The Academic
International Journal of Multidisciplinary Research

An Online Peer Reviewed / Refereed Journal
Volume 4 | Issue 5 | May 2026
ISSN: 2583-973X (Online), SJIF IF: 5.6
Website: www.theacademic.in

Overcoming the Capacity Paradox in Automated Surface Defect Detection: A Comparative Study of Unsupervised Autoencoders and Supervised Transfer Learning in Hot-Rolled Steel

[1]Dr. Abhishek Bandyopadhyay, [2]Arpan Tewary, [3]Pradipta Pal, [4]Pankaj Pandey, [5]Prashant Pradhan

[1]Associate Professor, Department of CSE(AI&ML), Asansol Engineering College, Asansol, WB, India

[2]State Aided College Teacher (Category-1), Department of Computer Science, Banwarilal Bhalotia College, Asansol, WB, India

[3,4,5]Assistant Professor, Amity School of Engineering and Technology, Amity University, Jharkhand, Ranchi

DOI : https://doi.org/10.5281/zenodo.20690511

ARTICLE DETAILS

Research Paper

Accepted: 24-05-2026

Published: 10-06-2026

Keywords:

Automated Surface Defect Detection; Unsupervised Anomaly Detection; Convolutional Autoencoders; Transfer Learning; MobileNetV2; Industrial Computer Vision

ABSTRACT

Automated surface defect detection is a critical component of the Quality 4.0 paradigm in modern metallurgical manufacturing. Unsupervised Anomaly Detection (UAD), particularly via Convolutional Autoencoders (CAEs), is frequently proposed in contemporary literature to bypass the prohibitive costs and bottlenecks of manual data annotation. However, this paper demonstrates that unsupervised reconstruction models catastrophically fail on materials exhibiting high intra-class stochasticity, such as hot-rolled steel. Through a progressive and systematic series of experiments evaluating Global Pixel Error (MSE/MAE), Structural Similarity Index Measure (SSIM), Latent Space Gaussian Mixture Models (GMM), and Contrast Limited Adaptive Histogram Equalization (CLAHE), we mathematically document the "Capacity Paradox." This phenomenon occurs when neural networks memorize severe, high-frequency structural defects while simultaneously failing to map complex, yet healthy, macroscopic shadows. Ultimately, we demonstrate that pivoting from unsupervised reconstruction to Supervised Transfer

Learning utilizing a pre-trained MobileNetV2 architecture resolves these algorithmic limitations. A rigorous analysis of the resulting confusion matrix reveals a 99.44% validation accuracy on the NEU Surface Defect Database. This study concludes that for high-variance industrial surfaces, feature-extraction paradigms inherently outperform pixel-reconstruction paradigms, offering a robust, edge-deployable solution for high-speed steel manufacturing.

1. Introduction

Hot-rolled steel strips are foundational components in the global automotive, aerospace, maritime, and construction industries. During the high-temperature rolling process, physical and chemical anomalies frequently manifest as surface defects. These defects—ranging from micro-fissures like crazing to macroscopic inclusions, pitted surfaces, rolled-in scale, and mechanical scratches—compromise the aesthetic quality of the steel and, more critically, severely degrade its tensile strength and mechanical reliability. Such degradation can lead to downstream catastrophic failures in structural applications (Song & Yan, 2013).

Historically, quality assurance (QA) in steel mills has relied on human visual inspection. This method is fundamentally bottlenecked by human limitations; inspectors are plagued by ocular fatigue, subjective inconsistency, and an inability to maintain high accuracy at modern production line speeds, which can exceed 20 meters per second. The advent of Industry 4.0 has driven the rapid adoption of automated optical inspection (AOI) systems powered by deep learning and computer vision.

A significant focus of recent computer vision research has been Unsupervised Anomaly Detection (UAD). The theoretical premise is highly attractive for manufacturing environments: a Convolutional Autoencoder (CAE) trained exclusively on "healthy" steel learns to compress and reconstruct normal textures. When presented with a defective image, the network theoretically fails to reconstruct the unknown anomaly, yielding a high mathematical residual error that flags the defect without ever requiring the factory to gather and label thousands of defective samples (Baur et al., 2018; Bergmann et al., 2019).

