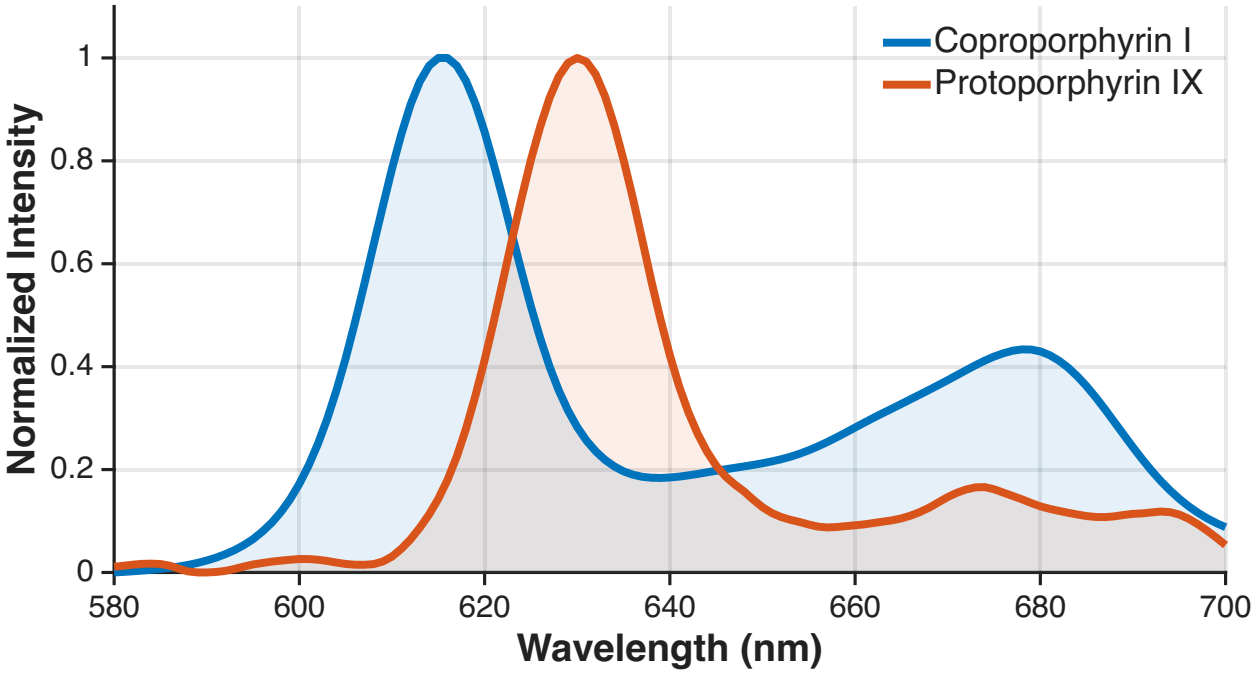
CHALMERS
UNIVERSITY OF TECHNOLOGY



AI-Based Spectral Analysis for Bacterial Biomarker Detection in Wound Diagnostics

Translating Non-Linear Optical Spectroscopy into Resource-Constrained Edge Architectures

Master's thesis in Biomedical Engineering

KORNEL KAMINSKI

OSCAR ERIKSSON

DEPARTMENT OF ELECTRICAL ENGINEERING
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se

Master's thesis 2026

AI-Based Spectral Analysis for Bacterial Biomarker Detection in Wound Diagnostics

Translating Non-Linear Optical Spectroscopy into Resource-Constrained Edge Architectures

KORNEL KAMINSKI
OSCAR ERIKSSON

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Electrical Engineering
Division of Signal Processing and Biomedical Engineering
Chalmers University of Technology
Gothenburg, Sweden 2026

AI-Based Spectral Analysis for Bacterial Biomarker Detection in Wound Diagnostics
Translating Non-Linear Optical Spectroscopy into Resource-Constrained Edge Architectures
KORNEL KAMINSKI
OSCAR ERIKSSON

Supervisor: Sofia Dall'Orso, Odinwell AB
Examiner: Xuezhi Zeng, Department of Electrical Engineering

Master's Thesis 2026
Department of Electrical Engineering
Division of Signal Processing and Biomedical Engineering
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Normalized fluorescence emission signatures of Coproporphyrin I (CpI, blue) and Protoporphyrin IX (PpIX, orange), showcasing the characteristic spectral peak separation detailed in Section 2.1.

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2026

AI-Based Spectral Analysis for Bacterial Biomarker Detection in Wound Diagnostics
Translating Non-Linear Optical Spectroscopy into Resource-Constrained Edge Architectures
KORNEL KAMINSKI
OSCAR ERIKSSON
Department of Electrical Engineering
Chalmers University of Technology

Abstract

The management of chronic wound infections presents a significant clinical challenge, often exacerbated by diagnostic delays and the limitations of subjective clinical assessment [1]. While point-of-care fluorescence imaging systems offer non-invasive visualization, precise biochemical quantification remains severely obstructed by complex, non-linear optical phenomena such as the Inner Filter Effect (IFE). Consequently, translating high-fidelity spectroscopy into cost-effective, edge-deployable diagnostic devices without a catastrophic loss of classification accuracy remains a major engineering bottleneck. Here, we establish a dual-track roadmap for clinical device engineering by systematically evaluating optical biomarker diagnostics for Coproporphyrin I (CpI) and Protoporphyrin IX (PpIX) across two distinct hardware paradigms: a continuous high-resolution miniature spectrometer and a constrained, discrete low-resolution photodiode sensor. For the high-resolution data streams, Convolutional Neural Networks (CNNs) and Spectral Transformers demonstrated a robust capacity to resolve non-linear optical variations, identifying optimal classification boundaries within controlled experimental datasets and serving as a scalable framework for future *in vivo* integration. Conversely, for the highly constrained low-resolution sensor array, the research isolated a physical indistinguishability limit stemming from coarse spectral resolution. This barrier was systematically overcome through domain-aware feature engineering and the realignment of diagnostic targets into a pragmatic 5×5 triage matrix. This physical-feature transformation enabled a minimalist Decision Tree architecture to achieve diagnostic parity, yielding a 98% classification accuracy. Concurrently, hardware profiling of the structurally pruned low-resolution convolutional models validated the embedded "race to sleep" paradigm [2]; the quantized 8-bit Integer (INT8) micro variant of the low resolution model executed edge inference in just 0.85 ms while consuming an ultra-low 0.17 mJ of energy on an ESP32-S3 microcontroller. Ultimately, this physics-first approach provides a transparent, certifiable foundation for immediate low-resolution edge deployment, while the deep-learning frameworks serve as a foundational proof-of-concept, demonstrating the diagnostic potential of these algorithms once comprehensive real-world datasets become available.

Keywords: Porphyrin Biomarkers, Spectral Transformers, 1D Convolutional Neural Networks, Post-Training Quantization, Quantization-Aware Training (QAT), Edge AI, Bacterial Autofluorescence, Optical Diagnostics.

Acknowledgements

We express our sincere gratitude to Odinwell and our supervisor there, Sofia Dall'Orso, for collecting and providing the experimental datasets utilized throughout this study, as well as for her insightful guidance and support over the course of this project.

Additionally, we would like to thank our examiner, Xuezhi Zeng, for her constructive feedback and academic review during the preparation and refinement of this manuscript.

Kornel Kaminski & Oscar Eriksson, Gothenburg, May 2026

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

1CH	1-Channel
2CH	2-Channel
3CH	3-Channel
AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under the Curve
CART	Classification and Regression Trees
CFU	Colony-Forming Unit
CNN	Convolutional Neural Network
CpI	Coproporphyrin I
CPU	Central Processing Unit
DRAM	Data Random Access Memory
DRS	Dynamic Reliability Switch
DT	Decision Tree
ESP-IDF	Espressif IoT Developement Framework
FP32	32-bit Floating-point
GAP	Global Average Pooling
GPIO	General Purpose Input Output
HR	High-Resolution
IFE	Inner Filter Effect
INT8	8-bit Integer
IRAM	Instruction Random Access Memory
LDO	Low-Dropout Regulator
LED	Light Emitting Diode
LOBO	Leave-One-Batch-Out
LR	Low-Resolution
MAC	Multiply-Accumulate
MDR	European Medical Device Regulation
MHSA	Multi-Head Self-Attention
ML	Machine Learning
MMU	Memory Management Unit

MTL	Multi-Task Learning
NIR	Near Infra-Red
PCB	Printed Circuit Board
PIL	Processior-in-the-Loop
PpIX	Protoporphyrin IX
PPK2	Power Profiler Kit II
PSRAM	Pseudo-Static Random Access Memory
PTQ	Post-Training Quantization
QAT	Quantization-Aware Training
RAM	Random Access Memory
ROI	Region of Interest
RTOS	Real-Time Operating System
SG	Savitzky-Golay
SIMD	Single Instruction Multiple Data
SiP	System-in-Package
SNV	Standard Normal Variate
SoC	System-on-Chip
SPI	Serial Peripheral Interface
SRAM	Static Random Access Memory
TFLM	TensorFlow Lite for Microcontrollers
UART	Universal Asynchronous Receiver-Transmitter
USB	Universal Serial Bus
VDD	Voltage at Drain
VIS	Visible
XAI	Explainable AI
XIP	Execute-in-Place

Contents

List of Figures

List of Tables

1
Introduction

This thesis investigates the design, optimization, and deployment of an embedded machine learning framework for the real-time optical quantification of bacterial pathogens in chronic wounds. Specifically, the research focuses on translating complex spectral autofluorescence data into actionable diagnostic metrics using resource-constrained edge devices. To establish the foundational context for this work, the following sections outline the clinical, physical, and computational motivations driving the study. Section 1.1 details the clinical challenge of chronic wound infections and the diagnostic principles of bacterial fluorescence. Section 1.2 then reviews current commercial imaging solutions, identifying the fundamental optical and quantitative limitations this research seeks to overcome. Finally, Section 1.3 explicitly defines the overarching goals and methodology of the project, while Section 1.4 establishes the strategic boundaries defining the scope of the study.

1.1 Background

The management of chronic and non-healing wounds represents a formidable challenge to global healthcare systems, burdening patients with diminished quality of life and imposing substantial economic costs on the system [1]. A central complicating factor in the pathology of these wounds is bacterial infection, which disrupts the normal wound healing process and can lead to severe systemic complications, including sepsis and amputation. Current diagnostic standards often rely on subjective clinical assessment or microbiological cultures that require 48-72 hours to incubate, creating a critical time gap where infections can escalate unchecked [4].
To bridge this diagnostic gap, Odinwell is developing an advanced optical sensor system that is capable of measuring the bacterial load in the wound. The sensor can collect light spectral data, and by leveraging bacterial autofluorescence, it can quantify the fluorescence emitted by bacterial biomarkers (e.g porphyrins), hence bacterial load [5]. While this sensor system captures detailed fluorescence spectra, the signals are complex, high-dimensional, and often corrupted by environmental noise and non-linear optical effects. Extracting reliable diagnostic information from these signals requires sophisticated signal processing that exceeds the capabilities of traditional linear regression methods .
This thesis proposes a high-performance Artificial Intelligence (AI) framework to process this spectral data. By utilizing high-capacity Deep Learning architectures, such as 1D Convolutional Neural Networks (CNNs) and Spectral Transformers, we aim to analyze the signal with greater precision. However, these complex models im-

pose significant computational demands that are often incompatible with the power and latency constraints of clinical edge devices [6]. Therefore, a core component of this research is the computational challenge of optimizing these models for real-time inference. This project will not only develop accurate diagnostic models but will also conduct rigorous research into model compression, quantization-aware training, and latency optimization, ensuring the solution is scalable and deployable on resource-constrained edge hardware.

1.2 State-of-the-art

The clinical adoption of fluorescence-based diagnostics has seen significant growth through the introduction of point-of-care imaging systems. Currently, the market is led by devices such as the MolecuLight i:X and DX systems and the Designs for Vision REVEAL FC. These devices utilize violet/blue light excitation (typically around 405 nm) to trigger the autofluorescence of bacterial biomarkers [7]. Specifically, MolecuLight systems provide real-time visual feedback by identifying red fluorescence, associated with porphyrin-producing bacteria (e.g., Staphylococcus aureus), and cyan fluorescence, associated with pyoverdine-producing bacteria like Pseudomonas aeruginosa [8].

However, a fundamental distinction exists between these current market solutions and the proposed research. Commercial devices like the MolecuLight are primarily qualitative or semi-quantitative imaging tools; they assist clinicians in "seeing" bacterial presence to guide debridement and sampling, but they do not provide precise molar quantification or nuanced spectral differentiation between similar porphyrin species [5]. The REVEAL FC serves a similar role in surgical and wound environments, focusing on the visualization of high bacterial loads ($> 10^4$ Colony-Forming Unit (CFU)/g) to improve standard clinical assessments which are traditionally subjective [7].

The primary limitation of these existing systems lies in their reliance on broad-band filter-based imaging. While effective for visualization, they are often confounded by the Inner Filter Effect (IFE) and the non-linear "Inverse Problem" described previously, where tissue scattering and high biomarker concentrations distort the perceived signal. For example, the spectral overlap between Coproporphyrin I (CpI) and Protoporphyrin IX (PpIX), with peaks near 616 nm and 630 nm respectively, is difficult to resolve with standard imaging hardware.

This research aims to transcend these visual-qualitative limitations by leveraging high-resolution spectral data. By applying Deep Learning architectures such as 1D-CNNs and Transformers, we seek to move beyond simple "presence/absence" detection toward accurate biochemical quantification. Furthermore, while commercial devices often utilize dedicated, high-power processing hardware, this thesis addresses the optimization challenge of bringing this level of analytical precision to low-cost, edge devices through Post-Training Quantization (PTQ) and Quantization-Aware Training (QAT). This would allow for a new class of diagnostic tools that are not only more precise than current imaging standards but also significantly more accessible for continuous monitoring in resource-constrained settings.

1.3 Purpose and goal

The primary purpose of this study is to design, implement, and rigorously optimize a machine learning pipeline for the spectral quantification and classification of bacterial biomarkers (CpI and PpIX). To bridge the gap between theoretical optical physics and deployable digital health technology, the research is structured around two synergistic tracks.

The Biomedical Engineering Track investigates the transition from high-resolution continuous spectroscopy to cost-effective, discrete low-resolution sensor arrays. This involves evaluating the absolute limits of deterministic signal processing and benchmarking them against advanced deep learning architectures, specifically comparing 1D-CNNs against Spectral Transformers for capturing long-range optical dependencies [9]. Crucially, this track explores the "Inverse Problem" of optical quenching (the IFE), systematically evaluating how domain-specific feature engineering, channel attention mechanisms, and the realignment of diagnostic targets into a pragmatic Clinical Triage matrix can overcome these hardware constraints. Ultimately, this track tests the hypothesis that injecting rigorous physical domain knowledge, specifically through engineered optical ratios and geometric gradients, can mathematically restructure a highly non-linear physical input space. Rather than attempting to entirely replace non-linear modeling with pure theoretical equations, this feature-space transformation explicitly decouples known optical variances (such as distance-induced attenuation and background substrate scattering). By pre-linearizing these predictable physical behaviors, the geometric complexity of the network's decision boundaries is significantly reduced, thereby enabling lightweight, edge-deployable machine learning architectures to achieve high classification accuracy.

Concurrently, the Embedded Systems & Edge Optimization Track focuses on the computational viability of deploying these diagnostic models directly to resource-constrained microcontrollers. Recognizing that clinical utility demands instantaneous, low-power feedback, this track evaluates architectural pruning, topological depth reduction, and 8-bit Integer (INT8) quantizaion using QAT/PTQ [6, 10]. By conducting hardware-in-the-loop profiling on an ESP32-S3 microcontroller, the study empirically quantifies the precise trade-offs between diagnostic accuracy, inference latency, memory allocation, and active energy consumption on bare-metal systems.

The anticipated benefits of this research extend to multiple stakeholders. Industrially, Odinwell will gain a validated software architecture capable of processing data from their sensors, which will serve as the basis for development of their digital health platform. Clinically, practitioners will benefit from a tool that provides objective, quantitative data about bacterial load, helping to reduce unnecessary antibiotic use and possibly improving wound healing outcomes [1]. Finally, the scientific community will receive a valuable benchmark comparison of CNNs versus Transformers for spectral regression tasks, contributing to the broader field of

computational chemical analysis.

1.4 Delimitations

The scope of this research is defined by several strategic boundaries intended to maintain academic focus and technical feasibility. Primarily, this study is delimited to spectral data acquired under standardized, laboratory-controlled conditions using synthetic specimens. The research does not extend to *in-vitro* bacterial cultures, clinical trials, or *in-vivo* diagnostics within a hospital environment; consequently, confounding biological variables, such as blood interference, tissue oxygenation and complex micro-biome diversity, are excluded from the current analysis.

Quantitatively, the study focuses on the differentiation and quantification of isolated biomarkers, specifically PpIX and CpI. The investigation is delimited to samples containing only a single biomarker species; the evaluation of mixed-biomarker samples is outside the scope to ensure a clear baseline for morphological differentiation. Furthermore, while the analytical framework aims to map the relationship between spectral features and chemical load, the experimental data is restricted to four discrete concentration levels for training and validation. The study does not aim to determine exact molarity across a continuous range beyond these predefined experimental tiers, nor does it address the physical complexities of samples outside this controlled concentration ladder.

Regarding model performance, establishing a formal clinical or regulatory diagnostic accuracy threshold is outside the scope of this study. Because the algorithms are trained and evaluated on synthetic laboratory data rather than true biological tissue, absolute accuracy metrics cannot be directly extrapolated to real-world patient outcomes. Instead, the target accuracy of this study is defined relatively: the deep learning architectures must substantially outperform the established deterministic signal-processing baseline to justify their computational overhead. Furthermore, to validate the hardware-software co-design pipeline, the study targets a high degree of mathematical parity during deployment; the optimized edge models must demonstrate strict quantization resiliency, retaining near-peak classification accuracy after being subjected to INT8 precision scaling and architectural pruning.

From a computational and hardware perspective, the project evaluates the software backend and high-performance optimization of predictive models. This assumes the functional calibration of the proprietary optical sensors provided by Odinwell, as the physical engineering of sensor hardware is not part of the thesis. While the study rigorously investigates the computational trade-offs of 8-bit integer quantization and profiles inference metrics on an ESP32-S3 microcontroller, the development of final, production-ready commercial firmware is not a central objective. Finally, the algorithmic exploration is explicitly delimited to high-capacity deep learning architectures suitable for sequential spectral data, specifically 1D-CNN and Spectral Transformers, alongside a deterministic Decision Tree baseline, excluding an exhaustive comparative analysis of all other traditional machine learning algorithms.

1.5 Societal, Ecological and Ethical Considerations

1.5.1 Societal Considerations

The primary potential societal impact of this research lies in improving the standard of care for patients with chronic and non-healing wounds. By providing a stepping stone for later development of real-time, objective bacterial quantification, which would reduce the clinical "time gap" that currently leads to escalating infections and systemic complications. Improved diagnostics can lead to more targeted treatment plans, potentially reducing the over-prescription of broad-spectrum antibiotics and contributing to the global effort against antibiotic resistance [11].

1.5.2 Ethical Considerations

As this study was strictly delimited to synthetic specimens and laboratory-controlled data provided by Odinwell, no human or animal subjects where involved in the research. Consequently, the project did not require ethical approval from a central board for clinical trials.

1.5.3 Ecological Considerations

The ecological impact of the project is considered from two perspectives:

- **Edge Efficiency:** A core objective of the HPC & Edge Optimization Track to maximize "accuracy-per-watt". By optimizing high-capacity models for low-power microcontrollers through quantization, we reduce the energy footprint required for advanced medical analysis at the point of care.
- **Waste Reduction:** Industrially, accurate wound diagnostics can reduce the disposal of clinical waste associated with ineffective dressing changes and redundant microbiological cultures.

2

Theory

This chapter establishes the theoretical and mathematical foundations required to interpret the spectral data and classification logic presented in this study. It transitions from the physical mechanisms of bacterial fluorescence to the mathematical frameworks used for signal isolation and non-linear pattern recognition.

2.1 Fluorescence Spectroscopy

Fluorescence is an optical phenomenon where a substance absorbs light at a high-energy, short wavelength and re-emits it at a lower-energy, longer wavelength [5, 12]. In this study, bacterial biomarkers (porphyrins) are excited using a 405 nm light source. The emitted photons form a characteristic spectral observable profile $I(\lambda)$, where intensity is a function of wavelength, presented in Figure 2.1.

2.1.1 Biomarker Spectral Signatures

The primary porphyrins evaluated are: Coproporphyrin I (CpI), that exhibits a broader emission profile with primary peak at $\lambda \approx 615\,\text{nm}$, and Protoporphyrin IX (PpIX), which characterizes by a sharp emission peak at $\lambda \approx 630\,\text{nm}$.

2.1.2 The Inner Filter Effect (IFE)

In dilute solutions with low optical density, the relationship between fluorophore concentration C and observed emission intensity I_{obs} is linear. However, as the biomarker load increases, the IFE introduces severe non-linearity. This phenomenon is primarily driven by secondary IFE (self-absorption), where emitted photons are re-absorbed by adjacent fluorophores within a highly concentrated environment before they can escape the sample. As derived from the foundational principles of sample geometry and deviations from the Beer-Lambert law [13], the observed intensity can be modeled as the product of linear fluorescence production and an exponential attenuation factor:

$$I_{obs} = k \cdot C \cdot 10^{-\epsilon \cdot C \cdot L}$$

Here k is a constant encompassing the instrumental geometry and quantum yield, ϵ is the molar absorptivity, and L is the optical path length the emission travels through the sample. Because the exponential attenuation term ($10^{-\epsilon \cdot C \cdot L}$) mathematically dominates the linear production term (C) at high optical densities, a decrease in absolute peak intensity can paradoxically indicate a significantly higher biomarker

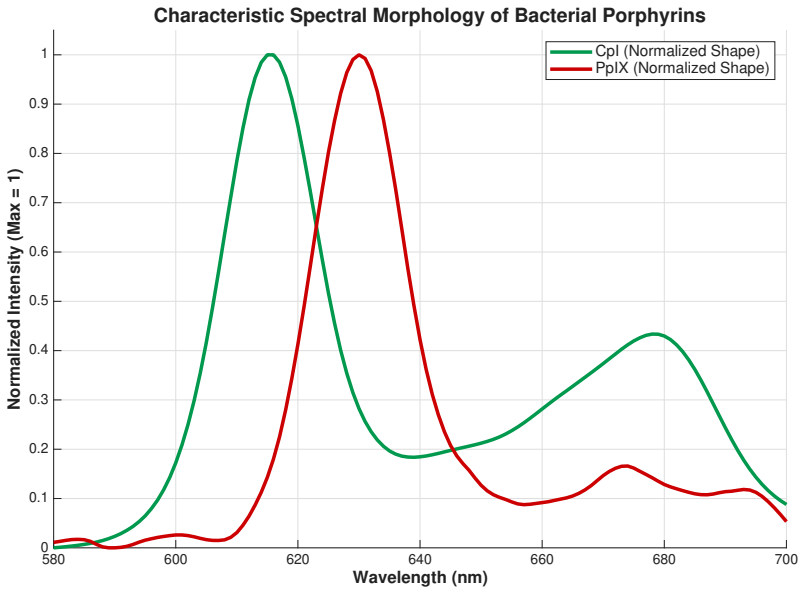


Figure 2.1: Characteristic fluorescence emission spectra of bacterial porphyrins, CpI (green) and PpIX (red). The curves demonstrate the distinct spectral morphology observed for biomarker speciation: CpI exhibits a primary emission peak at approximately 615 nm, whereas PpIX is characterized by a redshifted peak at 630 nm. Both signals are normalized to a maximum intensity of 1.0 to facilitate morphological comparison independently of absolute magnitude.

load. This establishes a critical optical "Inverse Problem", necessitating the use of relative morphological features, such as spectral red-shifting, rather than simple peak heights to accurately quantify high-density infections.

2.2 Mathematical Foundations of Signal Processing

2.2.1 Area Under the Curve (AUC) and Numerical Integration

In spectral analysis, the total energy emitted within a specific band is represented by the Area Under the Curve (AUC). Theoretically, this is the definite integral of the intensity function:

$$AUC = \int_{\lambda_{start}}^{\lambda_{end}} I(\lambda) d\lambda$$

Because spectral data is collected as discrete wavelength bins λ_i, the study utilizes the Trapezoidal Rule for numerical approximation. The area between two adjacent bins is calculated as:

$$AUC \approx \sum_{i=1}^{n-1} \frac{I(\lambda_i) + I(\lambda_{i+1})}{2} (\lambda_{i+1} - \lambda_i)$$

This integral provides a more robust proxy for biomarker load rather than single-pixel peak height, as it captures the total photon count across the entire emission envelope.

2.2.2 Standard Normal Variate (SNV) Transformation

To isolate spectral shape from variations in absolute magnitude, often caused by fluctuating sensor-to-tissue distances or excitation Light Emitting Diode (LED) intensity, the study employs the Standard Normal Variate (SNV) transformation. For a given spectrum x, each spectral data point x_i is centered and scaled by its own mean $\bar{x}$ and standard deviation s:

$$z_i = \frac{x_i - \bar{x}}{s}$$

where z_i is the normalized output value, x_i is the intensity at a specific spectral position, $\bar{x}$ denotes the arithmetic mean of the entire spectrum, and s represents the standard deviation of the spectral data. This transformation effectively maps the data into a unit-invariant space, forcing subsequent models to rely exclusively on relative spectral gradients and morphological features rather than absolute power levels [14].

2.2.3 Linear Detrending

The visible spectrum is often superimposed on a macroscopic scattering "slope" cause by the physical medium. This baseline is modeled as a first-order linear function $y = m\lambda + b$. By calculating the slope m between two non-fluorescent anchor points, this linear trend can be element-wise subtracted from the signal to flatten the spectral floor and isolate biological features.

2.3 Decision Logic in Biological Scales

2.3.1 Geometric vs. Arithmetic Boundary

Bacterial growth and serial dilutions follow a logarithmic scale ($10^1, 10^2, 10^3...$) [15]. When defining boundaries between concentration tiers (C4-C1), the arithmetic mean ($\frac{C_a+C_b}{2}$) is mathematically biased towards the higher value. To create equidistant "fences" in a log-biological space, the study utilizes the Geometric Mean:

$$T_{boundary} = \sqrt{C_a \cdot C_b}$$

where T_{boundary} represents the calculated operational decision threshold, while C_a and C_b denote the absolute baseline concentration values of two adjacent severity tiers. This mathematical framework ensures that the decision boundary is physically centered on the logarithmic midpoint, preventing "concentration bleeding" where lower-tier samples are erroneously upgraded to higher severity classes.

2.4 Deep Learning Theory

2.4.1 1D Convolutional Neural Networks (CNN)

While traditional Artificial Neural Networks (ANNs) utilize fully connected layers where every input feature interacts with every hidden unit, this dense architecture is highly susceptible to overfitting and computational inefficiency when processing high-dimensional sequential data [16]. To address these limitations, CNNs introduce a specialized architecture optimized for processing grid-like topologies, such as time-series data or 1D spectral arrays [17]. The fundamental operation in a 1D-CNN is the discrete convolution of a spectral signal x with a learnable weight matrix, known as a kernel k. Mathematically, the 1D convolution operation $(x * k)$ at a specific positional index t is defined as:

$$(x * k)_t = \sum_{i=0}^{K-1} x_{t+i} \cdot k_i$$

where t represents the positional index along the spectral sequence (corresponding to a specific wavelength or discrete detector bin), K is the spatial length of the kernel, and i iterates over this kernel dimension.

This operation acts as a highly efficient local feature extractor. Instead of learning a unique parameter for every single wavelength across the entire spectrum, the 1D-CNN employs sparse interactions by restricting the receptive field of the kernel to a narrow contiguous window of adjacent wavelengths [16]. As the kernel slides across the wavelength axis, computing the dot product at each position, it produces a transformed feature map that highlights specific morphological traits.

Crucially, this sliding mechanism inherently enforces parameter sharing. Because the exact same kernel weights are applied across the entire length of the spectrum, the network develops translation equivariance [16]. In the context of optical diagnostics, this means that if a kernel learns to detect the specific curvature or rising slope of a bacterial porphyrin peak, it can identify that exact spectral geometry regardless of where it appears on the detector array. This is particularly advantageous for spectroscopy, as biological emission peaks may experience minor wavelength shifts due to environmental factors, substrate variations, or hardware calibration drift [17].

By stacking multiple convolutional layers in a deep hierarchy, the network progressively learns increasingly complex representations. Early layers typically isolate low-level structural features, such as localized gradients, scattering baselines, or signal noise. Deeper layers subsequently aggregate these local variations to recognize high-level macro-geometric phenomena, such as the complete overlapping fluorescence profile of CpI and PpIX. This hierarchical feature extraction allows the 1D-CNN to autonomously build a robust mathematical model of the target fluorophores without requiring explicit, manually engineered thresholds.

2.4.2 Scaled Dot-Product Attention

Unlike traditional convolutional networks that process spatial features within a localized receptive field, Spectral Transformers utilize a mechanism known as Self-Attention to evaluate global dependencies across an entire input sequence simultaneously [18]. Conceptually, Self-Attention allows the model to examine a single wavelength bin and mathematically determine which other distant regions of the spectrum are most critical for providing diagnostic context [19]. To achieve this, the input spectral sequence is transformed into three distinct matrices Query (Q), Key (K), and Value (V) through learned linear projections. This mechanism functions similarly to a high-dimensional information retrieval system. The Query matrix represents the current spectral element being evaluated, establishing the specific morphological features the model is attempting to contextualize. The Key matrix serves as a set of corresponding identifiers for all other elements across the spectrum, signaling the specific features they contain. Finally, the Value matrix holds the actual underlying intensity or magnitude of those respective spectral elements.

The attention mechanism calculates how well the Queries match the Keys to determine the exact proportion of focus to allocate to each Value. This global relationship is mathematically defined by the Scaled Dot-Product Attention formula [18]:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

The mechanics of this equation begin with the dot product of the Query and Key matrices (QK^T), which computes a raw similarity score between every possible pair of elements in the sequence. A high score indicates a strong mathematical correlation between two distant regions of the spectrum, such as the primary fluorescence peak of a porphyrin and its corresponding secondary emission tail. This raw score is subsequently divided by a scaling factor, $\sqrt{d_k}$, where d_k represents the mathematical dimensionality of the Key vectors. This scaling step is critical because high-dimensional spaces can yield massive dot products, which would otherwise cause the subsequent softmax function to saturate [18]. Such saturation would push the network's gradients toward zero, effectively halting the optimization process during backpropagation.

Following this stabilization, the softmax function transforms the raw similarity scores into a normalized probability distribution, ensuring that the attention weights for any given element sum to exactly one. Finally, these normalized probabilities are multiplied by the Value (V) matrix. This concluding operation ensures that highly correlative spectral elements are mathematically amplified for the subsequent neural layer, while irrelevant background noise or non-actionable baseline scattering is systematically suppressed.

3
Methods

This chapter details the experimental framework utilized to develop, optimize, and physically deploy the diagnostic algorithms for bacterial biomarker quantification. Translating raw optical fluorescence into a reliable, edge-computed diagnostic requires a rigorous pipeline that spans physical data collection, mathematical signal processing, and low-level hardware compilation. Therefore, the methodology presented here is structured sequentially to reflect the full engineering lifecycle of an embedded machine learning system.

The chapter begins by outlining the standardized acquisition of dual-stream spectral data, detailing how the autofluorescence of CpI (produced by pathogens such as *S. aureus*) and PpIX (produced by pathogens such as *A. baumannii*, *E. coli*, and *K. pneumoniae*) was captured across both high- and low-resolution sensor architectures [20]. Following a description of the data preprocessing and feature engineering pipelines, a deterministic signal processing baseline is established to define the absolute mathematical limits of traditional classification. With this performance floor defined, the methodology transitions into the design and training of the deep learning architectures, detailing the evolution from standard Convolutional Neural Networks to highly constrained, domain-aware edge models. Finally, the chapter covers the hardware-software co-design phase, explaining the architectural pruning, quantization strategies, and physical power-profiling techniques required to transition these neural networks from theoretical workstation software into highly efficient firmware for low-power microcontrollers.

3.1 Data acquisition

All spectral data utilized in this study was acquired using Odinwell's proprietary optical sensor platform. The acquisition process was designed to capture the autofluoresence response of bacterial biomarkers under controlled excitation. Specifically, excitation light was delivered to the sample via a fiber coupled to a 405 nm LED. The resulting fluorescence emission was collected simultaneously by two independent return fibers, each routing the optical signal to a separate sensor.

