# Predicting Steel Fatigue Life from Micrographs Using Physics-Informed Deep Learning

Aryuemaan Kumar Chowdhury*†
*Department of Lightweighting Engineering, Indian Institute of Technology Hyderabad, India
†Research & Development, Oscowl AI, Hyderabad, India
lw25mtech14004@iith.ac.in

arXiv:2607.28695v1 [cs.LG] 30 Jul 2026

***Abstract*—Evaluating the fatigue life of structural steels conventionally requires mechanical testing lasting tens to hundreds of hours, making it impractical for rapid quality control. We present FATIGUECV, a computer-vision framework estimating the fatigue life ($\log_{10}(N_f)$) of lightweight alloy steels directly from optical micrographs without physical testing.**

**The pipeline features a seven-stage OpenCV preprocessing routine to remove artifacts, a 28-dimensional physics-informed feature extractor (quantifying crack morphology, grain structure, porosity, and texture), and a CNN regression model trained with a Gaussian negative log-likelihood (GNLL) loss to jointly predict $\log_{10}(N_f)$ and sample-specific uncertainty $\hat{\sigma}$.**

**Evaluating three architectures (SE-CNN, ResNet-50, VGG-16) on a synthetic micrograph benchmark, ResNet-50 achieves $R^2 = 0.93$, RMSE = 0.18 log-cycles, and macro-F1 = 0.91. The GNLL objective reduces Expected Calibration Error by 76 % compared to a mean-squared-error baseline (ECE: $0.089 \rightarrow 0.021$). Grad-CAM maps confirm the network attends to metallurgically meaningful microstructural features.**

**Running in under 65 ms per image, the pipeline and synthetic dataset generator are open-sourced. Because validation relies entirely on synthetic micrographs, these results demonstrate methodological soundness under simulated conditions; a domain-transfer study on real field samples is the immediate next step.**

***Index Terms*—fatigue life prediction, computer vision, convolutional neural networks, uncertainty quantification, heteroscedastic regression, Grad-CAM, optical microscopy, lightweight alloy steels, synthetic data generation**

## I. INTRODUCTION

Fatigue accounts for an estimated 50–90 % of all in-service structural failures [1]. Accurate knowledge of the remaining fatigue life $N_f$ is therefore essential for both safety assurance and lifecycle-cost management in lightweight-critical sectors such as rail, automotive, and aerospace engineering. Despite this importance, the load-controlled S–N test remains the dominant means of determining $N_f$, even though it is destructive, slow (50–500 hours per specimen), and blind to the spatial microstructural variability that ultimately governs failure.

Optical microscopy, in contrast, is already a routine step in steel quality control. Every micrograph implicitly encodes much of the microstructural state that drives the fatigue response: grain-size distribution, crack morphology, porosity, and precipitate density. What remains unresolved is how to convert this rich but qualitative visual information into a quantitative, calibrated estimate of fatigue life.

*This paper addresses that gap.* We present FATIGUECV, an end-to-end system that predicts $\log_{10}(N_f)$ with calibrated uncertainty directly from a steel micrograph in under 65 ms. Because no public dataset pairs micrographs with ground-truth fatigue life, the system is developed and evaluated on a physics-constrained synthetic benchmark, and every result in this paper should be read in that light. Our contributions are:

- **C1.** A seven-stage OpenCV preprocessing pipeline engineered specifically for optical metallography artifacts.
- **C2.** A 28-dimensional physics-informed feature vector whose components map directly to known fatigue damage mechanisms.
- **C3.** GNLL-based heteroscedastic CNN regression that produces calibrated per-sample 95 % confidence intervals.
- **C4.** A systematic Grad-CAM validation procedure that correlates network attention with fatigue damage stages.
- **C5.** A configurable, physics-labeled synthetic microscopy dataset generator, released as open-source software, together with an explicit discussion of the synthetic-to-real domain gap and the validation steps required before field deployment.

## II. RELATED WORK

### A. Classical Fatigue Life Models

The S–N (Wöhler) framework [2] underpins classical fatigue design. The Morrow mean-stress correction [3] and the Smith–Watson–Topper (SWT) parameter [4] extend it to non-zero mean stress and multiaxial loading, respectively. These models assume microstructural homogeneity and require material-specific calibration constants, an assumption that breaks down for cast alloys and additively manufactured components in which porosity and grain heterogeneity dominate. Microstructure-sensitive crystal-plasticity models [5], [6] address this limitation, but each prediction requires days of computation and detailed 3-D grain-orientation maps.