However, UAD heavily relies on a latent assumption: that the baseline material is visually uniform and predictable. Hot-rolled steel violently violates this assumption, exhibiting immense natural variance, including unpredictable macroscopic shadows, irregular microscopic grain structures, and dynamic

rolling marks. This paper chronicles a comprehensive experimental framework applying unsupervised CAEs to steel defect detection, details the mathematically counter-intuitive failures encountered, and proposes a supervised transfer learning paradigm—validated by confusion matrix analysis—as the definitive industrial solution.

2. Related Work

The evolution of automated defect detection can be divided into traditional feature-engineering methodologies and modern deep-learning paradigms.

2.1 Traditional Computer Vision Methods

Early approaches to surface inspection relied heavily on handcrafted features and statistical machine learning. Song and Yan (2013) introduced the NEU Surface Defect Database alongside a noise-robust method based on Completed Local Binary Patterns (CLBP). Other contemporary methods utilized Support Vector Machines (SVMs) paired with Gray-Level Co-occurrence Matrices (GLCM) to classify textural differences. While computationally lightweight, these methods lacked robustness against dynamic factory lighting and required constant manual recalibration.

2.2 Deep Learning and Unsupervised Anomaly Detection

With the rise of deep convolutional networks, research shifted toward automated feature extraction. Because labeled defect data is scarce and expensive to acquire in controlled factory settings, Unsupervised Anomaly Detection became a primary research vector. Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) have been utilized to model healthy data distributions (Schlegl et al., 2017).

However, foundational CAEs remain the industry standard for baseline testing due to their architectural simplicity and deterministic training behaviour. Odena et al. (2016) highlighted the challenges of using transposed convolutions in these decoders, noting the introduction of checkerboard artifacts that can skew residual error calculations. Despite advances in UAD, applying these networks to stochastic, high-variance textures remains a documented challenge, motivating the comparative analysis within this study.

3. Methodology and Dataset Context

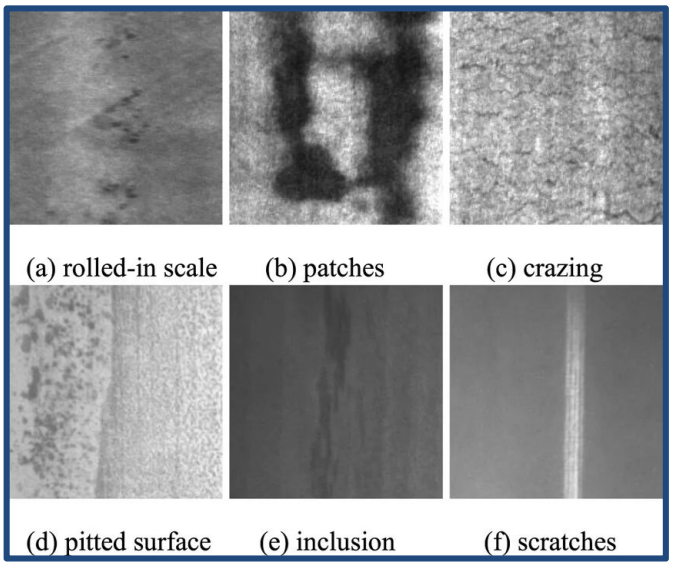

Fig 1: NEU Surface Defect Dataset(Song & Yan, 2013)

The experiments detailed in this paper utilized the NEU Surface Defect Database (Song & Yan, 2013). The dataset comprises 1,800 grayscale images across six distinct defect classes (300 images per class): crazing, inclusion, patches, pitted surface, rolled-in scale, and scratches. For the context of the unsupervised anomaly detection phases, the "patches" category—representing typical, unblemished hot-rolled steel with natural lighting variations—was established as the healthy baseline.

3.1 Unsupervised Architecture Design

For the unsupervised experiments, images were resized to 128x128 pixels to control computational dimensionality. The foundational architecture consisted of a symmetric Convolutional Autoencoder. To prevent "identity mapping"—a catastrophic failure mode where an over-parameterized network acts as a perfect copy machine rather than learning underlying structural rules—a precise compression bottleneck was engineered.