This parallel configuration generated a synchronized dual-stream dataset, enabling a direct comparative analysis between two distinct sensor types. The first is a high-resolution sensor, which resolves the fluorescence signal with sufficient detail to persevere fine spectral features, serving as a high-fidelity benchmark. The second

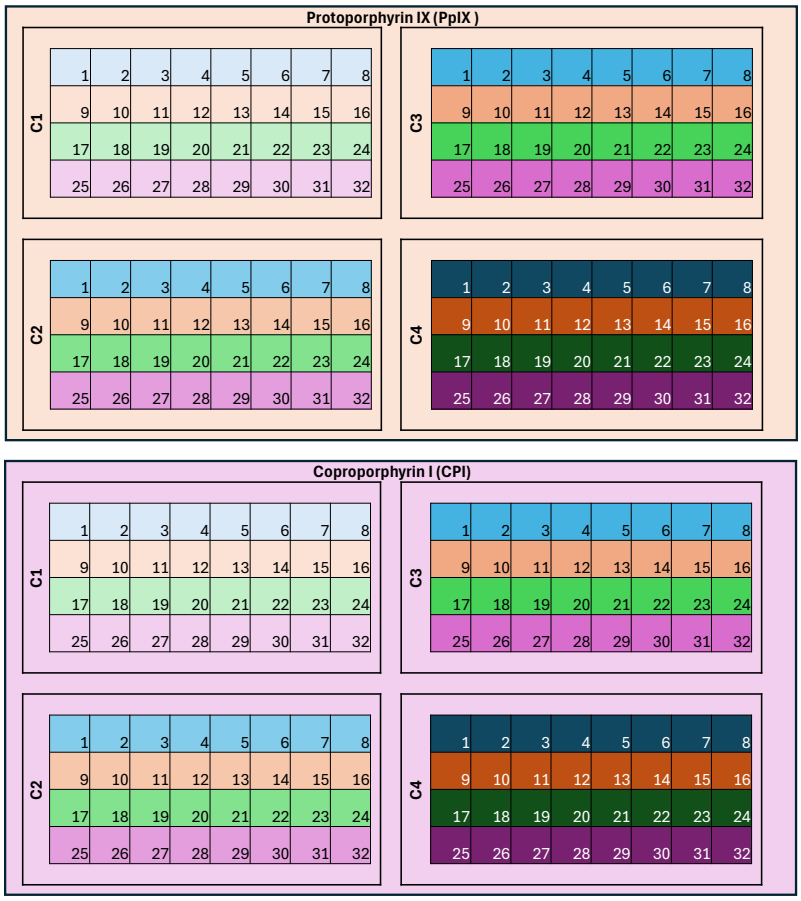


Figure 3.1: Schematic representation of the sample preparation process. Two biomarkers (PpIX, top; and CpI, bottom) were prepared at four concentrations (C1, C2, C3, C4). In total, 32 samples were generated across four distinct batches, with eight replicates per batch.

is a low-resolution sensor, which is designed to minimize the hardware footprint as well as the production cost. While the low-resolution sensor offers a much coarser spectral resolution (14 bins vs. 288 bins in the high-resolution sensor), it was hypothesized to retain sufficient morphological information for accurate biomarker quantification.

For both sensor pipelines, each data point consisted of a measurement triplet: a Background frame (LED off), followed by a Reference frame and a Sample frame. The Reference and Sample phases shared the exact same hardware capabilities but allowed for independent parameter settings, providing the flexibility to capture a secondary signals under different excitation or timing conditions within the same event.

To ensure the dataset encompassed the necessary experimental variability, a hierarchical sampling strategy was employed. The physical samples consisted of standardized filter papers that had absorbed a fixed volume of solution containing one of two bacterial biomarkers: PpIX and CpI. The sampling protocol was stratified across four distinct concentration levels to span a broad but relevant biological range. Within each concentration level, four independent productions batches were used to account for inter-batch variations in sample preparation.

From each batch, eight individual specimens are extracted, and each specimen was then subjected to ten consecutive measurement repeats to capture sensor noise and short-term temporal stability. Repeated specimens were produced to account for preparation and solution distribution variability (see Figure 3.1).

This hierarchical structure resulted in a total of 2560 active sampling events, calculated as follows:

$$N_{\text{total}} = N_{\text{bio}} \times N_{\text{conc}} \times N_{\text{batch}} \times N_{\text{specimen}} \times N_{\text{repeat}}$$

Substituting the experimental values:

$$2560 = 2 \times 4 \times 4 \times 8 \times 10$$

Here, N_{bio} is defined as the number of unique biomarkers (PpIX and CpI), N_{conc} as discrete concentration levels per biomarker, N_{batch} as independent production batches to account for preparation variance, N_{specimen} as the individual physical filter-paper specimens extracted per batch, and N_{repeat} as the consecutive measurement repetitions per specimen to capture sensor noise.

Additionally, 1280 measurements were taken from control samples where no biomarker was present. These control measurement were be taken across 16 different specimens.

Since each event was recorded by both the high-resolution and low-resolution sensors, this collection strategy yielded a perfectly synchronized dataset, ideal for benchmarking the trade-off between hardware fidelity and diagnostic accuracy. To ensure

strict procedural standardization, the entire data collection campaign was conducted externally by Odinwell personnel.

3.2 Data Preprocessing and Feature Engineering

Prior to network ingestion, the raw spectral data was processed through a multi-stage preprocessing pipeline designed to decouple genuine biological features from systematic experimental artifacts. This sequential workflow, visually outlined in Figure 3.2, serves as a structural roadmap for the data transformation process, systematically isolating the underlying optical signatures from hardware-induced noise.

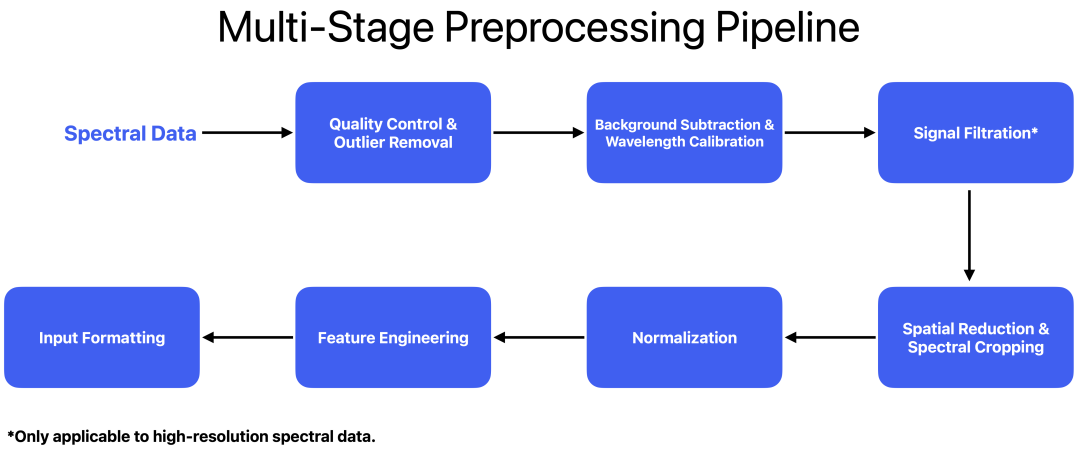


Figure 3.2: Overview of the multi-stage preprocessing pipeline. The flowchart illustrates the sequential transformation of raw spectral data into formatted neural network inputs.

3.2.1 Quality Control and Outlier Removal

Prior to any signal transformation, the raw dataset was subjected to a quality control filter to identify and remove invalid acquisition events. This step targeted four specific failure modes.

First, the data was checked for sensor saturation. For the low-resolution sensor, samples flagged by the sensor itself for digital or analog saturation were immediately discarded. For the high-resolution sensor, saturation around the excitation peak (405 nm) was permissible due to limited dynamic range, but measurements where saturation extended into the fluorescence emission region were discarded.
Second, the pipeline identified improper sensor contact by rejecting acquisition events where the intensity of the backscattered excitation light fell below a minimum threshold, indicating poor optical coupling.
Third, the background measurement was analyzed for excessive intensity; samples

exceeding a safety threshold were rejected due to severe ambient light contamination.
Finally, an inter-batch variance validation was performed to ensure the integrity of the cross-validation protocol. This involved comparing spectral profiles of identical concentrations across different batches to identify significant outliers, ensuring the models were evaluated on consistent chemical signatures rather than experimental anomalies.

3.2.2 Background Subtraction and Wavelength Calibration

To eliminate the influence of stray ambient light and establish a proper zero point, the background measurement (recorded with the excitation LED turned off) was element-wise subtracted from the corresponding active measurements for both sensor types. The empirical effect of this subtraction step on a representative high-resolution continuous frame is illustrated in Figure 3.3, demonstrating how subtracting the ambient baseline successfully isolates the true underlying autofluorescence emission peaks from external optical contamination.

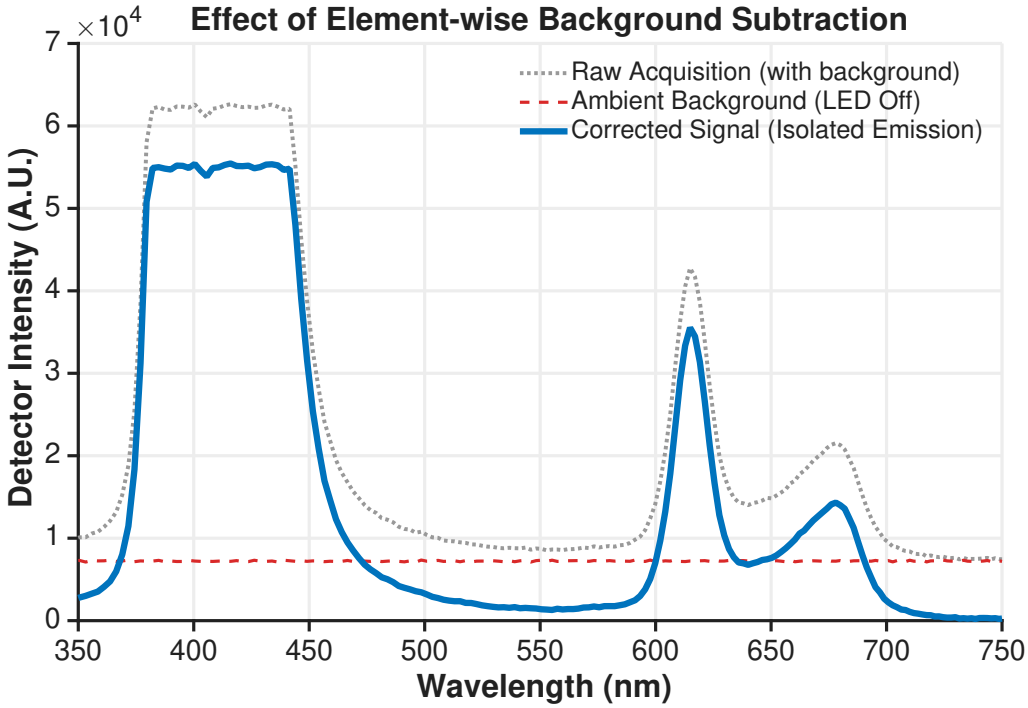


Figure 3.3: Empirical verification of the background subtraction step on a single spectrometer acquisition. Element-wise subtraction removes the underlying ambient stray light baseline (dashed red) from the uncorrected raw signal (dotted grey) to cleanly isolate the true biological fluorescence peaks (solid blue).

For the high-resolution sensor, the raw data consists of intensity values mapped to detector pixel indices. To ensure the model learns physically meaningful spectral features rather than hardware-specific artifacts, we utilize sensor-specific calibration data to map each pixel to its corresponding wavelength. The spectral data is then interpolated onto a standardized, linearly spaced wavelength grid. This ensures that

the i-th feature in the input vector always corresponds to the same physical wavelength across all samples, decoupling the model from specific hardware variances. Unlike the high-resolution sensor, the low-resolution sensor utilizes a standardized spectral mapping that does not require unit-specific calibration. Consequently, the discrete wavelength calibration step is omitted for this data stream, as the sensor output is directly mapped to a fixed, hardware-defined wavelength grid.

3.2.3 Signal Filtration and Morphological Preservation

To isolate the underlying biological fluorescence from high-frequency stochastic noise, signal smoothing was evaluated as the first stage of the preprocessing pipeline. The multifaceted impact of this filtration stage across distinct signal-to-noise ratio (SNR) regimes is empirically demonstrated in Figure 3.4 via a dual-concentration matrix mapping true dataset extremes.

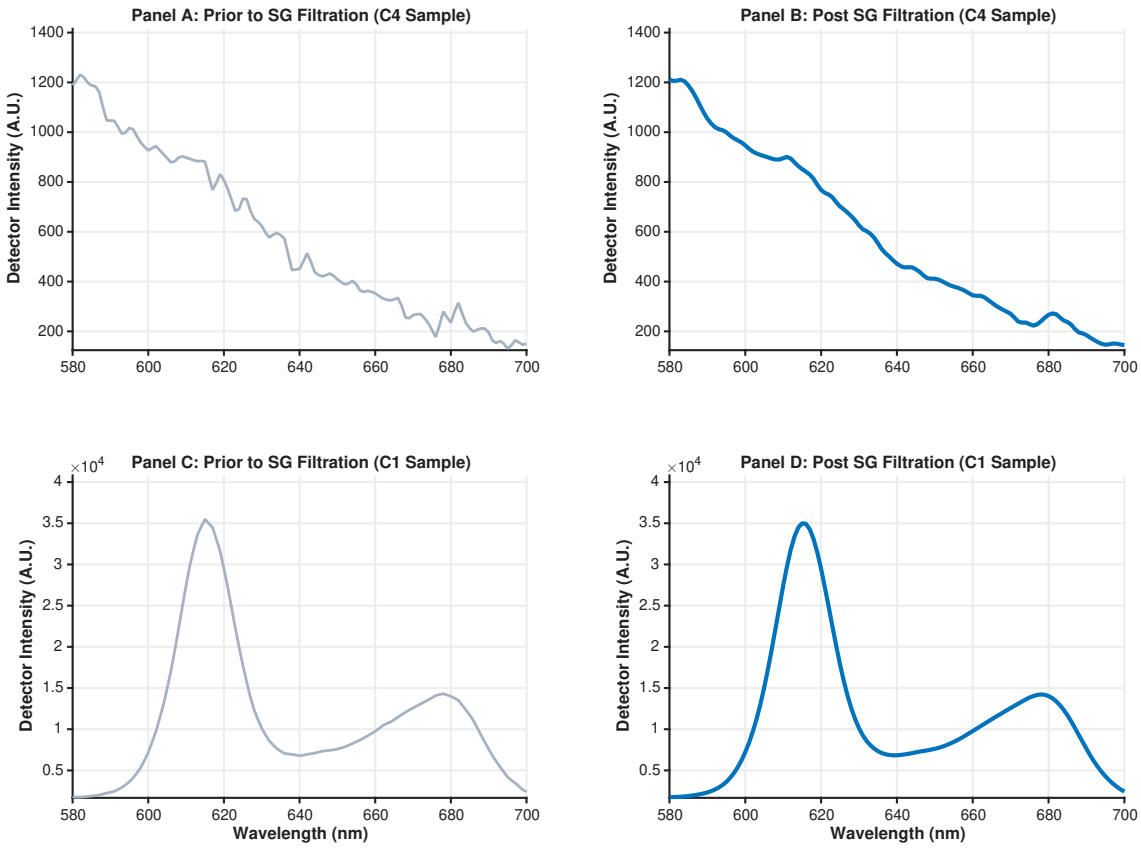


Figure 3.4: Multi-panel matrix verification of Savitzky-Golay (SG) filtration ($k = 3, f = 13$) on continuous high-resolution spectra. Panels A and B (Top Row) detail a low-concentration (C4) sample, showcasing baseline envelope stabilization under low SNR conditions. Panels C and D (Bottom Row) detail a high-concentration (C1) sample, demonstrating signal smoothing while strictly preserving the sharp, narrow 5–10 nm peak envelopes, absolute amplitudes, and diagnostic centroid positions without broadening or apex truncation.

For the continuous spectrum data stream generated by the high-resolution sensor, a

SG polynomial filter was implemented. This specific technique was selected over traditional moving average filters because it utilizes a localized least-squares polynomial fit to suppress high-frequency detector noise while strictly preserving the morphological integrity of the underlying signal [21]. Maintaining the exact width, height, and centroid of these localized emission profiles is critical for accurate biomarker speciation. Empirically, the filter was configured with a third-order polynomial ($k = 3$) and a frame length of 13 data points ($f = 13$). This specific configuration was determined to optimally dampen hardware noise without artificially broadening the narrow 5–10 nm emission peaks characteristic of the target porphyrins.

The physical efficacy of this configuration is demonstrated across the absolute boundaries of the diagnostic range in Figure 3.4. At the lower diagnostic boundary (Panels A and B), where the raw autofluorescence peak amplitude drops significantly and approaches the instrument's electronic noise floor, the transition visually confirms the filter's capacity to stabilize high-frequency roughness and smooth the baseline without compromising the faint biological signal.

Conversely, evaluating a high-concentration sample (Panels C and D) validates the morphological preservation capabilities of the architecture under high-amplitude conditions. Because this emission profile features intense, narrow geometric gradients peaking near 3.5×10^4 A.U., standard linear smoothing techniques would distort the signal by truncating peak heights and artificially expanding the peak base. The tracking confirms that the implemented polynomial architecture smooths the spectral envelope without inducing rounding artifacts at the apex or dimensional shifting along the steep margins of the primary porphyrin peak.

In contrast to the continuous data track, the low-resolution sensor maps optical data directly to 14 discrete, wide-band physical photodetectors. Because this data stream consists of a sparse vector rather than a continuous spectral gradient, polynomial smoothing filters were mathematically inapplicable. Consequently, high-frequency filtration was omitted from the low-resolution pipeline to prevent the introduction of synthetic interpolation artifacts that could distort the raw bin relationships.

3.2.4 Spatial Reduction and Spectral Cropping

Following the filtering, a spatial reduction strategy was applied to condense the input tensors, ensuring the neural networks optimized solely on actionable biochemical information while discarding sensor artifacts and uninformative regions.

For the high-resolution data, the spectra were truncated to a biologically relevant window bounded between 470 nm and 750 nm. This cropping served two essential functions. First, it intentionally discarded the saturated 405 nm excitation region and the uninformative Near Infra-Red (NIR) domain, forcing the models to focus exclusively on the visible fluorescence emission. Second, it performed a necessary mathematical correction for the preceding filtration stage; because the SG filter requires a span of neighboring data points to compute its polynomial fit, it inherently

introduces mathematical artifacts at the extreme ends of a data array. By filtering the wide-spectrum signal first and cropping second, these boundary artifacts were naturally excluded from the training data.

A parallel spatial reduction strategy was evaluated for the low-resolution sensor data. Unlike the high-resolution spectrometer, this hardware consists of discrete photodetector channels with varying spectral sensitivity profiles. Spatial reduction for this data stream involved the selective exclusion of specific photodetectors rather than a wavelength-gradient slice. Specifically, two of the 14 channels feature ultra-wide, broadband responsivity designed to capture overall visible light intensity. Depending on the specific architecture being utilized (described in Section 3.5.3), these broadband detectors were either retained to provide global brightness context or explicitly discarded to force the network to focus strictly on the relative relationships between the narrow-band channels, a configuration later formalized as the "Smart Router" architecture.

3.2.5 Normalization Pipelines

A fundamental challenge in point-of-care optical diagnostics is decoupling the underlying biological fluorescence from extrinsic physical variables. During *in vivo* acquisitions, fluctuations in sensor-to-tissue distance, applied optical pressure, local fluid saturation, and excitation light intensity inherently alter the absolute photon count reaching the detector. If this macroscopic variance is not mathematically neutralized, deep learning architectures risk overfitting to unstable environmental intensity shifts rather than extracting the true biological spectral morphology. Therefore, to systematically isolate the invariant biochemical signatures of the biomarkers from these extrinsic hardware and environmental factors, distinct normalization pipelines were engineered and evaluated for each sensor architecture.

For the continuous data generated by the high-resolution spectrometer, three distinct preprocessing pipelines were evaluated.
The first approach, Source Normalization, sought to scale the spectral data against the total incident energy delivered by the 405 nm excitation source. However, because the intense LED backscatter caused the optical detector to saturate (clip), the true excitation could not be directly quantified from the sample alone. To bypass this hardware limitation, the pipeline utilized the synchronized Reference frame to mathematically reconstruct the missing peak. First, a scaling coefficient was derived by calculating the ratio between the mean intensity of the sample and the mean intensity of the reference within a strictly unsaturated spectral window (355–375 nm). The reference spectrum was smoothed via moving average filter and multiplied by this coefficient to generate a reconstructed, unclipped excitation profile tailored to the specific acquisition event. The total incident energy was then quantified by calculating the definite integral (Area Under the Curve) of this reconstructed profile from the lowest wavelength up to 500 nm. Finally, the raw biological sample spectrum, after undergoing SG filtration, was divided by this integrated energy value.

The empirical effect of this energy rescaling within the primary emission viewport (580–700 nm) across three consecutive experimental acquisitions is visualized in Figure 3.5. Because the reference frame and the leakage wings track hardware-level excitation stability rather than local biological changes, this method standardizes the data scale from raw digital integers into bounded energy metrics (nm^{-1}), while accurately preserving real inter-frame emission variances, such as target photobleaching or thermal detector drift. However, it assumes a perfectly linear intensity correlation between the reference and sample phases, making the global scaling factor highly vulnerable to transient optical misalignments or reference-frame instability.

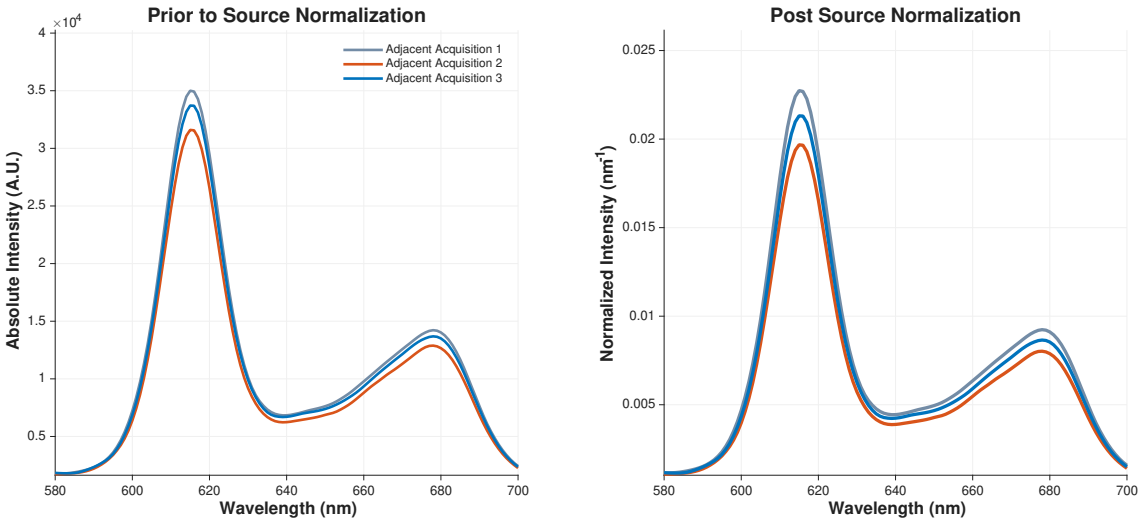


Figure 3.5: Performance evaluation of the Source Normalization pipeline within the 580–700 nm diagnostic window across three adjacent experimental acquisitions: (Left) uncorrected raw profiles displaying absolute intensity variance, and (Right) the post-normalization rescaling where reference peak reconstruction converts the vectors onto an energy-standardized scaling baseline (nm^{-1}) while preserving pristine inter-frame biological variances.

To circumvent the computational overhead and instability of peak reconstruction, the Wing Normalization was engineered. Instead of attempting to extrapolate the missing maximum intensity, this method utilized the unsaturated spectral "wings" immediately adjacent to the saturated 405 nm peak. By calculating the mathematical integral (Area Under the Curve) of the spectrum strictly bounded between 355 nm and 375 nm, the algorithm derives a robust, unclipped proxy for the total incident energy, as visually demonstrated in Figure 3.9. Following SG filtration, the entire emission spectrum is subsequently divided by this integral.

The empirical performance of this internal rescaling transformation focused within the diagnostic window is presented in Figure 3.6. Much like Source Normalization, dividing by the constant excitation wing area standardizes the data into a standardized spectral density framework (nm^{-1}) suitable for model optimization, without artificially eliminating localized kinetic intensity differences. Furthermore, this normalization strategy was adopted as the standard for the deterministic baseline established in Section 3.4, ensuring methodological continuity between the initial

signal processing models and the subsequent deep learning architectures.

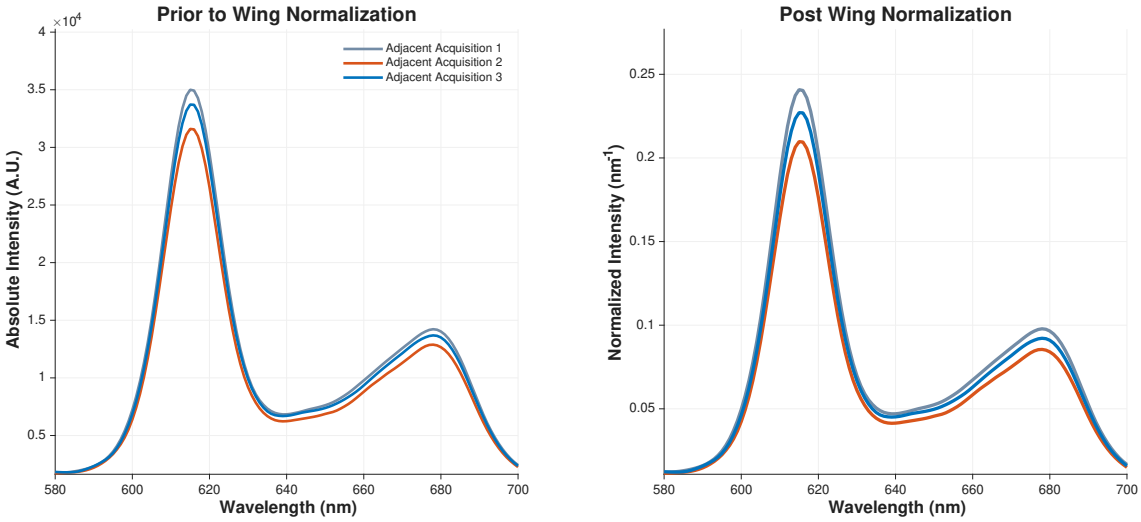


Figure 3.6: Performance evaluation of the Wing Normalization pipeline within the 580–700 nm diagnostic window across three adjacent experimental acquisitions: (Left) uncorrected raw profiles split by micro-environmental handling fluctuations, and (Right) the post-normalization output transforming absolute data counts into a standardized spectral density framework (nm^{-1}) while maintaining relative amplitude spacing.

The third high-resolution approach evaluated was the SNV transformation. This pipeline, like the previous, applies a localized, internal standardization. To ensure mathematical stability and prevent hardware noise from skewing the statistical baseline, the raw emission spectrum was first smoothed using a SG polynomial filter and then spatially cropped to the 470–750 nm range. The SNV mathematical transformation was executed strictly on this filtered, relevant window. By calculating the mean intensity of this isolated region, subtracting it from each data point, and dividing the result by the region's standard deviation, every sample was forced to possess a zero mean and a unit variance ($\mu = 0, \sigma = 1$).

The unique capacity of this transformation to achieve complete geometric profile alignment across consecutive exposures is demonstrated in Figure 3.7. By calculating statistical parameters directly inside the active data window, the transformation completely strips away multiplicative light scattering effects, absolute photon count variances, and additive baseline shifts. This forces the downstream deep learning models to optimize exclusively on pure geometric morphology and relative peak configurations of the fluorescence features, entirely decoupled from overall brightness changes.

Due to the discrete, 14-channel nature of the low-resolution sensor, continuous integration techniques could not be applied. Instead, three hardware-specific scaling pipelines were evaluated. The operational logic and algorithmic data flow for each discrete transformation are visually structured in Figure 3.8, detailing how specific hardware bins are dynamically isolated to act as normalization anchors.

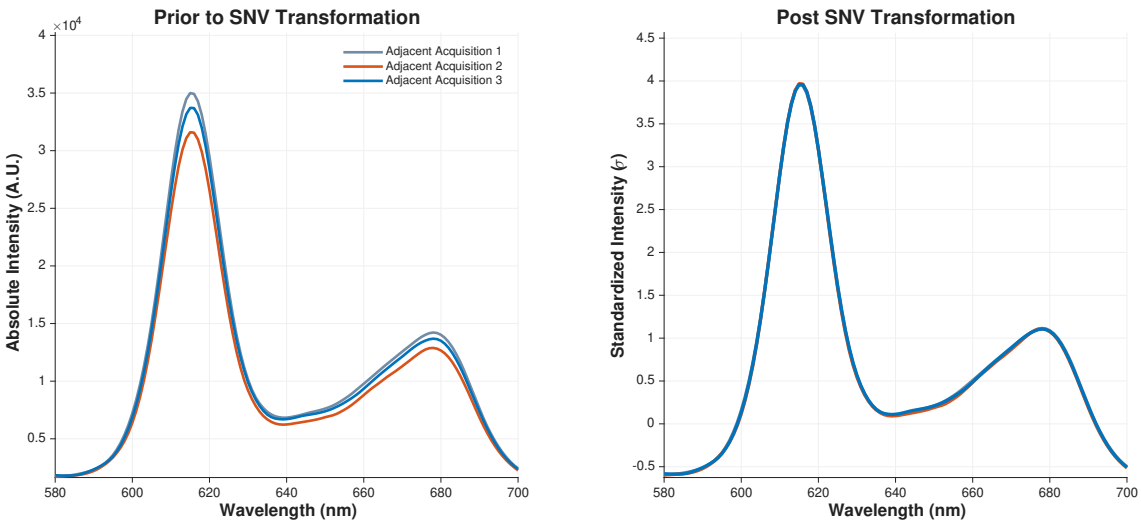


Figure 3.7: Performance evaluation of the SNV transformation pipeline within the 580–700 nm diagnostic window across three adjacent experimental acquisitions: (Left) uncorrected raw profiles displaying baseline separation, and (Right) the post-transformation output demonstrating complete geometric shape convergence (σ), forcing fluctuating spectral traces to lock into a single biological signature.

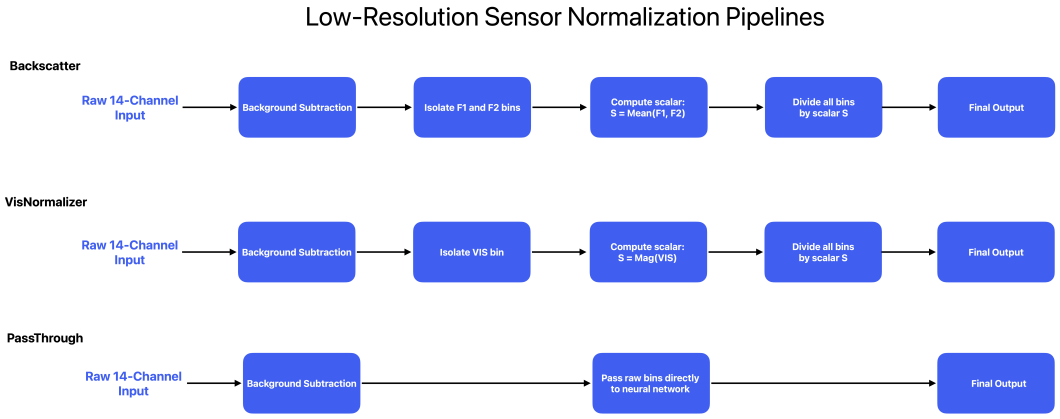


Figure 3.8: Schematic block diagram of the data preprocessing and feature engineering pipelines for the discrete 14-channel low-resolution sensor array. Parallel architectural tracks illustrate the PassThrough control branch, the broadband relative colorimetric transformation (VisNormalizer), and the excitation-driven dynamic alignment channel (Backscatter).

The first approach, the PassThrough pipeline, served as the absolute control. Aside from performing a basic element-wise substraction of the ambient background frame to remove stray light, the raw physical bin counts counts were passed directly to the neural network without any structural scaling. This pipeline was established to evaluate whether the advanced convolutional architectures could implicitly learn to map and ignore intensity fluctuations without explicitly programmed mathematical guidance.

The second approach, the VisNormalizer pipeline, scaled the spectrum against a dedicated broad-spectrum visible light detector engineered into the sensor hardware (the Visible (VIS) bin). By dividing the magnitudes of the narrow-band channels by the magnitude of this overall broadband channel, the 14-channel array is mathematically converted into a relative colorimetric profile. While this effectively neutralizes overall brightness fluctuations, it couples the normalization constant to the fluorescence emission itself, meaning highly fluorescent samples will inherently scale themselves down.

To mitigate the influence of biological emission on the scaling factor, the Backscatter pipeline was implemented as a targeted physical strategy. This pipeline explicitly leverages the intense 405 nm excitation light reflecting off the sample surface. It calculates the mean intensity of the two bins positioned over the excitation backscatter region (the F1 and F2 bins) and divides the entire discrete sequence by this average. Primarily, the intensity of this backscattered light is strongly proportional to macroscopic physical variables, such as sensor-to-target distance and optical contact pressure, allowing the method to dynamically standardize the sequence against hardware handling variances. However, it must be rigorously acknowledged that this backscatter is not perfectly independent from the underlying biological signal. The proportion of 405 nm light reflected back to the sensor is also modulated by the intrinsic optical scattering and absorption properties of the target tissue, including the density and composition of the bacterial load itself. Despite this inherent physical coupling, utilizing the excitation-dominated bins as a scaling denominator serves as a highly robust, pragmatic proxy to preserve the relative morphological shape of the fluorescence peaks across diverse acquisition environments.

