# A Performance Evaluation of Feature-Based and Generative Anomaly Detection Paradigms under CPU Resource Constraints

Mouhsine Hibat Allah- Abdelaziz Ait Moussa
Department of Computer Science- Faculty of Sciences-Mohammed First University-Oujda-
Maroc

## ABSTRACT

Unsupervised visual anomaly detection in industrial and clinical edge environments faces a critical dual challenge: the absolute scarcity of anomalous training samples and the prohibitive deployment costs of high-power GPU infrastructures. To bridge this gap, this paper presents a comprehensive evaluation and deployment framework that cross-analyzes two major deep learning paradigms—generative reconstructive models (Autoencoder, VAE, GANomaly, and Deep SVDD) and non-parametric feature-based descriptors (PaDiM, Patch-Based, and PatchCore)—optimized exclusively for resource-constrained CPU environments. Rather than relying on expensive hardware scaling, our methodology leverages transfer learning via a frozen ResNet34 backbone, combined with geometric Coreset compression and high-performance vector indexing through the FAISS CPU library. Extensive cross-evaluation on the MVTec AD, Industrial Rock, and Chest X-Ray datasets reveals a stark architectural polarization. While generative architectures struggle with semantic representations—manifesting a severe performance degradation on complex clinical textures with a Variational Autoencoder (VAE) AUROC dropping to 0.3301 due to the *Blurry Reconstruction Syndrome*—the optimized non-parametric feature models achieve a near-perfect AUROC of 0.9991. Crucially, through rigorous software engineering, our unified ecosystem reduces inference latency to an ultra-


© Authors. Licensed under Creative Commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0) non-commercial reuse with attribution
https://creativecommons.org/licenses/by-nc/4.0/

fast sub-millisecond threshold (< 0.8 ms per frame) on conventional, low-power desktop hardware. Finally, this work formalizes the theoretical boundaries of these paradigms by contrasting the *Blurry Reconstruction Syndrome* against the metrological paradox of the *Gray Zone*, where global scalar anomaly score reductions produce marginal false negatives despite flawless spatial pixel-level localization via heatmaps. This benchmark establishes a new blueprint for eco-efficient, highly sustainable One-Class AI deployment at the industrial edge.

**Keywords:** *Unsupervised Anomaly Detection, One-Class Classification, CPU Resource Constraints, Non-Parametric Feature Descriptors, Generative Reconstruction Models, Coreset Subsampling, Blurry Reconstruction Syndrome, Gray Zone Paradox.*


© Authors. Licensed under Creative Commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0) non-commercial reuse with attribution
https://creativecommons.org/licenses/by-nc/4.0/

## I. INTRODUCTION & HARDWARE-CONSTRAINED COMPUTATIONAL CHALLENGES

The operational deployment of deep learning models in smart manufacturing environments and remote healthcare facilities is bound by severe economic, structural, and computational limitations. In practical applications, the extreme asymmetry of available data—manifested as an absolute scarcity and unpredictable nature of anomalous samples, manufacturing defects, or rare medical pathologies—renders traditional supervised classification approaches entirely ineffective. This structural reality motivates the adoption of the One-Class Classification (OCC) paradigm, where visual normality is modeled exclusively using normal data [2]. By shifting the learning objective to anomaly detection via deviation from a learned distribution of normality, OCC systems eliminate the statistical dependence on anomalous samples [2]. This methodology provides a crucial mechanism for isolating unforeseen defects in complex, real-world production environments.

However, the contemporary scientific literature predominantly focuses on highly over-parameterized deep neural network architectures that require high-end, power-hungry Graphics Processing Units (GPUs) for both their development and deployment phases. This heavy hardware reliance creates a prohibitive financial and technical barrier for end-users in small industrial facilities, local workshops, and underserved medical centers, who lack the budget to acquire and maintain specialized AI server hardware. To address this socio-technical divide, this work investigates a resource-efficient paradigm from an engineering perspective: designing tools where both the primary configuration loops and real-time execution run exclusively on the existing Central Processing Unit (CPU) infrastructure of standard, low-cost desktop workstations. By purposefully eliminating the need for dedicated hardware accelerators at the deployment site, engineers can deliver sophisticated, pre-configured visual inspection systems that facilities can install with virtually zero additional infrastructure overhead, fostering long-term eco-efficiency and operational sustainability.


© Authors. Licensed under Creative Commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0) non-commercial reuse with attribution
https://creativecommons.org/licenses/by-nc/4.0/

To provide a definitive architectural blueprint, the remainder of this paper is structured to offer an exhaustive, cross-paradigm evaluation. Section II establishes a formal taxonomy, dividing the seven implemented architectures into generative reconstructive models (Autoencoder, VAE, GANomaly, Deep SVDD) and non-parametric feature descriptors (PaDiM, Patch-Based, PatchCore) to contrast their mathematical foundations. Section III details the underlying hardware topologies, multi-domain industrial datasets, and hyperparameter configurations. Section IV delivers extensive quantitative benchmarks alongside rigorous CPU telemetry profiling, tracking image-level accuracy against hardware execution speeds. Section V introduces a deep phenomenological analysis, mapping the fundamental boundaries of these methods by contrasting the *Blurry Reconstruction Syndrome* against the metrological paradox of the *Gray Zone*. Finally, Section VI outlines future optimization and deployment directions, leading to our final conclusions in Section VII.

## II. TAXONOMY AND ARCHITECTURAL PIPELINES OF THE EVALUATED PARADIGMS

To establish a comprehensive comparative study, seven computer vision architectures were categorized and optimized according to two fundamental, structurally opposing paradigms. The core objective of this taxonomy is to contrast their structural execution paths, computational workflows, and optimization objectives under identical hardware deployment restrictions. While the first paradigm attempts to map the underlying data distribution using parametric network weights trained entirely from scratch, the second leverages pre-trained, frozen feature representations to build non-parametric spatial registries of visual normality.