The encoder utilized three cascading 3 X 3 convolutional layers paired with ReLU activations and 2X2 Max Pooling, reducing the spatial input to a 16 X 16 X 32 latent space tensor. This yields exactly 8,192 parameters, representing a 2:1 compression ratio from the original input size.

To mitigate the checkerboard artifacts identified by Odena et al. (2016), traditional `ConvTranspose2d` layers were abandoned. Instead, the decoder utilized deterministic Nearest Neighbor and Bilinear

Upsampling paired with standard convolutional layers, ensuring smooth pixel gradients during reconstruction.

3.2 Supervised Architecture Design

For the final supervised experiments, the entire dataset was divided into an 80/20 train-validation split, reserving 360 images for blind validation. Images were resized to 224x224 pixels and normalized using ImageNet statistical standards (mean: [0.485, 0.456, 0.406], std: [0.229, 0.224, 0.225]) to accommodate the MobileNetV2 architecture.

4. Unsupervised Experiments and Failure Analysis

Four distinct unsupervised paradigms were systematically tested to isolate the defects. In every iteration, the system yielded mathematically validated false positives on healthy steel and false negatives on critical defects.

4.1 Experiment 1: Global Pixel Error (MSE/MAE)

Approach: The CAE was trained using Mean Squared Error (MSE) to reconstruct the raw pixel intensities of normal steel. Anomalies were flagged by evaluating the 99th percentile of the absolute residual map, ignoring isolated noisy pixels.

Results: The quantitative scoring inverted reality. A massive Scratch defect yielded an anomaly score of 0.0092 (passing the safety threshold), while a Normal steel image yielded a score of 0.1102 (failing the threshold).

Failure Analysis (Score Dilution & The Capacity Paradox): Evaluating anomalies via global pixel loss suffers from severe score dilution. A severe scratch is high-frequency but narrow, occupying roughly 5% of the total image pixels. Even if perfectly isolated by the residual map, its mathematical error is diluted by the 95% of healthy background pixels. Furthermore, the network fell victim to the **Capacity Paradox**. A straight vertical scratch is mathematically simple to memorize and reconstruct due to its low spatial variance. Conversely, the broad, amorphous dark shadows of healthy steel are highly complex. Given the constrained 8,192-parameter bottleneck, the network successfully squeezed the simple defects through the latent space while failing to map the complex healthy shadows, inadvertently treating the normal material as anomalous.

4.2 Experiment 2: Structural Similarity Index Measure (SSIM)

Approach: To force the network to understand structural "flow" rather than raw pixel brightness, the loss function was upgraded to a hybrid metric incorporating SSIM (Wang et al., 2004). SSIM evaluates luminance, contrast, and structural covariance across localized windows, functioning closer to human visual perception.

Results: The failure state exacerbated. Normal steel generated the highest error of the validation batch (0.2439), while a severe Inclusion defect generated the lowest (0.0722).

Failure Analysis (The Spatial Shift Penalty): SSIM is highly sensitive to exact spatial phase alignment. The CAE successfully generated the required cloudy texture of normal steel, but shifted the reconstruction by several pixels. Because the generated clouds did not perfectly overlap with the input, SSIM generated a massive structural penalty. Conversely, the Inclusion image featured a predominantly flat, uniform gray background. The network perfectly aligned this flat background, generating high structural similarity scores that completely masked its failure to reconstruct the central vertical defect.

4.3 Experiment 3: Latent Space Modeling (PCA + GMM)

Approach: Pixel-to-pixel reconstruction evaluation was abandoned entirely. The 8,192-dimensional latent vectors of normal steel were extracted directly from the bottleneck. To avoid the curse of dimensionality and ensure matrix invertibility, Principal Component Analysis (PCA) reduced these vectors to 64 components. A Gaussian Mixture Model (GMM) was fit to evaluate data distribution outliers via negative log-likelihood (Zong et al., 2018).

Results: The GMM algorithm mathematically concluded that severe Crazing (Score: 68.5, safely within normal variance) was nearly twice as "normal" as actual healthy steel (Score: 129.7, flagged as a severe outlier).