3.2.6 Derivative Feature Extraction for High-Resolution Sensor

In the high-resolution spectrometry pipeline, multi-channel input tensors were engineered to assist the deep learning models in identifying subtle spectral shifts that may be obscured in raw intensity plots. Because the 288-channel spectrometer provides a high spectral sampling density, it allows for the calculation of robust gradients that represent the underlying physical change in fluorescence independently of absolute magnitude. Three distinct input configurations were generated for evaluation. The baseline 1-Channel (1CH) configuration utilized only the normalized spectral data. The 2-Channel (2CH) configuration stacked the normalized spectra with its first derivative, highlighting the rate of change and the exact locations of

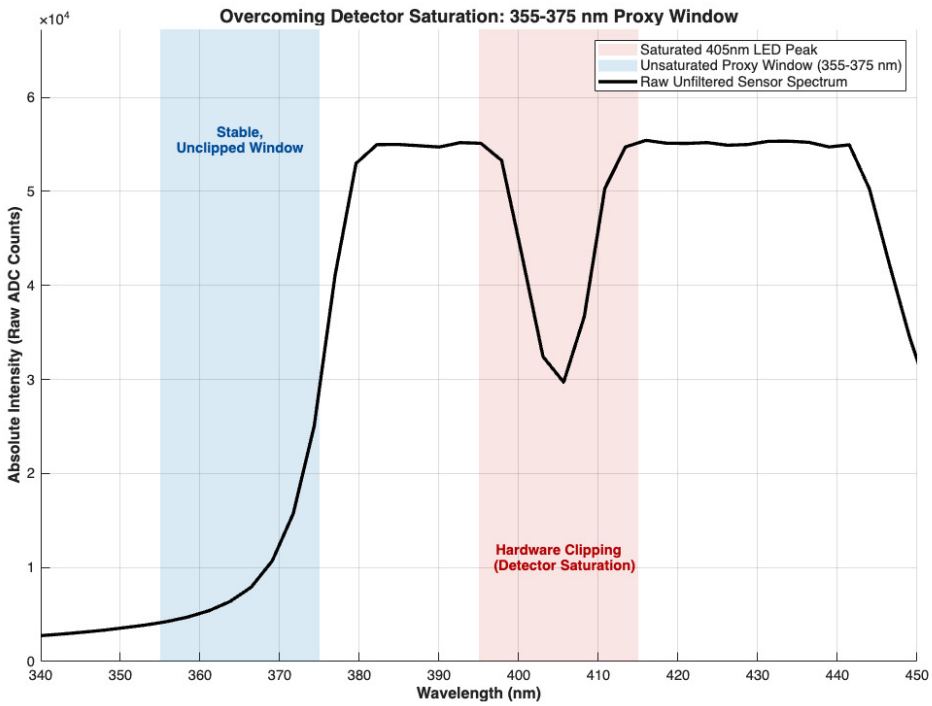


Figure 3.9: Visualization of raw high-resolution sensor data demonstrating optical detector saturation (clipping) at the primary 405 nm excitation peak. To mathematically bypass this hardware limitation, both the Source Normalization and Wing Normalization pipelines utilize the stable, unsaturated 355–375 nm spectral window (highlighted) as a robust proxy for incident energy scaling, thereby decoupling the biological fluorescence from physical variations in optical coupling.

spectral slopes. The final 3-Channel (3CH) configuration incorporated the spectral data alongside both the first and second derivatives, theoretically exposing hidden overlapping peaks and concave boundary features. These derivative configurations were utilized exclusively for the high-resolution models to assess the impact of local geometric gradients on classification accuracy.

3.2.7 Low-Resolution Sensor Architecture and Spectral Bins

The discrete low-resolution evaluations in this study were conducted utilizing a highly integrated 14-channel multi-spectral sensor. Operating across a range of approximately 380 nm to 1000 nm, the device employs high-precision optical interference filters deposited directly onto a 5×5 CMOS photodiode array [3]. This architecture physically divides the incoming light into 11 discrete channels within the visible spectrum, one NIR channel, one broadband clear channel (VIS), and a dedicated flicker detection (FD) channel. Understanding the specific central wavelengths (λ_p) and Full Width at Half Maximum (FWHM) of these integrated filters is critical, as they dictate the physical limitations of the spectral resolution and form the foundational inputs for all subsequent downstream feature engineering. Based on their relevance to autofluorescence diagnostics, these channels can be functionally categorized into four distinct continuous optical bands.

The first functional category comprises the excitation and violet-blue band, consisting of the F1, F2, FZ, and F3 channels centered at 407 nm, 424 nm, 450 nm, and 473 nm, respectively. Within the context of this diagnostic pipeline, the F1 and F2 bins are of paramount importance. Because they encompass the 405 nm wavelength of the primary excitation LED, they are utilized by the Backscatter Normalization pipeline to dynamically quantify surface-level optical coupling prior to biological emission.

Transitioning to longer wavelengths, the green, yellow, and broad-orange band includes the F4, F5, FY, and FXL channels, which are centered at 516 nm, 546 nm, 560 nm, and 596 nm, respectively. The FXL channel is particularly notable due to its wide FWHM of 93 nm. Rather than targeting a highly specific biological peak, FXL functions as a broad baseline proxy for general tissue scattering in the orange spectrum. This stable macroscopic scattering profile makes it the ideal mathematical denominator for calculating the Normalized Primary Red emission ratio.

The third category represents the primary biological acquisition region, comprising the red and deep-red fluorescence band. This band contains the F6, F7, and F8 bins, centered at 636 nm, 687 nm, and 738 nm, respectively. These bins are strategically aligned to isolate critical bacterial signatures, as F6 bin captures the primary emission peak of CpI, while the adjacent F7 bin detects the secondary, red-shifted emission tail characteristic of PpIX.

Finally, the sensor utilizes near-infrared and broadband reference channels to establish deep-tissue and ambient macroscopic baselines. The NIR channel, centered

at 855 nm, operates entirely outside the visible fluorescence window, providing a stable, isolated measurement of ambient infrared interference. Complementing this is the unfiltered VIS channel, which lacks a specific interference filter, allowing it to capture the total macroscopic brightness of the visible spectrum. This broadband intensity measurement is explicitly utilized as the scaling denominator in the VisNormalizer pipeline to neutralize overall brightness fluctuations.

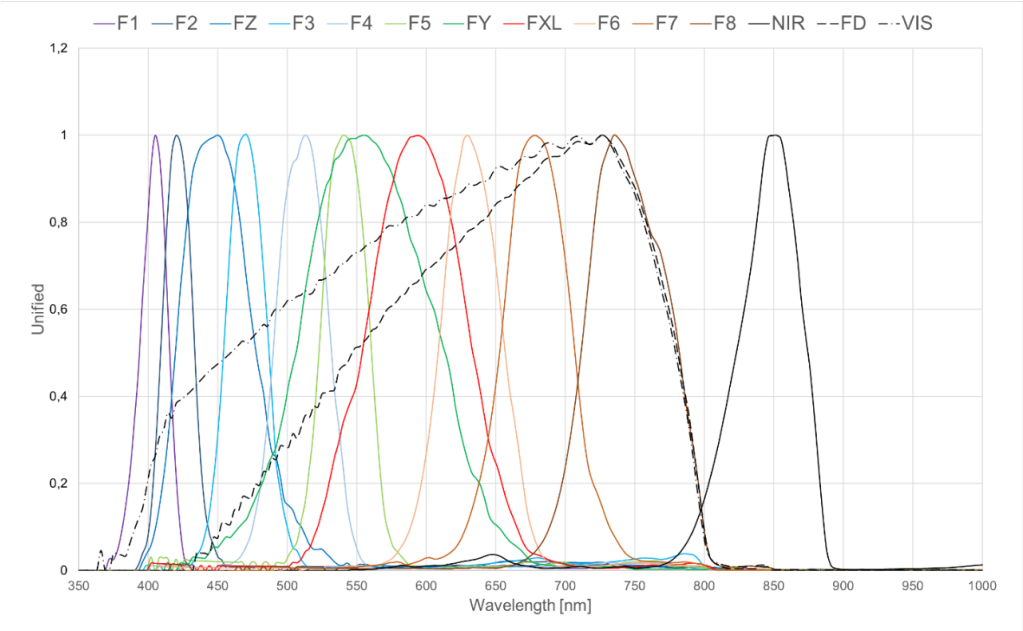


Figure 3.10: Typical spectral responsivity of the 14-channel sensor. The graph illustrates the normalized Gaussian profiles and Full Width at Half Maximum (FWHM) of the integrated interference filters, demonstrating the physical spectral overlap that necessitates advanced mathematical decoupling and feature engineering. Reproduced from [3].

3.2.8 Domain-Specific Feature Engineering for Low-Resolution Sensor

Conversely, the discrete and heavily constrained nature of the 14-channel sensor necessitated a more stable approach to feature engineering, focused on global ratios rather than derivatives. Prior to architectural processing for the low-resolution hardware, seven domain-specific physical features were engineered to combat the variance introduced by the inverse-square law. In the context of optical biopsy, absolute photon counts are highly vulnerable to distance fluctuations; therefore, mathematical neutralization was achieved by standardizing biological emission bands against internal spectral baselines.

The most critical of these engineered variables were the four Normalized Emission Ratios, which utilized the optical characteristics to decouple biochemical signatures from extrinsic hardware handling. Specifically, the Normalized Primary Red emission was calculated by dividing the 636 nm bin (F6) by the wide-band 596 nm orange

bin (FXL), effectively isolating the primary fluorescence peak of the porphyrins from general scattering. In a similar manner, the Normalized Deep-Red ratio (F7/FXL) was constructed using the 687 nm bin to capture the secondary emission tail characteristic of PpIX. To address deeper tissue interactions, the Far-Red Boundary (F8/NIR) and the Violet Backscatter Index (F1/NIR) were established. The former provided a stable ratio of surface absorption to deep-tissue baseline scattering, while the latter contrasted intense surface-level excitation backscatter against the deep-penetrating 855 nm Near-Infrared channel, serving as a powerful proxy for optical coupling.

In addition to these ratios, three macro-geometric slopes were calculated to capture the broader structural morphology of the spectrum independently of absolute brightness. The Red-to-NIR Scattering gradient tracked the transition from the primary fluorescence peak to the deep-tissue baseline, while the Near-Infrared Gradient measured the structural slope across the red-edge boundary (NIR-FD). Finally, a Mean Gradient was computed as the average first derivative across the entire 14-channel array to provide a scalar representation of global spectral volatility. These seven engineered features were concatenated into a secondary input vector for the hybrid deep learning architectures and later served as the exclusive inputs for the simplified Decision Tree algorithms.

3.2.9 Input Formatting

Finally, the preprocessed data were cast into the precise tensor formats required for neural network training and evaluation. Instead of relying on the raw, fixed lengths of the sensor outputs, the final tensor dimensions were dynamically defined by the preceding cropping and feature engineering stages.

For the high-resolution data stream, the spatial dimension of the tensor was dictated by the specific wavelength cropping boundary (e.g., the retained bins between 470 nm and 750 nm), resulting in a finalized spatial sequence length of L_{high}. Furthermore, to integrate the derivative feature extraction, these inputs were structured as multi-channel 1D tensors with a shape of (L_{high}, C). In this configuration, the channel depth $C \in \{1, 2, 3\}$ corresponded directly to the evaluated derivative configuration (1CH, 2CH, or 3CH). This multi-dimensional structure allowed the convolutional kernels to simultaneously process the spectral magnitude and its mathematical rate of change.

Conversely, the formatting for the low-resolution sensor maintained a consistent 14-channel input foundation across all experimental iterations to ensure a standardized benchmark. To accommodate the dual-branch architectures utilized in the later stages of the ablation study, the data were structured as a multi-modal input consisting of two distinct tensor objects. The primary object was a 1D spatial array containing the 14 raw hardware bins, intended for the convolutional branches. While the specialized "Smart Router" architecture physically cropped the first two bins from the sequence to prevent excitation blinding, this was handled as an internal

architectural operation rather than as a change in the underlying data delivery. The second object was a 1D feature vector containing the seven engineered physical ratios and macro-geometric slopes, computed dynamically from the source 14-channel data. This bifurcated formatting allowed the hybrid models to simultaneously leverage the spatial pattern recognition of the CNN and the explicit physical constraints provided by the dense branch, maintaining a uniform data pipeline across the entire ablation path.

By structuring the datasets into these strictly defined tensor shapes, the high-fidelity continuous models and the heavily constrained discrete edge models could be independently trained and systematically benchmarked against one another.

3.3 Data partitioning and LOBO cross-validation

To ensure the developed models generalize to unseen samples rather than memorizing specific sample noise patterns, the dataset is partitioned using a hierarchical Leave-One-Batch-Out (LOBO) strategy. This approach is mandatory for clinical validity, as it evaluates the system's robustness against inter-batch variability in sample preparation, sensor state and environmental conditions.

3.3.1 Prevention of data leakage

A naive random shuffle of the dataset would induce severe data leakage due to the hierarchical nature of the data collection process.

First repeated measurements from a single physical specimen share identical morphological characteristics, such as local fiber distribution in the filter paper and specific biomarker density variations. If these repeats where split across training and test sets, the model could achieve artificially high accuracy by "memorizing" the unique spectral fingerprint of a specific specimen, learning the noise and local artifacts rather than the generalized biochemical features.

Second, although all batches where produced in a single continuous session, systematic variations in sample preparation remain a confounding factor. Minor inaccuracies during the manual dilution and pipetting process can result in slight concentration offsets specific to a given batch. If samples from a single preparation batch are present in both the training and testing phases, the model runs the risk of learning these batch-specific preparation artifacts rather than the underlying spectral response of the biomarkers.

To prevent both forms of leakage, the data is partitioned according to the following strict isolation rules:

1. **Specimen-Level Grouping:** All 10 repeated measurements belonging to a specific specimen are treated as an indivisible unit, ensuring that a single physical sample never spans multiple partitions.

2. **Batch-Level Isolation:** Entire production batches are kept disjoint. The training and testing sets are constructed such that they share no common preparation origin, forcing the model to generalize across inter-batch preparation variances.

3.3.2 LOBO protocol

To strictly control for experimental batch effects, the dataset is partitioned utilizing a LOBO cross-validation protocol, a methodology proven to prevent machine learning models from memorizing physical preparation artifacts in spectroscopic data [22]. The dataset is divided into four disjoint folds based on the production batch. The model evaluation follows an iterative process:

- **Training and Validation:** In each iteration, three full batches are utilized for model development Within these batches, a specimen-grouped 80/20 split is applied to separate the Training Set from Validation Set, ensuring hyperparameter tuning is conducted on data temporally distinct from the final set.
- **Independent Testing:** The one of the four batches from each concentration is strictly held out as the Test Set. This batch, selected at random, represents "unseen" data from a completely different experimental session, serving as final unbiased benchmark for the system's ability to resolve spectral residuals.

3.3.3 Stratification and balanced sampling

To ensure each fold provides a representative training signal, the partitioning maintains strict stratification across concentration levels. Control samples are equally distributed across the folds to preserve the class balance between positive biomarker detection and negative baseline.

3.4 The Deterministic Baseline

Before implementing high-dimensional machine learning models, a deterministic signal processing baseline was established using MATLAB. The structural sequence of this algorithmic pipeline is visualized in Figure 3.11. The purpose of this baseline was to evaluate the absolute physical limits of traditional, threshold-based classification and to formally establish a performance floor to justify the necessity of AI for this diagnostic application

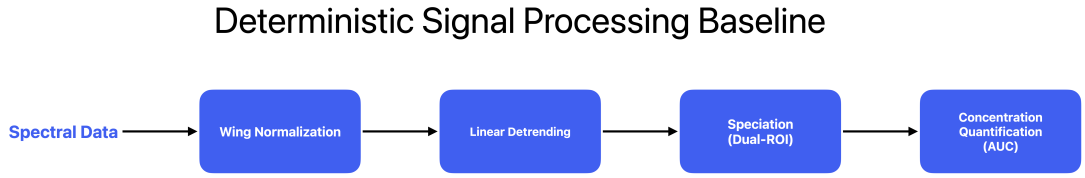


Figure 3.11: Sequential architecture of the deterministic signal processing baseline.

3.4.1 Preprocessing and Feature Extraction

The deterministic evaluation adhered to the Wing Normalization procedure detailed in Section 3.2.5, ensuring that signal intensities were standardized against total incident energy prior to feature extraction. To further isolate the biological signal, this was followed by a linear detrending operation applied within the 580–670 nm range. Detrending was a critical step in this traditional pipeline to flatten the spectral baseline, effectively decoupling the localized fluorescence peaks from the broad-band scattering slopes induced by the physical substrate.

Speciation, defined as the identification of the specific biomarker fluorophore, was achieved via a Dual-Region of Interest (ROI) peak-extraction algorithm. This method scanned the detrended spectrum to identify the maximum intensity within the expected CpI emission window (614–620 nm) and compared it against the maximum intensity found within the PpIX window (625–635 nm). The biomarker type was determined by whichever window yielded the higher relative peak magnitude. In the initial iteration of the model, concentration quantification was determined by calculating the total AUC between 600 nm and 750 nm, providing a single scalar value intended to represent the total pathogen load.

3.4.2 Evolution of the Thresholding Algorithms

Two variations of thresholding logic were developed to systematically test the absolute limits of traditional 1D classification.

3.4.2.1 The Standard Baseline

The standard deterministic approach utilized median-based thresholds optimized for average variance between classes. This baseline calculated class boundaries by finding the simple mathematical midpoint (arithmetic mean) between the AUCs of adjacent concentration classes. For example, the threshold separating a C3 infection from a C4 infection was placed exactly halfway between their respective median AUC values. While providing a functional starting point, this method assumed a linear relationship between signal and concentration, which does not reflect the physics of optical biopsy.

3.4.2.2 The Upgraded Baseline

An upgraded baseline was developed to address the mathematical trade-offs exposed by the first iteration. This model optimized the core signal processing math to reflect biological and logarithmic realities. First, the safety threshold for blank controls was adjusted to a 2.5-Sigma limit applied directly to the detrended peak height rather than the total AUC, mathematically isolating the biological emission bump from the physical LED scattering slope.

Second, the global AUC integration window was replaced with species-specific optical windows. The CpI integration was restricted to 600–650 nm, while the PpIX

integration was restricted to 610–710 nm. Crucially, these distinct AUC calculations were not used to compare the two biomarkers against one another; rather, they were utilized strictly within a single biomarker class to differentiate its specific concentration tiers. By tailoring the integration boundaries, the algorithm successfully eliminated irrelevant high-wavelength sensor noise that had previously skewed the severity calculations. Lastly, the algorithm replaced arithmetic means with geometric means to calculate the boundaries between concentration classes. Because biomarker concentrations scale logarithmically through serial dilution, utilizing geometric means prevented the artificial threshold skewing that previously caused adjacent concentration classes to bleed into one another. Together, these refinements established the theoretical maximum classification performance of a deterministic thresholding model.

3.5 Model Design and Training

To resolve the subtle spectral signatures of the biomarkers, different types of deep learning architectures where developed and tested for both the high-resolution and low-resolution sensors.

3.5.1 Shared Multi-Task Classification Heads

To efficiently utilize the complex hierarchical features extracted by the core convolutional backbone, all evaluated architectures across both sensor types employed a shared Multi-Task Learning (MTL) structure [23]. In this design, the final processed data often referred to as latent representations was directed into two parallel decision-making branches, allowing the model to perform two distinct but biologically related diagnostic tasks simultaneously. By jointly optimizing for multiple objectives, MTL acts as a structural regularizer, forcing the network to learn generalized, robust features that benefit all tasks rather than overfitting to a single target [23].

The first branch was dedicated to identifying the specific type of biomarker present in the sample, categorizing it as either a Control, PpIX, or CpI. This task was achieved using a standard softmax activation function applied to the latent vector. The softmax function mathematically normalizes the outputs into a probability distribution for mutually exclusive categories, ensuring the model confidently selects only one distinct biomarker species per inference event.

The second branch was tasked with quantifying the concentration severity of the identified biomarker. In traditional machine learning classification, categories are treated as completely independent entities without any inherent spatial order. Consequently, predicting a “low” concentration when the true answer is “high” would normally be mathematically penalized exactly the same as predicting a “medium” concentration, completely ignoring the magnitude of the error. To overcome this flaw and capture the naturally ordered scale of biomarker concentrations, this secondary branch utilized a cumulative ordinal regression framework rather than standard categorical one-hot encoding [24]. Under this method, ascending concentration levels are represented as a sequence of cumulative binary thresholds. For example,

if the diagnostic matrix contains four distinct concentration tiers, the third highest severity is encoded as the target array $[1, 1, 1, 0]$. This array is then evaluated using a sigmoid activation function, which calculates the independent probability of the sample surpassing each consecutive severity threshold [24]. By calculating the binary cross-entropy across these step-by-step probabilities, the network receives a mathematically heavier penalty for predictions that deviate further from the true concentration magnitude. This specialized ordinal architecture forces the deep learning models to respect the continuous, cumulative physical reality of biomarker concentration, rather than treating sequential concentrations as isolated and unrelated bins.

3.5.2 Deep Learning Architectures for the High-Resolution Sensor

For the high-resolution sensor, three architectural paradigms were evaluated to determine the optimal method for processing continuous spectral sequences. A visual comparison of these three models is presented in Figure 3.12.

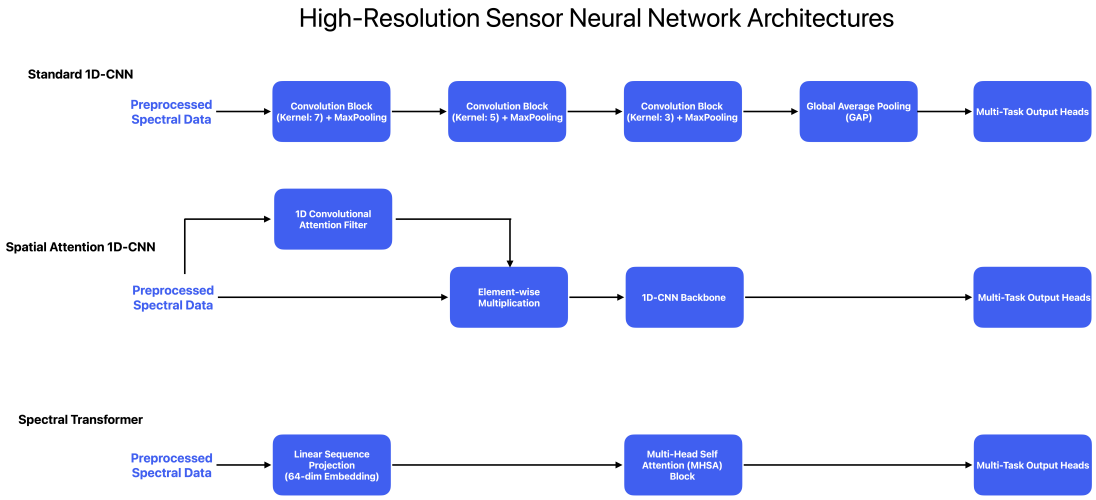


Figure 3.12: Structural overview of the three high-resolution deep learning architectures evaluated: (Top) the standard 1D-CNN baseline utilizing Global Average Pooling, (Middle) the Spatial Attention 1D-CNN featuring dynamic wavelength suppression, and (Bottom) the Spectral Transformer utilizing 64-dimensional embeddings and Multi-Head Self Attention.

The primary model as a Standard 1D-Convolutional Baseline consisting of a three-stage convolutional hierarchy (kernel sizes 7, 5, and 3). This backbone was aggressively downsampled via MaxPooling and evaluated using Global Average Pooling (GAP) to minimize parameter count while maintaining translation invariance. Building upon this, a Spatial Attention 1D-CNN was developed. In this configuration, a 1D-convolutional filter generated a directly against the input spectrum, enabling the network to dynamically suppress uninformative wavelengths. Finally, a Spectral Transformer was implemented to evaluate sequence modeling. This archi-

tecture utilized a 1D Vision Transformer that linearly projected the sequence into a 64-dimensional embedding space, processing it through a Multi-Head Self-Attention (MHSA) block to capture global optical dependencies.

3.5.3 Deep Learning Architectures for the Low-Resolution Sensor

Given the extreme spatial constraints of the 14-channel sensor, an incremental architectural ablation study was executed to systematically build and validate an optimal embedded network. The evolution proceeded through four distinct structural steps.

Initially, a Naive Baseline was established using a standard 1D-CNN that processed the uncropped 14-channel input array. This single-branch architecture relied entirely on the network's spatial inductive bias without external feature injection and evaluated the full 9×9 classification matrix. This 9×9 configuration represented the absolute theoretical limit of the dataset, encompassing all discrete experimental labels: one blank control alongside four distinct severity tiers each for CpI and PpIX. To explicitly integrate physical optical constraints without corrupting the CNNs spatial sequence, the second step introduced Domain Knowledge Injection via a dual-branch architecture. The primary convolutional branch processed the 14-channel array, while a secondary dense branch processed the seven dynamically engineered physical features and ratios, concatenating the latent representations prior to the output heads.

In the third step, Advanced Architecture and Explainable AI were integrated by engineering a "Smart Router" and a Dynamic Reliability Switch (DRS), as visualized in Figure 3.13. To prevent the massive 405 nm excitation backscatter from blinding the CNN, the router physically cropped the first two hardware bins from the convolutional input, while continuing to compute the seven physical ratios from the full array. A global Channel Attention mechanism (the DRS) was prepended to the CNN to dynamically mute unreliable hardware bins. This architecture was strictly evaluated on the unmerged 9×9 matrix to securely isolate architectural performance gains from any clinical label adjustments.
In the final step, representing Clinical Pragmatism, the fully optimized DRS architecture was trained and evaluated on the reconfigured 5×5 Clinical Triage matrix. This simplified matrix was generated by restructuring the raw labels to reflect actionable clinical goals: consolidating the blank controls and non-actionable trace infections into a single "Negative / Trace" baseline, and grouping the remaining active concentrations into broader severity tiers. This final step successfully reconciled the mathematical boundaries of the AI with the physical limitations of the edge hardware.

3.5.4 Interpretability Baseline: Decision Tree Classification

To establish an absolute computational floor for embedded deployment and to validate the sufficiency of the engineered physics features, a minimalist machine learning

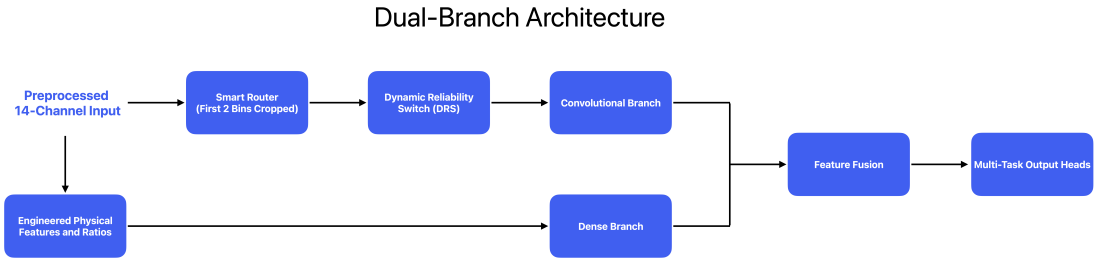


Figure 3.13: Architecture of the optimized dual-branch network for the low-resolution sensor. The model dynamically routes preprocessed data to extract physical ratios while simultaneously cropping excitation bins prior to convolutional processing. The branches are fused before passing into the shared multi-task classification heads.

baseline was developed utilizing a standard Decision Tree (DT) classifier [25]. This model was primarily intended to test the hypothesis that high-dimensional spatial pattern recognition, such as that performed by CNNs, may be redundant if the underlying biological signatures are already effectively captured by lower-dimensional engineered ratios. Consequently, the DT was explicitly denied access to the raw 14-channel spectral array and was instead trained exclusively on the seven dynamically engineered features defined in Section 3.2.8, comprising four optical ratios and three macro-geometric slopes. Prior to training, these feature vectors underwent standard scaling to ensure a normalized baseline for comparative analysis.

The implementation utilized the Classification and Regression Trees (CART) algorithm, which constructs the decision model by mathematically evaluating how well each engineered feature separates the data into distinct diagnostic categories [26]. At each branching point (or node), the algorithm calculates a metric known as Gini impurity. A statistical measure of how often a randomly chosen sample would be misclassified if it were labeled randomly according to the distribution in that subset. The algorithm then selects the specific feature threshold that maximizes the "purity" of the resulting sub-groups, effectively isolating distinct pathogen signatures at each consecutive branch [25]. To maintain diagnostic depth while rigorously mitigating the risk of overfitting on specific hardware noise profiles, the tree's expansion was explicitly constrained to a maximum depth of eight layers, with a minimum requirement of five samples per terminal leaf node. Furthermore, algorithmic class weighting was applied during training to account for the class distribution shift induced by the clinical target reconfiguration. By consolidating the blank controls (C0) and trace infections (C4) into a single "Negative / Trace" category within the 5×5 matrix, this non-actionable class became significantly more populous than the specific active infection stages. Class weighting was therefore utilized to mathematically penalize misclassifications of the minority classes more heavily, ensuring the model maintained high diagnostic sensitivity for the specific pathogen categories despite the dominance of the consolidated negative baseline.

Beyond the reduction in computational overhead, the inherent transparency of the DT facilitated a rigorous feature importance analysis. By tracking the total reduc-

tion in Gini impurity contributed by each specific feature across the entire tree, the relative diagnostic weight of each engineered ratio could be mathematically quantified [26]. This provided a robust, interpretable method to verify that the classification logic aligned with established biological porphyrin emission characteristics rather than stochastic sensor noise, fulfilling critical regulatory requirements for algorithmic explainability.

3.6 Embedded Edge Optimization and Deployment

Following the baseline evaluations, the highest-performing models from both the high-resolution and low-resolution pipelines were selected for embedded optimization. This process focused on adapting the neural networks to the strict computational, memory, and power limitations of embedded microcontrollers.

3.6.1 Architectural Pruning and Network Scaling

To systematically identify the minimal parameter footprint required to maintain diagnostic accuracy, a series of scaled-down architectural variants were engineered. These variants were created by applying both fractional pruning (reducing the width of the network layers) and topological depth reduction (removing entire network layers).

For both the high-resolution and low-resolution models, four specific structural variants were evaluated:

- **Base:** The original, unconstrained architecture utilized during the baseline evaluations, serving as the upper performance benchmark.
- **Lite:** A fractional scaling variant where the channel depths of all convolutional filters, attention bottlenecks, and dense layers were halved. For example, the high-resolution convolutional filter progression was reduced from 64-128-128 to 32-64-64.
- **Mini:** A more aggressive fractional scaling variant where the base channel depths were reduced by a factor of four. This variant was designed to test the absolute minimum informational width the networks required before suffering feature collapse.
- **Micro:** A topological depth-reduction variant designed to minimize latency. Instead of simply narrowing the network, the depth of the convolutional backbone was fundamentally altered. For the high-resolution models, the three-stage convolutional hierarchy was reduced to two stages. For the low-resolution models, the convolutional backbone was reduced to a single *Conv1D* layer, and all *MaxPooling* operations were entirely removed to prevent aggressive spatial decimation on the already constrained input.

3.6.2 Weight Quantization Strategies

Standard neural networks operate using 32-bit Floating-point (FP32) arithmetic, which is computationally expensive and memory-intensive for embedded microcontrollers. To adapt the pruned models to the severe hardware constraints typical of edge devices, the network weights and activations required conversion to INT8. This precision reduction effectively compressed the mathematical size of the models by a factor of four. Crucially, this compression was essential not only for accelerating inference latency and reducing active Random Access Memory (RAM) consumption, but also for minimizing the non-volatile storage footprint required to persist the serialized model weights on the microcontroller's limited flash memory. To achieve this translation and determine the optimal deployment pipeline, two distinct quantization methodologies were evaluated and compared.

The first methodology evaluated was PTQ [27]. Under this standard approach, the neural networks were fully trained to convergence utilizing FP32 precision. Following training, a representative subset of the spectral dataset was passed through the model to calibrate the internal activation ranges, and the continuous weights were subsequently rounded to static INT8 values. While PTQ is computationally inexpensive to implement, this rigid post-training rounding process inherently introduces quantization noise [27]. On highly sensitive spectral data, this mathematically induced noise can severely degrade the network's learned classification boundaries.

To mitigate this potential performance degradation, QAT was employed as the secondary, advanced optimization methodology [28]. Unlike PTQ, QAT simulates the informational loss of low-precision arithmetic directly within the training loop. By injecting "fake quantization" nodes into the computation graph, the forward pass of the network simulated discrete INT8 mathematical bounds, while the backward pass and gradient updates were maintained in high-precision FP32 utilizing a straight-through estimator [28, 29]. This simulated noise forced the model to actively adapt its weights to the discrete INT8 grid during the optimization process itself. This ensures that the final quantized models retain high morphological fidelity to the original spectral signals while successfully satisfying the storage, memory, and latency constraints of edge hardware.

