, and let $\hat{y}_i \in \{1, \dots, C\}$ denote the predicted class label for the i -th window, where C is the number of scenarios. The final predicted scenario $\hat{y}$ for the entire acquisition is defined as:

$$\hat{y} = \arg \max_{c \in \{1, \dots, C\}} \sum_{i=1}^M \mathbb{I}(\hat{y}_i = c)$$

where $\mathbb{I}(\cdot)$ is the indicator function, equal to 1 if its argument is true and 0 otherwise.

From an operational perspective, this aggregation strategy exploits the temporal persistence of application behavior. Even when some windows are incorrectly classified due to noise or partial loss of discriminative information, the correct behavioral pattern generally remains dominant throughout the acquisition. As a result, majority voting stabilizes the final decision by emphasizing consistent predictions over isolated local errors.

In noisy environments, majority voting therefore acts as an implicit decision-level denoising mechanism. Importantly, this robustness improvement is achieved without modifying the underlying feature representation, preprocessing pipeline, or classifier architecture. Instead, robustness emerges from temporal aggregation and behavioral consistency over time.

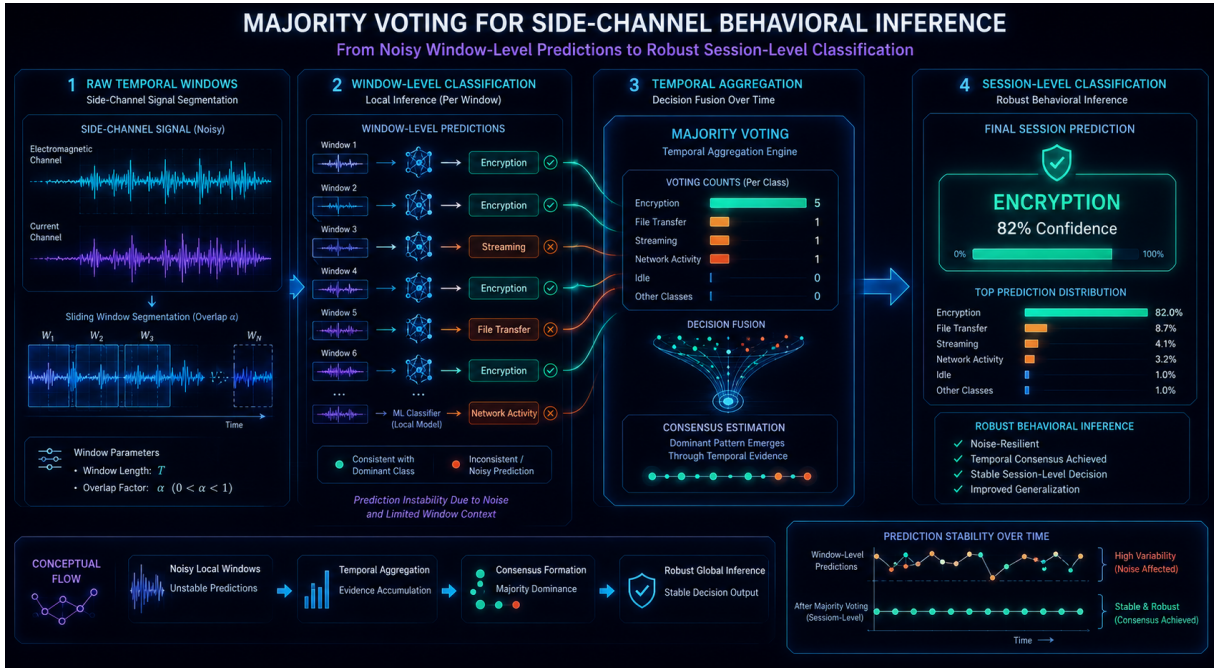


Figure C.2: Column-wise majority voting strategy for acquisition-level behavioral inference. Temporal windows extracted from the side-channel acquisition are classified independently, producing a sequence of local predictions that may include unstable or noisy decisions. The majority voting mechanism aggregates these predictions and selects the most frequent class as the final acquisition-level scenario, improving robustness and reducing the impact of transient misclassifications.