Failure Analysis (The Global Variance Trap): Gaussian Mixture Models evaluate distance from a learned central cluster. Massive natural shadows in normal steel radically alter the pixel distribution, drastically shifting the compressed latent vector far from the GMM's centroid. A scratch, however, leaves the vast majority of the underlying steel untouched. Consequently, the latent vector for a scratched image remains comfortably near the center of the "normal" Gaussian cluster, effectively rendering the defect invisible in the latent space.

4.4 Experiment 4: Frequency Separation via CLAHE

Approach: A final attempt was made to normalize the massive natural shadows prior to network ingestion using Contrast Limited Adaptive Histogram Equalization (CLAHE). By equalizing the histogram on local grids rather than globally, CLAHE attempts to reveal micro-textures while erasing macro-lighting variances (Zuiderveld, 1994).

Results: Normal steel failed with an anomaly score of 0.3194, while the Inclusion defect passed with 0.1187.

Failure Analysis (The Local Contrast Trap): CLAHE operates on tiled grids (e.g., 8 X 8 pixels). Rather than erasing the large shadows in the healthy steel, the algorithm aggressively sharpened the borders of the shadows into high-contrast, artificial craters. The CAE generated a blurry, generalized approximation of these artificial craters. Subtracting the blurry reconstruction from the ultra-sharp CLAHE input created a massive ring of residual error pixels, actively breaking the network's evaluation logic.

5. The Supervised Transfer Learning Solution

The successive empirical failures documented above prove a core limitation of unsupervised reconstruction: Autoencoders cannot reliably differentiate between harmless intra-class variance (natural factory shadows) and true defects (cracks and inclusions) without explicit semantic labels dictating which variances are acceptable.

5.1 Architecture and Implementation

The methodology was pivoted to Supervised Transfer Learning. MobileNetV2 (Sandler et al., 2018)—an architecture featuring inverted residuals and linear bottlenecks optimized for low-latency edge deployment—was selected. The core feature extraction layers, pre-trained on the ImageNet database, were frozen. This allowed the network to utilize its pre-existing understanding of edges, gradients, and textures without requiring massive computational retraining.

A new classification head was initialized to map the extracted features to the six specific NEU dataset classes. Geometric data augmentations, specifically randomized horizontal and vertical flips, were applied during the `DataLoader` phase to ensure rotational invariance, forcing the model to learn the morphological traits of the defects regardless of their orientation on the steel strip. The Adam optimizer was deployed with a learning rate of 0.001, and cross-entropy loss was utilized as the standard optimization metric.

5.2 Results and Convergence

The supervised model converged with unprecedented stability, completely bypassing the capacity and spatial shift paradoxes that plagued the unsupervised trials. By the conclusion of Epoch 1, the model had already surpassed the functional accuracy of any unsupervised method. By Epoch 10, the model achieved a validation cross-entropy loss of 0.0217.