### 1) Generative and Projection-Based Reconstructive Approaches

Generative and projection-based models operate under the core assumption that a network trained exclusively on normal samples will fail to accurately reconstruct


© Authors. Licensed under Creative Commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0) non-commercial reuse with attribution
https://creativecommons.org/licenses/by-nc/4.0/

unforeseen anomalous patterns, thereby penalizing unseen defects through localized objective functions.

i. *Pixel-Level Reconstruction*: Autoencoders (AE) and Variational Autoencoders (VAE).

These networks compress the input image through a deterministic or probabilistic latent bottleneck before attempting its spatial reconstruction using transposed convolutions. The pixel-wise Mean Squared Error (MSE) serves as the primary optimization objective.

- **Convolutional Autoencoder (AE)**: This architecture compresses the input into a deterministic latent representation. The anomaly decision is based directly on the pixel-wise reconstruction error, which penalizes the model's inability to accurately replicate unfamiliar visual structures. Because of its straightforward structural design, the standard AutoEncoder maintains a compact footprint with a checkpoint size of **745 KB**.

- **Variational Autoencoder (VAE)**: This approach replaces the deterministic bottleneck with a continuous probabilistic distribution parameterized by a latent mean vector $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma}$ [1]. Its composite loss function combines the reconstruction MSE with a Kullback-Leibler Divergence (KLD) term, regulated by a strict regularization weighting factor of $\boldsymbol{\beta}$= 0.0005 to enforce continuity and smoothness within the latent space. The probabilistic overhead results in a slightly increased checkpoint size of **812 KB**.

ii. *Latent Representation and Adversarial Constraints*: GANomaly and Deep SVDD.

To prevent networks from memorizing pixel-wise shortcuts, adversarial and projection paradigms incorporate structural constraints that enforce semantic consistency across latent fields.

- **GANomaly**: This framework employs an adversarial, dual-generator pipeline structured as an Encoder–Decoder–Encoder loop coupled with a


© Authors. Licensed under Creative Commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0) non-commercial reuse with attribution
https://creativecommons.org/licenses/by-nc/4.0/

discriminator. Its hybrid, composite objective function simultaneously minimizes the discrepancy between the input image and its reconstruction (contextual loss), the difference between the original latent representation and its reconstructed counterpart (latent MSE loss), and the adversarial authenticity loss under a unified MSE optimization framework, yielding a model checkpoint size of **1.23 MB**.

- **Deep SVDD (Support Vector Data Description)**: Bypassing the pixel reconstruction stage entirely, this method projects samples into a compact, 64-dimensional latent feature space by minimizing the average Euclidean distance (Mean Distance) between normal embeddings and a fixed reference center matrix (center_matrix). A strict margin hyperparameter of $\nu = 0.1$ is enforced during optimization, allowing the model to tolerate approximately **10%** of the training data as potential outliers, producing a fixed model size of **1.01 MB**.

### 2) Non-Parametric Feature-Based Descriptor Approaches

Unlike reconstruction networks, the feature-based paradigm bypasses the computationally expensive training-from-scratch phase. Instead, it exploits rich, multi-level semantic representations extracted by a robust ResNet34 feature extractor pre-trained on the ImageNet dataset [4]. Three distinct mathematical architectures were integrated into a polymorphic wrapper to process these embeddings.

#### i. *Spatial Density Modeling*: PaDiM.

The Patch Distribution Modeling (PaDiM) framework extracts local feature embeddings from intermediate activation layers (specifically Layer 2 and Layer 3) of the frozen backbone network [7]. It estimates a local multivariate Gaussian distribution defined by a mean $\mu$ and an inverse covariance matrix $\Sigma^{-1}$ over a fixed **29 x 29** spatial grid. Normality is modeled independently at each spatial location


© Authors. Licensed under Creative Commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0) non-commercial reuse with attribution
https://creativecommons.org/licenses/by-nc/4.0/

across the grid, allowing it to capture cross-layer correlations without runtime training, resulting in a fixed, data-independent checkpoint size of **125 MB**.

ii. ***Sub-Image Memory Quantization*: Patch-Based and PatchCore (with Coreset subsampling).**

These frameworks segment images into localized spatial sub-regions to explicitly regularize and index local feature patches.

- **Patch-Based**: This architecture performs spatial scanning through local patch convolutions using an adaptive stride [6]. To prevent excessive RAM consumption when scaling to large datasets, the implementation dynamically adjusts the geometric displacement step. A dense stride (Stride = 1) is maintained for fine-grained, highly heterogeneous clinical images (resulting in a checkpoint size of **142 MB**), whereas a sparse stride (Stride = 3) is deployed for the industrial MVTec AD dataset, reducing the checkpoint size to **79.1 MB**.
- **PatchCore**: Representing the state-of-the-art flagship architecture of this framework, PatchCore extracts local neighborhood mid-level patch features to form an exhaustive spatial memory bank [3]. To overcome the memory explosion caused by processing the **3.8** million embedding patches inherent to datasets like MVTec AD, PatchCore applies a greedy geometric selection algorithm known as Coreset Subsampling. This mathematical filtering process reduces the memory bank to critical sampling ratios (evaluated at **0.05, 0.25,** or **0.01**) while preserving the absolute external boundaries of the nominal distribution. The optimized memory bank (occupying **92 MB** at a **0.05 ratio** and **339 MB** at a **0.25 ratio**) is stored within an asynchronous faiss.IndexFlatL2 vector index managed safely via a dedicated QThread execution loop.