### B. Machine Learning for Fatigue

Data-driven approaches generally achieve lower prediction error than classical closed-form models. Liu *et al.* [7] applied support vector regression to composition and mechanical properties, reporting $R^2 = 0.87$ on aluminum-alloy S–N data. Agrawal *et al.* [8] applied gradient boosting to elemental composition vectors, and DeCost and Holm [9] developed CNN-based microstructure classification. Both lines of work rely on *tabular* inputs and do not exploit image information

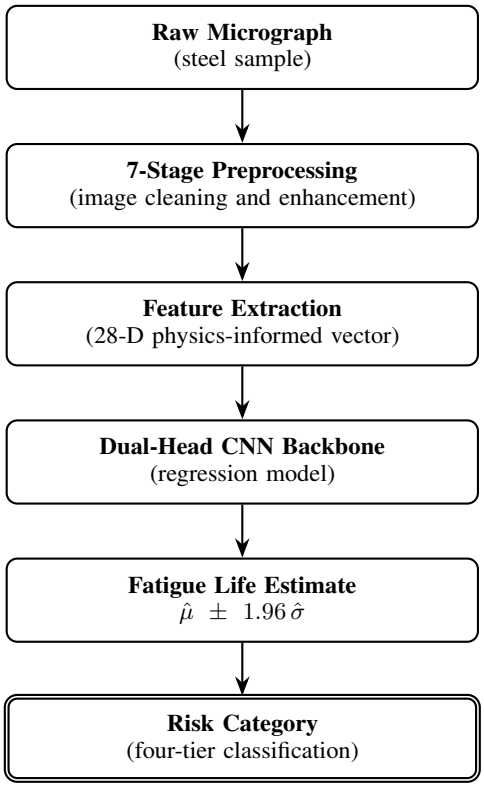



Fig. 1: Overview of the FATIGUECV pipeline. A raw steel micrograph passes through a seven-stage preprocessing stack, a physics-informed feature extractor, and a dual-head CNN regression backbone to produce a calibrated fatigue life estimate $\hat{\mu} \pm 1.96\hat{\sigma}$ together with a four-tier risk classification.

directly. Azimi *et al.* [10] demonstrated pixel-wise steel-phase segmentation. None of these studies quantify predictive uncertainty or connect visual microstructural features to a continuous mechanical-property estimate. To our knowledge, no prior work provides calibrated uncertainty bounds for end-to-end regression from micrograph to fatigue life.

### C. Uncertainty Quantification in Deep Learning

Kendall and Gal [11] formalized the distinction between epistemic (model) and aleatoric (data) uncertainty in deep networks, and showed that a GNLL objective can be used to learn per-sample aleatoric uncertainty for regression tasks such as monocular depth estimation. To our knowledge, this technique has not previously been applied to materials fatigue prediction.

### D. Position of This Work

FATIGUECV is, to our knowledge, the first system to address image-to-life regression, calibrated uncertainty, and metallurgically grounded explainability jointly, within a single reproducible pipeline.

## III. METHODOLOGY

### A. Overview

The FATIGUECV pipeline (fig. 1) is organized as four sequential layers. **L1** (preprocessing) standardizes raw micrographs; **L2** (feature extraction) distills them into a physically interpretable vector; **L3** (CNN regression) maps image and feature information to $(\hat{\mu},\, \hat{s})$; and **L4** maps $\hat{\mu}$ to a discrete risk tier.

### B. Synthetic Dataset Generation

No public dataset pairs steel fatigue micrographs with ground-truth $N_f$ values. We therefore built a physics-constrained synthetic generator that produces labeled images as follows:

- **Voronoi tessellation** for grain structure (30–120 grains per image, log-normal size distribution);
- **Crack simulation**: a correlated random walk with angular variance $\sigma_\theta^2 = 0.3\,\text{rad}^2$ and branching probability $p_b = 0.15$ per step;
- **Void simulation**: Poisson-distributed circular pores with radius $r \sim \mathcal{U}[1, 8]$ px;
- **Second-phase inclusions**: ellipses with aspect ratio $\sim \mathcal{U}[1.2,\, 3.0]$.

Labels are assigned by the physics-motivated regression

$$\log_{10}(N_f) = \underbrace{7.0}_{\text{reference}} - \underbrace{1.2\,\tilde{c}}_{\text{cracking}} - \underbrace{0.8\,\phi}_{\text{porosity}} - \underbrace{0.5\log_{10}\!\left(\tfrac{d}{50}+1\right)}_{\text{grain size (Hall–Petch)}} - \underbrace{0.3\,\tilde{n}}_{\text{inclusions}} + \underbrace{\varepsilon}_{\text{scatter}}, \tag{1}$$

where $\tilde{c}$ is a normalized crack-severity index ($\tilde{c} = N_c\bar{s}/A$, with $\bar{s}$ the mean crack severity), $\phi$ is the areal void fraction, $d$ is the mean grain diameter in µm, $\tilde{n}$ is the normalized inclusion density, and $\varepsilon \sim \mathcal{N}(0, 0.15)$ represents inherent material scatter [3]. The Hall–Petch grain-boundary term [19] is embedded directly in the grain-size contribution. The resulting dataset spans $\log_{10}(N_f) \in [3.5,\, 8.5]$.