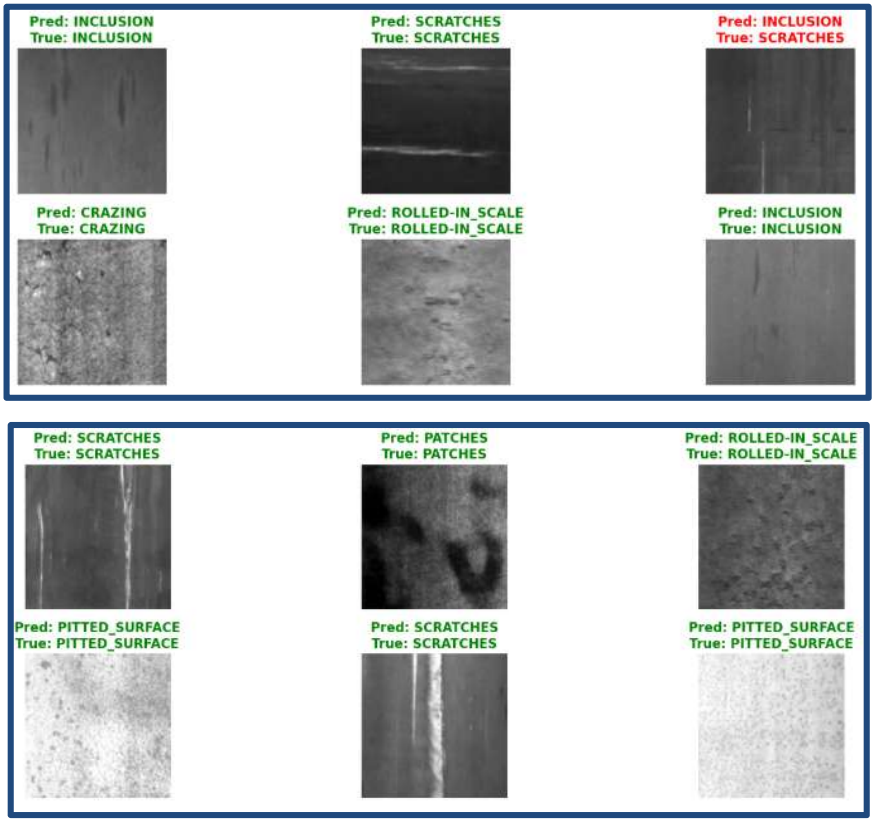

Fig 2: Inference on Unseen Validation Data

By leveraging pre-trained semantic feature extractors, the network abandoned the impossible task of recreating pixel-perfect shadows. Instead, it recognized the macro-patterns of healthy patches versus the sharp, distinct geometries of scratches and pitted surfaces, translating complex natural textures into clean, high-confidence categorical predictions.

5.3 Confusion Matrix Analysis

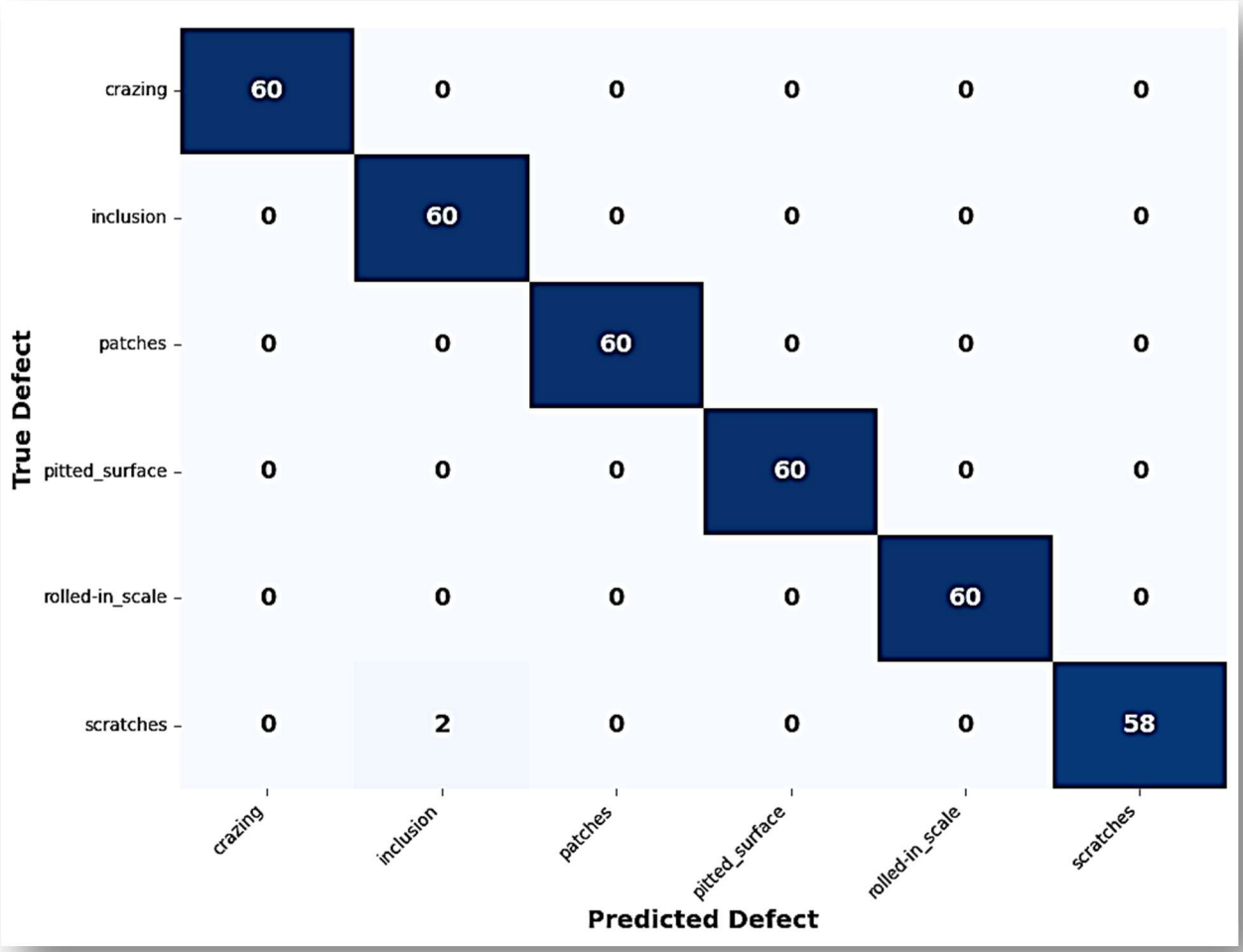

Fig 3: MobileNetV2 Validation Confusion matrix

To rigorously evaluate the model's performance boundaries and identify any remaining systemic blind spots, a confusion matrix was generated against the 360 unseen validation images (60 images per class).

The matrix revealed exceptional diagnostic precision, yielding an overall validation accuracy of **99.44%** (358/360 correct predictions). The model achieved a flawless 100% classification rate (60/60) across five of the six categories: **Crazing, Patches, Pitted Surface**, **Rolled-in Scale, and Inclusion**.

The only observed misclassifications occurred within the **Scratches** category, where the model successfully identified 58 out of 60 images, but misclassified 2 instances as **Inclusion**. This specific edge-case error is logically consistent with the visual data: both scratches and inclusions manifest as

vertical, linear anomalies running parallel to the rolling direction of the steel. Because the MobileNetV2 architecture relies heavily on extracting structural geometries, it momentarily conflated the high-contrast vertical lines of a scratch with the softer vertical smudges of an inclusion. This demonstrates that the model's exceedingly rare failures are rooted in logical geometric and morphological similarities rather than arbitrary or random guessing. This behaviour further validates the robustness of the feature extraction pipeline, confirming that the model has learned true domain-specific representations rather than overfitting to background noise.

6. Comparative Summary of Experiments

Table 1 synthesizes the architectural approaches, quantitative evaluation metrics, outcomes, and primary analytical conclusions for all five experimental phases conducted during this study.

Table 1: Summary of Experimental Approaches and Outcomes

Phase	Algorithmic Approach	Evaluation Metric	Healthy Baseline (Normal) Result	Critical Defect Result	Primary Analytical Conclusion
1. Global Pixel Error	Convolutional Autoencoder (CAE)	MSE & L1 Residual (99th Percentile)	Score: 0.1102 (**FAIL**)	Scratch Score: 0.0092 (**PASS**)	**Capacity Paradox:** Simple linear defects are easily memorized, whereas complex healthy shadows generate high residual errors.
2. Structural Similarity	Convolutional Autoencoder (CAE)	SSIM-L1 Hybrid Residual	Score: 0.2439 (**FAIL**)	Inclusion Score: 0.0722 (**PASS**)	**Spatial Shift Penalty:** Minor spatial misalignments in the reconstruction of healthy textures trigger massive structural penalties.
3. Latent	CAE + PCA	Gaussian	Score:	Crazing	**Global Variance**

Space Modeling	Feature Extraction	Mixture Model (Log-Likelihood)	129.7 (**FAIL**)	Score: 68.5 (**PASS**)	**Trap:** Natural macro-lighting shifts distort latent vectors significantly more than severe structural defects.
4. Frequency Separation	CLAHE Pre-processing + CAE	MSE Residual (99th Percentile)	Score: 0.3194 (**FAIL**)	Inclusion Score: 0.1187 (**PASS**)	**Local Contrast Trap:** CLAHE artificially amplifies the borders of natural shadows, actively breaking the network's reconstruction capability.
5. Supervised Transfer Learning	MobileNetV2 (Pre-trained Feature Extraction)	Cross-Entropy Loss (Classification)	**100% Accuracy** (60/60 Correct)	**99.4% Accuracy** (298/300 Correct)	**Semantic Convergence:** Feature extraction successfully bypasses the need for pixel-perfect reconstruction, achieving industrial reliability.