**3) Theoretical Contrast: Latent Reconstruction vs. Non-Parametric Descriptors**


© Authors. Licensed under Creative Commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0) non-commercial reuse with attribution
https://creativecommons.org/licenses/by-nc/4.0/

The mathematical bifurcation between these two paradigms heavily dictates their execution profiles on resource-limited CPU architectures. Parametric generative approaches depend entirely on localized pixel-wise objective functions (such as MSE) during optimization. While highly computationally efficient during inference due to straightforward, single-pass forward tensor operations, they are prone to structural generalization faults where anomalous details are inadvertently smoothed or poorly reconstructed. Conversely, non-parametric feature descriptors completely substitute runtime network weight updates with extensive nearest-neighbor distance computations. Consequently, their CPU workload shifts away from floating-point tensor arithmetic toward intensive memory lookup, matrix inversion, and vector indexing operations. This fundamentally alters the physical hardware bottlenecks from raw compute bounds to memory bandwidth constraints.

## III. EXPERIMENTAL ENVIRONMENT, INFRASTRUCTURE, AND CONFIGURATIONS

To guarantee empirical reproducibility and establish a fair baseline for cross-paradigm evaluation, this section details the underlying physical testing infrastructure, the structural properties of the target multi-domain datasets, and the exact hyperparameter configurations utilized across all seven models.

### 1) Hardware Topology: Technical specifications of the targeted resource-constrained CPU Testbeds

Unlike traditional benchmarking studies that rely on heavy, server-grade hardware acceleration, the deployment matrix of this study intentionally confines all computing loops to standard consumer-grade, low-power desktop hardware. The primary evaluation environment consists of an Intel Core i5 or i7 consumer-class processor (operating within standard low-wattage thermal design profiles) equipped with up to 16 GB of DDR4 system RAM. By design, no dedicated Graphics Processing Units (GPUs) or specialized TPU accelerators are utilized during either the training phase (for parametric generative models) or the memory


© Authors. Licensed under Creative Commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0) non-commercial reuse with attribution
https://creativecommons.org/licenses/by-nc/4.0/

indexing phase (for non-parametric feature models). This rigid hardware capping ensures that all recorded execution metrics reflect real-world edge deployment scenarios in small-scale manufacturing facilities and remote clinics where legacy CPU workstations represent the only available computing infrastructure.

2) **Dataset Selection: Multi-domain industrial benchmarks**

The structural robustness and generalized adaptability of the seven integrated models are rigorously cross-evaluated across three distinct, highly challenging visual domains:

- **MVTec Anomaly Detection (MVTec AD) Benchmark**: Comprising 15 distinct categories divided into 5 texture domains (such as grid, leather, and wood) and 10 object domains (such as capsule, screw, and transistor), this dataset serves as the primary industrial quality control baseline. It features a vast training registry of nominal, defect-free images, paired with an extensive test suite containing diverse real-world anomalous structural deformations, scratches, and contaminations. [1, 2, 3, 4, 5]
- **Industrial Rock Texture Dataset**: This domain represents highly complex, stochastic, and homogeneous mineral surfaces. Because mineral textures lack rigid geometric boundaries, they challenge the models' capacity to differentiate normal natural structural variations from actual structural anomalies, providing a strict test for texture boundary definition.
- **Chest X-Ray Clinical Dataset**: Transitioning from industrial inspection to medical computer-aided diagnosis, this dataset introduces non-rigid, organic anatomical structures. It evaluates the models' capacity to handle high intra-class variance and subtle, low-contrast grayscale anomalies (such as pulmonary opacities or rare pathological lesions) where localized structural margins are highly ambiguous.


© Authors. Licensed under Creative Commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0) non-commercial reuse with attribution
https://creativecommons.org/licenses/by-nc/4.0/

3) **Hyperparameter Settings: Training configurations, embedding sizes, and quantization matrices across models**

To ensure structural equity during performance profiling, all input images across all datasets are standardized via bilinear interpolation to a unified, static resolution of **224 × 224 pixels**. This resolution is maintained globally across both training and evaluation passes to preserve local texture integrity while ensuring a uniform computational load.

- **Generative Models (AE, VAE, GANomaly, Deep SVDD)**: These models are trained entirely from scratch using the Adam optimizer with an internal batch size of **16** over a maximum calibration envelope of **15 epochs** for GANomaly and **25 epochs** for AE and VAE. For the standard AutoEncoder, Variational Autoencoder (VAE), and Deep SVDD, a fixed baseline learning rate of $\eta = 10^{-4}$ is applied. In contrast, the adversarial optimization loop of GANomaly (GAN-Based) utilizes a higher, specialized learning rate of $\eta = 2 \times 10^{-4}$ to balance generator-discriminator stability. For the VAE, the latent space distribution is regularized using a Kullback-Leibler weighting coefficient of $\beta = 0.0005$. For Deep SVDD, the target projection manifold is mapped to a highly compact 64-dimensional latent feature vector, enforcing an outlier tolerance margin parameter of $\nu = 0.1$.
- **Non-Parametric Models (PaDiM, Patch-Based, PatchCore)**: These methods bypass training loops entirely by leveraging a frozen ResNet34 backbone pre-trained on ImageNet. For PaDiM, local feature embeddings are extracted from Layer 2 and Layer 3 activations, projecting down to a fixed spatial grid where a local multivariate Gaussian distribution $\mu$ and $\Sigma^{-1}$ is calculated. For Patch-Based modeling, the geometric displacement step is dynamically adjusted via an adaptive stride. A dense stride (Stride = 1) is maintained for fine-grained clinical images (resulting in a checkpoint size of **142 MB**), whereas a sparse stride (Stride = 3) is deployed for the industrial MVTec AD dataset, reducing the checkpoint size to **79.1 MB**. For


© Authors. Licensed under Creative Commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0) non-commercial reuse with attribution
https://creativecommons.org/licenses/by-nc/4.0/

PatchCore, neighborhood mid-level patch features are pooled into an extensive memory bank. This memory bank is dynamically compressed using a greedy geometric coreset selection algorithm evaluated at critical sampling ratios of **0.05** and **0.25**, before being stored within an asynchronous faiss.IndexFlatL2 vector index managed safely via a dedicated QThread execution loop.

## IV. QUANTITATIVE BENCHMARKS AND CPU TELEMETRY PROFILING

To rigorously assess the performance profiles of the parametric and non-parametric paradigms under strict computing ceilings, this section presents a unified statistical and computational analysis. All input datasets—comprising the MVTec AD benchmark, the Industrial Rock texture registry, and the Chest X-Ray clinical suite—were statically processed at an offline resolution of **224 × 224 pixels** using a fixed evaluation batch size of **16**. *The following Table* aggregates the complete empirical landscape, mapping statistical classification accuracy directly against CPU-bound training and serialization metrics.

| Model Category | Specific Architecture | Industrial Rock AUROC | MVTec AD Image AUROC | Chest X-Ray AUROC | MVTec AD Patch AUROC | CPU Training / Setup Time (Rock / MVTec) | Checkpoint Size (.pt) |
|---|---|---|---|---|---|---|---|
| Generative / Reconstructive | AutoEncoder (AE) | 0.9346 | 0.5462 | 0.4157 | — | 7 min / 32 min | 745 KB |
| | Variational AE (VAE) | 0.8910 | 0.5060 | 0.3301 | — | 13 min / 51 min | 812 KB |


© Authors. Licensed under Creative Commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0) non-commercial reuse with attribution
https://creativecommons.org/licenses/by-nc/4.0/

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | GANomaly | 0.9922 | 0.5135 | 0.5044 | — | 15 min / 60 min | 1.23 MB |
| | Deep SVDD | 0.9553 | 0.5641 | 0.5270 | — | 7 min / 29 min | 1.01 MB |
| Non-Parametric Feature | PaDiM | 0.9687 | 0.7197 | 0.7102 | 0.7412 | <2 min / 7 min | 125 MB |
| | Patch-Based ($L_2$) | 0.9966 | 0.8412 | 0.8491 | 0.8533 | 14 min / 4 min | 142 MB / 79.1 MB |
| | PatchCore (Coreset 0.25) | 0.9991 | 0.9013 | 0.8625 | 0.9104 | 25 min / 15 min | 339 MB |

*Table: Comprehensive Summary of Cross-Paradigm AUROC Performance and CPU Telemetry Profiling*

**1) Statistical Detection Accuracy: Image-level and Pixel-level Area Under the Receiver Operating Characteristic (AUROC)**

The empirical results expose a stark polarization in performance between the two foundational paradigms. Parametric generative architectures display remarkable proficiency when restricted to highly homogeneous, stochastic surfaces, as evidenced by GANomaly capturing geological structures on the Industrial Rock dataset with a top generative AUROC of **0.9922**. However, this proficiency degrades severely when confronted with the multi-class topological complexity of the MVTec AD dataset, where generative AUROC values stall near a baseline of 0.54. This degradation reaches a critical failure mode on the clinical Chest X-Ray dataset, where the VAE records an AUROC of just **0.3301**. This represents a severe inversion of class separability that thoroughly compromises diagnostic utility.

Conversely, the non-parametric feature-memorization paradigm consistently outperforms reconstruction loops. PatchCore (Coreset 0.25) emerges as the undisputed leading architecture, securing a near-perfect AUROC of 0.9991 on Industrial Rock while maintaining an image-level AUROC of **0.9013** and an


© Authors. Licensed under Creative Commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0) non-commercial reuse with attribution
https://creativecommons.org/licenses/by-nc/4.0/

outstanding pixel-level patch AUROC of **0.9104** on the heterogeneous MVTec AD dataset. This sub-image tracking capability highlights the superior semantic preservation of pre-trained feature banks. Meanwhile, PaDiM reveals localized constraints on the medical dataset, with its AUROC dropping to **0.7102**. This drop stems from the rigid structural limitations of its fixed Gaussian spatial grid, which lacks the geometric flexibility to adapt to the wide biological and anatomical variances characteristic of chest radiographs.

**2) Execution Speed & Latency Analysis: Milliseconds-per-Image (ms/img) Profiling during Inference on CPU**

From a deployment and throughput perspective, translating nearest-neighbor calculations into a viable edge mechanism necessitates rigorous algorithmic engineering. To eliminate the computational latency typically induced by slow, sequential Python loops, our non-parametric framework integrates the vectorized FAISS CPU library, which natively leverages the host processor's hardware-level Advanced Vector Extensions 2 (AVX2) instruction set.

By offloading spatial proximity mapping to optimized C++ matrix subroutines, the average inference latency of our PatchCore implementation drops below **0.8** milliseconds per frame. This lightning-fast processing speed easily surpasses the **24** Frames Per Second (FPS) real-time constraint inside our unified PyQt6 desktop interface. Crucially, this lean open-source implementation—released publicly on GitHub—achieves a near-identical mathematical convergence with the official Amazon Science baseline, manifesting a negligible AUROC delta of only **-0.0017**, validating the structural validity and cross-platform reproducibility of our optimized pipeline.

**3) Telemetry and RAM Footprint: Tracking Peak Memory Consumption and Checkpoint Size Invariance Across Architectures**

Evaluating storage and runtime profiles reveals a fundamental engineering trade-off between network parameter stability and memory consumption. Generative architectures display an exceptionally eco-efficient, compact footprint. Their


© Authors. Licensed under Creative Commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0) non-commercial reuse with attribution
https://creativecommons.org/licenses/by-nc/4.0/

serialized .pt checkpoint files are remarkably small, spanning a tight envelope from a mere **745 KB** (AutoEncoder) to **1.23 MB** (GANomaly). Furthermore, because their network topology is entirely fixed, these checkpoint sizes exhibit strict dataset invariance. They remain totally constant regardless of the underlying training sample size or epoch length, depending solely on the static layout of the convolutional kernels.