As discussed in Chapter 4, column-wise majority voting significantly improves acquisition-level accuracy and reliability, particularly under moderate and high noise conditions. The obtained results confirm that side-channel information is inherently cumulative over time and that reliable behavioral inference can emerge even when individual temporal windows exhibit limited discriminative power.

This strategy is particularly suitable for operational monitoring scenarios in cyber-physical and IoT systems, where stable long-term behavioral identification is often more relevant than isolated instantaneous predictions. Furthermore, the computational simplicity and interpretability of majority voting make it compatible with lightweight embedded deployments and real-time monitoring pipelines.

Appendix D

Implementation and Reproducibility Notes

All experiments were implemented in Python using standard scientific and machine learning libraries, including:

- NumPy and SciPy for numerical computation and signal processing;
- scikit-learn for classical machine learning models, feature selection, and evaluation;
- TensorFlow/Keras for deep learning experiments involving the CNN+LSTM architecture.

Experiments were executed on a workstation-class PC. GPU acceleration was not required for classical machine learning models, while the CNN+LSTM architecture optionally benefited from GPU support during training.

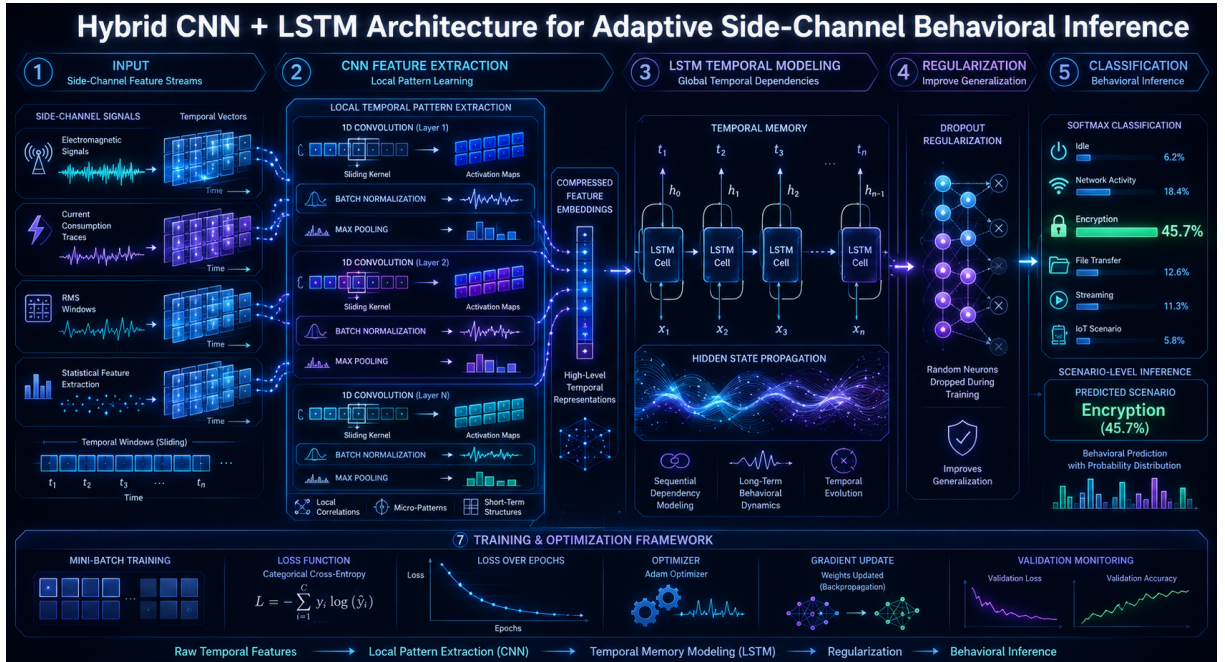


Figure D.1: Hybrid CNN+LSTM architecture adopted for side-channel behavioral inference. Sequences of feature vectors extracted from electromagnetic emissions, current consumption traces, RMS windows, and statistical descriptors are first processed by one-dimensional convolutional layers to capture local temporal patterns. The resulting high-level embeddings are then passed to an LSTM layer to model longer-term dependencies across consecutive windows. Dropout regularization and a final softmax layer produce scenario-level probability estimates.