While PTQ and QAT yield varying diagnostic accuracies due to their distinct optimization processes, both methodologies ultimately compile down to identical network architectures for deployment. Because both utilize the exact same INT8 data structures and target computational graphs, the choice between PTQ and QAT has no impact on physical memory footprint or execution time. This structural equivalence means that any empirical hardware metrics gathered during the benchmarking phase would apply uniformly to a model's architecture, regardless of the specific quantization technique utilized to generate it.

3.6.3 Inference Runtime

The optimized INT8 models were deployed using TensorFlow Lite for Microcontrollers (TFLM), a lightweight machine learning framework designed specifically for executing inference on bare-metal systems [30]. As the core execution engine, TFLM is responsible for loading the serialized FlatBuffer model, instantiating the computational graph, and orchestrating the sequential flow of data through the network layers. During runtime, it maps the preprocessed spectral arrays to the input tensors, triggers the corresponding mathematical operations layer by layer, and extracts the final classification predictions.
To integrate this execution pipeline into the target hardware environment, the *esp-tflite-micro* component was utilized [31]. Rather than altering the core interpreter logic of the machine learning framework, this vendor-specific port serves as an interface between the base TensorFlow distribution and the Espressif IoT Developement Framework (ESP-IDF). It functions as a build-system wrapper that intercepts TFLMs standard operational calls, directing the neural network computations away from generic *C++* reference implementations and toward the *ESP-NN* kernel library. The *ESP-NN* library provides hand-optimized assembly routines tailored for the *Xtensa LX7* architecture found in the *ESP32-S3*, covering operations such as *Conv1D*, *DepthwiseConv*, and *Softmax* [32]. These kernels are designed to utilize the processor's Single Instruction, Multiple Data (Single Instruction Multiple Data (SIMD)) vector extensions, enabling the parallel execution of multiple INT8 Multiply-Accumulate (MAC) operations at once. This integration of hardware-specific kernels reduced the inference latency and computational energy footprint of the quantized models on the embedded device.

3.7 Emebedded performance evaluation

To then evaluate the computational suitability of the developed models for edge environments, we utilized a Processior-in-the-Loop (PIL) methodology where the optimized models where deployed to representative hardware and benchmarked. This phase was not intended to validate a final commercial hardware design, but rather to benchmark the intrinsic inference characteristics (latency, energy, and memory usage) of the neural networks on a representative low-power architecture.
To ensure relevance to resource-constrained edge devices, this evaluation is strictly limited to the quantized INT8 models. The FP32 baseline models are excluded from this phase and will be evaluated solely on workstation hardware.

3.7.1 Hardware platform

The hardware chosen to be our benchmarking target was a *ESP32-S3-WROOM-1-N16R8* module. This integrated module features the *ESP32-S3* System-on-Chip (SoC), a dual-core *Xtensa LX7* Central Processing Unit (CPU) running at 240 MHz, with a hierarchical memory architecture specifically structured as follows [33]:

- **512 KB Internal Static Random Access Memory (SRAM):** The primary high-speed on-chip memory pool. During the initial boot sequence, the

hardware bootloader and ESP-IDF startup routines partition this physical SRAM into three distinct functional regions:
  - *L1 Cache (up to 96 KB):* A reserved segment managed by the Memory Management Unit (MMU). It is dynamically divided into an Instruction Cache (up to 32 KB) and a Data Cache (up to 64 KB) to transparently accelerate execution and data retrieval from the slower external buses.
  - *Instruction Random Access Memory (IRAM):* A dedicated segment accessed via the CPUs instruction bus, utilized to execute time-critical firmware routines and hardware interrupts with zero wait-states.
  - *Data Random Access Memory (DRAM) and Application Heap:* The remaining SRAM accessed via the CPUs data bus. After provisioning static variables and FreeRTOS task stacks, the residual DRAM forms the primary application heap. This specific region offers the lowest data access latency and is prioritized for high-frequency dynamic memory allocations.
- **8 MB Octal Serial Peripheral Interface (SPI) Pseudo-Static Random Access Memory (PSRAM):** A volatile SPI-attached memory utilized to expand the system heap. Though integrated into the SoC package via a System-in-Package (SiP) design, it relies on an Octal SPI interface, providing extra capacity for large data structures albeit at a higher access latency than internal SRAM.
- **16 MB Quad SPI Flash:** Non-volatile, module-level SPI storage utilized to persist the application firmware and static read-only data payloads across power cycles.

3.7.2 Memory placement strategy

In the context of the TFLM runtime, the memory footprint of a deployed neural network is fundamentally divided into two distinct components. The first component is the serialized FlatBuffer, a static, read-only payload containing the structural schema and the trained network weights [30, 34]. The second component is the tensor arena, a dynamically allocated memory pool utilized exclusively during runtime to store intermediate layer activations, computation buffers, and the input/output tensors. Because all evaluated INT8 models successfully fit entirely within the available internal SRAM, it was possible to conduct a full-factorial benchmark of the memory hierarchy.

To isolate and quantify the specific latency and energy penalties introduced by the external buses, all six possible combinations of weight and arena placement were explicitly tested for every optimized architecture:

- **Static Weight Placement:** The read-only FlatBuffer was evaluated in three distinct locations: mapped directly from non-volatile Flash memory, loaded into the external PSRAM at boot, or copied entirely into the internal high-speed SRAM.
- **Dynamic Tensor Arena Allocation:** The read/write computation pool was evaluated in two locations: allocated within the external PSRAM or allocated strictly within the internal SRAM.

Benchmarking the models with both components strictly inside the internal Static

Random Access Memory (SRAM) established the absolute peak efficiency baseline. Conversely, offloading the weights or the tensor arena to Flash and Pseudo-Static Random Access Memory (PSRAM) provided a direct, empirical measurement of the processor's wait-state penalty when retrieving data over the slower SPI interfaces.

3.7.3 Latency and throughput measurement

To measure the real-time performance of the models, we will utilize the *ESP32-S3*'s internal high-resolution hardware timer (*esp_timer_get_time()*), which offers microsecond-level precision [33]. This will be used to measure the inference latency (t_{inf}) which is defined as the wall-clock time required to execute a single forward pass. Timestamps are captured strictly before and after the inference call, excluding data loading and pre-processing overhead.

3.7.4 Power consumption

To precisely quantify the energy consumption of the embedded neural networks, a Nordic Semiconductor Power Profiler Kit II (PPK2) was utilized. This instrument features an advanced analog measurement unit with a high dynamic range (from single μA to 1 A) and an integrated 8-channel logic analyzer [35]. Its auto-ranging capability allows measurement resolution to scale dynamically (from 100 nA to 1 mA), ensuring both micro-ampere idle currents and milli-ampere computational peaks are captured with optimal fidelity.

Figure 3.14: Bottom view of the modified ESP32-S3-DevKitC-1 clone board used. A trace on the lower right (marked by a red arrow) has been severed to isolate the main 3.3V Voltage at Drain (VDD) rail (which powers the ESP32 and external General Purpose Input Output (GPIO) pins) from the onboard 5V-to-3.3V Low-Dropout Regulator (LDO) and all other onboard 3.3V consumers.

To ensure accurate chip-level power profiling, a physical hardware modification was performed on the ESP32-S3 development board. Specifically, a Printed Circuit Board (PCB) trace connecting the onboard 5V-to-3.3V LDO to the ESP32-S3's VDD input pins was mechanically severed as seen in Figure 3.14). This split the development board into two independent power domains. When connected to a host machine via Universal Serial Bus (USB), the onboard LDO continued to safely power peripheral components-such as the USB-to-Universal Asynchronous Receiver-Transmitter (UART) bridge and status LEDs allowing for uninterrupted serial logging. Simultaneously, the ESP32-S3 chip itself was powered exclusively by the PPK2 via the external 3.3V header pins. This ensured that the quiescent current of the LDO and the active draw of the serial communication chips were completely excluded from the recorded neural network power measurements. While the physical USB data lines remained connected to the SoC, potentially introducing minute reverse-leakage currents through the host machine's pull-up resistors, this artifact was deemed mathematically negligible. Such data-line leakage typically operates in the low micro-ampere (μA) range, representing an insignificant margin of error when integrating the tens of milliamperes (mA) of active computational current recorded during inference.

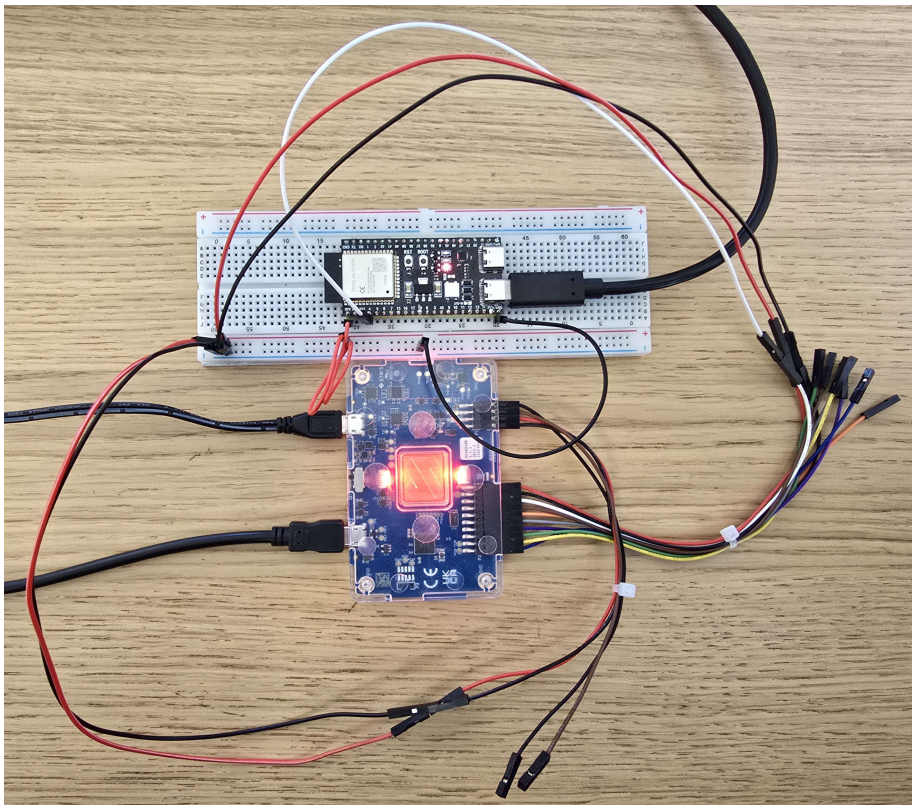


Figure 3.15: Experimental setup for power profiling the modified ESP32-S3 using a Nordic nRF PPK2. The PPK2 supplies power to the isolated 3.3V rail of the ESP32 while simultaneously logging current consumption. An ESP32 GPIO pin is connected to the PPK2's logic inputs for measurement synchronization. To support this, the logic reference voltage is backfed from the PPK2's power output.

The modified ESP32-S3 was conncted to the PPK2 via a breadboard as seen in Figure 3.15. The PPK2 was configured in "Source Meter" mode, supplying a stable 3.3 V directly to the isolated external VDD pins of the microcontroller. Because the PPK2's logic analyzer requires a reference voltage to establish logic-high and logic-low thresholds, the 3.3 V power output was routed across the breadboard and looped back into the logic reference input of the PPK2. While the external development board regulator was bypassed, the ESP32-S3's internal power domains and regulators remained fully active, ensuring the measurements accurately reflected the true energy cost of the silicon.
To achieve exact temporal alignment between the neural network execution and the analog power measurement, a hardware synchronization strategy was implemented. A dedicated GPIO pin on the ESP32-S3 was toggled high immediately before the TFLM inference call and driven low strictly upon its completion, with this signal wired directly to a PPK2 digital input. Data acquisition was managed using the companion software, *nRF Connect for Desktop* (running the Power Profiler application), configured in "scope mode." . The digital input was utilized as a hardware trigger, prompting the software to initiate a capture with a pre-configured negative time offset (to record the pre-event baseline) and a total sampling duration manually adjusted to enshure the entire inference event was captured. During data extraction within the software interface, a measurement window was manually delineated to match the exact duration the digital input remained high. The software automatically integrated the measured current strictly over this user-defined active window to output the total charge transferred (in Coulombs). The exact energy per inference (in Joules) was subsequently calculated by multiplying this calculated charge by the stable 3.3 V supply voltage.

4

Results

This chapter presents the experimental findings for the spectral classification of CpI and PpIX. The evaluation follows an ablation study format, beginning with a high-resolution laboratory baseline and progressing through various algorithmic optimizations for low-resolution, cost-effective hardware.

4.1 Evaluation of the Deteministic Baseline: The Limits of Signal Processing

The deterministic MATLAB models (described in Section 3.4) were evaluated to establish a performance baseline for 1D threshold-based classification prior to the introduction of machine learning. The algorithm was evaluated through two distinct iterations to document its performance across different parameter tunings.

4.1.1 Results of the Standard Baseline

The initial standard deterministic approach utilized median-based thresholds optimized for average variance between classes. As shown in Figure 4.1, this model achieved a global accuracy of 81.06%. Analysis of the confusion matrix identified three primary areas of misclassification. First, within the blank Control group, 139 out of 320 samples were misclassified as trace infections. Second, regarding high-concentration classifications, 47 out of 160 (29.375%) PpIX C3 samples were incorrectly upgraded to the C3 category. Third, a significant proportion (66.25%) of CpI trace infection were misclassified as control samples.

4.1.2 Results of the Upgraded Baseline

The final deterministic baseline incorporated linear detrending for the Control threshold (2.5 Sigma), species-specific AUC integration windows, and geometric means for concentration boundaries. As shown in Figure 4.2, this model achieved a peak of accuracy of 96.69%. The algorithm almost perfectly separated the Control samples from the trace (C4) infections and resolved the majority of the concentration bleeding. The remaining misclassifications were strictly localized to 21 samples of PpIX C3 (upgraded to C2).

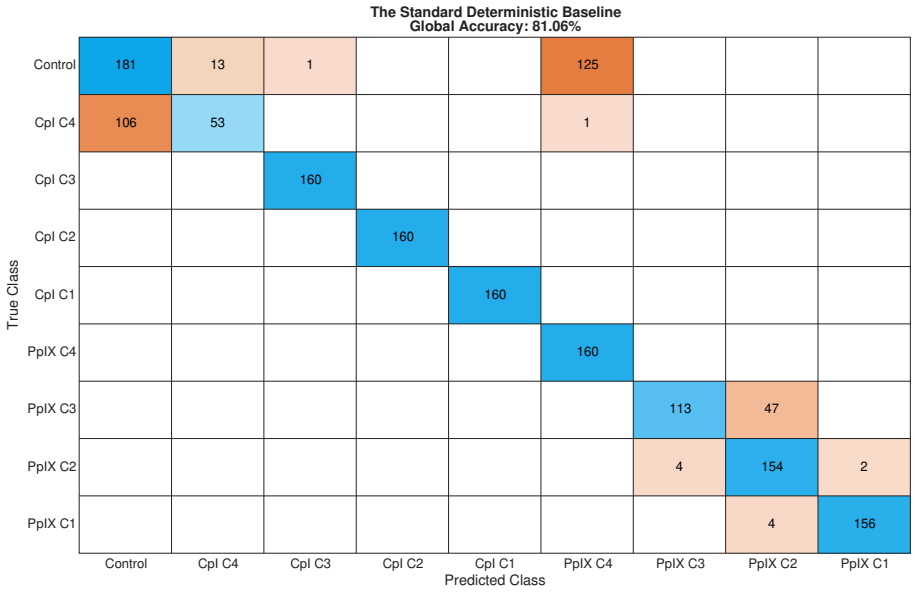


Figure 4.1: Confusion matrix for the standard deterministic baseline utilizing median-based thresholds, achieving an overall accuracy of 81.06%.

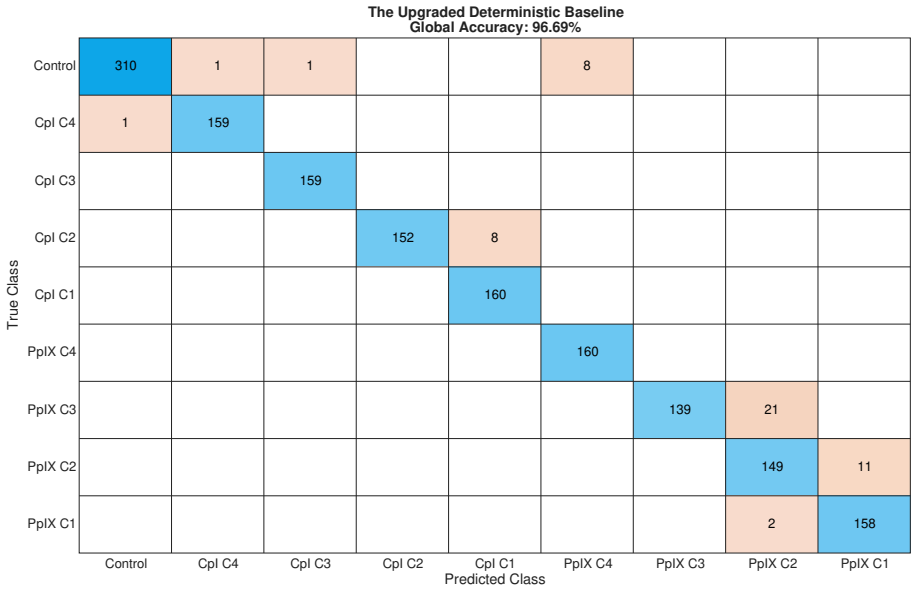


Figure 4.2: Confusion matrix for the upgraded baseline utilizing detrended peak isolation and geometric mean boundaries, achieving a final accuracy of 96.69%.

4.2 High-Resolution Performance Baseline

To establish a high-fidelity performance ceiling and verify the detectability of bacterial fluorophores, a 288-channel high-resolution spectrometer was utilized. The objective of this baseline was to prove that machine learning could overcome the deterministic limitations identified in Section 3.4 (such as noise overlap and concentration-based quenching) when provided with rich spectral data.

4.2.1 Deep Learning Preprocessing Strategies

To evaluate the baseline 1CH 1D-CNN architecture's response to different input scaling methods, the models were trained and evaluated across three distinct preprocessing pipelines: Source Normalization, Wing Normalization, and the SNV. The global accuracy outcomes of this ablation are presented in Figure 4.3.

The evaluation revealed a significant variance in classification performance depending on the applied normalization technique. Source Normalization yielded the lowest performance, achieving a global classification accuracy of 83%. In contrast, Wing Normalization resulted in a substantial improvement, increasing the global accuracy to 97%. The highest classification performance was observed when applying the SNV transformation, which achieved a global accuracy of 99%.

Based on these objective performance metrics, Source Normalization was excluded from further experiments. Subsequent evaluations regarding the impact of multi-channel feature engineering proceed exclusively utilizing the SNV and Wing Normalization pipelines.

4.2.2 Feature Engineering and Architectural Attention

Following the preprocessing evaluation, the impact of multi-channel feature engineering was assessed. The 1D-CNN models were trained using progressively complex input tensors: a single-channel raw spectrum (CH1), a dual-channel input appending the first derivative of the spectrum (CH2), and a tri-channel input appending both the first and second derivatives (CH3). This progression was evaluated for both SNV and Wing Normalization pipelines. Furthermore, to assess architectural robustness, these configurations were tested both with and without the integration of a global Channel Attention mechanism. The complete performance trajectories are visualized in Figure 4.4.

For standard 1D-CNN architectures lacking an attention mechanism, the SNV pipeline demonstrated highly stable performance, yielding global accuracies of 99% for CH1, 98% for CH2, and 98% for CH3. In contrast, the Wing Normalization pipeline exhibited greater performance variance as input complexity increases; it achieved 97% accuracy on CH1, peaked at 98% on CH2, and subsequently dropped to 95% upon introducing of the second derivative in CH3.

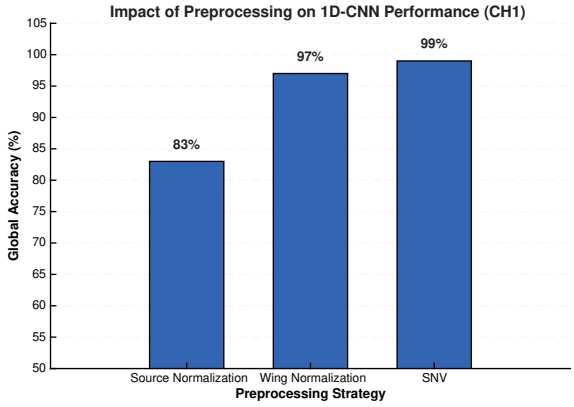


Figure 4.3: Global classification accuracy of the baseline 1-channel (CH1) 1D-CNN architecture across three distinct preprocessing pipelines. Performance is benchmarked using Source Normalization (83%), Wing Normalization (97%), and the SNV transformation (99%). Data labels represent the macroscopic accuracy achieved on the independent test set.

The introduction of the Channel Attention mechanism fundamentally altered the performance trajectories for specific pipelines. With attention enabled, the SNV pipeline maintained a constant global accuracy of 98% across all three channel configurations. Conversely, the Wing Normalization pipeline with attention recorded its lowest global accuracy of 90% on the single channel input (CH1), but exhibited a sharp increase in performance on the multi-channel inputs, achieving a peak accuracy of 99% on CH2 and 98% on CH3. While the global accuracy trajectories demonstrate the macroscopic impact of feature engineering, examining the absolute classification boundaries remains critical. To provide deeper insight into the exact class-wise error distribution of these standard convolutional architectures, the full confusion matrices for both the optimal and sub-optimal 1D-CNN configurations are included in Appendix A for comprehensive reference.

4.2.3 Spectral Transformer Performance

To evaluate the efficacy of global self-attention mechanisms on the spectral dataset, a Spectral Transformer architecture was implemented and subjected to the same multi-channel feature ablation (CH1, CH2, CH3). The model was evaluated across both the SNV and Wing Normalization preprocessing pipelines. The global accuracy outcomes for these configurations are presented in Figure 4.5.

For single-channel raw spectral inputs (CH1), the Transformer exhibited its lowest classification performance. The SNV pipeline achieved a global accuracy of 79%, while the Wing Normalization pipeline reached 91%.

Upon the introduction of the first derivative in the dual-channel configuration (CH2), both pipelines registered a substantial increase accuracy. The SNV pipeline improved by 18 percentage points to reach 97%, and the Wing Normalization pipeline peaked at a global accuracy of 98%.

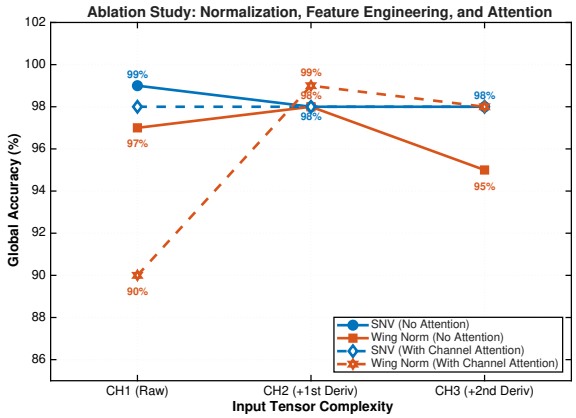


Figure 4.4: Global classification accuracy across varying input channel complexities (CH1, CH2, CH3). The graph plots the performance of both SNV and Wing Normalization pipelines, contrasted across standard architectures (solid lines) and architectures augmented with a Channel Attention mechanism (dashed lines).

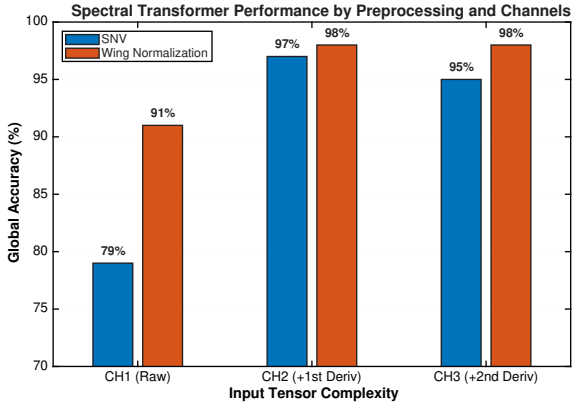


Figure 4.5: Global classification accuracy of the Spectral Transformer architecture across increasing input channel complexities (CH1, CH2, CH3). Performance is contrasted between the Wing Normalization (orange) and Standard Normal Variate (blue) preprocessing pipelines.

The addition of the second derivative in the tri-channel configuration (CH3) did not yield further performance gains. The Wing Normalization pipeline maintained its peak accuracy of 98%, whereas the SNV pipeline experienced a slight performance degradation, decreasing to 95%.

Detailed confusion matrices representing the exact class-wise distribution, alongside the internal attention weight maps for both the lowest-performing (SNV CH1) and highest-performing (Wing Normalization CH2) Transformer configurations, are provided in Appendix B for comprehensive reference.

4.3 Low-Resolution Model Optimization

To evaluate the diagnostic viability of the 14-channel low-resolution sensor, an incremental ablation study was conducted. The objective of this evaluation was to establish the hardware's baseline performance on a complex 9-class classification task, and subsequently measure the impact of domain-specific feature engineering, architectural attention mechanisms, and clinical target realignment. The models were evaluated across three distinct preprocessing pipelines: Backscatter, PassThrough, and VisNormalizer. The complete global accuracy trajectory across all optimization steps is visualized in Figure 4.8.

4.3.1 Baseline Evaluation and Feature Augmentation

The baseline evaluation utilized a standard 1D-CNN architecture processing the raw 14-channel spectral input without explicit feature engineering or optical bin cropping. Across the three preprocessing pipelines, the raw baseline model achieved a global accuracy of 87% for Backscatter, 86% for PassThrough, and 81% for VisNormalizer.

Analysis of the class-wise performance for the baseline Backscatter configuration, presented in Figure 4.6, revealed two primary zones of misclassification: severe bleeding between the blank Control (C0) and Trace (C4) classes, and overlap between the Moderate (C3) and High (C2) concentration classes.

When evaluated using the feature-augmented parallel architecture (described in Section 3.2.8) global accuracies improved to 89% for Backscatter, 83% for PassThrough, and 89% for VisNormalizer. While the macroscopic accuracy showed minimal variance compared to the baseline, the class-level matrices indicated a resolution of the C0 and C4 noise floor misclassifications, isolating the remaining errors primarily to the C2 and C3 quenching overlap.

4.3.2 Architectural Optimization and Attention Mechanisms

The optimized architecture incorporating the DRS was subsequently evaluated, resulting in global accuracies of 90% for Backscatter, 85% for PassThrough, and 88% for VisNormalizer. The internal attention weights, visualized in Figure 4.7, confirm that the DRS successfully learned to independently adjust trust multipliers across the hardware based on the detected fluorophore species. Specifically, for CpI, the lowest attention weights were observed on channels 3, 13 and 14 (FZ, VIS and FD), while the highest weights were concentrated on bins 8 and 10-12 (FXL, F7, F8 and NIR). In contrast, the weighting strategy for PpIX was significantly more distributed; while the overall attention profile remained flatter, the model maintained a localized emphasis on the broad-band FXL channel (bin 8).

Despite this successful implementation of the attention mechanism and the dynamic suppression of noisy optical bins, the global classification accuracy plateaued at 90%

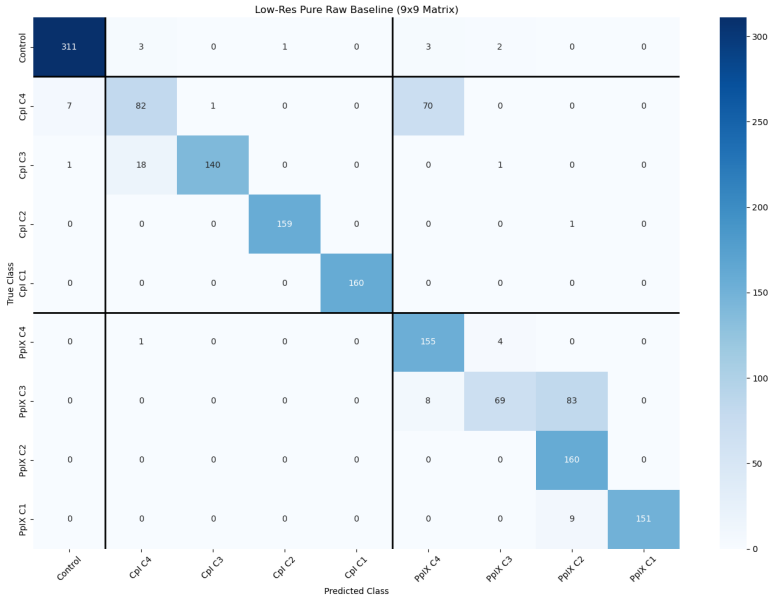


Figure 4.6: Confusion matrix for the baseline 1D-CNN processing the raw 14-channel Backscatter dataset (9×9 unmerged matrix). Significant misclassifications are observed between the Control/Trace boundary (top left) and the C2/C3 quenching boundary (center).

for the Backscatter pipeline. Class-wise evaluation confirmed that the persistent errors remained concentrated at the physical boundary between the C2 and C3 classes.

4.3.3 Clinical Matrix Reconfiguration

As outlined in the final step of the ablation study, the fully optimized DRS architecture was evaluated on the reconfigured 5×5 Clinical Triage matrix. This label realignment resulted in a substantial increase in classification performance across all three data variants. The global accuracy surged to 98% for the Backscatter pipeline, 97% for PassThrough, and 98% for VisNormalizer.

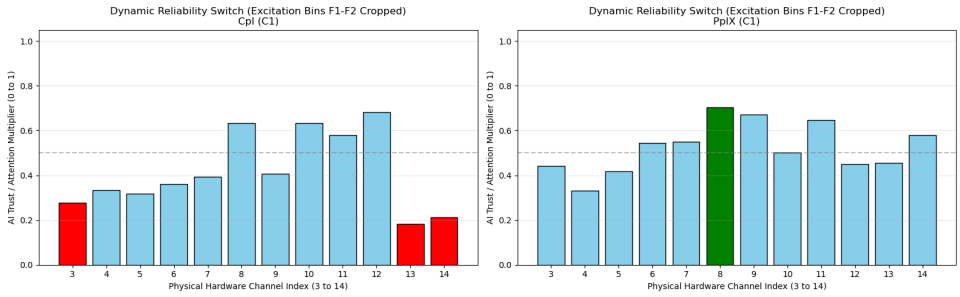


Figure 4.7: Explainable AI (XAI) output demonstrating the Dynamic Reliability Switch (Channel Attention weights) applied to the cropped optical bins. The algorithm dynamically adjusts trust multipliers across specific hardware channels based on the detected fluorophore species.

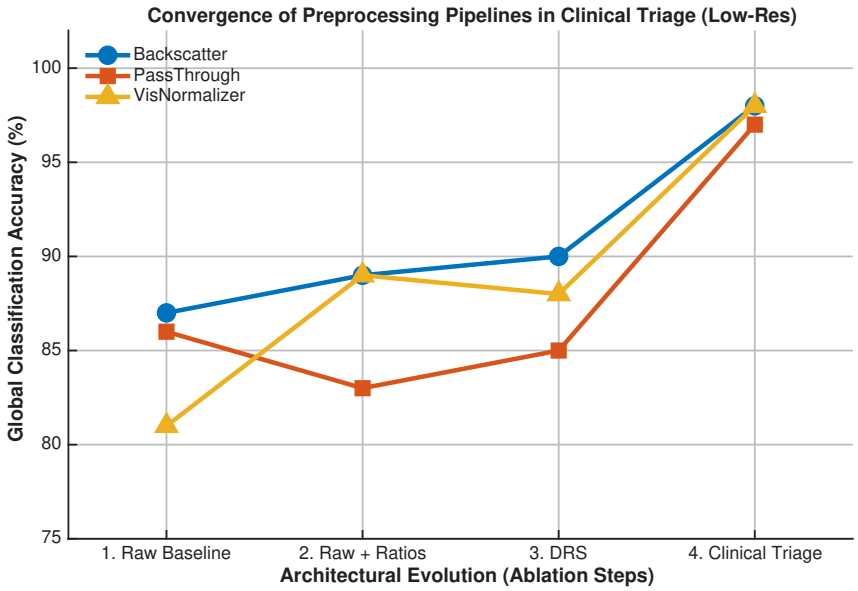


Figure 4.8: Global classification accuracy trajectory across the four distinct optimization steps: Raw Baseline, Feature Augmentation (Ratios), Architectural Optimization (using DRS), and Clinical Triage Alignment. The data demonstrates the eventual convergence of all three preprocessing pipelines.