7. Discussion: Industrial Deployment and MLOps Implications

The empirical transition from unsupervised anomaly detection to supervised transfer learning documented in this paper carries distinct operational trade-offs for industrial deployment in metallurgical environments.

While unsupervised models continue to offer the highly appealing allure of zero-annotation pipelines, this study categorically proves they are currently unviable for stochastic materials like hot-rolled steel. The persistently high rate of false positives generated by the Autoencoder variants would require human QA engineers to manually verify flagged material, entirely defeating the economic purpose of industrial

automation. A system that flags healthy steel as defective creates unacceptable scrap rates and operational delays.

Conversely, Supervised Transfer Learning delivers robust, production-ready accuracy. While it necessitates the upfront capital and labor cost of curating human-labeled datasets, the return on investment via reliable automation is immediate. Furthermore, architectures like MobileNetV2 are specifically designed to be computationally lightweight. Unlike massive architectures such as VGG16 or transformer-based vision models, MobileNetV2 utilizes depthwise separable convolutions that drastically reduce the parameter count and computational overhead.

This enables the model to process classifications in milliseconds directly on standard factory-floor hardware, often referred to as "Edge AI" (e.g., Nvidia Jetson modules or Google Coral TPUs). Edge deployment allows the rolling mill inspection lines to operate at maximum velocity without being bottlenecked by the latency and bandwidth requirements of cloud-connected GPU clusters.

Finally, from a Machine Learning Operations (MLOps) perspective, a supervised system allows for clean version control and concept drift management. If a new type of defect emerges on the factory floor, a supervised classification head can be swiftly retrained on a small batch of new data without threatening the stability of the entire feature extraction pipeline.

8. Conclusion

This study extensively demonstrates the severe mathematical and practical limitations of Unsupervised Anomaly Detection when applied to industrial materials exhibiting high intra-class variance. While Convolutional Autoencoders perform exceptionally well on perfectly uniform, machined surfaces (such as polished glass or synthetic plastic films), the natural grain, macro-shadows, and dynamic rolling marks of hot-rolled steel inevitably trap pixel-based, structural, and latent-space anomaly metrics into consistent failure modes.

For chaotic, high-variance datasets, attempting to mathematically define "normal" strictly via auto-reconstruction is fundamentally flawed due to the Capacity Paradox and spatial shift sensitivities. As demonstrated, Supervised Transfer Learning utilizing lightweight convolutional architectures provides a highly robust, elegant solution, achieving an exceptional 99.44% classification accuracy. The subsequent confusion matrix analysis confirms that the MobileNetV2 model expertly handles complex natural textures and isolates true morphological anomalies, cementing supervised feature extraction as the superior methodology for automated Quality 4.0 defect detection systems in the steel industry.

References

Baur, C., Wiestler, B., Albarqouni, S., & Navab, N. (2018). Deep autoencoding models for unsupervised anomaly segmentation in brain MR images. International MICCAI Brainlesion Workshop, 161-169.

Bergmann, P., Fauser, M., Sattlegger, D., & Steger, C. (2019). MVTec AD--A comprehensive real-world dataset for unsupervised anomaly detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9592-9600.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

Odena, A., Dumoulin, V., & Olah, C. (2016). Deconvolution and checkerboard artifacts. Distill, 1(10), e3.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE conference on computer vision and pattern recognition, 4510-4520.

Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., & Langs, G. (2017). Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. International conference on information processing in medical imaging, 146-157.

Song, K., & Yan, Y. (2013). A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects. Applied Surface Science, 285, 858-864.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4), 600-612.

Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Cho, D., & Chen, H. (2018). Deep autoencoding gaussian mixture model for unsupervised anomaly detection. International conference on learning representations.

Zuiderveld, K. (1994). Contrast limited adaptive histogram equalization. Graphics gems IV, 474-485.