D.1 CNN+LSTM Architecture

A hybrid Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) architecture was implemented to model temporal dependencies across sequences of side-channel feature vectors. The objective of this model was to evaluate whether a deep learning architecture could jointly exploit short-term local patterns and longer-term temporal dependencies emerging from consecutive observation windows.

As shown in Figure D.1, the input of the model consists of temporal sequences of feature vectors derived from the side-channel acquisition pipeline. Each sequence is composed of consecutive windows extracted from electromagnetic emissions and current consumption traces. These windows are represented through RMS values and statistical descriptors, forming a structured temporal representation of the observed system behavior.

The first stage of the architecture is the CNN feature extraction block. One-dimensional convolutional layers operate along the temporal dimension of the input sequence, applying sliding kernels over consecutive windows. This allows the network to identify local temporal correlations, short-term fluctuations, and micro-patterns that may characterize specific behavioral scenarios. The use of one-dimensional convolutions is particularly suitable in this context because the input is not an image-like structure, but an ordered temporal sequence of feature vectors.

Each convolutional layer is followed by batch normalization and max-pooling. Batch normalization stabilizes the distribution of intermediate activations during training, improving convergence and reducing sensitivity to local variations in the feature space. Max-pooling compresses the learned activation maps, reducing dimensionality and emphasizing the most relevant local responses. This also helps mitigate the effect of localized noise and improves the generalization capability of the model.

The high-level representations produced by the CNN block are then passed to the LSTM temporal modeling stage. While the CNN captures local patterns over short sequences of windows, the LSTM is responsible for modeling longer-term dependencies and temporal evolution. Through its recurrent memory mechanism, the LSTM maintains hidden states across time steps and learns how information evolves throughout the sequence. This enables the model to capture global behavioral dynamics that may not be evident from isolated windows.

This architectural design reflects a local-to-global learning strategy. The CNN component extracts compact local temporal features, while the LSTM component integrates these features over time to infer the broader behavioral scenario. In this way, the model can exploit both short-term side-channel variations and longer temporal dependencies associated with application-level behavior.

Before the final classification stage, dropout regularization is applied to reduce overfitting. Dropout randomly disables a subset of neurons during training, forcing the network to learn more robust and distributed internal representations. This is particularly important in side-channel datasets, where the number of acquisitions may be limited and the risk of learning acquisition-specific artifacts is non-negligible.

The final layer of the model is a fully connected softmax classifier. It produces a probability distribution over the considered scenarios, allowing the model to assign each input sequence to the most likely behavioral class. The predicted scenario corresponds to the class associated with the highest softmax probability.

The model was trained using categorical cross-entropy as the loss function and the Adam optimizer for parameter updates. Training was performed through mini-batch gradient descent for a fixed number of epochs, using a validation split to monitor performance during training. This setup allowed the CNN+LSTM model to be evaluated as a supervised deep learning baseline for temporal side-channel inference.

D.2 Role of the CNN+LSTM Model

The CNN+LSTM architecture was employed exclusively as a static supervised baseline and for comparative analysis. No architectural adaptation, online retraining, or reinforcement learning interaction was applied to the deep learning model.

During the reinforcement learning experiments presented in Chapter 6, the CNN+LSTM classifier remained fixed, while adaptation was restricted to the feature-representation level. This design choice ensured a fair comparison between static deep learning approaches and the proposed feature-aware adaptive control framework.

For consistency with the supervised learning setup, CNN+LSTM predictions were evaluated at the window level without temporal aggregation. This choice was made to assess the intrinsic sequence mod-

eling capability of the architecture rather than the effect of decision-level fusion mechanisms such as majority voting.