The resulting 5×5 confusion matrix for the optimal Backscatter pipeline is presented in Figure 4.9. While the Backscatter pipeline is visualized here as the primary representative model, the corresponding class-wise confusion matrices for the PassThrough and VisNormalizer pipelines, spanning both the raw 9×9 baseline and the reconfigured 5×5 clinical triage task, are documented in Appendix C to ensure complete methodological transparency. Furthermore, the data trajectory (Figure 4.8) reveals a convergence among the preprocessing pipelines; once the target classes were aligned with the physical separability of the signals, the system's sensitivity to the initial normalization method was drastically reduced.

Crucially, this simplification of the target classes altered the macroscopic behavior of the network's attention mechanisms. The internal attention weights for the 5×5 architecture, visualized in Figure 4.10, reveal a drastically flattened trust profile compared to the granular 9×9 model. When evaluating the realigned clinical triage targets, the DRS multipliers for both CpI and PpIX hover consistently near the neutral 0.5 baseline. For the CpI target, the network maintains a relatively distributed weighting strategy, electing to strongly suppress only bin 13, which corresponds to the broad-band VIS channel. For the PpIX target, the attention profile is entirely distributed, lacking any extreme prioritization or suppression spikes across the 12-channel sequence.

4.3.4 Algorithmic Simplification and Feature Importance

To evaluate the sufficiency of the engineered features, a Decision Tree was trained as a secondary baseline. The algorithm achieved a global accuracy of 98%, demon-

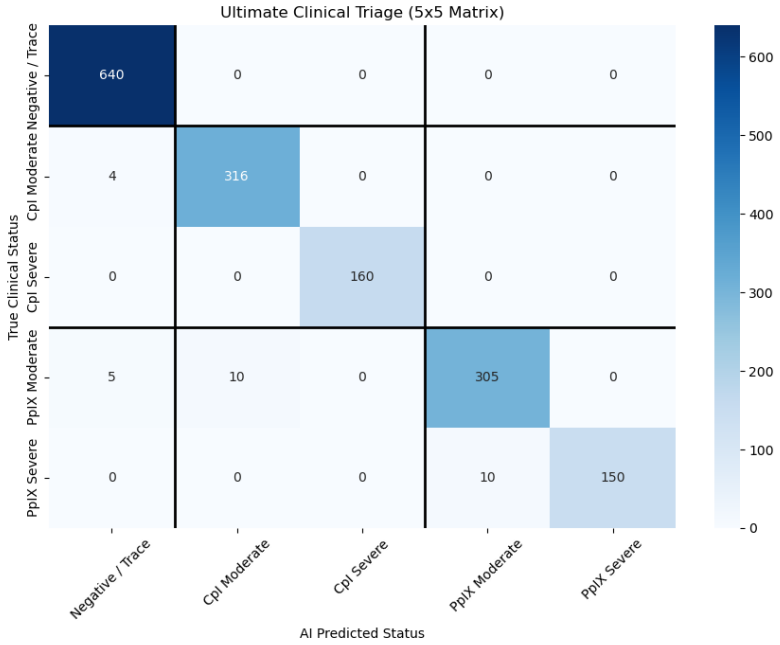


Figure 4.9: Final 5×5 Clinical Triage confusion matrix utilizing the fully optimized DRS architecture and the Backscatter pipeline. Merging the physically constrained classes yielded a global diagnostic accuracy of 98%.

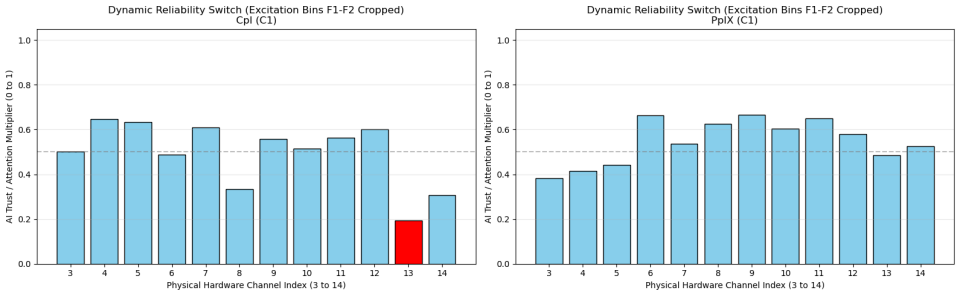


Figure 4.10: XAI output from the DRS architecture optimized for the 5×5 Clinical Triage matrix. Compared to the highly granular 9×9 model, the attention weights exhibit a significantly flatter, more distributed profile.

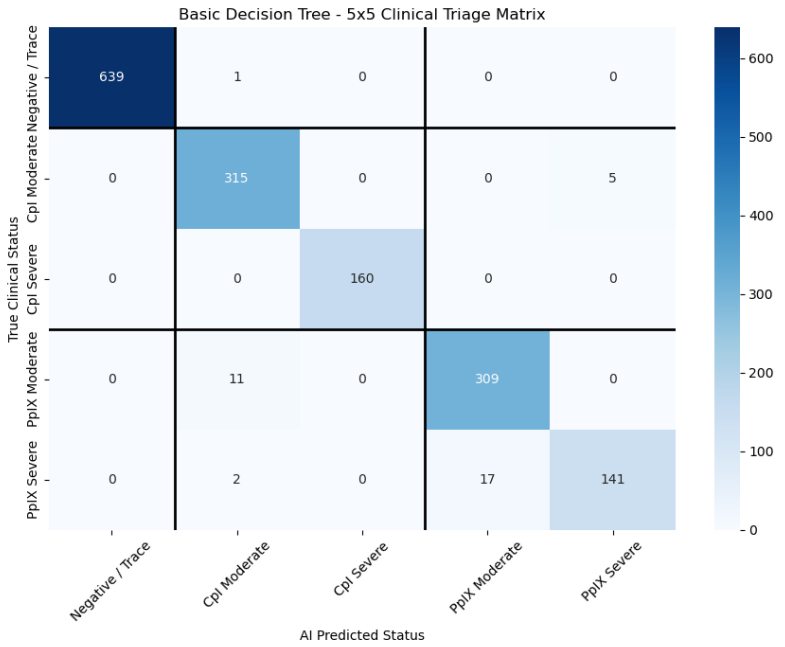


Figure 4.11: Confusion matrix for the basic Decision Tree operating exclusively on the 7 engineered physics features. The algorithm achieves 98% accuracy on the 5×5 Clinical Triage matrix.

strating performance parity with the attention-augmented deep learning models. As visualized in Figure 4.12, Gini feature importance analysis indicated that the classification logic was overwhelmingly driven by two specific engineered variables: the Normalized Deep-Red (F7/XL) and the Normalized Primary Red (F6/XL).

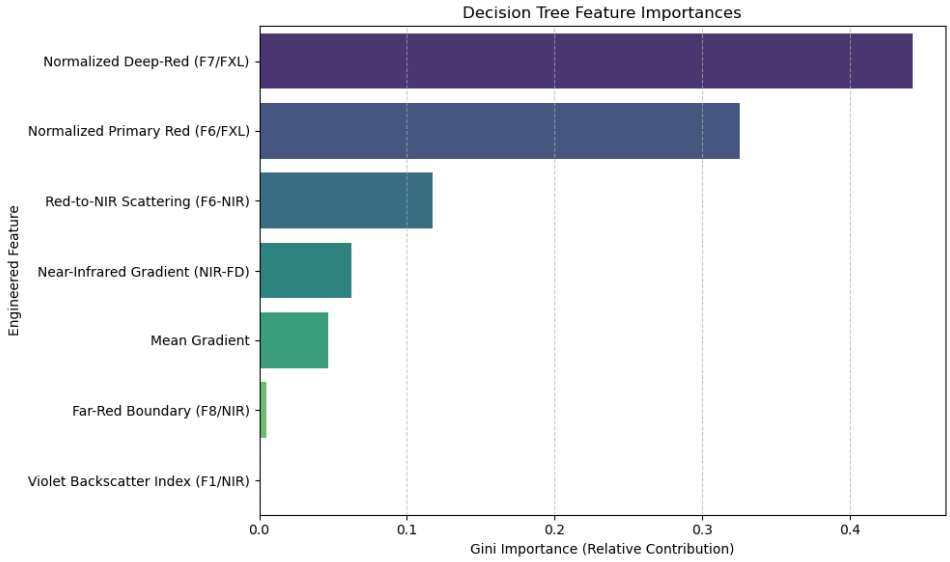


Figure 4.12: Gini feature importance extraction from the Decision Tree classifier. The classification logic is overwhelmingly dominated by the Normalized Deep-Red and the Normalized Primary Red variables.

4.4 Architectural Compression for Edge Deployment

Based on these results four of the best preforming models where chosen to undergo architectural compression to allow them to be evaluated on representative embedded hardware.
As outlined in Section 3.6.1, the network architectures were systematically scaled into four structural variants (*base*, *lite*, *mini*, and *micro*) utilizing variations of fractional width reduction and topological depth reduction. The following subsections detail the performance of these specific variants when subjected to differing input channel dimensions and quantization techniques.

4.4.1 High-Resolution Standalone Model Optimization

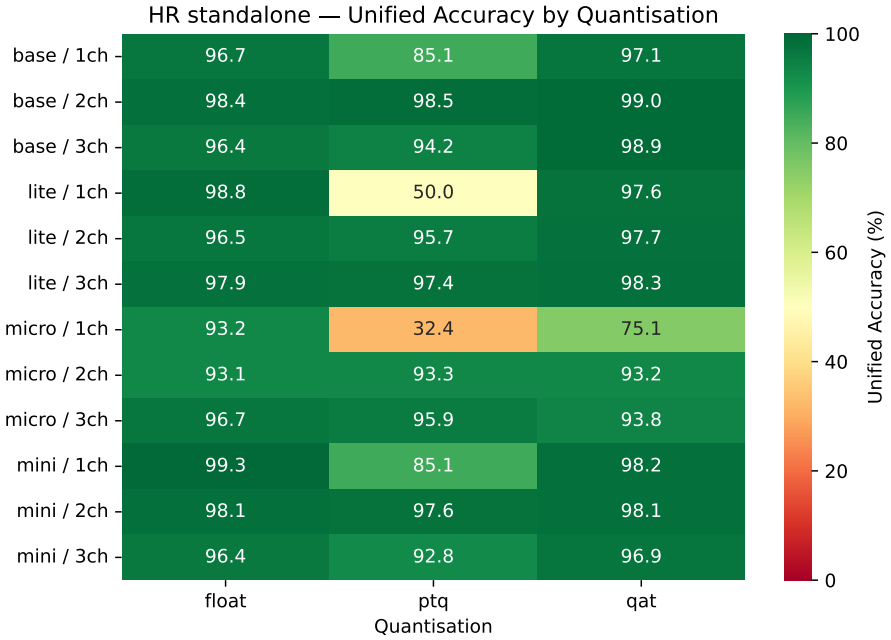


Figure 4.13: Unified accuracy of the High-Resolution Standalone model variants across varying architectural scales, input channel depths, and quantization schemes.

The unified accuracy across all evaluated parameters for the High-Resolution (HR) Standalone model is presented in Figure 4.13. In standard 32-bit floating-point precision, the High-Resolution Standalone architecture maintained its performance during structural pruning.
The smallest *mini* variant, utilizing a single input channel (1ch), achieved a unified accuracy of 99.3% which is higher than all other variants of the standalone model including the baseline.
However, applying INT8 PTQ to the single-channel variants resulted in a decrease in accuracy. The *base* 1ch model's accuracy decreased to 85.1%, the *lite* 1ch model to 50.0%, and the topologically reduced *micro* 1ch model to 32.4%.

Expanding the input tensor to include derivative channels (2ch and 3ch) mitigated this precision loss during the quantization process. For instance, the *lite* variant's accuracy increased from 50.0% (1ch) to 95.7% when provided with two input channels under PTQ. Alternatively, QAT preserved the accuracy of the single-channel models without requiring expanded input tensors, resulting in 97.6% for the *lite* 1ch variant and 98.2% for the *mini* 1ch variant.

4.4.2 High-Resolution Attention Model Optimization

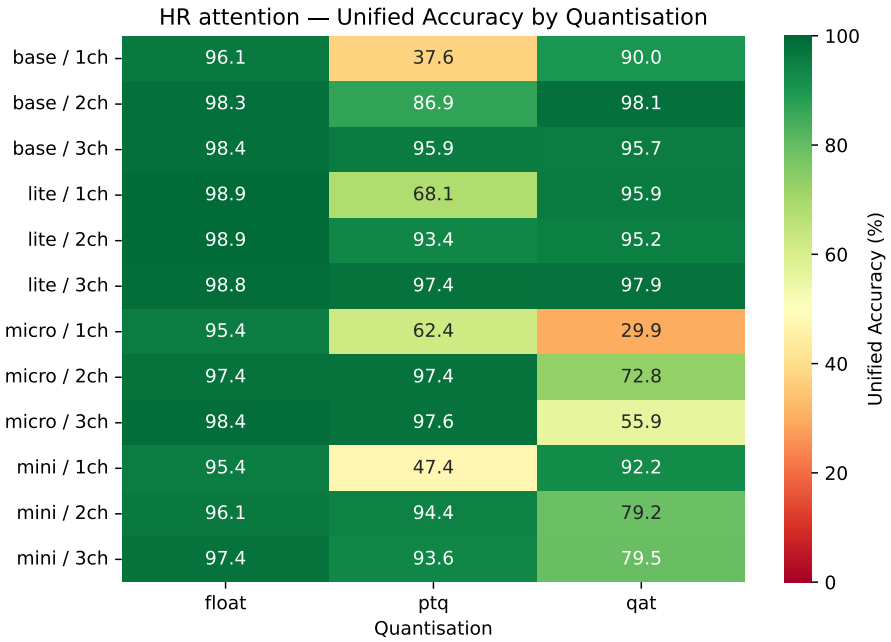


Figure 4.14: Unified accuracy of the High-Resolution Attention model variants across varying architectural scales, input channel depths, and quantization schemes.

The unified accuracy for the HR Attention model, presented in Figure 4.14, exhibited a more pronounced sensitivity to INT8 quantization compared to the Standalone model. In standard floating-point precision, the architecture maintained accuracy across structural pruning. The *lite* variant yielded a higher accuracy than the *base* model in the single-channel configuration, reaching 98.9% compared to the base model's 96.1%.
When applying INT8 PTQ, the single-channel configurations experienced a greater reduction in accuracy than their Standalone counterparts.
Specifically, the accuracy of the 1ch *base* and *mini* models decreased to 37.6% and 47.4%, respectively (compared to 85.1% for both corresponding Standalone variants).
Increasing the input depth to include the derivative channels (2ch and 3ch) improved the performance of the PTQ models, allowing the *lite* PTQ model to increase from 68.1% (1ch) to 97.4% (3ch).
While QAT improved accuracy for the 1ch models (recording 90.0% for the *base* 1ch model and 95.9% for the *lite* 1ch model), it showed convergence instability when

applied to the depth pruned *micro* architecture, yielding accuracies of 29.9% and 55.9% for the 1ch and 3ch configurations, respectively. The data indicates that the *lite* architecture, combined with QAT and a single input channel, provides an equilibrium that maintains 95.9% unified accuracy while reducing tensor dimensions and filter widths.

4.4.3 Low-Resolution DRS Model Optimization

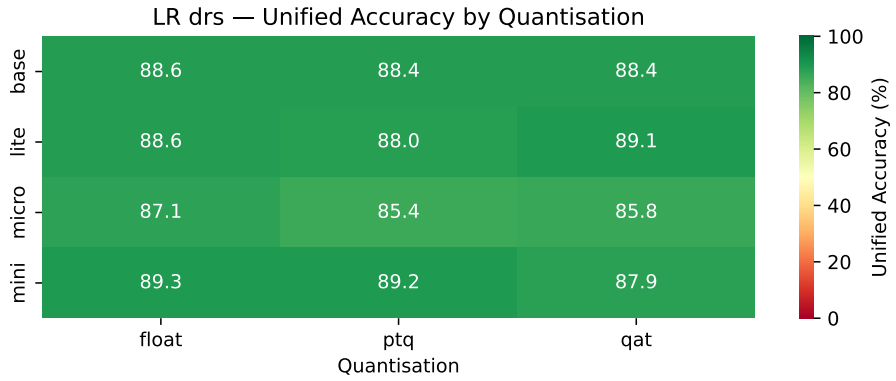


Figure 4.15: Unified accuracy of the Low-Resolution DRS model variants comparing standard floating-point precision against INT8 quantization techniques.

The unified accuracy for the Dynamic Reliability Switch Low-Resolution model are presented in Figure 4.15. In contrast to the high-resolution architectures, the Low-Resolution (LR) DRS models maintained consistent accuracy across INT8 quantization and structural pruning.
Across the four architectural variants (*base*, *lite*, *mini*, and *micro*) and all three precision states (Float, PTQ, QAT), the unified accuracy remained bounded between 85.4% and 89.3%. For example, the floating-point *base* model achieved 88.6%, while the structurally reduced *mini* model utilizing PTQ recorded an 89.2% accuracy.

4.4.4 Low-Resolution Clinical Triage Model Optimization

This stability during quantization was also observed in the Low-Resolution Clinical Triage target as seen in Figure 4.16. The *base* floating-point model achieved a 98.6% accuracy. Applying fractional pruning combined with PTQ yielded minor variations, the *mini* PTQ variant maintained a 97.4% accuracy.
Applying QAT to the models resulted in performance comparable to the float baseline, with the *mini* QAT model reaching 98.2%. These results indicate that the network can be reduced to the *mini* architecture using either PTQ or QAT with minimal variation in diagnostic accuracy compared to the floating-point baseline.

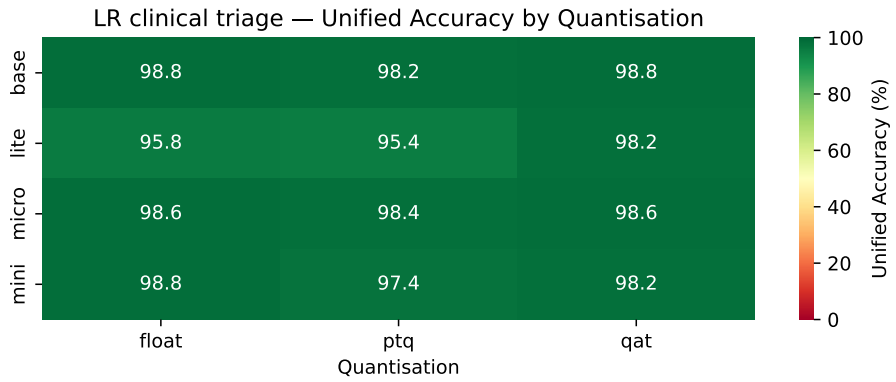


Figure 4.16: Unified accuracy of the Low-Resolution Clinical Triage model variants comparing standard floating-point precision against INT8 quantization techniques.

4.5 Embedded Hardware Performance Evaluation

Following the architectural compression phase, a subset of the finalized INT8 models was deployed to the ESP32-S3 microcontroller to evaluate their physical execution characteristics. This section details the empirical measurements of memory utilization, inference latency, and the specific impact of the microcontroller's memory hierarchy on computational throughput and energy consumption.

4.5.1 Memory Footprint and Hierarchy Penalties

The static memory footprint of the optimized architectures is presented in Figure 4.17. As defined in Section 3.7.2, this total footprint is fundamentally divided into the static weights and the dynamic Tensor Arena. Due to INT8 quantization and structural pruning, the total combined memory requirements for all evaluated models successfully fit within the available internal SRAM of the ESP32-S3. However, by leveraging the microcontroller's Execute-in-Place (XIP) capabilities to map the read-only weights directly from non-volatile Flash, the absolute minimum RAM requirement for deployment is strictly limited to the size of the Tensor Arena. For example, while the largest model (HR Attention *base*) possesses a combined footprint of 194 KB, executing the weights from Flash reduces its RAM consumption to just 82 KB.

The data illustrates an architectural shift in memory distribution as the networks are scaled down. For the uncompressed *base* architectures, the model weights constituted the majority of the footprint (e.g., 110 KB weights vs. 81 KB arena for the HR Standalone *base*). However, applying aggressive topological depth reduction altered this ratio. In the HR Standalone *micro_3ch* variant, the dynamic Tensor Arena (27 KB) exceeded the static weight footprint (17 KB).

To quantify the latency penalties introduced by offloading these components to external SPI buses, a subset of models was evaluated across the full-factorial memory placement grid (Figure 4.18). Allocating both components directly in the internal SRAM (SRAM+SRAM) provided the peak performance baseline.

The data reveals a critical relationship between model size and the efficiency of

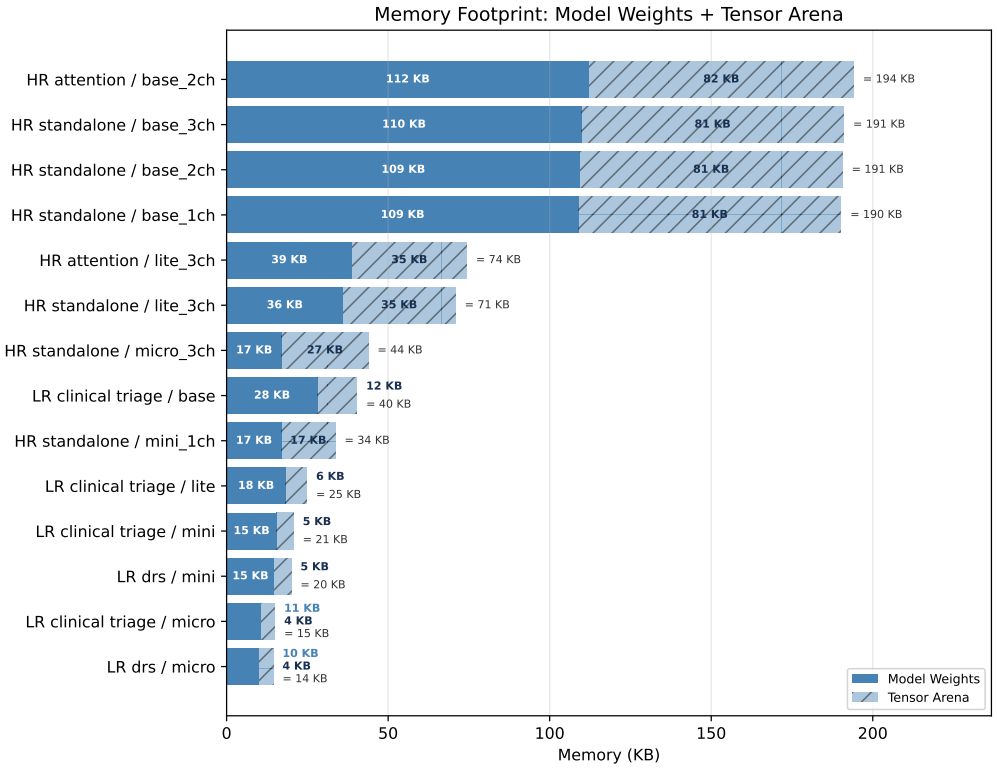


Figure 4.17: Static memory footprint of the optimized models, separated by read-only Weights (FlatBuffer) and dynamic read/write Tensor Arena requirements.

the ESP32-S3's hardware cache. For the large HR Standalone *base_3ch* model, executing weights from Flash to save SRAM (the FLASH+SRAM configuration) resulted in a latency of 211.3 ms, a 2.68x penalty compared to its 78.9 ms baseline in SRAM+SRAM. Moving the Tensor Arena to the external PSRAM (FLASH+PSRAM) caused the latency to surge further to 399.1 ms. Similarly, the HR Attention *base_2ch* model exhibited massive latency spikes when weights were placed in PSRAM (PSRAM+SRAM), jumping to 181.3 ms compared to its 80.6 ms baseline.
Conversely, for the structurally pruned architectures, memory placement had virtually no impact on computational throughput. The HR Attention *lite_3ch* hovered consistently around ≈ 40 ms across all configurations, while the highly compressed HR Standalone *mini_1ch* remained perfectly flat at ≈ 18.5 ms regardless of where its components were allocated.

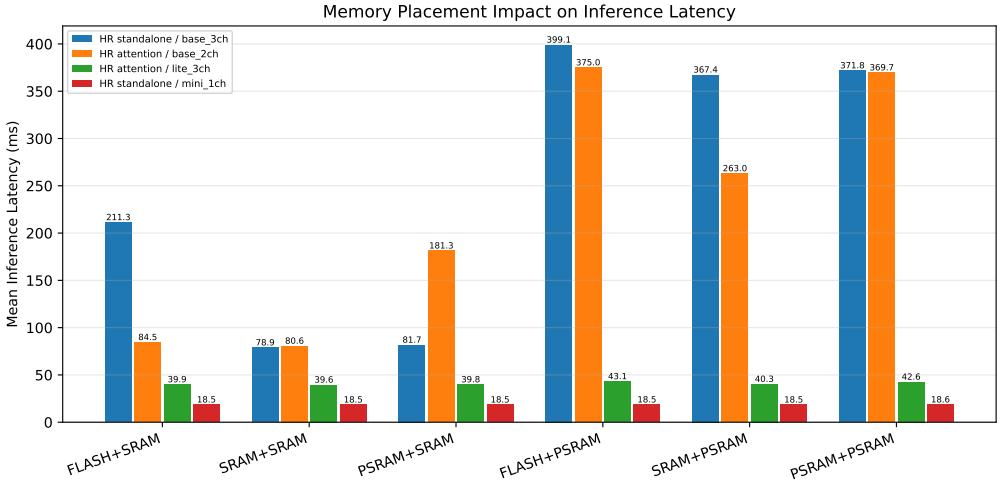


Figure 4.18: Impact of memory placement configurations on mean inference latency. Data is formatted as [Weights Location] + [Tensor Arena Location].

4.5.2 Inference Latency and Pareto Efficiency

The standard execution latency (FLASH+SRAM configuration) for the fully optimized models is detailed in Figure 4.19. The uncompressed HR Standalone *base_3ch* model required 211.3 ms per inference. Applying width reduction to the *mini_1ch* variant reduced this execution time to 18.5 ms. The LR models demonstrated shorter execution times, ranging from 2.7 ms for the *base* variant down to 0.9 ms for the *micro* variants. Additionally, expanding the input from a single raw channel (1CH) to include first and second derivatives (3CH) resulted in a minimal latency increase, for the HR Standalone *base* architecture, the execution time shifted from 211.2 ms to 211.3 ms.

To identify the most efficient deployment architectures, unified diagnostic accuracy was plotted against inference latency to establish a Pareto frontier (Figure 4.20). The dashed Pareto curve highlights the models that provide the maximum achievable accuracy for a given latency budget.

The frontier is bounded by two distinct architectural clusters. On the far left, the LR Clinical Triage models define the lowest latency boundary, achieving $\approx 98.8\%$ accuracy with an execution time of ≈ 1.5 ms (noting that this specific model evaluates five clinical classes rather than the standard nine). For the full 9-class diagnostic target, the HR Standalone *mini_1ch* variant emerges as an optimal configuration on the frontier. It captures a substantial reduction in latency (18.5 ms) while maintaining a unified accuracy of 98.2%, sacrificing less than 1% diagnostic performance compared to the highest-performing *base_2ch* model on the far right of the curve. Conversely, architectures such as the HR Standalone *micro_3ch* (75.1% accuracy at 24.4 ms) fall significantly below the Pareto boundary.

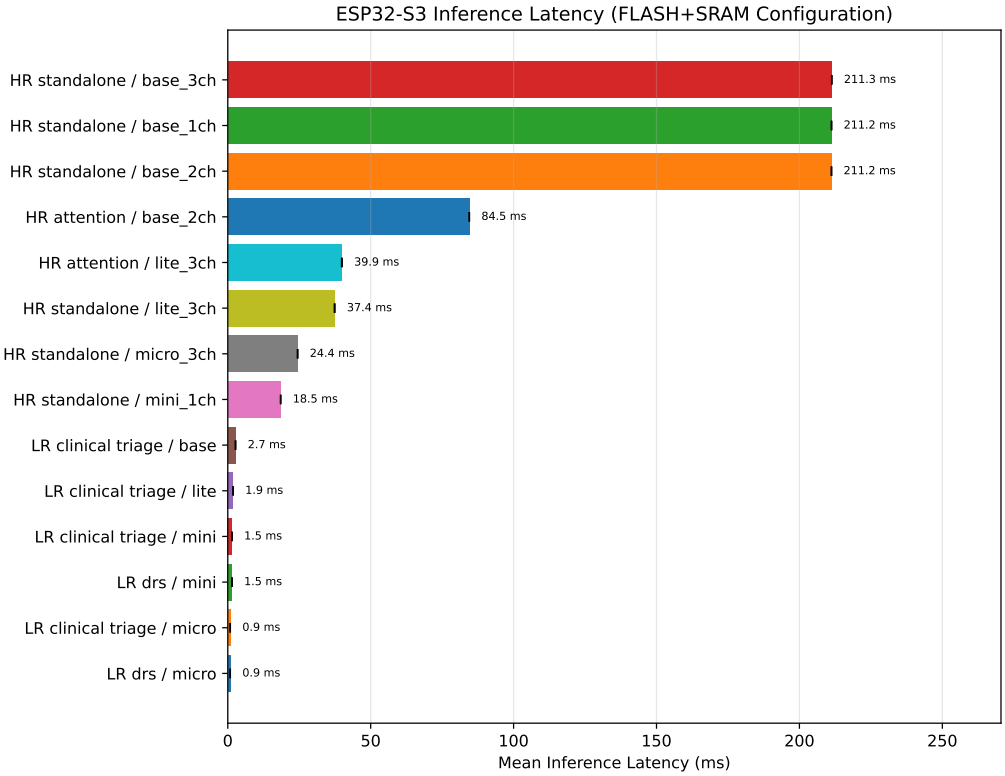


Figure 4.19: Mean inference latency of subset of INT8 architectures on the ESP32-S3, utilizing the FLASH+SRAM memory configuration.

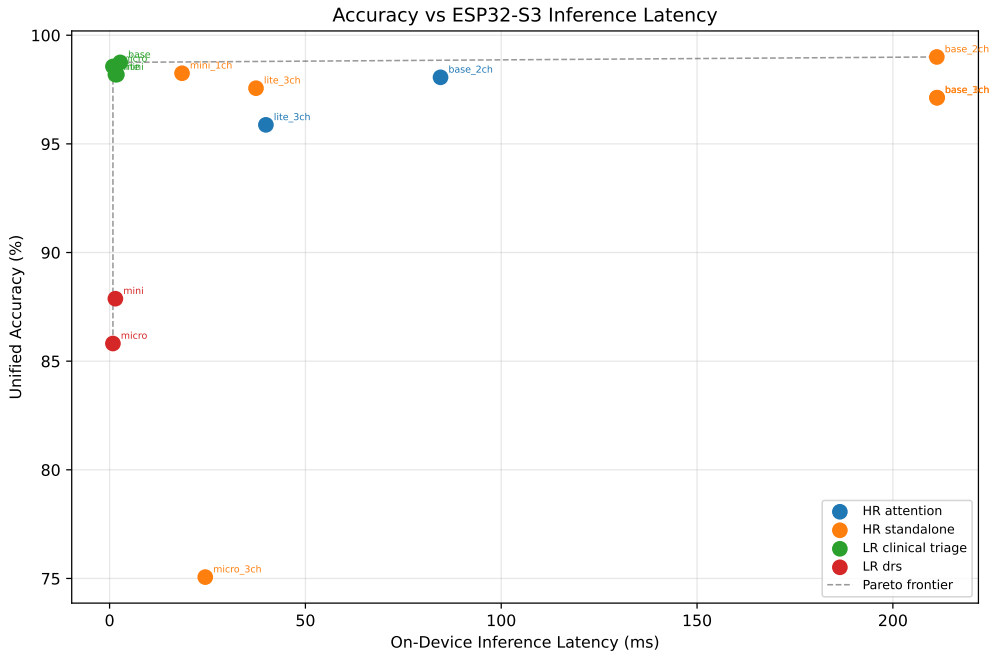


Figure 4.20: Pareto frontier illustrating the optimal trade-off boundary between unified diagnostic accuracy and on-device inference latency across all evaluated edge models.

4.5.3 Energy Consumption and Power Profiling

Physical power profiling was conducted for a targeted subset of model and memory configurations to quantify the absolute energy per inference. By measuring the exact charge transfer during the hardware-triggered inference window and multiplying it by the stable 3.3 V supply voltage, the total energy consumption was calculated. The summarized results are presented in Table 4.1, while the raw PPK2 power oscillograms for all listed configurations are provided in Appendix D.

Table 4.1: Empirical power profiling results for selected models and memory configurations on the ESP32-S3 (3.3V Supply). Energy per inference is calculated as Charge × Voltage.