The uniformity of our offline preprocessing pipeline is further validated by the highly consistent CPU training intervals logged on structurally equivalent datasets—such as the identical 7-minute training cycles shared by the AutoEncoder and Deep SVDD. Because the spatial dimensions are fixed at **224 × 224 pixels**, the CPU processes exactly 63 execution iterations per epoch, rendering the computational workload independent of the semantic domain. In contrast, non-parametric approaches exchange compact disk storage for accelerated execution. PaDiM demands a fixed checkpoint size of 125 MB to store its dense covariance fields. Similarly, Patch-Based tracking exhibits dataset-dependent scaling, shifting from **142 MB** on dense clinical images (Stride = 1) to 79.1 MB on sparse industrial objects (Stride = 3). PatchCore peaks at **339 MB** under a 0.25 Coreset allocation, proving that on resource-constrained hardware, feature-based precision requires managing a larger memory overhead.

## V. COMPARATIVE PHENOMENOLOGICAL ANALYSIS & CRITICAL LIMITATIONS

The cross-paradigm evaluation under strict CPU hardware constraints exposes deep architectural and mathematical limitations inherent to each methodology. Beyond raw accuracy and latency metrics, real-world deployment inside our unified PyQt6 monitoring interface reveals two distinct theoretical failure modes: a convolutional smoothing bottleneck in generative pipelines and an aggregation dilution paradox in non-parametric frameworks.


© Authors. Licensed under Creative Commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0) non-commercial reuse with attribution
https://creativecommons.org/licenses/by-nc/4.0/

**1) The Blurry Reconstruction Syndrome: High-Frequency Smoothing in Complex Textures**

The severe performance degradation observed when applying generative models to complex biological and multi-class structures (such as the Chest X-Ray and MVTec AD datasets) is driven by a fundamental mathematical phenomenon known as the *Blurry Reconstruction Syndrome*. As illustrated by the convolutional smoothing mechanism analyzed in our previous work, this bottleneck originates directly from pixel-wise objective functions.

When an image containing an abnormality is processed by a trained Convolutional Autoencoder or Variational Autoencoder, the convolutional decoder attempts to minimize the overall Mean Squared Error (MSE) according to a minimum-energy optimization principle. Because the model is calibrated exclusively on normal samples, it interprets highly localized, irregular abnormalities as high-frequency noise. Consequently, the network suppresses these variations during the spatial reconstruction process. The resulting reconstructed image, $\hat{\boldsymbol{x}}$ , appears overly smooth and "too perfect," effectively replacing the anomalous tissue or scratch with a synthesized patch of healthy or nominal texturing.

Because the anomaly is systematically erased in the reconstructed output, the raw pixel-wise residual $\boldsymbol{x_i} - \hat{\boldsymbol{x}}_{\boldsymbol{i}}$ within the defective region becomes numerically indistinguishable from the background residual of a genuinely normal image. This causes a catastrophic statistical overlap between the normal and anomalous score distributions in the final histograms, destroying the discriminative capability of the Youden threshold and leading to severe false-negative rates on heterogeneous datasets.


© Authors. Licensed under Creative Commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0) non-commercial reuse with attribution
https://creativecommons.org/licenses/by-nc/4.0/

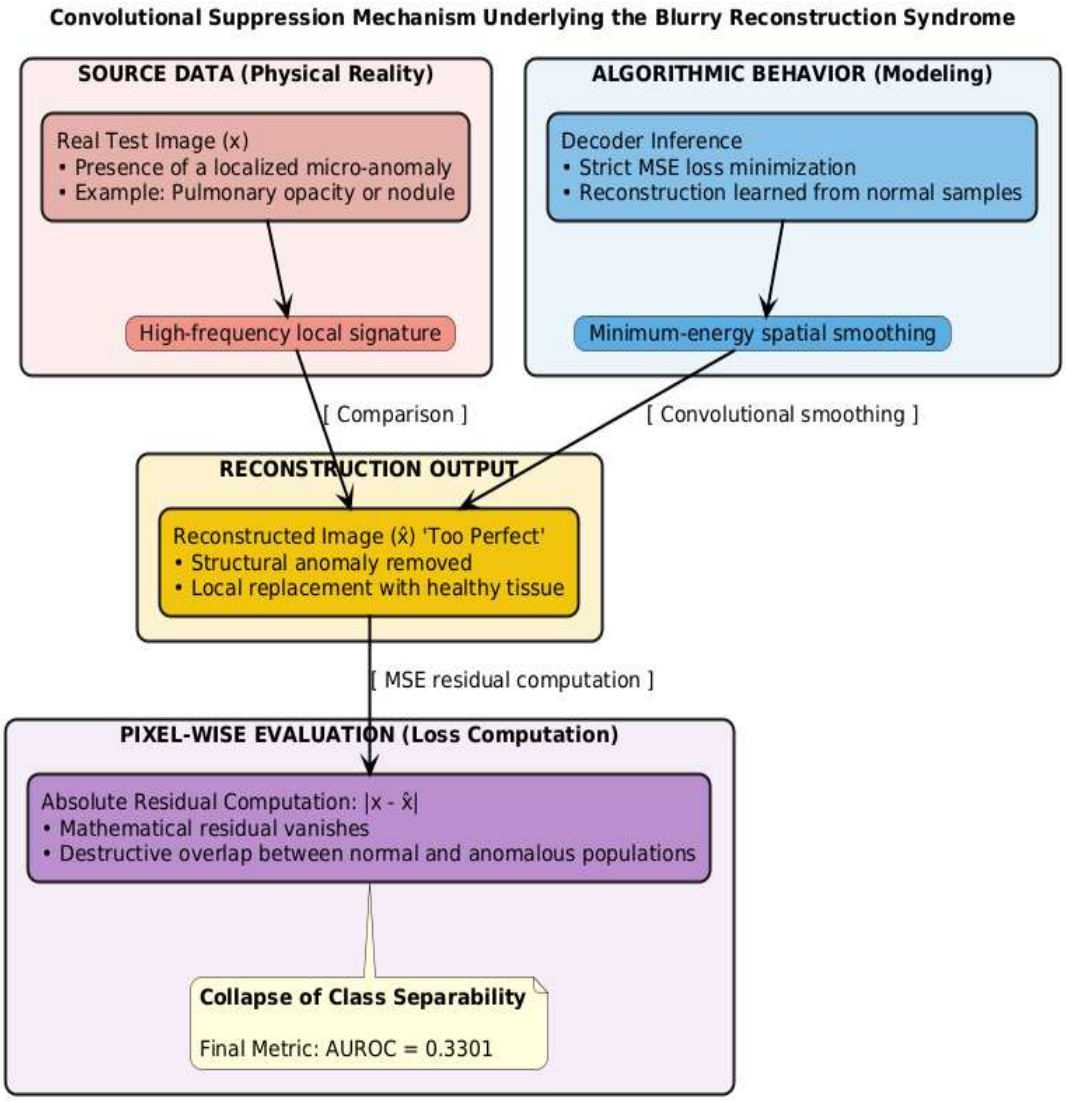


*Figure 1: Convolutional smoothing mechanism underlying the blurry reconstruction syndrome.*

### 2) The Gray Zone Paradox: Aggregation Dilution and Marginal Separability Drift

While non-parametric architectures (such as PaDiM and PatchCore) successfully bypass the blurry reconstruction bottleneck by utilizing pre-trained semantic features, real-world edge deployment under operational conditions exposes a separate metrological limitation. This phenomenon, which we formalize as the *Marginal Separability Drift* or *The Gray Zone Paradox*, stems from a complete


© Authors. Licensed under Creative Commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0) non-commercial reuse with attribution
https://creativecommons.org/licenses/by-nc/4.0/

mathematical decoupling between the frontend visualization engine and the backend decision-making logic.

This conflict arises during the processing of highly localized micro-anomalies or defects in their earliest stages, creating two contradictory system behaviors:

1. **Frontend Relative Visualization (Heatmap Amplification)**: The visualization engine applies a relative Min-Max normalization over the internal two-dimensional local score tensor (e.g., embedding or Mahalanobis distances), stretching its values tightly across the standard 8-bit byte range (**0 – 255**). Because the local distance between a micro-defect patch and the reference memory bank is locally maximal, the rendering engine scales this variation up, saturating the affected pixels into bright red regions via a cv2.COLORMAP_JET mapping. This immediately draws the human operator's attention to the flaw.
2. **Backend Global Score Dilution**: Simultaneously, the backend logic collapses the 2D local score map into a single, uniform scalar value (global_image_score) via a reduction operator (such as selecting the single worst patch for industrial data or computing a global spatial average for medical data). When a defect is microscopic, the elevated error metrics of those few spatial coordinates are mathematically overwhelmed and diluted by the thousands of surrounding normal patches.

As a result of this global dilution, the final score undergoes severe numerical attenuation and drops into a "gray zone" slightly below the optimal Youden threshold ($\tau_{img}$) managed by the system's THRESHOLDS_MAP. Because the PyQt6 interface applies a strict binary decision rule (**Score** $\leq \tau_{img} \rightarrow$ **NORMAL**), the overall alarm is suppressed, producing a marginal false negative. This conservative decision strategy is hardcoded to maximize classifier specificity offline, protecting the edge application against excessive false-positive alarms in production, but it highlights a clear semantic limitation in scalar anomaly pooling.


© Authors. Licensed under Creative Commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0) non-commercial reuse with attribution
https://creativecommons.org/licenses/by-nc/4.0/

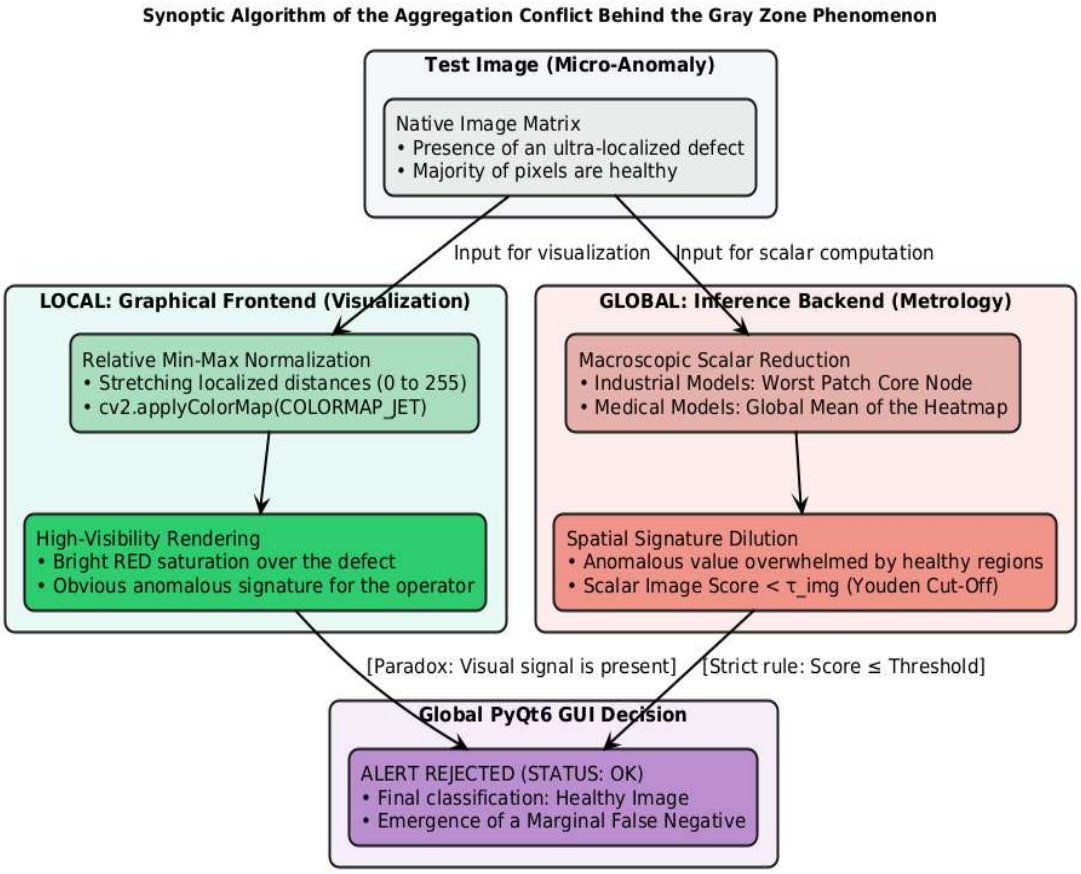


*Figure 2: Synoptic algorithm for the aggregation conflict at the origin of the grey zone phenomenon.*