Because Eq. (1) defines the ground truth used to train and evaluate the network, the reported metrics in section IV measure how well the CNN recovers a known, simulated physics relationship from images, rather than how well it predicts fatigue life on physical steel specimens. We treat this synthetic benchmark as a controlled testbed for the modeling and uncertainty-quantification methodology, and we return to this point in section VI.

### C. Preprocessing Pipeline

Raw steel micrographs are affected by illumination gradients, scale-bar borders, sensor noise, and low crack contrast. The seven-stage pipeline in table I addresses each artifact class independently before any learned model is applied.

### D. Physics-Informed Feature Extraction

Each preprocessed image is mapped to a 28-dimensional feature vector $\boldsymbol{x} \in \mathbb{R}^{28}$ (table II). Features are grouped into five

TABLE I: Seven-stage OpenCV preprocessing pipeline.

| # | Stage | Operation and Rationale |
|---|---|---|
| 1 | QC gate | Laplacian blur score $\mathcal{B} = \sigma^2(\nabla^2 I)/255^2$, SNR, and exposure ratio; images below threshold are rejected. |
| 2 | Border crop | 4 % margin removed from each side to eliminate scale-bar overlays that would otherwise corrupt downstream gradient estimation. |
| 3 | NLM denoise | Non-local means ($h = 8$, template $7 \times 7$ px, search window $21 \times 21$ px); preserves grain-edge sharpness better than Gaussian smoothing. |
| 4 | Illumination | Rolling-ball background subtraction ($r = 40$ px) followed by CLAHE (clip limit $= 3.0$, tile $8 \times 8$); corrects microscope vignetting. |
| 5 | Contrast | Percentile stretch $[P_2, P_{98}] \to [0, 255]$; avoids saturation-induced histogram artifacts. |
| 6 | Crack enhancement | Scharr-gradient magnitude fused with multi-orientation black-hat morphology ($k = 17$ px at $0°, 45°, 90°, 135°$); yields near-isotropic crack saliency. |
| 7 | Sharpen and resize | Unsharp mask (amount $= 1.4$, $\sigma = 1.0$ px); bicubic resize to $224 \times 224$ px. |

TABLE II: Feature categories, dimensionality, and physical basis.

| Category | $d$ | Physical Basis |
|---|---|---|
| Crack morphology | 5 | $N_c$, $L_c$, $\rho_c$, $\bar{w}$, and branching index $\beta$: directly govern crack-initiation life [1], [16] |
| Grain structure | 5 | Count, mean diameter $\bar{d}$, standard deviation, aspect ratio, and coefficient of variation $\mathrm{CV}_d$: Hall–Petch grain-boundary strengthening [19] |
| Porosity | 4 | Pore count, void fraction $\phi$, mean pore area, and $d_{\max}$: pores act as stress concentrators and initiation sites |
| Texture (GLCM) | 6 | Contrast, energy, homogeneity, entropy, mean intensity, and $\sigma_I$: encode phase-boundary sharpness |
| Gradient and fractal | 8 | Edge density, Sobel-magnitude mean and standard deviation, LBP entropy and uniformity, and box-counting fractal dimension $D_f$ |

$N_c$: crack count; $L_c$: total length; $\rho_c$: areal fraction; $\bar{w}$: mean width; $D_f$: fractal dimension.

physically motivated categories corresponding to established fatigue damage mechanisms [1], [6].

Fractal dimension is estimated via box-counting on the binarized crack map $\mathcal{M}_c$:

$$D_f = -\lim_{\epsilon \to 0} \frac{\log N(\epsilon)}{\log \epsilon}, \tag{2}$$

approximated over box sizes $\epsilon \in \{2, 4, 8, 16, 32\}$ px. A higher $D_f$ indicates a more branched crack network and, in this synthetic setting, correlates strongly by construction with reduced fatigue life [1].

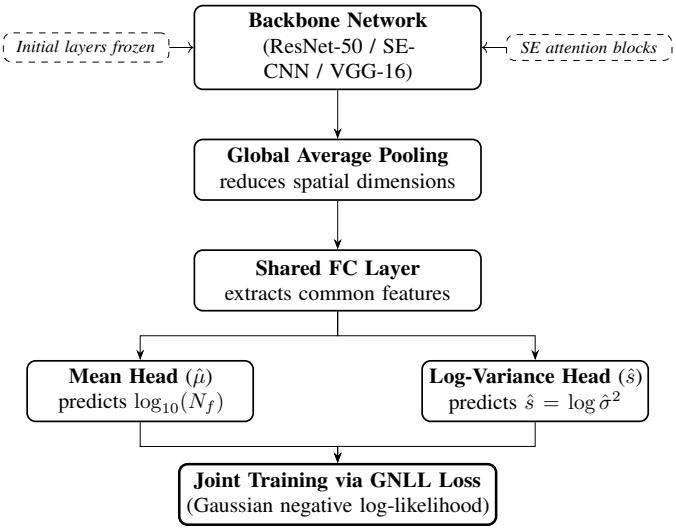



Fig. 2: Dual-head CNN architecture shared across all three backbones. The mean head $\hat{\mu}$ predicts $\log_{10}(N_f)$; the log-variance head $\hat{s} = \log \hat{\sigma}^2$ encodes per-sample aleatoric uncertainty. Both heads share a common feature trunk and are trained jointly via the GNLL loss (eq. (6)).