Model	Memory Config.	Latency	Avg. Current	Charge	Energy
HR Standalone *base_3ch*	FLASH+PSRAM	399.3 ms	60.47 mA	24.15 mC	**79.70 mJ**
HR Standalone *base_3ch*	FLASH+SRAM	211.2 ms	64.88 mA	13.71 mC	**45.24 mJ**
HR Standalone *base_3ch*	SRAM+SRAM	78.7 ms	73.79 mA	5.81 mC	**19.17 mJ**
HR Standalone *mini_1ch*	FLASH+SRAM	18.5 ms	69.24 mA	1.28 mC	**4.22 mJ**
LR Clinical Triage *micro*	FLASH+SRAM	0.85 ms	61.78 mA	52.52 µC	**0.17 mJ**

The empirical data highlights the variance in energy consumption across different memory placement strategies for the HR Standalone *base_3ch* architecture. Executing this model entirely within the internal memory (SRAM+SRAM) yielded the highest average current draw (73.79 mA) but the lowest energy footprint (19.17 mJ) among its configurations due to its 78.7 ms execution time. Conversely, offloading the Tensor Arena to external memory (FLASH+PSRAM) lowered the average current to 60.47 mA but increased the total energy per inference to 79.70 mJ, representing a 4.1x energy penalty compared to the SRAM baseline.
When evaluating the structurally pruned architectures under the standard FLASH+SRAM memory configuration, the energy requirements scaled down significantly. The uncompressed HR Standalone *base_3ch* model required 45.24 mJ per inference. In comparison, the pruned HR Standalone *mini_1ch* variant consumed 4.22 mJ, achieving a 90.6% reduction in active energy consumption. Finally, the LR Clinical Triage *micro* model recorded the lowest overall energy consumption of the evaluated subset, requiring 0.17 mJ (173.3 μJ) per inference.

5

Discussion

The evaluation of the optical diagnostic pipelines developed in this study provides critical insights into the physical boundaries of spectral signal processing. By systematically transitioning from a traditional 1D deterministic algorithm to advanced machine learning architectures, this research exposes the fundamental differences between laboratory-optimized mathematics and clinically viable diagnostics. The following sections first critically analyze the performance of the deterministic baseline, deconstructing both its mathematical successes in a controlled environment and its inevitable failure points when confronted with the chaotic physics of in vivo wound diagnostics. Building upon these identified limitations, the discussion then evaluates the capacity of deep learning models, specifically 1D-CNNs and Spectral Transformers, to successfully resolve complex, non-linear optical phenomena such as quenching and polymicrobial shadowing. Finally, the chapter examines the critical engineering trade-offs between diagnostic granularity, algorithmic transparency, and hardware efficiency required for successful deployment in resource-constrained edge environments.

5.1 The Limitations of 1D Signal Processing: Evaluating the Deterministic Baseline

The deterministic MATLAB models, as evaluated in Section 4.1, were established to quantify the absolute performance ceiling of traditional, threshold-based classification. While iteratively optimizing mathematical parameters successfully raised the diagnostic accuracy from 81.06% to a peak of 96.69%, a rigorous analysis of this progression reveals that this ceiling is heavily constrained by biological variance and sensor noise. Ultimately, the deterministic baseline serves as mathematical proof that scalar thresholding cannot resolve the non-linear realities of severe bacterial infections.

5.1.1 The Specificity Crisis and Arithmetic Boundaries

The initial Standard Baseline's inadequate performance (81.06%) highlights a fundamental specificity crisis inherent to basic optical thresholding. The misclassification of 139 out of 320 blank Control samples as trace infections demonstrates that simple median-based thresholds are insufficient to separate the physical noise floor of the sensor hardware from the low-magnitude fluorescence emitted by trace bacterial concentrations (C4). In a clinical point-of-care setting, an algorithm exhibiting this

magnitude of a False Positive rate would be medically unviable, as diagnostic uncertainty is a primary driver of the over-prescription of antibiotics and the triggering of unnecessary interventions [11]. The Standard Baseline failed primarily because it attempted to apply a rigid absolute intensity limit to a signal inherently obscured by the macroscopic scattering wave of the physical medium.

Furthermore, a significant source of error in this initial approach was concentration bleeding, specifically the algorithmic upgrading of nearly 30% of PpIX C3 samples into the more severe C2 category. This mathematical failure demonstrates the structural flaw of using arithmetic means to define diagnostic boundaries for biological dilutions. Because bacterial loads scale logarithmically through serial dilution and proliferation, the variance within a given concentration class follows a log-normal distribution, resulting in an asymmetric, heavy-tailed spread of intensity values [15]. By defining the classification threshold as the simple arithmetic midpoint between two medians, the model created an unbalanced boundary positioned far too close to the lower concentration [15]. This rendered the algorithm hypersensitive to natural biological fluctuations in fluorophore expression, fundamentally lacking the statistical buffer zone required to robustly resolve adjacent infection severities.

5.1.2 Mathematical Optimization and the Deterministic Illusion

To address the failures of the standard model, the Upgraded Baseline incorporated rigorous physics-informed processing, surging to an exceptional global accuracy of 96.69%. The implementation of linear detrending allowed the model to anchor the spectral floor at 580 nm and 670 nm, effectively peeling away the broadband scattering slope to isolate the biological fluorescence bump beneath. Paired with species-specific AUC integration windows, this technique successfully decoupled the biological signal from irrelevant high-wavelength sensor static. Furthermore, by replacing the arithmetic boundary with a geometric mean, the algorithm artificially shifted the decision threshold downward in logarithmic space. This successfully threaded the statistical needle between the C2 and C3 distributions, resolving the concentration bleeding that plagued the initial iteration.

However, a critical analysis of these optimizations reveals that this peak performance is largely a mathematical artifact of the highly controlled laboratory environment. In this clean setting, the filter paper substrate provides a highly uniform, predictable baseline characterized by a linearly decreasing optical scattering slope. The deterministic model actively exploits this uniformity via linear detrending, effectively flattening the background to uncover the minute fluorescence peaks characteristic of trace biomarker concentrations. Yet, much like this baseline manipulation, the resolution of the C2 and C3 concentration boundary was achieved not by modeling the underlying physics, but by mathematically side-stepping them. By restricting the AUC calculation strictly to the primary emission peaks, the algorithm intentionally excised the heavily red-shifted spectral tail. While this successfully bypassed the mathematical distortion caused by the early stages of optical quenching, it ef-

fectively ignored the physical reality of the phenomenon.

Crucially, real-world biological substrates, such as highly heterogeneous *in vivo* tissue beds, do not present these uniform, linear optical backgrounds [36]. Human tissue exhibits complex, wavelength-dependent absorption and scattering properties that non-linearly distort both the excitation light and the resulting fluorescence emission [37]. Ultimately, the success of this deterministic model relies entirely on absolute scalar thresholds and rigid geometric assumptions, namely, that the background scattering baseline remains predictably linear. While these mathematical workarounds are optimal for serialized laboratory dilutions, they represent an illusion of efficacy when extrapolated to the optical complexity of authentic clinical applications.

5.1.3 Clinical Incompatibility: Environmental Stochasticity and the Inner Filter Effect

Translating this strictly threshold-based logic to an *in vivo* wound bed introduces insurmountable points of failure. The primary vulnerability lies in the model's absolute reliance on static intensity thresholds. In a clean laboratory setting, the sensor-to-target distance and optical coupling remain perfectly constant, allowing a specific integrated energy value to reliably correlate with a specific bacterial density. However, clinical environments introduce extreme stochasticity. Variables such as fluctuating practitioner-applied pressure, the presence of highly absorptive hemoglobin or exudate, and the heterogeneous scattering properties of necrotic tissue will drastically alter the absolute number of photons returning to the sensor. A localized drop in global intensity, caused by a mere millimeter increase in sensor distance, would trigger a cascading failure across the static geometric boundaries, causing the deterministic algorithm to systematically under-diagnose the severity of the infection.

Secondly, the deterministic architecture is inherently defenseless against the non-monotonic nature of severe optical quenching. The algorithm's thresholding logic dictates a strict linear assumption: an increase in bacterial concentration must consistently yield a proportionally higher integrated AUC. However, as the biofilm density approaches clinical severity (C1), the secondary IFE dictates that self-absorption overtakes photon production. At this critical juncture, the absolute measured intensity crashes, paradoxically dropping below the integrated volumes of lower-density (C2 and C3) infections. Confronted with a critically dense, highly quenched biofilm, a rigid thresholding model is mathematically forced to misclassify the severe infection as a resolving, low-level trace signal, introducing a catastrophic degree of clinical risk.

5.1.4 Polymicrobial Blinding and Spectral Shadowing

Finally, the deterministic speciation logic is fundamentally incapable of resolving the complex biological realities of polymicrobial infections. The baseline algorithm

utilized a mutually exclusive, Dual-ROI comparison, determining speciation based strictly on whichever fluorophore exhibited the highest relative peak magnitude. While functional for isolated mono-species laboratory assays, this rigid approach severely oversimplifies the *in vivo* clinical environment. In a chronic wound, a biofilm operates as a complex, highly diverse community of microorganisms encased in a protective polymeric matrix, creating an environment that inherently prevents healing [38]. Within these intricate polymicrobial ecosystems, it is highly common for numerous distinct pathogen species to thrive simultaneously, resulting in the concurrent presence of both CpI and PpIX fluorophores within the exact same optical field of view [38].

Because the CpI fluorophore generally exhibits a significantly higher quantum yield than PpIX, and its broad emission tail extends physically into the PpIX detection window, a mixed biological signal results in severe spectral shadowing. In such a polymicrobial scenario, the overwhelming absolute intensity of the CpI component forces the deterministic Dual-ROI logic to completely suppress the secondary pathogen signature. By relying on absolute peak height rather than the differential morphology of the combined spectral curve, the algorithm is effectively blinded. Consequently, this limitation causes the deterministic model to fail entirely in alerting the clinician to the presence of a severe co-infection, masking critical diagnostic data required for targeted antibiotic therapies.

5.1.5 The Imperative for Morphological Machine Learning

Ultimately, the 96.69% accuracy achieved by the Upgraded Baseline serves not as a finalized diagnostic solution, but as an empirical quantification of the limits of traditional 1D signal processing. The irreducible misclassifications remaining in the model represent fundamental physical overlaps where deterministic separation is impossible without generating secondary errors. Overcoming dynamic baseline shifts, the non-linear intensity inversions of the Inner Filter Effect, and polymicrobial spectral shadowing requires an architecture capable of looking entirely beyond absolute scalar volumes.

This physical ceiling directly necessitates the transition to the machine learning pipelines detailed in this study. Overcoming the limitations of optical diagnostics requires algorithms capable of evaluating structural curve morphology independently of global brightness. By deploying CNNs and Spectral Transformers paired with SNV normalization, the diagnostic pipeline is forced to optimize entirely on spatial geometric patterns, engineered physical ratios, and dynamic channel attention. This architectural leap successfully decouples the diagnostic output from the rigid constraints of absolute intensity, untangling the biological overlap and transitioning the system from a laboratory instrument into a clinically robust triage tool.

5.2 High-Resolution Architectural Ablation: Shape vs. Magnitude

The transition to deep learning on the high-resolution dataset provided a definitive benchmark for how neural networks interpret spectral data. The primary discovery in this stage was the overwhelming superiority of the SNV transformation. By standardizing each individual spectrum to possess a zero mean and unit variance, the SNV pipeline forced the 1D-CNN to ignore absolute photon counts, which are prone to hardware fluctuations, and optimize exclusively on the geometric morphology of the fluorescence peaks. The resulting 99% accuracy across almost all channel configurations proves that the "Inverse Problem" of bacterial quantification is solved more effectively through shape analysis than through absolute magnitude.

5.2.1 Feature Engineering and the Role of Spectral Derivatives

The multi-channel feature ablation (CH1, CH2, CH3) exposed the varying sensitivities of convolutional and attention-based architectures to local gradients. For the standard 1D-CNN, the addition of the first derivative (CH2) provided a marginal performance boost in the Wing Normalization pipeline, as it highlighted the exact locations of spectral slopes that the pooling layers might otherwise decimate. However, the introduction of the second derivative (CH3) frequently led to a performance plateau or slight degradation. This suggests that while second-order gradients theoretically expose hidden overlapping peaks, they simultaneously amplify high-frequency hardware noise, which, introduces a degree of stochasticity that outweighs the morphological gains.

5.2.2 The Spectral Transformer and the Locality Challenge

A significant finding was the initial struggle of the Spectral Transformer on raw single-channel data (CH1), particularly in the SNV pipeline (79%). This performance deficit compared to the 1D-CNN (99%) is likely due to the Transformer's lack of an innate inductive bias for local spatial relationships. Unlike a CNN, which uses local kernels to sweep for peaks, the Transformer uses global self-attention to relate every wavelength to every other wavelength simultaneously [18]. On a raw 1D signal, this global view can be "too broad", causing the model to lose focus on the narrow 5–10 nm peaks that define the fluorophores. Interestingly, the accuracy surged to 97–98% upon the introduction of the first derivative (CH2). By providing the Transformer with local gradient context in the second channel, the architecture was finally able to resolve the "locality" of the signal, proving that for 1D spectral tasks, Transformers require higher-order features to achieve parity with convolutional backbones.

5.3 The Limits of Architectural Optimization in Low-Resolution Hardware

The evaluation of the 14-channel low-resolution sensor highlighted a clear distinction between architectural explainability and macroscopic performance gains. The implementation of the DRS and Channel Attention mechanisms was intended to provide the network with the capacity to prioritize stable biological signals over hardware noise. The resulting attention weight maps (Figure 4.7) confirm that the model learned to dynamically adjust its hardware trust; however, this weighting was highly species-dependent rather than uniform. For CpI, the network aggressively suppressed uninformative boundary channels, specifically FZ, VIS, and FD (bins 3, 13, and 14), while heavily prioritizing the core biological emission window located in bins 8 and 10 through 12 (FXL, F7, F8, and NIR). Conversely, for PpIX, the network applied a much flatter, distributed weighting strategy across the sensor array, though it notably maintained a specific focus on the FXL broadband channel (bin 8).

Despite this successful, dynamic extraction of physics-logical features, the impact on global accuracy was marginal. For the Backscatter pipeline, the transition from feature-augmented models to the attention-augmented DRS architecture yielded only a 1 percentage point (p.p.) increase in accuracy (from 89% to 90%).

This plateau suggests that while attention mechanisms enhance the interpretability of the model, providing a "physics-logical" weighting that dynamically adapts to the specific fluorophore, they cannot overcome the fundamental informational ceiling of the 14-channel sensor on a complex 9×9 task. The marginal nature of these gains indicates that the persistent misclassifications between Moderate (C3) and High (C2) concentrations were not due to the network "looking at the wrong bins", but rather due to a total physical overlap of the spectral signatures within the discrete sensor's sampling range. This establishes a physical "indistinguishability limit" created by the Inner Filter Effect. Ultimately, the DRS served more as a tool for XAI and sensor-fault tolerance than as a driver for raw diagnostic accuracy.

5.3.1 Convergence through Clinical Realignment

The most profound discovery in the low-resolution study was that the true diagnostic breakthrough was achieved through label realignment rather than architectural complexity. As shown in the accuracy trajectory, the global accuracy for the Backscatter pipeline surged from 90% to 98% only after the target matrix was compressed into the 5×5 Clinical Triage system. This reconfiguration merged the mathematically indistinguishable states (C2/C3) into a single "Moderate" tier and the noise-floor-limited states (C0/C4) into a "Negative/Trace" tier.

This jump in performance proves that the earlier 9×9 matrix reached the aforementioned physical indistinguishability limit, where no amount of deep learning or attention weighting could resolve the signal overlap. This hypothesis of physical sig-

nal overlap is visually corroborated by the shift in the network's XAI outputs. In the previous highly granular setup, the network aggressively prioritized and suppressed specific bins in a forceful attempt to mathematically untangle the overlapping C2 and C3 states. However, upon transitioning to the 5×5 matrix, the DRS multipliers drastically flattened out, hovering consistently near the neutral 0.5 baseline for both fluorophores. Because the 5×5 matrix grouped the mathematically indistinguishable concentration ranges into unified clinical tiers, the diagnostic problem became deterministically stable. Consequently, the neural network was no longer forced to rely on aggressive, high-variance feature modulation to achieve high classification accuracy.

Furthermore, this label realignment led to a macroscopic convergence among the preprocessing pipelines. Once the diagnostic targets were aligned with the physical separability of the 14-channel signals, the system's sensitivity to the initial normalization method (Backscatter vs. VisNormalizer vs. PassThrough) was almost entirely eliminated. This strongly suggests that for low-resolution hardware, the physical definition of the clinical problem is vastly more influential than the specific architecture of the neural network or its data scaling pipeline.

5.4 The Diagnostic Trade-Off: Algorithmic Complexity versus Clinical Pragmatism

The juxtaposition of the high-dimensional neural networks against the minimalist Decision Tree establishes a fundamental engineering trade-off for the future of embedded optical diagnostics. As demonstrated in this study, the primary advantage of Deep Learning architectures, specifically 1D-CNNs and Spectral Transformers, is their unparalleled ability to resolve highly non-linear optical phenomena. Where deterministic models critically fail due to the signal inversions of the Inner Filter Effect (quenching) and polymicrobial spectral shadowing, convolutional networks succeed by implicitly mapping high-dimensional spatial relationships. In essence, deep learning algorithms are capable of bypassing rigid physical boundaries by identifying latent proxy variables within the spectral curve to achieve high classification accuracy.

However, this computational power comes with severe developmental and regulatory caveats. Because neural networks lack an innate understanding of optical physics, their accuracy and clinical safety are strictly bounded by the variance of their training manifolds. In the foundational stages of this study, samples were acquired within a highly controlled laboratory environment where inter-sample variability remained exceptionally low; sequential samples within the same specimen and batch were essentially optically identical. Consequently, there is an inherent risk that deep learning models, operating as opaque "black boxes", may achieve high validation accuracy by memorizing the specific hardware noise floor that separates blank controls from trace infections, rather than isolating the true biological fluorescence. This phenomenon, often referred to as "shortcut learning", occurs when models op-

timize for latent environmental artifacts rather than invariant pathological features [39]. To prevent a CNN from simply overfitting to environmental laboratory noise, it must be forced to generalize across every possible clinical background. This necessitates an exhaustive, exponentially large training dataset capturing every conceivable variation of "nothing". Encompassing diverse skin tones, complex exudate compositions, and myriad synthetic dressing materials so the network can reliably identify the "something" of an active pathogen. For small-to-medium medical enterprises operating under constrained budgets, the temporal and economic burden of acquiring such a massive, highly variant clinical dataset is often prohibitive.

Furthermore, the inherently opaque nature of deep convolutional features presents a formidable barrier to clinical certification. Navigating the complex intersection of the European Medical Device Regulation (MDR) [40] and the recently enacted EU AI Act (Regulation (EU) 2024/1689) [41] requires rigorous proof of algorithmic safety, risk management, and transparency [42]. Under the AI Act, medical diagnostic software is frequently classified as a "High-Risk AI System", necessitating strict adherence to requirements for explainability and human oversight [41, 42]. These mandates are notoriously difficult to fulfill in high-dimensional deep learning systems, where the decision-making process is distributed across millions of non-linear parameters.

Conversely, this research demonstrates that if the diagnostic objective is realigned to match the physical separability of the optical signal, the computational requirements drop precipitously. Because the convergence of the preprocessing pipelines indicated a highly robust signal, it was hypothesized that the overlapping physical boundaries had been so effectively neutralized that the complex spatial pattern recognition capabilities of the 1D-CNN were computationally redundant. By accepting the physical reality of optical quenching and compressing the highly granular 9×9 classification matrix into a pragmatic 5×5 Clinical Triage system, the diagnostic problem was fundamentally transformed. Rather than forcing a neural network to untangle physically indistinguishable signals, the triage matrix grouped these overlapping states into actionable clinical buckets (Negative/Trace, Moderate, and Severe). This outcome mathematically proves that the previously observed struggles of the neural network on the 9×9 matrix were not due to architectural flaws, but rather the network attempting to forcefully untangle physically identical optical signals. Once these clinical boundaries were realigned with optical physics, the diagnostic problem transitioned from a highly non-linear spatial challenge to a deterministically separable algebraic task.

By manually extracting explicit, domain-aware physical features, specifically the Normalized Deep-Red (F7/XL) and the Normalized Primary Red (F6/XL) ratios, a simple and transparent Decision Tree was able to achieve diagnostic parity with the deep learning models. This finding strongly aligns with modern Explainable AI paradigms, which argue that for high-stakes medical decisions, inherently interpretable models should be prioritized over black-box algorithms whenever diagnostic parity can be achieved [43]. It demonstrates that while deep learning is a powerful

tool for overcoming hardware noise and physical signal overlap, a deeply domain-aware approach to feature engineering and clinical target alignment can completely eliminate the need for computationally expensive convolutions. By framing the diagnostic question correctly, the computational requirement is reduced from thousands of MAC operations down to a lightweight sequence of fundamental logic gates. This minimalist approach establishes an optimal foundation for extreme low-power edge deployment, requires significantly less training data, and operates with near-zero inference latency. Crucially, the explicit nature of a Decision Tree provides the exact "white-box" transparency favored by the EU AI Act's requirements for high-risk systems. Every node corresponds to a human-readable physical threshold, providing a clear audit trail for clinical validation.
Ultimately, the development of an embedded diagnostic device demands a strategic choice between two divergent engineering pathways. The first pathway pursues maximum diagnostic granularity; this requires deep learning architectures, massive and highly variable datasets, specialized edge-AI hardware, and a complex, expensive regulatory certification process. The second pathway prioritizes clinical pragmatism and hardware efficiency; this requires a profound understanding of optical physics by the engineering team to manually extract the governing features, yielding a transparent, easily certifiable algorithm that operates on ultra-low-cost microcontrollers. While this second pathway sacrifices fine-grained concentration tracking in favor of broader triage tiers, it provides a highly robust, explainable, and immediately deployable solution for point-of-care wound diagnostics.

5.5 Hardware-Software Co-Design

While the theoretical accuracy of a neural network on a high-performance workstation establishes its clinical validity, the practical viability of a point-of-care medical device is heavily dependent on its execution on constrained silicon. The deployment of the evaluated architectures to the ESP32-S3 microcontroller required bridging the gap between FP32 algorithmic development and INT8 bare-metal execution. The empirical data generated during the architectural compression phase revealed key insights into how neural networks physically adapt to mathematical constraints, highlighting the necessity of hardware-software co-design in edge AI.

5.5.1 Quantization Resiliency

The transition to INT8 precision exposed a structural vulnerability in networks trained exclusively on single-channel spectral data. As demonstrated in Figure 4.13 and Figure 4.14, PTQ caused a severe reduction in accuracy in the 1CH HR architectures. To classify overlapping fluorophores relying solely on these primary spectral intensities, a 1CH network must mathematically calculate rate-of-change internally, often by learning high-magnitude weight outliers. Because PTQ applies a uniform rounding process, these outliers stretch the quantization bins, compressing lower-magnitude weights to zero and impairing the network's learned derivative representations.

Expanding the input tensor to explicitly include calculated derivatives (the 2CH and 3CH configurations) stabilized the PTQ HR models by providing these slopes before the data entered the network. This allowed the model to naturally learn a more uniformly distributed weight matrix that quantized with minimal error, effectively providing a computationally lightweight alternative to the more complex training pipeline required for QAT. However, the empirical data indicates that domain-aware feature engineering does not serve as a complete replacement for QAT, but rather as a highly effective complementary optimization. In multiple instances, the combination of multi-channel inputs and QAT yielded the absolute highest quantized performance; for example, the HR Standalone *base_3ch* architecture achieved its peak accuracy of 98.9% only when both strategies were employed simultaneously.

A practical hardware insight from the benchmark data is that this feature engineering is highly efficient from the perspective of the neural network's overall execution time. Expanding the input from 1Ch to 3CH incurred an inference latency penalty of less than 0.1 ms on the ESP32-S3 as seen in Figure 4.19. This negligible penalty is fundamentally due to the architectural structure of deep convolutional networks. Expanding the input channel dimension only increases the number of MAC operations in the very first convolutional layer, leaving the computational load of all subsequent deep layers entirely unaffected. Consequently, the relative increase in total mathematical operations is inherently minuscule. This architectural efficiency is then further amplified by the *ESP-NN* library and the *Xtensa LX7* processor's SIMD vector extensions, which parallelize and process this minor operational increase almost instantaneously. While the explicit mathematical calculation of these derivatives does introduce a minor pre-processing overhead on the CPU prior to invoking the inference engine, this arithmetic cost is a highly favorable trade-off for the resulting quantization stability and performance improvement.

This quantization vulnerability was however entirely absent in the LR architectures. Both the LR DRS and LR Clinical Triage models maintained high accuracy parity between FP32 and INT8 PTQ representations. This inherent stability is a direct consequence of the LR hardware's discrete, wide-bin optical architecture. Because the LR data lacks the fine-grained continuous gradients captured by a high-resolution spectrometer, the neural network does not need to learn highly sensitive, high-magnitude filters to evaluate subtle morphological shifts. Furthermore, because the LR models process a drastically reduced input sequence while maintaining similar convolutional depths, they possess a higher mathematical redundancy per input feature. This architectural redundancy allows the LR models to distribute their learned representations across more low-magnitude weights rather than forcing a few specific filters to adopt extreme, high-magnitude values to process dense data. Consequently, the LR models develop a naturally uniform and constrained weight distribution during standard training, allowing them to map flawlessly onto the INT8 PTQ grid.

5.5.2 QAT Instability in Constrained Attention Mechanisms

While QAT proved effective with the larger convolutional architectures, its application revealed a critical instability when utilized with highly constrained attention mechanisms. As observed in the HR Attention model evaluations seen in Figure 4.14, QAT induced a severe accuracy collapse when applied to the aggressively pruned *mini* and *micro* variants. For instance, while the HR Attention *micro_3ch* model maintained a highly viable 97.6% accuracy under PTQ, its performance plummeted to 55.9% when subjected to QAT.

This pronounced degradation highlights a fundamental conflict between low-capacity architectures and simulated quantization noise. Attention mechanisms rely heavily on the Softmax function to generate probability distributions across the sequence length, an operation that is notoriously sensitive to minor variations in precision due to its exponential scaling [44]. During QAT, discrete INT8 mathematical bounds are simulated during the forward pass, injecting artificial noise into the training loop. In the uncompressed *base* models, the network possesses sufficient parameter redundancy to absorb this gradient noise and successfully realign its weights to the quantized grid [29]. However, when the network's capacity is drastically reduced via fractional or topological pruning, this simulated noise overwhelms the limited parameter space. The restricted attention heads fail to converge under the noisy gradients, leading to a collapse of the learned spatial relationships. Conversely, because PTQ simply rounds the pre-converged FP32 weights after the optimization manifold has already settled, it entirely bypasses this active gradient instability. This mathematical distinction effectively explains why the *micro* attention models successfully survived static PTQ but failed to converge under QAT.

5.5.3 Architectural Pruning

While the structurally unpruned *base* models successfully fit within the internal SRAM of the ESP32-S3 due to INT8 quantization, this specific hardware represents the higher end of microcontrollers in terms of both performance and energy consumption. For an future commercial deployment in a battery-operated, point-of-care medical device, the diagnostic algorithm must target highly energy-efficient silicon, which typically features a fraction of this available memory capacity. Furthermore, in a complete commercial system, the machine learning model will not execute in isolation. It must share the available memory pool with concurrent system processes, such as wireless communication stacks and user interface controllers. Therefore, the architectural pruning phase was not a strict necessity for the immediate test environment, but rather a forward-looking exploration to identify the a minimal parameter footprint required to maintain diagnostic accuracy under true real-world constraints.

The Pareto frontier seen in Figure 4.20 demonstrated that fractional width reduction (*mini*) consistently outperformed topological depth reduction (*micro*). The HR Standalone *mini* model achieved substantial latency reductions (18.5 ms) while sacrificing less than 1% accuracy, whereas the *micro* models fell notably below the

optimal boundary.
This performance divergence can be attributed to the mechanics of 1D signal processing. In a deep, narrow network (*mini*), the stacked convolutional layers maintain a large receptive field. The network retains the ability to process the entire spectral curve simultaneously, allowing it to evaluate broad macroscopic phenomena like the scattering baseline slope or the overarching shift of the IFE. Conversely, reducing the topological depth (*micro*) limits this receptive field. A shallow network can only process narrow, isolated segments of the spectrum at any given time, reducing its capacity to resolve complex biological overlaps. This indicates that for spectral diagnostics, maintaining architectural depth is beneficial, and computational speed may be better optimized by reducing layer width rather than removing layers entirely.

5.5.4 Memory Scaling

A notable finding from the memory footprint analysis detailed in Figure 4.17 was the shifting proportionality between static weights and dynamic memory as the models were compressed. In the structurally unpruned HR *base* models, the static, read-only weights constituted the larger share of the memory footprint, accounting for approximately 58% of the total size (e.g., 110 KB weights versus an 81 KB arena). Because the ESP32-S3 hardware supports XIP, these static weight matrices can be mapped directly from non-volatile Flash, successfully mitigating their impact on the internal SRAM.

However, as the high-resolution network is aggressively pruned, the reduction in parameter count significantly outpaces the reduction in intermediate activation buffers. Applying fractional width reduction (*lite* and *mini*) brought the weight-to-arena ratio down to an approximate 1:1 parity, as seen in the HR *mini_1ch* variant which requires roughly 17 KB for each respective component. This scaling disparity culminated in the topologically reduced HR *micro* variant, which was the only architecture where the dynamic Tensor Arena (27 KB) actively surpassed the footprint of the static weights (17 KB). This highlights a critical, well-documented bottleneck in edge Machine Learning (ML) design: while removing network layers rapidly reduces the total parameter count, the dynamic memory required to process a high-resolution input tensor does not scale down proportionally [45].

In stark contrast, the LR architectures exhibited a fundamentally different memory scaling behavior. Because the LR models process a significantly smaller spatial input tensor, their intermediate activation buffers are inherently minimal. Across all LR variants, the static weights consistently remained the dominant component of the overall memory footprint. For instance, the LR Clinical Triage *micro* variant requires only a 4 KB Tensor Arena compared to its 10 KB weight matrix. This confirms that when the physical input resolution is sufficiently low, the parameter count remains the primary memory bottleneck, allowing these models to leverage Flash-based XIP to achieve an exceptionally minimal RAM footprint.

Unlike the static weights, the Tensor Arena requires constant read/write access for

intermediate network calculations and must therefore reside entirely within volatile RAM [34]. While advanced platforms like the ESP32-S3 can theoretically offload this arena to external PSRAM, relying on external memory buses introduces severe latency and power penalties that are often prohibitive for true ultra-low-power devices [45]. Therefore, fitting the Tensor Arena entirely within the highly constrained internal SRAM remains the definitive target for edge deployment. Dedicating tens of kilobytes of this internal RAM exclusively to a neural network's feature maps severely constrains the memory budget available for the rest of the application firmware, such as communication stacks and sensor drivers. When scaling these architectures down to more constrained microcontrollers in the broader edge ecosystem where external RAM is rarely available and total SRAM is drastically reduced. This volatile memory requirement becomes the primary deployment bottleneck. Consequently, the dynamic footprint of the Tensor Arena is often the ultimate determining factor for whether a highly compressed machine learning model is physically viable on ultra-low-power silicon.

5.5.5 Energy Profiling and the "Race to Sleep"

Finally, the physical power profiling of the networks illustrated the importance of memory placement and the operational characteristics of edge energy consumption. While algorithmic efficiency is traditionally measured in total arithmetic operations, physical battery drain is fundamentally tied to memory access times and total execution duration.

The power profiling data seen in Table 4.1 demonstrated that for the HR Standalone *base_3ch* architecture, executing the network entirely from internal memory (SRAM+SRAM) drew the highest average active current (73.79 mA). However, because the processor operated efficiently without waiting for the SPI bus, it completed the inference in just 78.7 ms, yielding the lowest total energy footprint (19.17 mJ) among the tested memory configurations for that specific model. When the static weights were offloaded to external Flash (FLASH+SRAM), the inference latency increased to 211.3 ms, more than doubling the total energy consumption (45.24 mJ) due to the external weight fetch bottleneck. This penalty was severely exacerbated when the Tensor Arena was subsequently offloaded to external PSRAM (FLASH+PSRAM). While constant cache miss stalling lowered the average active current draw to 60.47 mA, the resulting 5x increase in execution time compared to the pure internal configuration led to a massive 4.1x penalty in total energy (79.70 mJ).

This phenomenon perfectly validates the "race to sleep" paradigm in embedded systems: it is vastly more energy-efficient to draw a higher current for a shorter duration than to draw a moderate current over an extended period caused by for example a bus bottleneck [2]. However, the empirical data reveals that this external memory penalty primarily afflicts the larger, structurally unpruned models. Because the highly compressed HR *mini* architectures possessed dramatically smaller memory footprints, they fit almost entirely within the ESP32-S3's internal data caches.

Consequently, the HR Standalone *mini_1ch* maintained a near-constant inference latency of roughly 18.5 ms regardless of whether its memory was mapped from internal SRAM, external Flash, or even external PSRAM. It must be acknowledged that this observed cache immunity is partially a byproduct of the looped benchmarking methodology, which inherently guarantees a warm cache. In a true point-of-care setting where diagnostic inferences are executed intermittently, the processor would encounter a cold cache, likely introducing a minor initial latency penalty as the weights are fetched for the first time. Despite this real-world caveat, the ultimate expression of the "race to sleep" principle is seen in the highly pruned LR Clinical Triage *micro* model, which completed inference in a mere 0.85 ms, consuming an almost negligible 0.17 mJ. This establishes that a primary objective of edge ML compression is not merely to reduce arithmetic, but to shrink the network's active footprint to fit within the silicon's internal cache geometry. By doing so, the system can bypass external bus latency, fully leverage the processor's maximum speed, and return to a low-power sleep state as rapidly as possible.