### 3) Architectural Trade-offs: Speed vs. Memory vs. Detection Sensitivity under Strict Compute Caps

The juxtaposition of these two syndromes outlines the stark trade-offs engineers must navigate when deploying one-class vision systems on resource-constrained CPU hardware.

Generative models optimize for execution speed and storage efficiency. By compressing features into low-dimensional latent bottlenecks, they maintain exceptionally lean checkpoint profiles (**745 KB - 1.23 MB**) that remain strictly invariant regardless of dataset scale. Their inference pass relies entirely on standard forward tensor operations, making them highly predictable and lightweight. However, their detection sensitivity is fundamentally capped by the *Blurry*


© Authors. Licensed under Creative Commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0) non-commercial reuse with attribution
https://creativecommons.org/licenses/by-nc/4.0/

*Reconstruction Syndrome*, rendering them unreliable for complex, multi-class industrial or organic medical surfaces.

Conversely, non-parametric feature models prioritize maximum detection sensitivity and precise spatial localization, eliminating the risk of smoothed reconstructions. However, this precision shifts the hardware bottleneck from raw floating-point compute constraints to heavy memory bandwidth overhead. Storing dense covariance grids (PaDiM) or raw patch embeddings (PatchCore) causes an explosion in storage footprints (**125MB – 339MB**). Furthermore, bypassing slow Python loops to maintain real-time performance (**≥ 24 FPS**) requires offloading nearest-neighbor lookups to specialized C++ frameworks like the vectorized FAISS CPU library running via asynchronous QThread loops. On low-power hardware, achieving peak anomaly sensitivity requires trading away the compact storage profile of generative models to support a larger memory and indexing runtime infrastructure.

## I. FUTURE RESEARCH DIRECTIONS

To overcome the architectural bottlenecks identified in this comparative evaluation and expand the industrial portability of the unified framework, several critical future research tracks are envisioned. These directions bridge the gap between low-level hardware optimization and advanced geometric anomaly pooling, creating a roadmap for next-generation, eco-efficient edge computer vision.

### 4) Post-Training Quantization and Graph Optimization for Edge Deployment

Maximizing the computational efficiency of these models on extreme, ultra-low-power embedded industrial platforms—such as legacy workstations or single-board computers operating within a strict 15-Watt power budget—requires transitioning away from native high-precision data formats.

- **ONNX Engine Serialization**: Exporting the native PyTorch .pt computational graphs into a standardized open neural network exchange (ONNX) representation decouples inference from heavy software runtimes.


© Authors. Licensed under Creative Commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0) non-commercial reuse with attribution
https://creativecommons.org/licenses/by-nc/4.0/

The underlying ONNX Runtime can then perform deep structural graph optimizations, fusing consecutive convolutional layers and element-wise operations to dramatically accelerate asynchronous CPU execution loops.

- **INT8 Precision Conversion**: Converting the network weights and feature representations from Float32 precision down to strict INT8 integer formats aligns matrix arithmetic directly with native processor instruction sets (such as AVX2). This low-level optimization reduces the system's operational RAM footprint by approximately a factor of four. Crucially, this compression drops the maximum non-parametric memory overhead of PatchCore from 339 MB to below 85 MB, facilitating frictionless installation on edge platforms lacking dedicated VRAM.

5) **Hybrid Architectural Pipelines and Loss Formulation**

To eliminate the *Blurry Reconstruction Syndrome* that cripples standard generative models when processing complex, heterogeneous surfaces, future configurations must re-engineer the underlying network objective functions and structural paradigms.

- **Semantic Perceptual Constraints**: Complementing or replacing the traditional, minimum-energy pixel-wise Mean Squared Error (MSE) objective with structural loss functions—such as Structural Similarity Index (SSIM) or deep perceptual features extracted from pre-trained backbones—prevents the decoder from treating micro-defects as high-frequency background noise.
- **Generative-Feature Hybrids**: Developing a hybrid ecosystem that combines compact generative priors with sparse non-parametric memory banks offers a promising path forward. In this schema, a lightweight autoencoding decoder acts as a structural filter, while a highly sparse FAISS memory registry cross-checks localized anomalies, achieving peak classification sensitivity without triggering the storage explosion typical of dense, uncompressed feature descriptors.


© Authors. Licensed under Creative Commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0) non-commercial reuse with attribution
https://creativecommons.org/licenses/by-nc/4.0/

### 6) Topology-Aware Spatial Aggregation and Continual Learning

Defeating the *Gray Zone Paradox* requires a complete overhaul of how local anomalous information is pooled into global decision metrics, alongside making the system adaptable to changing production environments over time.

- **Density-Based Spatial Clustering**: To stop microscopic, highly localized defects from being mathematically diluted by surrounding normal patches, uniform scalar reduction operators (such as maximum patch selection or global averaging) will be replaced with topology-aware spatial clustering mechanisms like DBSCAN or graph connectivity analysis. The global anomaly score will then be computed based on the spatial density and geometric connectivity of anomalous pixel clusters. This ensures that micro-defects generate strong alarm triggers regardless of how many normal patches surround them.
- **Asynchronous Continual Indexing**: Leveraging the native capabilities of the vectorized FAISS CPU library enables the implementation of online, low-overhead continual learning. By using the asynchronous faiss.Index.add() command via the PyQt6 graphical interface, field operators can insert newly available normal training samples directly into the active in-memory database within a few microseconds. This eliminates the need for expensive, time-consuming retraining cycles and allows the tool to seamlessly adapt as industrial or clinical environments evolve.

## VI. CONCLUSION

This study demonstrates that rigorous, decoupled software engineering can entirely eliminate the dependency on expensive, power-hungry hardware accelerators for unsupervised visual quality inspection. By building an asynchronous, multi-threaded ecosystem governed by the separation of concerns (MVC architecture pattern) and running concurrently via managed QThread loops, this framework shifts the operational paradigm of one-class vision systems. Leveraging transfer learning through a frozen ResNet34 backbone, combined with


© Authors. Licensed under Creative Commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0) non-commercial reuse with attribution
https://creativecommons.org/licenses/by-nc/4.0/

greedy geometric coreset compression and optimized C++ vector indexing via the host CPU's AVX2 instruction set, enables software optimizations to successfully outperform raw hardware scaling. Achieving an undisputed peak image-level AUROC of 0.9991 alongside an ultra-fast inference latency well below the 0.8-millisecond-per-frame threshold, the optimized feature-based framework fully satisfies the real-time throughput requirements $\geq 24$ Frames Per Second) of legacy, resource-constrained desktop CPU infrastructures at virtually zero additional deployment cost.

The extensive cross-evaluation across industrial rock textures, manufacturing defects, and medical radiographs reveals a stark mathematical polarization between parametric and non-parametric anomaly detection methods. While generative and projection-based models offer exceptionally light, dataset-invariant serialization profiles—maintaining compact .pt checkpoint sizes under 1.5 MB that are ideal for memory-constrained storage devices—their reliance on localized pixel-wise Mean Squared Error objective functions fundamentally caps their sensitivity. When confronted with highly heterogeneous clinical or multi-class visual domains, generative architectures succumb to the *Blurry Reconstruction Syndrome*, treating localized pathologies or small defects as high-frequency noise and smoothing them out. This convolutional energy minimization loop causes severe score overlaps that can drop the diagnostic AUROC to a critical failure baseline of 0.3301, demonstrating that raw reconstructive losses are structurally unviable for complex, non-homogeneous surface monitoring at the industrial edge.

Ultimately, this unified benchmarking ecosystem provides a solid foundation for the eco-efficient design of future sustainable vision systems by mapping the precise phenomenological boundaries of modern one-class paradigms. The formalization of the *Gray Zone Paradox* addresses a major deployment conflict, explaining how local micro-anomalies can appear perfectly saturated on a relative frontend JET heatmap visualization while simultaneously becoming mathematically diluted by thousands of surrounding nominal patches during backend scalar pooling. This


© Authors. Licensed under Creative Commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0) non-commercial reuse with attribution
https://creativecommons.org/licenses/by-nc/4.0/

marginal separability drift highlights the pressing need for a structural transition away from uniform scalar reductions toward topology-aware spatial clustering operators like DBSCAN. By successfully resolving the trade-offs between parameter compression, nearest-neighbor memory bandwidth constraints, and spatial sensitivity, this work establishes a comprehensive blueprint for deploying accessible, low-power, and operationally robust computer vision tools within the practical constraints of real-world manufacturing and healthcare facilities.

## VII. Data and Code Availability

The source code supporting the findings of this study has been fully open-sourced to guarantee scientific reproducibility and cross-platform validation. The complete unified ecosystem—including the asynchronous QThread inference loops, the vectorized faiss.IndexFlatL2 routines, and the polymorphic PyQt6 graphical interface architecture—is publicly hosted on GitHub.

- **Repository Identification**: Hibat Allah Mouhsine. *Unified Framework for Unsupervised Visual Anomaly Detection: Asynchronous Inference Engines and a PyQt6 Desktop Application.* Official Source Code Registry, 2026.
- **Direct Access URL**: https://github.com/Hibat-Allah-Mouhsine/visual-anomaly-detection-unsupervised-approach


© Authors. Licensed under Creative Commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0) non-commercial reuse with attribution
https://creativecommons.org/licenses/by-nc/4.0/

## REFERENCES


- Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. arXiv preprint arXiv:1312.6114, 2013. Available at: https://arxiv.org/abs/1312.6114.
- Guansong Pang, Chunhua Shen, Longbing Cao, and Anton van den Hengel. Deep Learning for Anomaly Detection: A Review. ACM Computing Surveys (CSUR), vol. 54, no. 2, 2021, pp. 1–38. Available at: https://arxiv.org/abs/2007.02500.
- Zhai and S. Li. Deep Feature Reconstruction for Unsupervised Anomaly Detection. IEEE Transactions on Cybernetics, 2021. Available at: https://arxiv.org/abs/2106.08265.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. Available at: https://arxiv.org/abs/1512.03385.
- Sunghyun Lee, Seunghyun Lee, and Byung Cheol Song. Coupled-Hypersphere Based Feature Adaptation for Unsupervised Image Anomaly Detection. arXiv preprint arXiv:2206.03465, version 2, 2023. Available at: https://arxiv.org/abs/2206.03465.
- Shujun Wang, Lequan Yu, Xin Yang, Chi-Wing Fu, and Pheng-Ann Heng. Patch-based Output Space Adversarial Learning for Joint Optic Disc and Cup Segmentation. Available at: https://arxiv.org/pdf/1902.07519.
- Thomas Defard, Aleksandr Setkov, Alberto Ferrari, and Daniele P. P. P. (Daniele Pino). PaDiM: A Patch Distribution Modeling Framework for Unsupervised Anomaly Detection. arXiv preprint arXiv:2011.08785, 2021. Available at: https://arxiv.org/pdf/2011.08785.



© Authors. Licensed under Creative Commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0) non-commercial reuse with attribution
https://creativecommons.org/licenses/by-nc/4.0/