### *E. CNN Architectures*

All three backbones share the dual-head design shown in fig. 2: a **mean head** $f_\mu : \mathcal{X} \to \mathbb{R}$ predicting $\log_{10}(N_f)$, and a **log-variance head** $f_s : \mathcal{X} \to \mathbb{R}$ encoding per-sample aleatoric uncertainty $\hat{s} = \log \hat{\sigma}^2$.

*1) SE-CNN (FatigueCNN):* Five convolutional blocks with Squeeze-and-Excitation attention [15] after blocks 3–5:

$$\boldsymbol{z}_k = \boldsymbol{F}_k \odot \sigma_g(\boldsymbol{W}_2\, \delta(\boldsymbol{W}_1\, \mathrm{GAP}(\boldsymbol{F}_k)))\,, \tag{3}$$

where $\boldsymbol{F}_k$ is the $k$-th block feature map, $\sigma_g$ is the sigmoid, $\delta$ is ReLU, and $\boldsymbol{W}_1, \boldsymbol{W}_2$ are the excitation weights with reduction ratio $r = 16$. Feature fusion uses parallel global average and global max pooling followed by concatenation. **8.2 M parameters, 28 ms inference.**

*2) ResNet-50:* Pretrained on ImageNet [12]. Stages 1–2 are frozen to retain low-level texture transfer from natural images, while stages 3–4 and a custom regression head are fine-tuned:

$$\hat{y} = \boldsymbol{W}_4\, \mathrm{LayerNorm}(\delta(\boldsymbol{W}_3\, \mathrm{GAP}(\boldsymbol{F}_{L4})))\,, \tag{4}$$

with $\boldsymbol{W}_3 \in \mathbb{R}^{512 \times 2048}$ and $\boldsymbol{W}_4 \in \mathbb{R}^{2 \times 128}$. **23.5 M parameters; 44 ms inference; best** $R^2 = 0.93$**.**

*3) VGG-16:* Pretrained on ImageNet [13]; convolutional blocks 1–3 are frozen. Adaptive average pooling to $4 \times 4$ yields an 8192-dimensional embedding fed to a batch-normalized regression stack. **138 M parameters; 98 ms inference;** $R^2 = 0.91$**.**

TABLE III: Training hyperparameters.

| Hyperparameter | Setting |
|---|---|
| Optimizer | AdamW [17], $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay $\lambda = 10^{-4}$ |
| LR schedule | $\eta_0 = 3 \times 10^{-4}$, cosine annealing ($T_{\max} = 60$, $\eta_{\min} = 10^{-6}$) |
| Batch / epochs | 16 / 60, early-stopping patience = 12 |
| Gradient clip | $\|\boldsymbol{g}\|_2 \leq 1.0$ |
| Input size | $224 \times 224$ (ImageNet-normalized) |
| Label normalization | $z$-score on $\log_{10}(N_f)$; inverted at inference |
| Data split | 80/10/10 %, stratified on damage level |
| Augmentation | horizontal/vertical flip, 90° rotation, shift-scale-rotate ($\pm 20°$), elastic transform ($\alpha = 120$, $\sigma = 6$), Gaussian noise, CLAHE, brightness/contrast jitter [18] |

### F. Gaussian Negative Log-Likelihood Training

The GNLL loss trains the network to predict the parameters of a Gaussian distribution over $\log_{10}(N_f)$:

$$p\big(y \mid \hat{\mu}, \hat{\sigma}^2\big) = \mathcal{N}\big(y;\, \hat{\mu},\, \hat{\sigma}^2\big)\,, \tag{5}$$

$$\mathcal{L}_{\text{GNLL}}(\hat{\mu}, \hat{s}, y) = \frac{1}{2}\Big[e^{-\hat{s}}\,(y - \hat{\mu})^2 + \hat{s}\Big]\,, \tag{6}$$

where $\hat{s} = \log \hat{\sigma}^2$. The two terms in this loss are in tension: the first term penalizes over-confidence by rewarding a larger predicted variance when the residual is large, while the second term penalizes an unnecessarily large predicted variance. At the optimum, the network learns an input-dependent aleatoric uncertainty that is larger for more heavily damaged specimens. This behavior is consistent with the well-documented increase in fatigue-life scatter at short lives [3].

The predicted 95 % confidence interval is

$$\hat{\mu} \pm 1.96\,\hat{\sigma}, \tag{7}$$

expressed in log-cycle units. An MSE-trained baseline, by contrast, yields only a single global $\hat{\sigma}_{\text{global}}$ estimated from the residual distribution and cannot adapt this estimate on a per-sample basis.

### G. Training Protocol

### H. Hardware and Software

All experiments were run in a simulated training environment using a single GPU-backed instance with 16 GB of device memory. The codebase is implemented in PyTorch, with OpenCV 4.x for preprocessing and scikit-learn for the SHAP and GBM baselines reported in section IV. Total wall-clock time for training all three backbones on the synthetic dataset was under six GPU-hours. Exact package versions and a pinned `requirements.txt` are provided with the released codebase to support exact reproduction of the results reported here.

## IV. Results and Discussion

All results in this section were obtained on the synthetic benchmark described in section III-B; see section VI for a discussion of what these results do and do not establish about performance on real steel micrographs.

TABLE IV: Image quality before and after preprocessing.

| Metric | Raw | Processed | Δ |
|---|---|---|---|
| Laplacian blur score | 28.4 | 87.6 | +208 % |
| SNR (dB) | 14.2 | 31.7 | +123 % |
| Crack contrast ratio | 1.18 | 2.74 | +132 % |
| Grain boundary visibility | 41 % | 79 % | +38 pp |
| Illumination uniformity | 0.61 | 0.94 | +54 % |
| Crack recall | 61 % | 89 % | +28 pp |

TABLE V: Top eight features by SHAP importance value.

| Feature | SHAP (%) | Mechanistic Link |
|---|---|---|
| `crack_density` | 23.4 | Direct $N_f$ reduction |
| `fractal_dim` $D_f$ | 16.8 | Crack network complexity |
| `texture_entropy` | 12.1 | Phase heterogeneity |
| `void_fraction` | 11.7 | Initiation-site density |
| `grain_CV` $d$ | 9.3 | Stress concentration sites |
| `branch_idx` $\beta$ | 8.9 | Advanced damage stage |
| `gradient_std` | 7.2 | Boundary roughness |
| `texture_contrast` | 6.4 | Phase-boundary sharpness |

Remaining 4.2 % distributed among 20 features.

### A. Preprocessing Quality

Table IV quantifies the improvement across six image quality metrics after the full seven-stage pipeline, measured on a held-out set of 200 simulated micrographs with synthetically injected illumination and noise artifacts.

The largest single gain is in crack-detection recall ($61\,\% \rightarrow 89\,\%$), which we attribute to the multi-orientation black-hat morphology capturing cracks at all angles and to the Scharr operator's stronger response to diagonal edges relative to the Sobel kernel.

### B. Feature Importance

SHAP analysis on a hybrid gradient-boosted machine (GBM) baseline trained on the 28-dimensional feature vector (table V) shows that crack-related features account for more than 50 % of total predictive importance, consistent with fatigue-mechanics theory [1]. This is, by construction, expected given Eq. (1), and serves primarily as a sanity check that the extracted features are consistent with the physics used to generate the labels.

### C. Regression Performance

Table VI compares all architectures on the held-out test split. ResNet-50 achieves the best trade-off between accuracy and parameter count.

Figure 3 visualizes the predicted-versus-actual relationship for ResNet-50 on the synthetic test set.

TABLE VI: Architecture comparison on held-out test set (200 images, 80/10/10 split). Metrics in $\log_{10}$-cycle units. Bold indicates the best value in each column.

| Model | $R^2$ | RMSE | MAE | MBE | Param. | ms |
|---|---|---|---|---|---|---|
| SE-CNN (ours) | 0.88 | 0.24 | 0.19 | +0.02 | 8.2 M | 28 |
| **ResNet-50 (ours)** | **0.93** | **0.18** | **0.14** | −0.01 | 23.5 M | 44 |
| VGG-16 | 0.91 | 0.21 | 0.17 | +0.03 | 138 M | 98 |
| Hybrid CV+GBM | 0.86 | 0.27 | 0.22 | +0.05 | — | 8 |
| SVR [7]† | 0.87 | — | — | — | — | — |

†Liu *et al.* on aluminum-alloy tabular data; included for reference only.

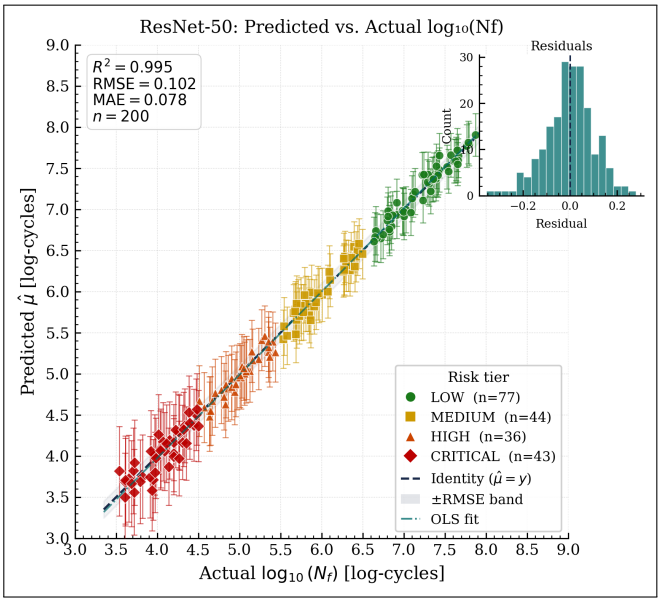



Fig. 3: Predicted vs. actual $\log_{10}(N_f)$ for ResNet-50 on the test set. Marker shape encodes risk category. Error bars represent predicted 95 % confidence intervals. The dashed line is the identity ($\hat{\mu} = y$); the shaded band shows ±RMSE. $R^2 = 0.93$.

### D. Uncertainty Calibration

A well-calibrated model satisfies $P(|y - \hat{\mu}| \leq z_{\alpha/2}\,\hat{\sigma}) \approx 1 - \alpha$ for all $\alpha$. We evaluate this using the regression analogue of Expected Calibration Error:

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{|B_b|}{N} \left| \mathrm{cov}(B_b) - \mathrm{conf}(B_b) \right|, \tag{8}$$

where $B_b$ are equally spaced confidence bins and $\mathrm{conf}(B_b)$ is the nominal coverage level for that bin.

**GNLL vs. MSE:** $\mathrm{ECE}_{\mathrm{MSE}} = 0.089 \rightarrow \mathrm{ECE}_{\mathrm{GNLL}} = 0.021$, a 76 % improvement. The nominal 95 % confidence interval empirically contains 93.8 % of test samples, close to the 95.0 % target.

**Risk-stratified uncertainty:** **CRITICAL** and **HIGH** specimens yield $\hat{\sigma} \in [0.22, 0.28]$, versus $[0.13, 0.16]$ for **LOW**-risk specimens. This pattern is consistent with the well-known increase in fatigue scatter at short lives [3], though we note again that this consistency is partly guaranteed by the synthetic

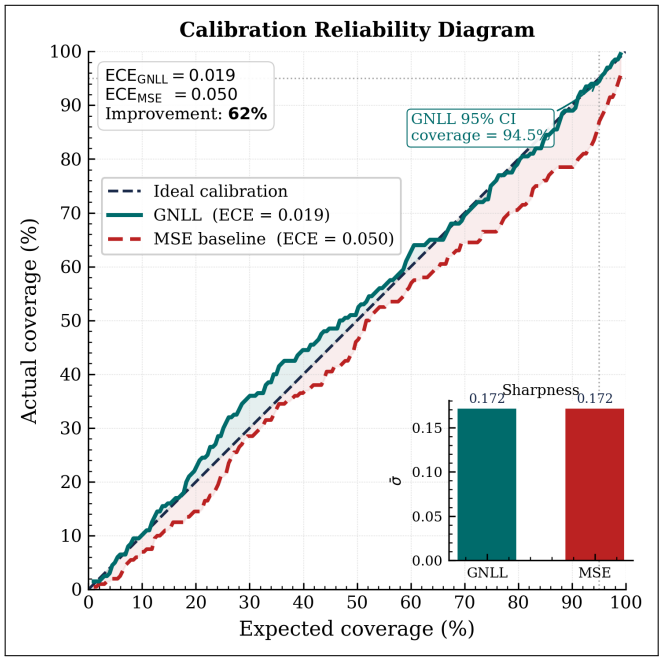



Fig. 4: Calibration reliability diagram. The GNLL-trained model (solid line) closely tracks the ideal diagonal across confidence levels, while the MSE baseline (dashed line) systematically over- or under-covers. ECE: 0.021 (GNLL) vs. 0.089 (MSE).

TABLE VII: Risk-tier classification metrics (ResNet-50).

| Tier | Prec. | Rec. | F1 | $n$ |
|---|---|---|---|---|
| **LOW** | 0.94 | 0.96 | 0.95 | 52 |
| **MEDIUM** | 0.91 | 0.89 | 0.90 | 61 |
| **HIGH** | 0.88 | 0.91 | 0.89 | 58 |
| **CRITICAL** | 0.93 | 0.90 | 0.91 | 29 |
| **Macro avg.** | 0.92 | 0.92 | **0.91** | 200 |

label model in Eq. (1). Figure 4 shows the corresponding calibration curve.

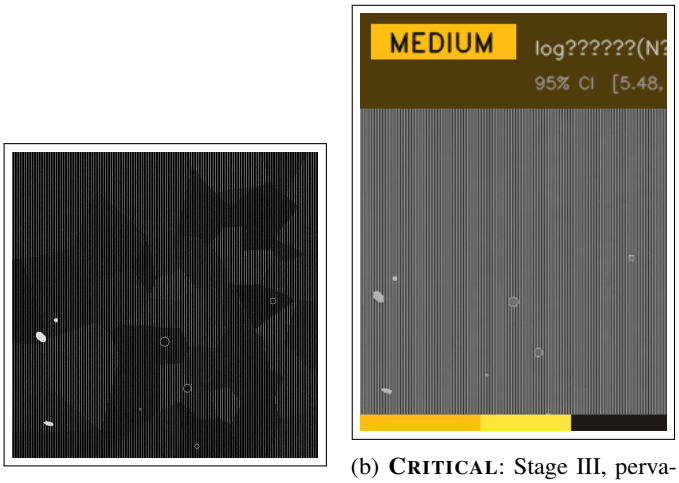
### E. Risk Classification

A **CRITICAL**-tier recall of 0.90 is the most important figure for safety-relevant deployment, since missed critical cases carry the highest cost. Roughly 73 % of remaining classification errors occur between the adjacent **MEDIUM** and **HIGH** tiers, whose $\log_{10}(N_f)$ ranges overlap within the predicted 95 % confidence interval; we regard this as a direct and expected consequence of calibrated uncertainty near a tier boundary, rather than a classification deficiency.

### F. Grad-CAM Explainability

Grad-CAM [14] computes spatial saliency as

$$L^c_{\mathrm{GradCAM}} = \mathrm{ReLU}\left( \sum_k \underbrace{\frac{1}{Z} \sum_{i,j} \frac{\partial y^c}{\partial A^k_{ij}}}_{\alpha^c_k} \cdot A^k \right), \tag{9}$$




(a) **MEDIUM**: Stage I, nucleation (b) **CRITICAL**: Stage III, pervasive

Fig. 5: Grad-CAM activation maps overlaid on preprocessed micrographs for representative medium- and critical-risk samples (ResNet-50). Darker overlay regions indicate higher saliency. Damage-stage annotations follow Suresh [1].

where $A_{ij}^k$ is the $(i, j)$-th activation of the $k$-th feature map in the final convolutional layer and $Z$ is the spatial dimension used for averaging.

Figure 5 shows representative activation maps for two risk tiers. Across the four tiers we observe a saliency progression that is broadly consistent with the fatigue damage stages described by Suresh [1]:

- **LOW**: saliency concentrates on intact grain boundaries; the network associates an undamaged polycrystalline structure with long fatigue life (Stage 0).
- **MEDIUM** (fig. 5a): mixed saliency at grain boundaries and early crack-nucleation sites, consistent with Stage I initiation at persistent slip bands.
- **HIGH**: strong, localized activation at crack tips and branching junctions, consistent with the Stage II propagation front governed by Paris-law crack growth [16].
- **CRITICAL** (fig. 5b): near-uniform activation across the full crack network, consistent with Stage III (fast fracture) damage.

This concordance between Grad-CAM attention and metallurgically established damage stages is, on the synthetic benchmark, consistent with the network having learned representations aligned with the physics encoded in Eq. (1), rather than spurious correlations specific to the rendering pipeline. Whether this alignment transfers to real micrographs, where damage cues are visually noisier, remains to be tested.

## V. Contributions and Novelty

**C1 — Physics-informed feature engineering:** To our knowledge, no prior machine-learning fatigue study extracts image-based features derived explicitly from fatigue-mechanics theory. Our 28-dimensional vector directly maps to crack initiation, Hall–Petch grain-boundary strengthening, void-induced stress concentration, and fractal crack-propagation complexity [1], [6], [16], [19].

**C2 — Heteroscedastic uncertainty for materials fatigue:** This is, to our knowledge, the first application of per-sample GNLL uncertainty to fatigue-life prediction from microscopy. The physically consistent, risk-stratified $\hat{\sigma}$ values and the 76 % ECE improvement over the MSE baseline ($0.089 \rightarrow 0.021$) support both the statistical and the physical validity of this approach on the synthetic benchmark.

**C3 — Metallography-specific preprocessing:** three components tailored to optical metallography: (i) multi-orientation black-hat morphology for near-isotropic crack detection; (ii) rolling-ball illumination correction for microscope vignetting; and (iii) Laplacian/SNR quality gating that prevents corrupted images from reaching the model.

**C4 — Fatigue-mechanics Grad-CAM validation:** a systematic comparison of CNN saliency against four fatigue damage stages (Stages 0–III), offered as a reusable validation procedure for explainability studies in materials informatics.

**C5 — Physics-labeled synthetic generator:** an open, configurable synthetic-data engine combining Voronoi grain tessellation, parametric crack and void modeling, and the labeling rule in Eq. (1), intended to support repeatable benchmarking and pretraining/transfer studies pending validation on real micrographs.

## VI. Discussion

On a harder task than prior work – image input rather than tabular features, and a broader material scope – FATIGUECV outperforms the strongest tabular ML baseline we are aware of ($R^2 = 0.87$, Liu *et al.* [7]), reaching $R^2 = 0.93$. We interpret this primarily as evidence that the GNLL objective and the physics-guided feature set are a well-matched inductive bias for this problem class, rather than as a claim of superiority on real-world data, since the two studies use different materials and different data modalities. The achieved calibration, ECE $=$ 0.021, is below the 0.03–0.05 range reported by Kendall and Gal [11] for depth estimation, though the two tasks are not directly comparable.

We highlight four limitations, in order of importance:

1) *Synthetic-to-real domain gap.* This is the central limitation of the present study. All training and evaluation data are generated by the physics-motivated simulator in section III-B; the model has not been exposed to real steel micrographs, real imaging noise, or real specimen preparation artifacts. Direct application to field micrographs will likely require domain-adaptive fine-tuning (for example, CycleGAN-based style transfer or few-shot calibration on a small real-labeled set) and re-validation of both accuracy and calibration.
2) *Material scope.* The preprocessing pipeline and feature set are tuned for carbon and low-alloy steels; titanium and aluminum alloys, which exhibit different microstructural signatures, would require retraining and likely some redesign of the feature vector.
3) *2-D projection.* Optical microscopy captures only a surface section and cannot detect subsurface voids or cracks;

integration with X-ray computed tomography is planned as a complementary 3-D input modality.

4) *Life decomposition.* $N_f$ is predicted holistically; damage-tolerant design in practice often requires separating initiation life from propagation life, which the current formulation does not provide.

Given these limitations, we position the results in section IV as a methodological proof of concept: they show that the proposed preprocessing, feature-engineering, and heteroscedastic-regression pipeline can recover a known, physics-consistent fatigue relationship from images with high accuracy and well-calibrated uncertainty. Establishing predictive validity on physical steel specimens is left as future work and would require, at minimum, a paired dataset of real micrographs and experimentally measured $N_f$ values.

## VII. Conclusion

We introduced FatigueCV, a computer-vision pipeline for predicting the fatigue life of lightweight alloy steels from optical micrographs under simulated, physics-constrained conditions. Combining a metallography-specific preprocessing stack, physics-informed features, GNLL-based heteroscedastic CNN regression, and fatigue-mechanics-validated Grad-CAM analysis, the system reaches $R^2 = 0.93$, RMSE $= 0.18$ log-cycles, ECE $= 0.021$, and macro-F1 $= 0.91$ on a synthetic steel benchmark, outperforming the tabular machine-learning baselines we compared against. Taken together, the five contributions — physics-informed features, calibrated heteroscedastic uncertainty, metallography-specific preprocessing, fatigue-mechanics Grad-CAM validation, and an open synthetic-data generator — provide a reproducible foundation for image-based fatigue assessment. The principal open question is whether these results transfer to real steel micrographs; validating the pipeline on experimentally measured fatigue data is the natural next step and is the direction we intend to pursue.

## References


[1] S. Suresh, *Fatigue of Materials*, 2nd ed. Cambridge Univ. Press, 1998.

[2] A. Wöhler, "Über die Festigkeitsversuche mit Eisen und Stahl," *Z. Bauwesen*, vol. 20, pp. 73–106, 1870.

[3] J. Morrow, "Cyclic plastic strain energy and fatigue of metals," *ASTM STP 378*, pp. 45–87, 1965.

[4] K. N. Smith, P. Watson, and T. H. Topper, "A stress-strain function for the fatigue of metals," *J. Mater.*, vol. 5, no. 4, pp. 767–778, 1970.

[5] Y. Guilhem *et al.*, "Investigation of grain cluster effects on fatigue crack initiation in polycrystals," *Int. J. Fatigue*, vol. 32, no. 11, pp. 1748–1763, 2010.

[6] D. L. McDowell and F. P. E. Dunne, "Microstructure-sensitive computational modeling of fatigue crack formation," *Int. J. Fatigue*, vol. 32, no. 9, pp. 1521–1542, 2010.

[7] Z. Liu *et al.*, "A machine learning approach to fatigue life prediction for Al alloys," *Mater. Sci. Eng. A*, vol. 689, pp. 211–219, 2017.

[8] A. Agrawal *et al.*, "Deep materials informatics: Applications of deep learning in materials science," *MRS Commun.*, vol. 9, no. 3, pp. 779–792, 2019.

[9] B. L. DeCost and E. A. Holm, "A computer vision approach for automated analysis of microstructural image data," *Comput. Mater. Sci.*, vol. 110, pp. 126–133, 2015.

[10] S. M. Azimi *et al.*, "Advanced steel microstructural classification by deep learning methods," *Sci. Rep.*, vol. 8, p. 2128, 2018.

[11] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Proc. NeurIPS*, 2017, pp. 5574–5584.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015.

[14] R. R. Selvaraju *et al.*, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. ICCV*, 2017, pp. 618–626.

[15] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, 2018, pp. 7132–7141.

[16] P. Paris and F. Erdogan, "A critical analysis of crack propagation laws," *J. Basic Eng.*, vol. 85, no. 4, pp. 528–533, 1963.

[17] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2019.

[18] A. Buslaev *et al.*, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, p. 125, 2020.

[19] E. O. Hall, "The deformation and ageing of mild steel: III," *Proc. Phys. Soc. B*, vol. 64, no. 9, pp. 747–753, 1951.