5.6 Summary of the Path Forward

The transition from high-fidelity spectrometry to cost-effective embedded sensing represents the critical "last mile" of optical bacterial diagnostics. However, as the empirical benchmarking data demonstrates, deploying these diagnostics to the edge requires more than simply converting floating-point algorithms into integer arithmetic. It demands a holistic hardware-software co-design strategy where the physical constraints of the microcontroller actively dictate the architectural evolution of the diagnostic algorithm.

For future commercialization, a portable point-of-care device will inevitably target silicon even more resource-constrained than the ESP32-S3 utilized in this study. To maintain viability on ultra-low-power architectures, engineers must prioritize shrinking the dynamic Tensor Arena so that it fits entirely within the limited internal SRAM [34, 45]. As proven by the energy profiling data, avoiding the massive latency penalties of external memory buses is the only way to successfully execute the "race to sleep" paradigm and preserve the battery life of a handheld medical device [2]. Consequently, utilizing lower-resolution hardware and aggressively pruned architectures provides the most immediate and energy-efficient path to physical deployment.

From an algorithmic perspective, this hardware reality splits the path forward into two distinct but complementary engineering trajectories. In the short term, the deeply domain-aware, physics-informed Decision Tree represents the optimal deployment candidate. By compressing the diagnostic task into a 5×5 Clinical Triage matrix and manually extracting engineered features, this minimalist approach completely bypasses the computational and memory bottlenecks of convolutional processing. Furthermore, its inherent "white-box" transparency provides a more clear, auditable logic trail [43], offering a strategic advantage when navigating the stringent explainability mandates of the MDR and the EU AI Act [40, 41, 42].

In the long term, the high-dimensional deep learning architectures developed in this study serve as a foundational proof-of-concept for the optical sensing platform. While these models demonstrate that neural networks can successfully navigate the complex non-linearities of fluorescence and the Inner Filter Effect, their real-world clinical deployment remains strictly gated by the need for more extensive data. Once this translational gap is bridged by training the networks on authentic clinical confounders, such as complex biological exudates and heterogeneous wound beds, the optimization pipelines established here will become crucial [39, 5]. Utilizing techniques like QAT and multi-channel feature injection will allow these networks to be safely compressed and stabilized for edge execution.

Ultimately, the methodologies evaluated in this study provide a highly scalable framework for the next generation of digital health platforms. By prioritizing clinical pragmatism, hardware efficiency, and an adherence to optical physics, rapid and objective bacterial quantification can be seamlessly integrated into embedded bedside devices, establishing a robust technological foundation for improved global wound management.

6

Conclusion

This research has systematically investigated the application of optical biomarker diagnostics across two distinct hardware paradigms: a high-resolution miniature spectrometer and a cost-effective, low-resolution photodiode-based sensor. By evaluating a progression of methodologies for both data streams, ranging from deterministic 1D signal processing to high-dimensional deep learning architectures, this study has illuminated the physical and mathematical boundaries of autofluorescence-based pathogen quantification. Ultimately, this comparative approach defines an optimal and highly practical pathway for future clinical device engineering.

The implementation of 1D-CNNs and Spectral Transformers demonstrated an exceptional capacity to resolve highly non-linear optical phenomena, achieving near-perfect accuracy on high-resolution laboratory datasets. These deep learning architectures are undeniably powerful tools, offering immense potential to outpace traditional signal processing algorithms in complex spectral pattern recognition [9, 19]. To fully realize this potential in a clinical setting, future development must focus on bridging the translational gap by massively expanding the training manifolds to prevent models from learning localized environmental shortcuts [39]. Because neural networks are highly susceptible to such dataset bias, they require vast and highly variable datasets to safely generalize across the stochastic reality of *in vivo* clinical scenarios [39]. Specifically, gathering real-world data encompassing diverse wound tissues, biological fluids, and optical confounders is a necessary next step to ensure algorithmic robustness and clinical efficacy [5]. Furthermore, transitioning these high-capacity models into certified medical devices will require ongoing research into AI explainability to confidently meet the stringent transparency requirements of the MDR and the EU AI Act [40, 41, 42].

While the deep learning architectures represent the future ceiling of diagnostic capability, this study simultaneously achieved a breakthrough through clinical target realignment and domain-aware engineering. As the research progressed to the low-resolution sensor, a fundamental "indistinguishability limit" of the discrete hardware was identified. By compressing the diagnostic matrix into a pragmatic 5×5 Clinical Triage system and manually extracting physics-informed features, the algorithm was tailored to strictly obey established optical physics. This strategic optimization proved that a highly efficient, minimalist Decision Tree can achieve robust diagnostic parity (98%) with the advanced deep learning models.

This finding establishes a highly advantageous diagnostic philosophy for the im-

mediate future of digital health platforms. By elegantly merging mathematically indistinguishable trace and control signals into a unified baseline, the physics-informed model provides a reliable, actionable triage tool. It guarantees that positive detections are explicitly grounded in biological reality with clear, interpretable decision boundaries. Crucially, the "white-box" transparency of this approach inherently satisfies the current regulatory demands for explainability, providing a clear and auditable pathway to clinical certification [43].

In conclusion, this research provides Odinwell with a robust, validated dual-track roadmap for the development of their optical sensing platform. Hand-crafted, physics-based algorithms offer a powerful, understandable, and certifiable foundation for immediate edge deployment, achieving remarkable accuracy with minimal computational overhead. Concurrently, the deep learning frameworks developed herein serve as a foundational proof-of-concept, establishing the architectural groundwork required to harness comprehensive, real-world clinical datasets as they become available. Through this strategic, physics-first approach to medical engineering, rapid and objective bacterial triage can be safely and swiftly delivered to the bedside, providing a scalable foundation for improved wound-care outcomes and enhanced global antibiotic stewardship.

Declaration of AI Usage

During the development of this project and the preparation of this manuscript, Artificial Intelligence (AI) tools were utilized as supplementary assistants to optimize workflow efficiency. The primary applications of AI included proofreading the report, correcting typographical errors, and refining the text to ensure a consistent, academic, and professional tone. Additionally, AI was employed as a technical resource for troubleshooting programming errors, debugging code, and engaging in exploratory discussions regarding potential methodological improvements.

All core concepts, analytical logic, and final conclusions remain the original work of the authors. The authors have rigorously reviewed all AI-assisted outputs and assume full responsibility for the accuracy, integrity, and originality of the content presented in this report.

Bibliography

[1] J. Johnson and G. Bohn, "Clinical and economic evidence supporting the value of fluorescence imaging of bacteria in wound care," *Journal of Market Access & Health Policy*, vol. 13, no. 4, 2025. [Online]. Available: https://www.mdpi.com/2001-6689/13/4/48

[2] D. H. Kim, C. Imes, and H. Hoffmann, "Racing and pacing to idle: Theoretical and empirical analysis of energy optimization heuristics," in *2015 IEEE 3rd International Conference on Cyber-Physical Systems, Networks, and Applications*, 2015, pp. 78–85.

[3] *TCS3448 Datasheet: 14-Channel Multi-Spectral Sensor*, ams-OSRAM AG, February 2025, version 1.01. [Online]. Available: https://look.ams-osram.com/m/1c24b057e65ee61e/original/TCS3448-14-Channel-multi-spectral-sensor.pdf

[4] C. Fontana, M. Favaro, M. Pelliccioni, S. Minelli, M. C. Bossa, A. Altieri, C. D'Orazi, F. Paliotta, O. Cicchetti, M. Minieri, C. Prezioso, D. Limongi, and C. D'agostini, "Laboratory automation in microbiology: Impact on turnaround time of microbiological samples in covid time," *Diagnostics*, vol. 13, no. 13, 2023. [Online]. Available: https://www.mdpi.com/2075-4418/13/13/2243

[5] M. Y. Rennie, D. Dunham, L. Lindvere-Teene, R. Raizman, R. Hill, and R. Linden, "Understanding real-time fluorescence signals from bacteria and wound tissues observed with the moleculight i:xtm," *Diagnostics*, vol. 9, no. 1, 2019. [Online]. Available: https://www.mdpi.com/2075-4418/9/1/22

[6] S. Saha and L. Xu, "Vision transformers on the edge: A comprehensive survey of model compression and acceleration strategies," *Neurocomputing*, vol. 643, p. 130417, Aug. 2025. [Online]. Available: http://dx.doi.org/10.1016/j.neucom.2025.130417

[7] R. DaCosta, "Portable real-time fluorescence imaging device," Canadian Patent CA2 891 990A1, May, 2014. [Online]. Available: https://patents.google.com/patent/CA2891990A1/en

[8] MolecuLight Inc., "MolecuLight i:X receives FDA 510(k) clearance for the device's ability to detect wounds," Press Release, August 2018. [Online]. Available: https://us.moleculight.com/news/moleculight-ix-receives-fda-510k-clearance-for-the-devices-ability-to-detect-wounds/

[9] F. Zeng, W. Peng, G. Kang, Z. Feng, and X. Yue, "Spectral data classification by one-dimensional convolutional neural networks," in *2021 IEEE Inter-*

national Performance, Computing, and Communications Conference (IPCCC), Oct 2021, pp. 1–6.

[10] B. Clark, "What is quantization aware training?" Nov 2025. [Online]. Available: https://www.ibm.com/think/topics/quantization-aware-training

[11] C. Llor and L. Bjerrum, "Antimicrobial resistance: Risk associated with antibiotic overuse and initiatives to reduce the problem," *Therapeutic advances in drug safety*, vol. 5, pp. 229–241, 12 2014.

[12] H. Lu, F. Floris, M. Rensing, and S. Andersson-Engels, "Fluorescence spectroscopy study of protoporphyrin ix in optical tissue simulating liquid phantoms," *Materials*, vol. 13, no. 9, 2020. [Online]. Available: https://www.mdpi.com/1996-1944/13/9/2105

[13] J. R. Lakowicz, *Principles of Fluorescence Spectroscopy*. Springer Science & Business Media, 2006.

[14] M. Zeaiter and D. Rutledge, "3.04 - preprocessing methods," in *Comprehensive Chemometrics*, S. D. Brown, R. Tauler, and B. Walczak, Eds. Oxford: Elsevier, 2009, pp. 121–231. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780444527011000740

[15] E. Limpert, W. A. Stahel, and M. Abbt, "Log-normal distributions across the sciences: Keys and clues: On the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into variability and probability—normal or log-normal: That is the question," *BioScience*, vol. 51, no. 5, pp. 341–352, 05 2001. [Online]. Available: https://doi.org/10.1641/0006-3568(2001)051[0341:LNDATS]2.0.CO;2

[16] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[17] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1d convolutional neural networks and applications: A survey," *Mechanical Systems and Signal Processing*, vol. 151, p. 107398, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0888327020307846

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.

[19] P. Fu, Y. Wen, Y. Zhang, L. Li, Y. Feng, L. Yin, and H. Yang, "Spectratr: A novel deep learning model for qualitative analysis of drug spectroscopy based on transformer structure," *Journal of Innovative Optical Health Sciences*, vol. 15, no. 03, p. 2250021, 2022. [Online]. Available: https://doi.org/10.1142/S1793545822500213

[20] X. Sun, Y. Wang, A. Du, M. Dong, Y. Wang, Y. Zhang, Y. Zhang, Y. Huang, X. Huang, Y. Liu, and J. Ni, "Autofluorescence properties of

wound-associated bacteria cultured under various temperature, salinity, and ph conditions," *BMC Microbiology*, vol. 25, no. 1, p. 511, 2025. [Online]. Available: https://doi.org/10.1186/s12866-025-04200-3

[21] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures." *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964. [Online]. Available: https://doi.org/10.1021/ac60214a047

[22] K. Cheveralls, I. Kolb, and D. G. Mets, "Leave-one-batch-out cross-validation reveals strong batch effects in Raman spectroscopy of yeast cultures," *The Stacks*, mar 26 2026. [Online]. Available: https://thestacks.org/publications/negative-data-raman-batch-effects-yeast/v1

[23] M. Crawshaw, "Multi-task learning with deep neural networks: A survey," 2020. [Online]. Available: https://arxiv.org/abs/2009.09796

[24] J. Cheng, "A neural network approach to ordinal regression," 2007. [Online]. Available: https://arxiv.org/abs/0704.1028

[25] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)*. Springer, 02 2009.

[26] L. Breiman, J. Friedman, C. Stone, and R. Olshen, *Classification and Regression Trees*. Taylor & Francis, 1984. [Online]. Available: https://books.google.se/books?id=JwQx-WOmSyQC

[27] A. Gholami, S. Kim, D. Zhen, Z. Yao, M. Mahoney, and K. Keutzer, *A Survey of Quantization Methods for Efficient Neural Network Inference*. Chapman & Hall/CRC, 01 2022, pp. 291–326.

[28] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2704–2713.

[29] M. Nagel, M. Fournarakis, R. A. Amjad, Y. Bondarenko, M. van Baalen, and T. Blankevoort, "A white paper on neural network quantization," 2021. [Online]. Available: https://arxiv.org/abs/2106.08295

[30] A. Jain, E. Kim, and R. Kuester, "tflite-micro," https://github.com/tensorflow/tflite-micro, Apr 2025. [Online]. Available: https://github.com/tensorflow/tflite-micro

[31] V. Dattu, S. Nalegave, S. Tipnis, T. Rezucha, I. Grokhotkov, espressif-bot, K. Sovani, B. Parhusip, T. Sebestik, Dmitry, Legend, and L. S. Vaz, "espressif/esp-tflite-micro," https://github.com/espressif/esp-tflite-micro, Nov 2025. [Online]. Available: https://github.com/espressif/esp-tflite-micro

[32] V. Dattu, K. Sovani, S. Nalegave, espressif-bot, I. Grokhotkov, and S. Tipnis, "espressif/esp-nn," https://github.com/espressif/esp-nn, Sep 2025. [Online]. Available: https://github.com/espressif/esp-nn

[33] Espressif Systems (Shanghai) Co., Ltd, *ESP32-S3-WROOM MCU Datasheet*, 2025, rev. 1.7. [Online]. Available: https://documentation.espressif.com/esp32-s3-wroom-1_wroom-1u_datasheet_en.pdf

[34] R. David, J. Duke, A. Jain, V. J. Reddi, N. Jeffries, J. Li, N. Kreeger, I. Nappier, M. Natraj, S. Regev, R. Rhodes, T. Wang, and P. Warden, "Tensorflow lite micro: Embedded machine learning on tinyml systems," 2021. [Online]. Available: https://arxiv.org/abs/2010.08678

[35] Nordic Semiconductor ASA, *Power Profiler Kit II Product Brief*, 2025, rev. 1.0. [Online]. Available: https://nsscprodmedia.blob.core.windows.net/prod/software-and-other-downloads/product-briefs/power-profiler-kit-ii-pbv10.pdf

[36] S. Jacques, "Optical properties of biological tissues: A review," *Physics in medicine and biology*, vol. 58, pp. R37–R61, 05 2013.

[37] R. Richards-Kortum and E. Sevick-Muraca, "Quantitative optical spectroscopy for tissue diagnosis," *Annual Review of Physical Chemistry*, vol. 47, no. Volume 47, 1996, pp. 555–606, 1996. [Online]. Available: https://www.annualreviews.org/content/journals/10.1146/annurev.physchem.47.1.555

[38] P. G. Bowler, B. I. Duerden, and D. G. Armstrong, "Wound microbiology and associated approaches to wound management," *Clinical Microbiology Reviews*, vol. 14, no. 2, pp. 244–269, 2001. [Online]. Available: https://journals.asm.org/doi/abs/10.1128/cmr.14.2.244-269.2001

[39] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, no. 11, p. 665–673, Nov. 2020. [Online]. Available: http://dx.doi.org/10.1038/s42256-020-00257-z

[40] European Union, "Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 april 2017 on medical devices," 2017, official Journal of the European Union, L 117/1. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2017/745/oj

[41] ——, "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 june 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," 2024, official Journal of the European Union, L, 2024/1689. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[42] M. Aboy, T. Minssen, and E. Vayena, "Navigating the eu ai act: implications for regulated digital medical products," *npj Digital Medicine*, vol. 7, 09 2024.

[43] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, pp. 206–215, 05 2019.

[44] Y. Bondarenko, M. Nagel, and T. Blankevoort, "Understanding and overcoming the challenges of efficient transformer quantization," 2021. [Online]. Available: https://arxiv.org/abs/2109.12948

[45] J. Lin, W.-M. Chen, Y. Lin, J. Cohn, C. Gan, and S. Han, "Mcunet: Tiny deep learning on iot devices," 2020. [Online]. Available: https://arxiv.org/abs/2007.10319

A

High-Resolution 1D-CNN Confusion Matrices

This appendix provides the comprehensive class-wise evaluation metrics for the high-resolution 1D-CNN architectures. While the macroscopic accuracy trajectories in the main text illustrate the overall impact of feature engineering and preprocessing, the confusion matrices presented here document the absolute classification boundaries, highlighting exactly where the models succeeded or failed.

Specifically, this section contrasts the absolute extremes of the 1D-CNN ablation study. It first details the sub-optimal configurations: the standard architecture utilizing Source Normalization, which failed to separate trace concentration boundaries, and the Channel Attention architecture utilizing single-channel Wing Normalization, which suffered a severe accuracy collapse. These are directly compared against the optimal configurations: the highly stable SNV baseline and the "rescued" dual-channel Wing Normalization attention model.

By mapping the exact class-wise error distributions, these matrices prove that the macroscopic accuracy fluctuations were not caused by random noise, but by specific, physically constrained optical boundaries that the optimized pipelines successfully overcame.

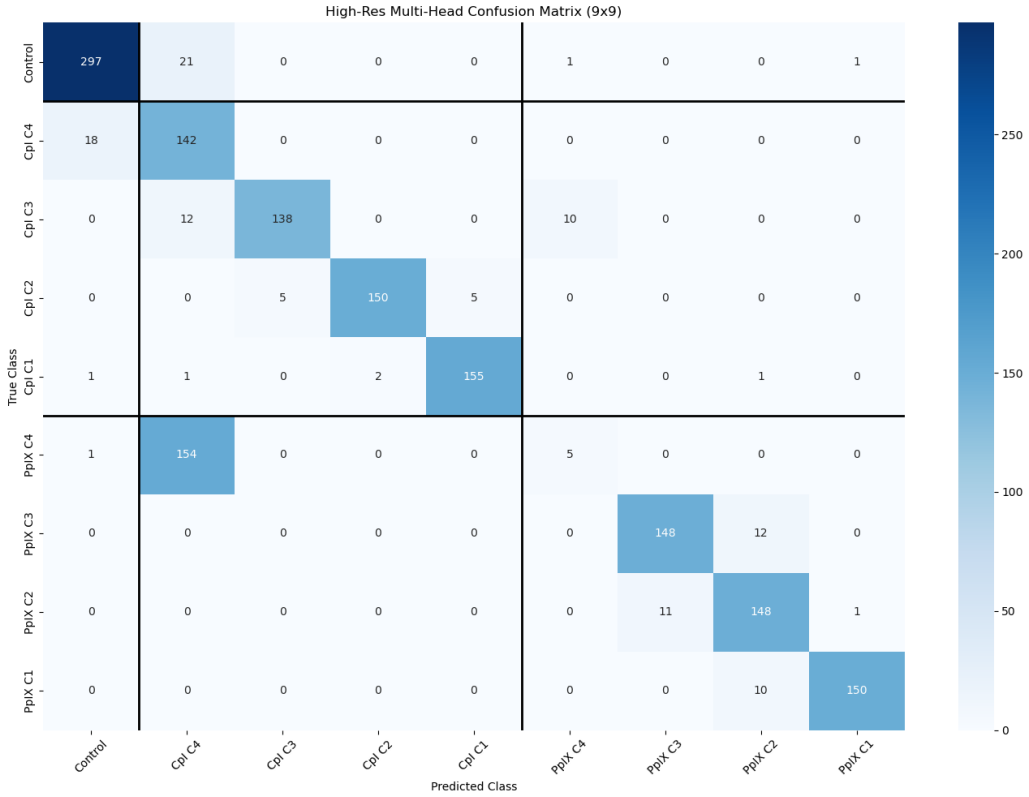


Figure A.1: Confusion matrix for the sub-optimal standard 1D-CNN utilizing Source Normalization (CH1). Correlating with its low 83% global accuracy, the matrix reveals a severe failure to distinguish trace infections. Notably, 154 PpIX C4 samples are critically misclassified as CpI C4, and substantial bleeding occurs at the baseline Control boundary.

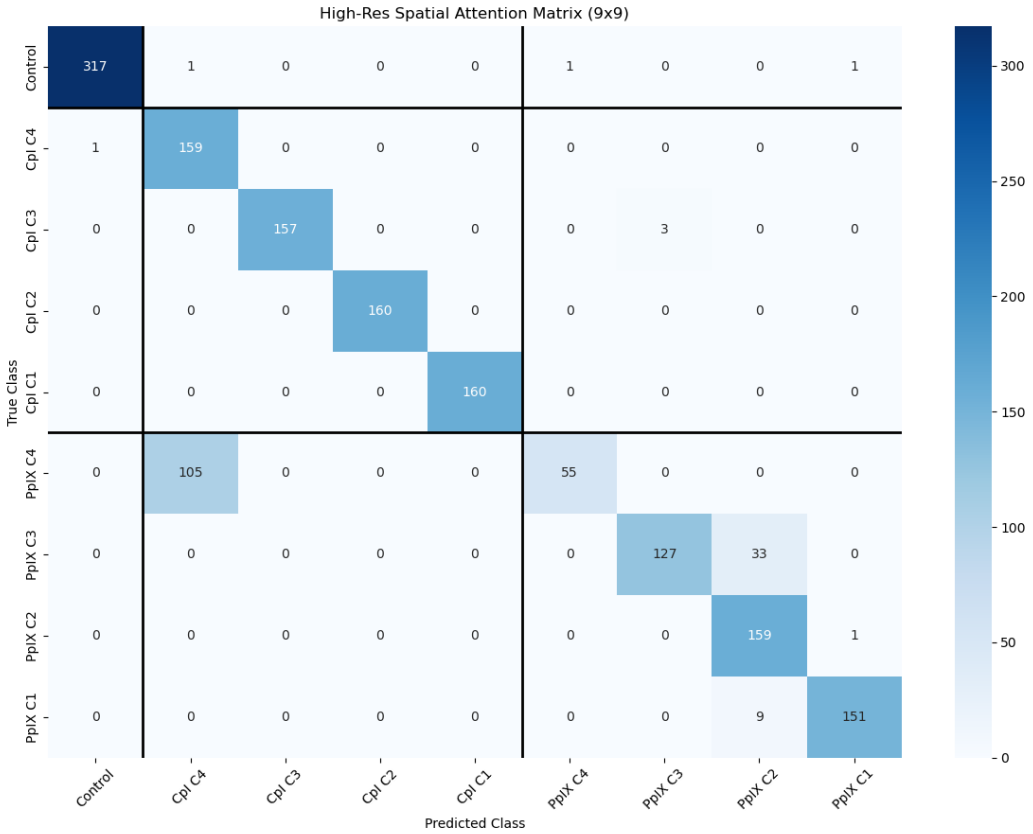


Figure A.2: Confusion matrix for the sub-optimal Channel Attention 1D-CNN utilizing Wing Normalization (CH1). The application of the attention mechanism to a single raw channel caused a severe performance degradation (90% global accuracy). The matrix exposes the network's inability to separate the biological species at trace concentrations, resulting in 105 PpIX C4 samples being incorrectly mapped to CpI C4.

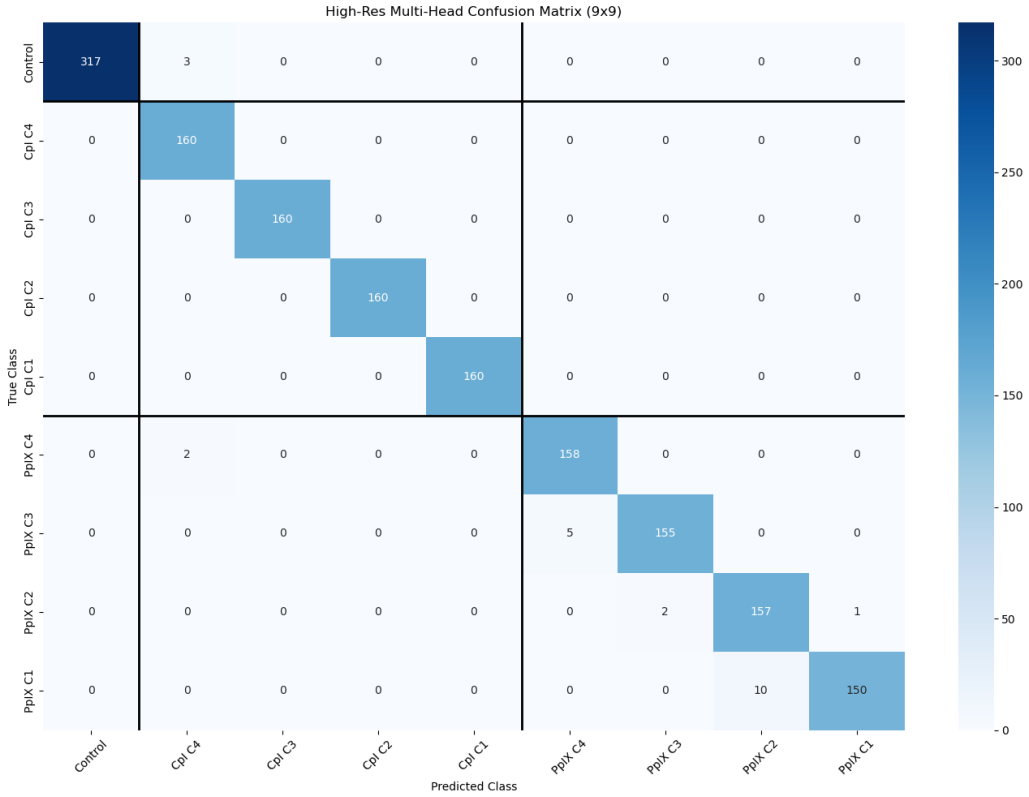


Figure A.3: Confusion matrix for the optimal standard 1D-CNN utilizing the SNV transformation (CH1). Achieving a peak baseline accuracy of 99%, this configuration successfully utilizes statistical standardization to isolate relative morphological features. The matrix demonstrates a near-perfect diagonal, completely resolving the trace concentration overlaps seen in the Source Normalization pipeline.

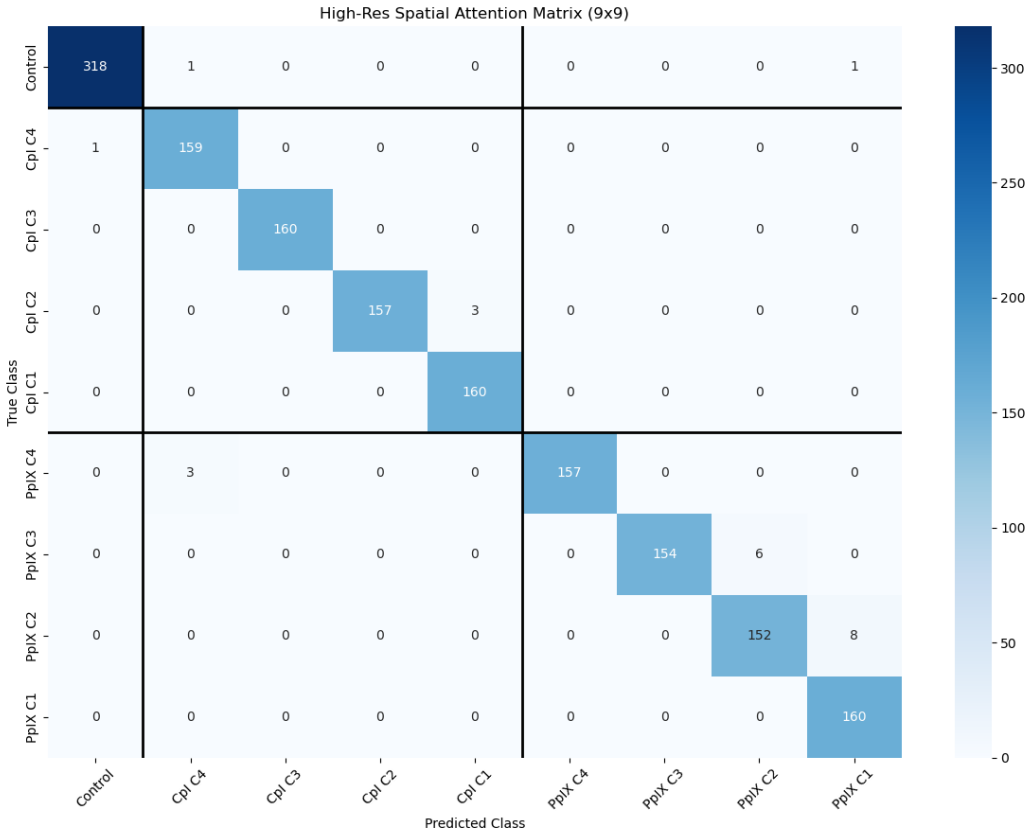


Figure A.4: Confusion matrix for the optimal Channel Attention 1D-CNN utilizing Wing Normalization (CH2). By expanding the input tensor to include the first derivative of the spectrum, the attention mechanism is provided with explicit shape-change features. This structural upgrade "rescues" the pipeline, eliminating the trace boundary failures observed in Figure A.2 and restoring the network to a 99% global accuracy.

B

Transformer Attention and Classification Performance

This appendix provides comprehensive evaluation metrics and interpretability visualizations for the Spectral Transformer models evaluated in this study. Specifically, it contrasts the absolute extremes of the Transformer architectural configurations: the lowest-performing pipeline utilizing SNV transformed single-channel data (CH1), and the highest-performing pipeline utilizing Wing Normalized dual-channel data (CH2).

To demystify the "black-box" nature of the Transformer and validate its physical feature extraction, self-attention maps were generated for the highest concentration class (C1) of each biomarker (CpI and PpIX). In these visualizations, the blue curve represents the standardized spectral intensity input presented to the network. Overlaid in red is the normalized attention score (scaled from 0 to 1). These attention spikes mathematically indicate exactly which wavelength indices the self-attention mechanism weighted most heavily when making its final classification decision.

A physically robust model will demonstrate sharp, localized attention weights directly overlapping with the true biological emission peaks and distinct morphological features of the porphyrins. Conversely, a poorly optimized model will exhibit erratic, scattered attention across irrelevant background indices, correlating directly with high off-diagonal error rates in its respective confusion matrix.

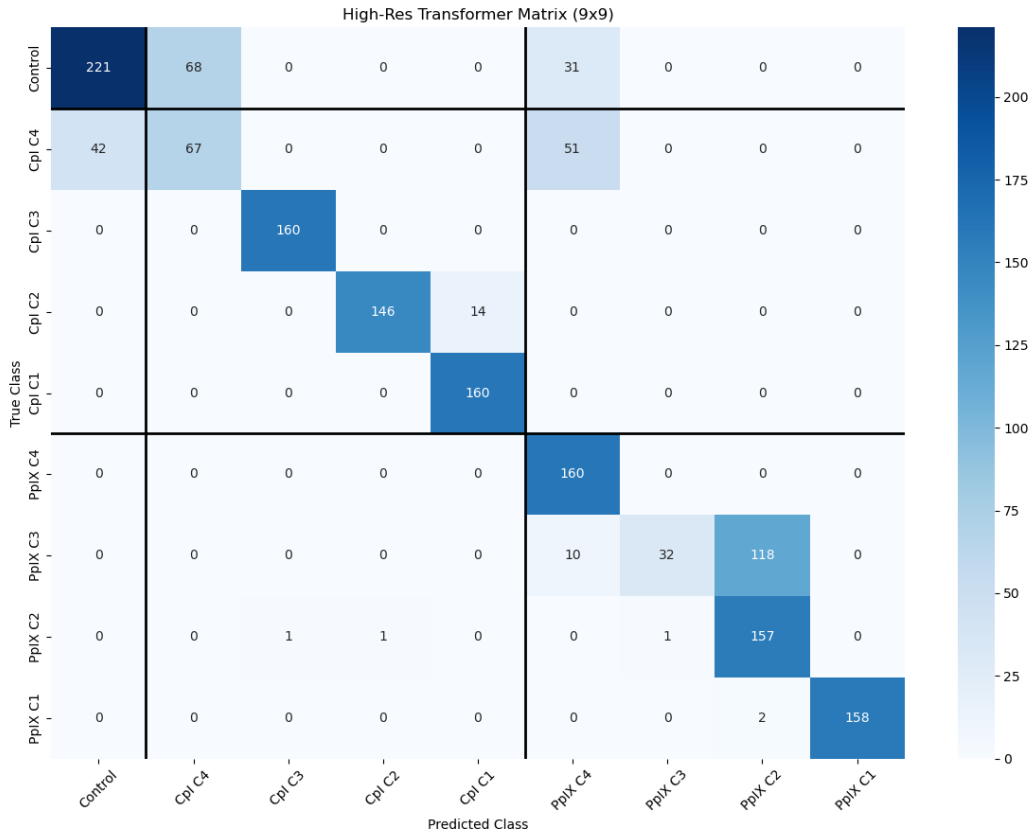


Figure B.1: Confusion matrix for the lowest-performing Transformer configuration (SNV CH1). The network struggles significantly with absolute boundary mapping, exhibiting severe cross-class confusion. Most notably, a large portion of baseline Control samples are erroneously classified as high-concentration CpI (C4), and adjacent concentration bands suffer from high bleed-over.

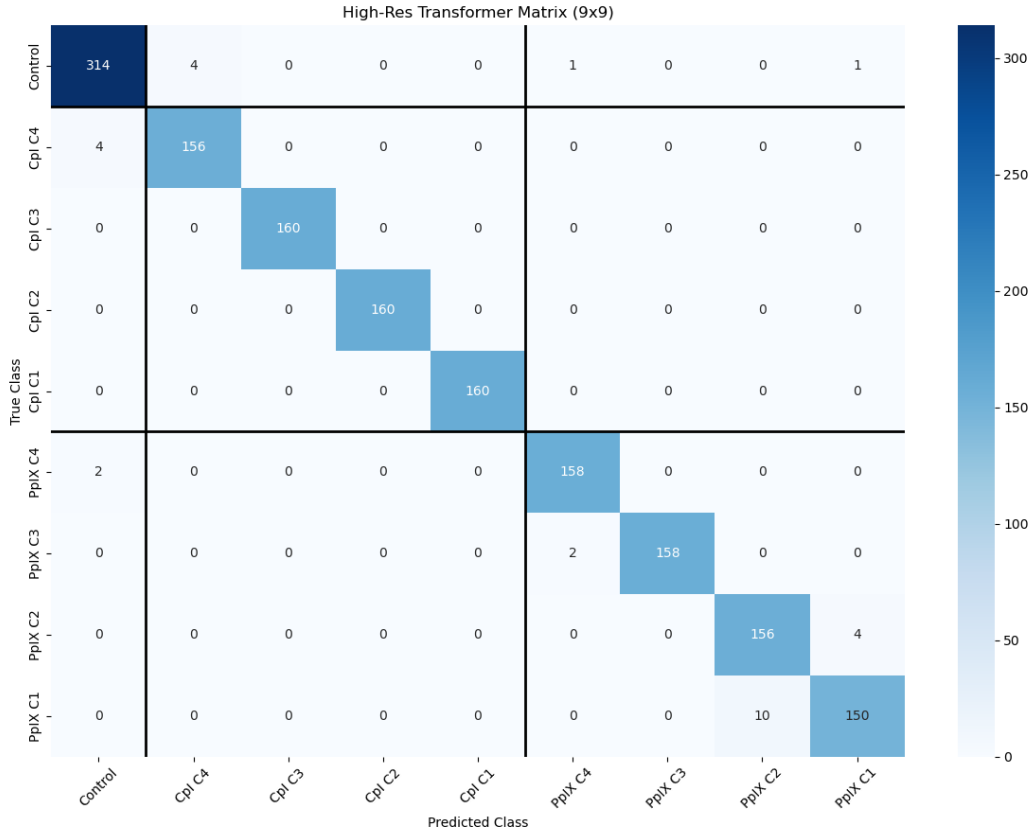


Figure B.2: Confusion matrix for the highest-performing Transformer configuration (Wing Normalization CH2). By preserving relative amplitude spacing and utilizing dual-channel data, the model achieves a near-perfect diagonal alignment, resolving the complex non-linear optical overlaps with exceptional precision.

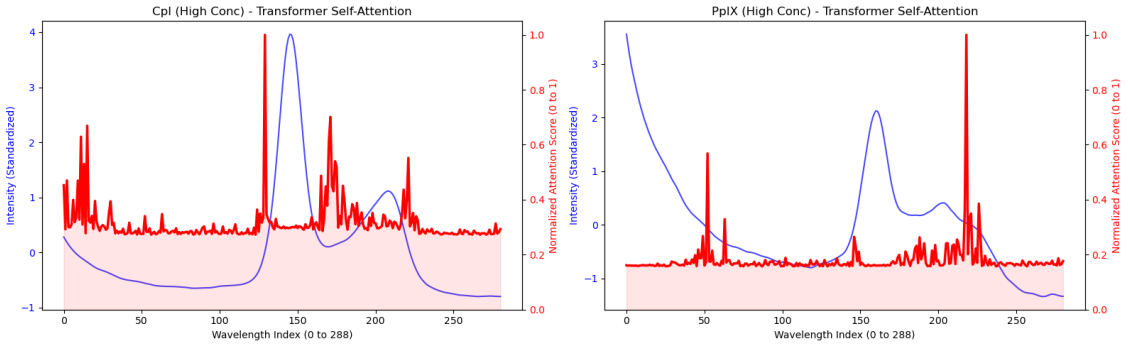


Figure B.3: Internal self-attention maps for the SNV CH1 Transformer. The red attention scores are highly erratic and dispersed across the entire spectral window, indicating that the network failed to converge on the true physical biomarkers (the blue peaks). This scattered attention directly explains the severe misclassifications observed in Figure B.1.

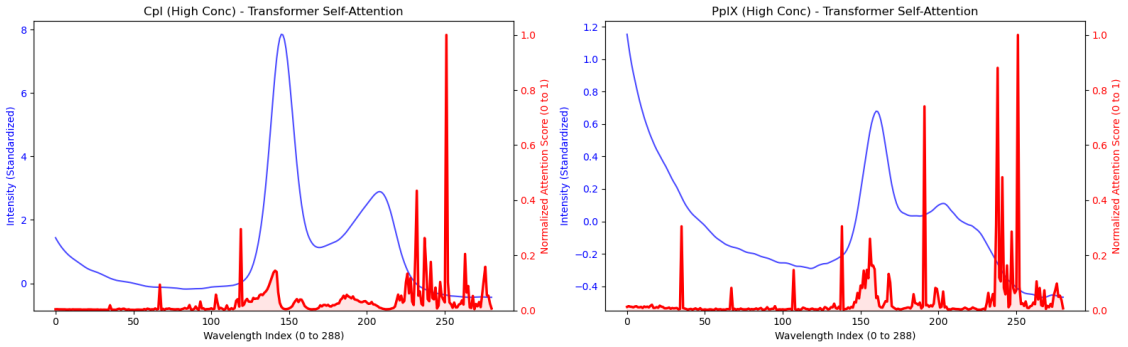


Figure B.4: Internal self-attention maps for the Wing Normalization CH2 Transformer. In stark contrast to the SNV pipeline, the attention mechanism here successfully isolates and heavily weights the mathematically distinct features of the spectra. The sharp red attention spikes align precisely with critical slopes, secondary shoulders, and primary emission peaks, proving that the deep learning architecture is triggering on genuine optical physics.

C

Alternative Low-Resolution Pipelines

This appendix provides the comprehensive class-wise evaluation metrics for the alternative preprocessing pipelines evaluated on the 14-channel low-resolution sensor. While the Backscatter pipeline was utilized as the primary representative model in the main text to illustrate the optimization trajectory, establishing complete methodological transparency requires documenting the exact behavior of the PassThrough and VisNormalizer pipelines.

Specifically, this section contrasts the raw 9×9 baseline architectures against the fully optimized Dynamic Reliability Switch (DRS) architectures applied to the realigned 5×5 Clinical Triage matrix.

By mapping the exact class-wise error distributions, these matrices physically demonstrate the convergence phenomenon detailed in Section 4.3.3. While the PassThrough and VisNormalizer pipelines initially exhibited unique failure modes and severe noise-floor vulnerabilities on the highly granular 9-class task, aligning the diagnostic targets with the sensor's physical separability limits allowed both pipelines to converge at a highly robust diagnostic parity (97% and 98%, respectively).

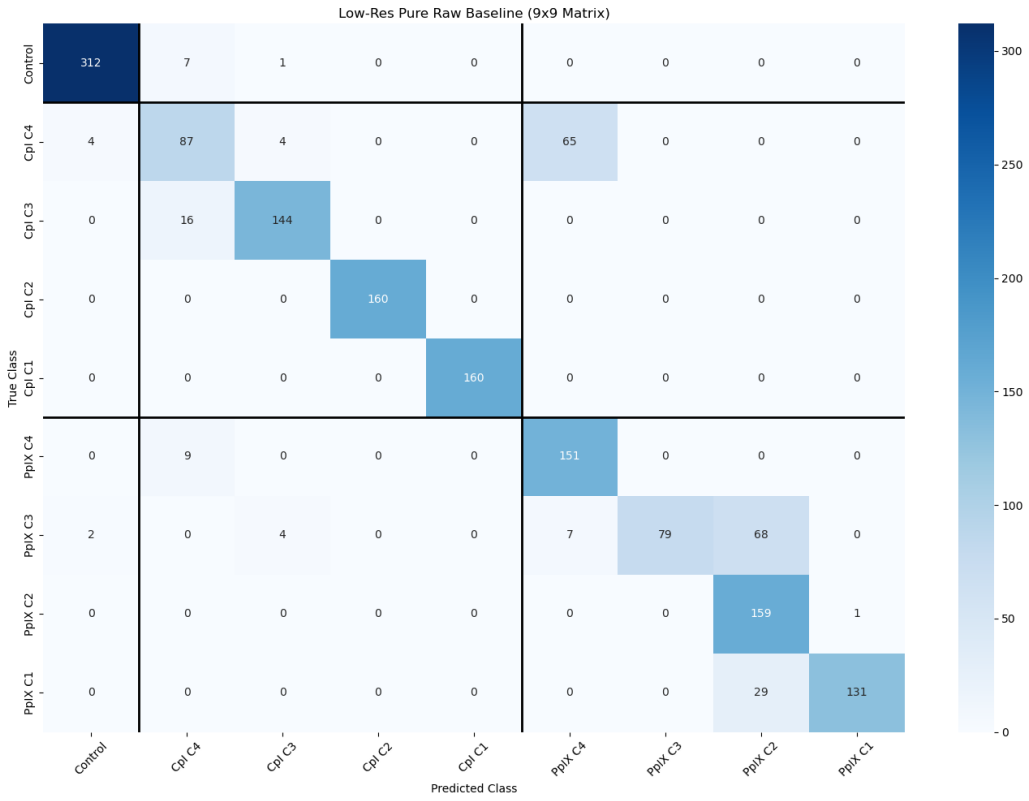


Figure C.1: Confusion matrix for the raw baseline 1D-CNN utilizing the PassThrough pipeline on the 9×9 task. Yielding an 86% global accuracy, the unscaled hardware counts present immediate physical boundary issues. The model struggles with absolute intensity variations, evident in the bleeding between Control and CpI C4, as well as the mid-to-high concentration quenching boundaries for both PpIX and CpI.

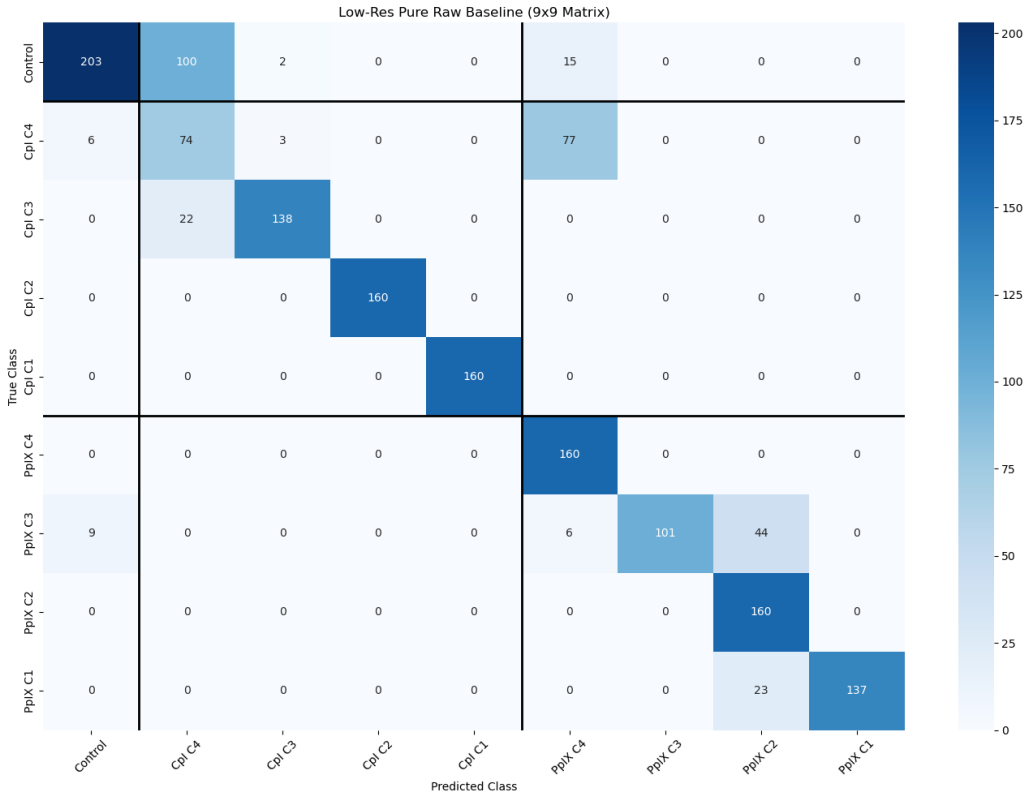


Figure C.2: Confusion matrix for the raw baseline 1D-CNN utilizing the VisNormalizer pipeline on the 9×9 task. Achieving the lowest baseline accuracy (81%), this matrix highlights the severe vulnerability of broadband-relative scaling to noise floor fluctuations. A critical failure is observed at the baseline boundary, where 100 discrete Control samples are erroneously classified as trace CpI (C4) infections.

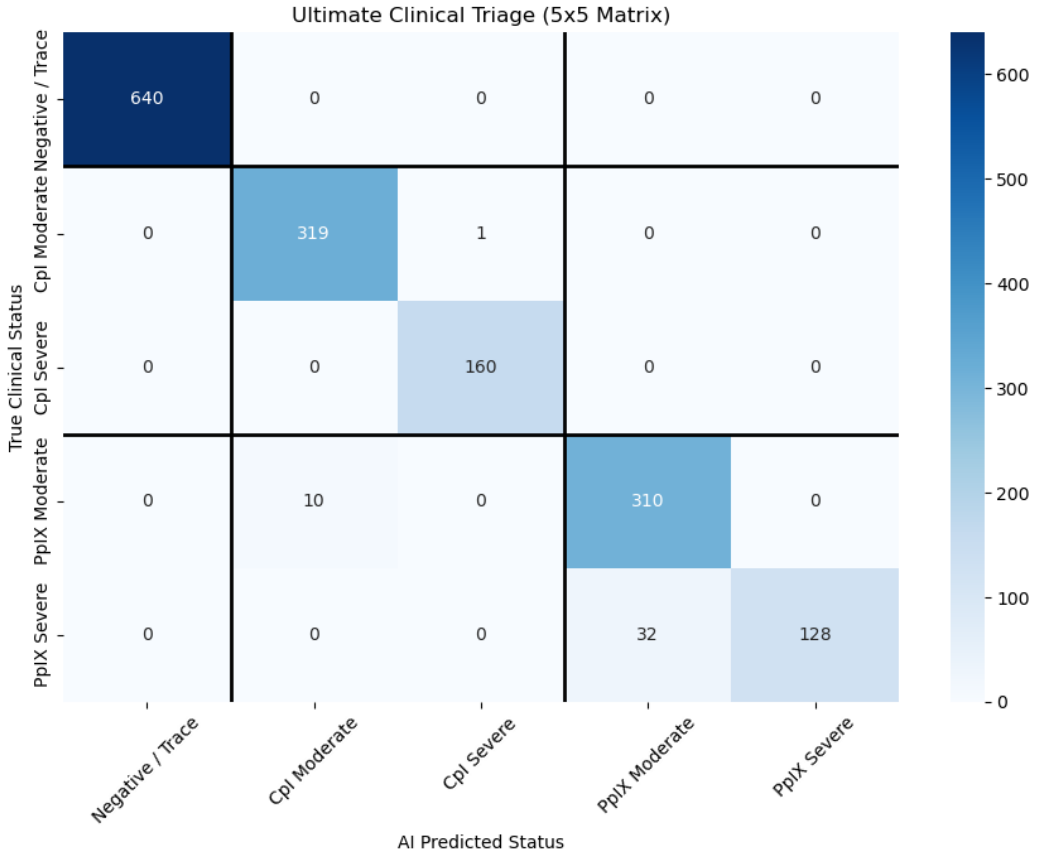


Figure C.3: Final 5×5 Clinical Triage confusion matrix utilizing the fully optimized DRS architecture and the PassThrough pipeline. By applying domain-aware channel attention and merging the physically constrained concentration classes, the absolute intensity errors observed in Figure C.1 are entirely resolved, elevating the global diagnostic accuracy to 97%.

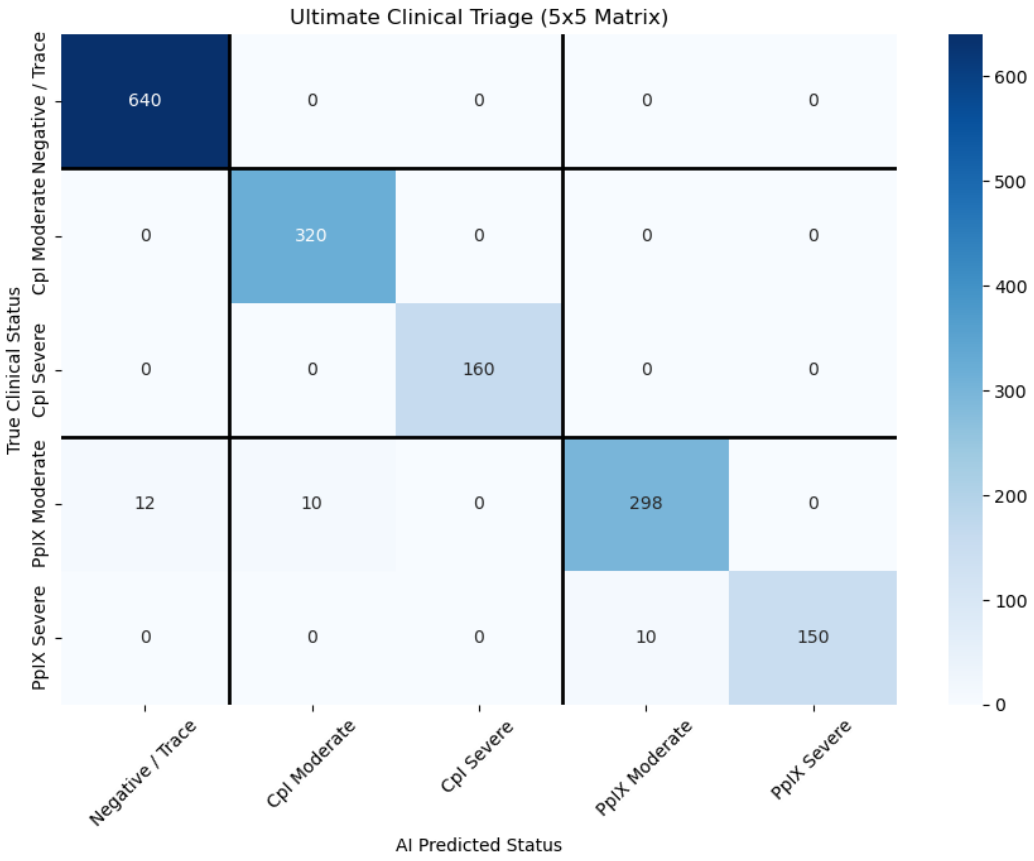


Figure C.4: Final 5×5 Clinical Triage confusion matrix utilizing the fully optimized DRS architecture and the VisNormalizer pipeline. The structural realignment completely eliminates the massive Control/Trace boundary failure previously seen in Figure C.2. Achieving a near-perfect 98% accuracy, this visually confirms that once clinical targets are appropriately aligned with hardware physics, the system's sensitivity to the underlying normalization method is virtually neutralized.

D

Raw Power Profiling Captures

This appendix provides the raw oscillograms captured during the physical hardware power profiling phase using the Nordic Semiconductor PPK2.
Each figure displays the instantaneous current draw (in milliamperes) of the ESP32-S3 over time. To ensure precise temporal alignment, the neural network execution was synchronized via a dedicated GPIO hardware trigger. This trigger signal is visible as the digital logic trace (Channel 7) at the bottom of each screenshot, which was driven high immediately prior to each inference call and pulled low strictly upon completion.
The shaded selection window in each plot was manually aligned to this active logic pulse. The companion software then automatically integrated the area under the analog current curve strictly within this active window to calculate the exact total charge transferred (in millicoulombs or microcoulombs) during the inference event.
A consistent artifact across the longer inference captures is the presence of perfectly periodic, sharp current spikes occurring exactly every 10 ms. These spikes represent the underlying Real-Time Operating System (RTOS) scheduler tick (configured to a 100 Hz tick rate), where the CPU briefly interrupts the neural network execution to update system timers and evaluate task states.

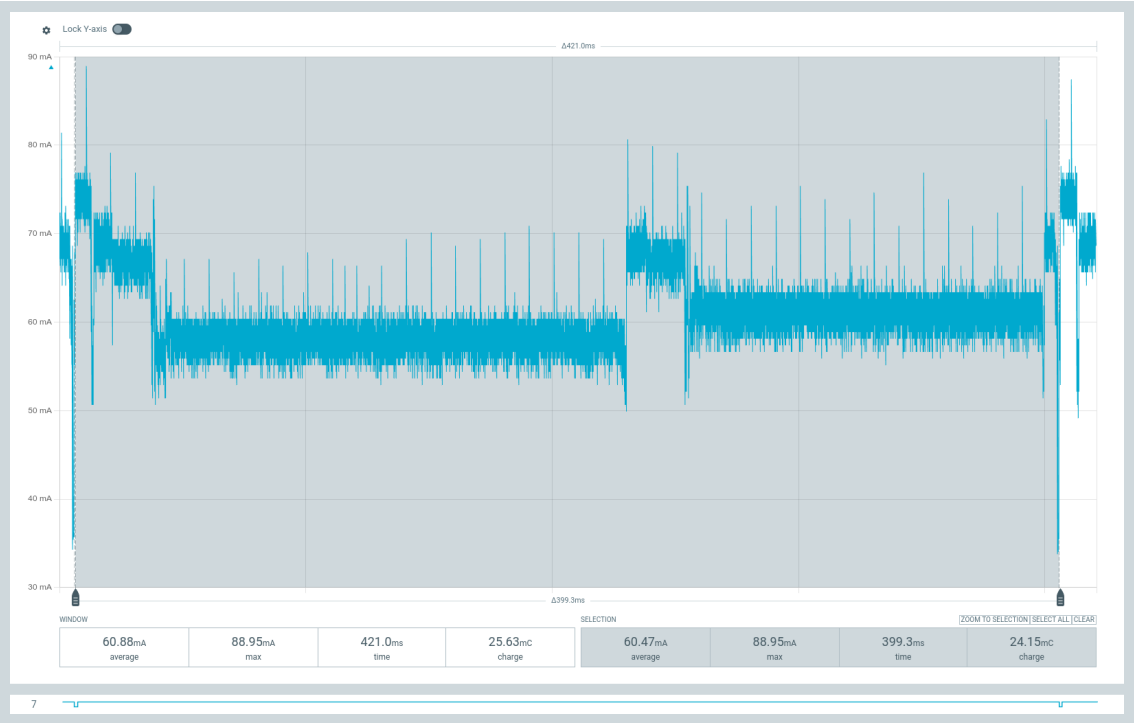


Figure D.1: Power profile of the HR Standalone *base_3ch* model using the FLASH+PSRAM memory configuration. The severe latency penalty of external memory access is evident in the extended 399 ms execution time. The lower average current (60.47 mA) indicates frequent CPU stalling as it waits to fetch weights and tensor arena data over the SPI buses.

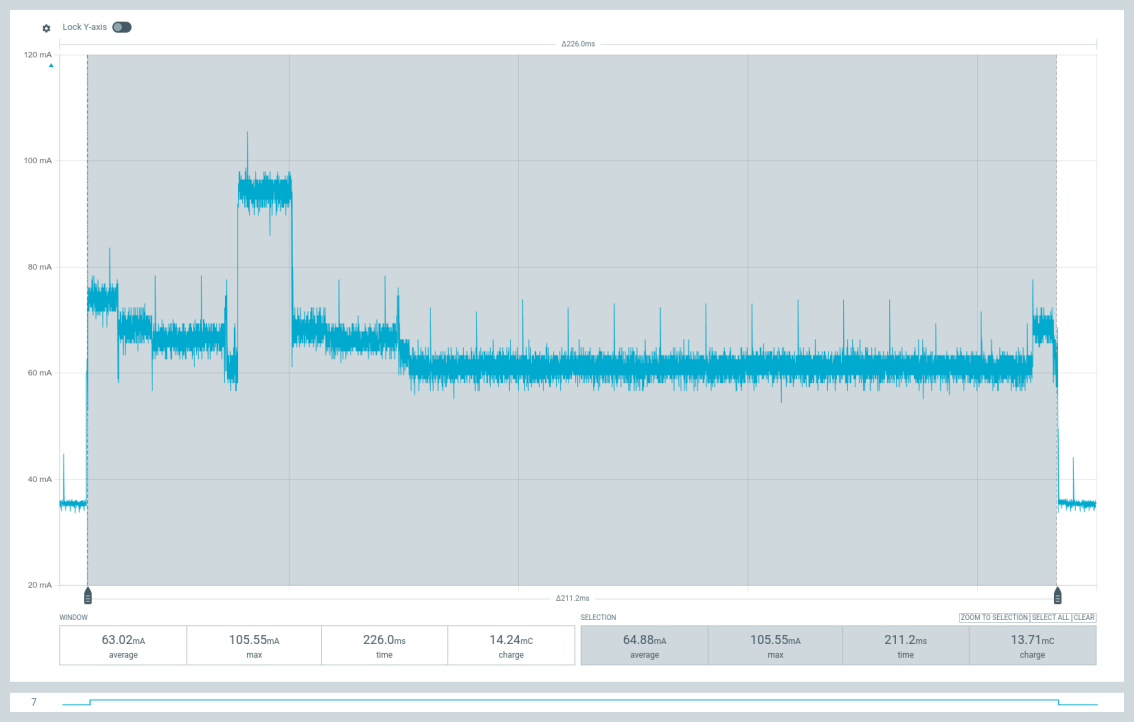


Figure D.2: Power profile of the HR Standalone *base_3ch* model using the FLASH+SRAM memory configuration. With the Tensor Arena localized to internal SRAM, the CPU wait-states are significantly reduced, improving the execution time to 211 ms.

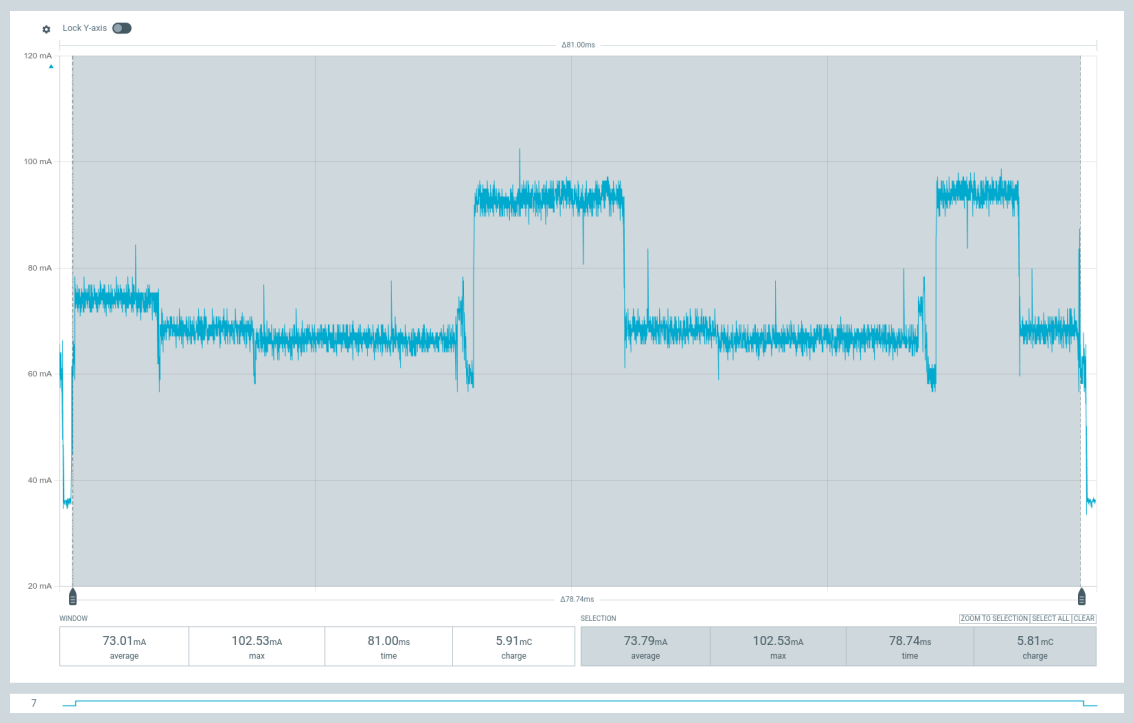


Figure D.3: Power profile of the HR Standalone *base_3ch* model using the SRAM+SRAM memory configuration. With all data localized internally, the CPU operates continuously at maximum computational efficiency without SPI-induced stalling. This results in the highest average current draw (73.79 mA) but drastically reduces the execution time to 81 ms, yielding the lowest overall energy footprint for this architecture.

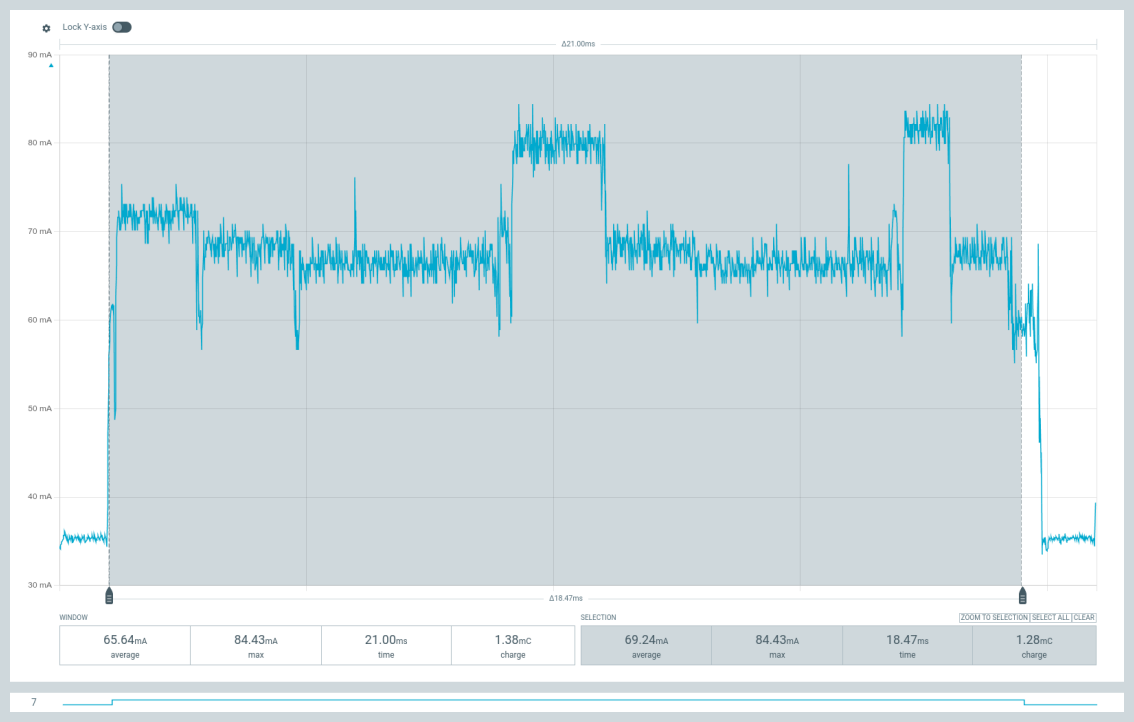


Figure D.4: Power profile of the Pareto-optimal HR Standalone *mini_1ch* model utilizing the realistic FLASH+SRAM deployment configuration. Due to aggressive structural pruning, the model fits almost entirely within the ESP32-S3's cache, allowing it to complete inference in 21 ms with minimal external bus penalty.

Figure D.5: Power profile of the LR Clinical Triage *micro* model utilizing the FLASH+SRAM configuration. The sub-millisecond execution time with inference taking < 1 ms) demonstrates the efficiency of this heavily optimized architecture, requiring only 137.66μC of charge.

DEPARTMENT OF ELECTRICAL ENGINEERING
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY