

Benchmarking a Modular Deep Learning Pipeline for High-Throughput ASC-Speck Image Classification in Drug Discovery

Sifeddine Elkardoudi^{1*}, Khalil Ladrham², Abdelilah Majdoubi³, Ahmed Eddaoui⁴, Mohamed Talea⁵

^{1,4,5}Information Processing Laboratory, Faculty of Sciences Ben M'sik, University Hassan II, Casablanca, Morocco.

²Intelligent Processing and Security of Systems, Faculty of Sciences, Mohammed V University, Rabat, Morocco.

³Microbiology and Immunology, Department of Microbiology, Infectiology and Immunology, Faculty of Medicine, University of Montreal, Montreal, Canada.

E-mail: ^{1*} sifeddine.elkardoudi-etu@etu.univh2c.ma, ² khalil_ladrham@um5.ac.ma, ³ majdoubi.abdelilah@gmail.com,

⁴ Ahmed_edaoui@yahoo.fr, ⁵ taleamohamed@yahoo.fr

Orcid ID: ^{1*} 0009-0001-7888-8471, ² 0009-0004-9107-2401, ³ 0000-0002-2053-4793, ⁴ 0000-0003-1321-5396,

⁵ 0000-0003-2364-0804

Abstract

Inflammasome drug screening using THP-1-ASC-GFP reporter cells produces fluorescence microscopy datasets that are difficult to assess manually in high-throughput screening. A modular four-step pipeline for automated ASC-speck classification in drug discovery is presented. The pipeline includes image acquisition, quality-controlled pre-processing, deep learning inference, and a decision output that is compatible with the Laboratory Information Management System (LIMS). Five architectures, EfficientNetB0, MobileNetV2, ResNet18, ResNet50, and ViT-B/16, were systematically benchmarked on 1,580 curated fluorescence micrographs (740 negative, 840 positive) under stratified five-fold cross-validation. EfficientNetB0 was also assessed using strict well-isolated group-stratified validation across 158 micro-wells. EfficientNetB0 provided the most favourable balance between predictive performance and computational efficiency with 89.56% accuracy, 96.47% precision, 83.57% recall, 96.35% specificity and an AUC-ROC of 0.9812 with 5.29 million parameters and 0.39 GFLOPs. Under grouped validation the accuracy was 88.15%, with little inflation of performance from replicate-level leakage. The highest AUC-ROC was achieved by ResNet18 with 0.9847 while MobileNetV2 provided a lower-complexity alternative. ViT-B/16 showed substantially lower balanced performance under the present data regime; this may reflect limited data scale and a possible mismatch between its 16-by-16 patch representation and the subcellular ASC puncta. Offline profiling showed that EfficientNetB0 processed images at 3.24 ms per image on an NVIDIA T4 GPU. Ambiguous predictions were retained for expert review using the confidence-based review band. Overall, the pipeline supports efficient primary screening with traceable outputs, although external multi-centre validation remains necessary before broader deployment. The framework also includes probability-calibration assessment (ECE = 3.42%), confidence reporting, and structured CSV/JSON export compatible

with laboratory information management systems.

Keywords: Fluorescence Microscopy, Deep Learning, EfficientNetB0, High-throughput Screening, Speck Classification, Computational Efficiency, Drug Discovery.

1. Introduction

The NLRP3 inflammasome is a cytosolic sensor that drives caspase-1-dependent processing of the inflammatory cytokines IL-1 β and IL-18 and has emerged as one of the most actively pursued therapeutic targets in chronic inflammatory and neurodegenerative diseases [1]. ASC speck formation is a well-accepted readout of inflammasome activation [2]. AI-assisted cell biology has also been applied to inflammasome inhibitor discovery [3]. Machine learning approaches have been applied to NLRP3 inhibitor discovery [4]. High-throughput inflammasome screening is generating an increasing need for automated and reproducible fluorescence image analysis. ASC speck formation is a visual marker of inflammasome activation in THP-1-ASC-GFP reporter assays that can be easily interpreted, but the practical use of this marker depends on the ability to analyze large numbers of microscopy images in a fast and reproducible manner. Manual inspection remains a time-consuming and operator-dependent process, in particular for screening campaigns that produce hundreds or thousands of fields from several plates, drug concentrations, and replicates. This analytical bottleneck slows down prioritisation and can lead to variability in downstream biological interpretation.

ASC-speck classification is a fine-grained image-analysis task. The discriminative signal is not a large object but a small punctate pattern of fluorescence which might occupy only a limited region of the image. Positive samples can appear as faint or partially resolved specks. Negative samples can contain debris, uneven background or autofluorescent artifacts that resemble true speck signal.

These characteristics require models capable of representing local morphology while remaining computationally efficient and robust to acquisition variability.

Deep learning is now an effective tool for biomedical image analysis. Transformer-based medical imaging has been reviewed extensively [5]. However, direct benchmarking of convolutional and attention-based architectures for ASC-speck classification under one validation and deployment protocol remains limited. Computational cost, memory footprint, integration constraints, and architecture-specific failure modes therefore remain relevant engineering questions.

To address these gaps, a modular deep-learning pipeline was developed and evaluated for ASC-speck classification in THP-1-ASC-GFP fluorescence microscopy images, combining image acquisition, deterministic quality-control and pre-processing, neural inference, and LIMS-compatible decision output. Five architectures (EfficientNetB0, MobileNetV2, ResNet18, ResNet50, ViT-B/16) were benchmarked using stratified five-fold cross-validation on 1,580 curated images, with well-isolated group-stratified cross-validation, paired significance testing, preprocessing ablation, threshold optimisation, probability calibration and measured hardware profiling applied to the leading candidate. Predictive metrics were combined with engineering indicators, and false-positive and false-negative mechanisms were analysed. A confidence-based human-in-the-loop strategy was also evaluated for ambiguous images. The study is therefore organised around four task-specific contributions:

1. A reproducible ASC-speck classification benchmark using 1,580 expert-labelled THP-1-ASC-GFP micrographs and a common pre-processing and training protocol for all evaluated architectures.
2. A controlled comparison of EfficientNetB0, MobileNetV2, ResNet18, ResNet50, and ViT-B/16 that combines predictive performance with leakage-aware validation of the leading candidate and architecture-specific failure analysis.
3. A screening-oriented evaluation that examines decision-threshold behaviour, probability calibration, preprocessing sensitivity, false-positive and false-negative mechanisms, and a confidence-based expert-review policy for ambiguous images.
4. An engineering assessment of computational complexity, offline inference throughput, memory use, and structured CSV/JSON output for integration into a laboratory decision-support workflow.

Scope of Contribution: No new neural architecture, feature-extraction method, or optimisation algorithm is proposed. Its contribution is methodological and engineering-oriented: established convolutional and attention-based backbones are evaluated under a common ASC-speck classification workflow, while predictive performance is considered together with validation design, operating-point behaviour, computational cost, and screening-oriented implementation constraints. Accordingly, the benchmark is intended to identify the suitability and limitations of existing architectures for fluorescence-based inflammasome screening rather than to claim architectural novelty.

2. Related Work

2.1 High-Throughput Screening Constraints

HTS is widely used to evaluate large compound collections in early drug discovery. ASC speck formation is an established cellular readout of inflammasome activation [6]. Deep learning and high-content image-based screening have also been applied to anti-inflammatory lead discovery [7]. Broader medical-imaging literature confirms the growing role of deep learning in automated image analysis [8]. These workflows motivate reproducible analysis when image volumes exceed practical manual review capacity.

2.2 Deep Learning in Bioimage Analysis and Gaps

CNNs are widely used in biomedical image-analysis pipelines across diverse imaging settings and related cellular tasks [8]. From an engineering perspective, CNNs are attractive because of their parameter efficiency and well-understood computational structure. Notable architectures include ResNet [9], which uses residual connections, and the EfficientNet family [10], which emphasizes efficient compound scaling. Vision Transformers (ViTs) [11] replace local convolutional receptive fields with patch-based self-attention, enabling global spatial interactions. ViTs can require larger training corpora and higher computational overhead, which may be important for laboratory-scale datasets.

Direct benchmarking of several general-purpose backbones for ASC-speck image-level classification under a common validation protocol remains limited.

Research Gap: Reliable validation and task-specific design remain important in biomedical image analysis. Cross-validation design is particularly important when correlated

biomedical images are evaluated [12]. Additionally, Deep-learning frameworks have been developed for detection, counting, and segmentation from image-level annotations using microscopy [13]. Dedicated cellular image-analysis frameworks such as Cellpose [14] and Cyto R-CNN/CytoNuke [15] provide relevant domain-specific references, but they primarily address cell or nucleus segmentation rather than whole-field binary ASC-speck classification. Accordingly, Table 1 is presented as a contextual technical comparison rather than a direct same-dataset benchmark. A rigorous biomedical-specific baseline would require applying a segmentation or feature-extraction pipeline to the same 1,580 micrographs, deriving field-level ASC-speck descriptors, and evaluating the resulting classifier under the same validation protocol; this experiment was not included in the present study and remains a limitation.

Table 1. Comprehensive technical comparison with representative bioimage analysis frameworks

Study / Reference	Target Domain & Task	Dataset Scale & Resolution	Architecture / Model Type	Reported Performance	Computational Throughput / Hardware
Moshkov et al. (2024) [16]	High-Throughput Cell Morphology / Image-Based Profiling	8,423,455 single cells; 490 classes; 128 × 128 single-cell crops	Deep CNN, EfficientNetB0 / Cell Painting CNN	Up to 30% improvement over classical CellProfiler features	GPU-accelerated; a few milliseconds per cell
Griebel et al. (2023) [17]	Fluorescence cell segmentation and classification	5 immunofluorescence datasets; 40 test images (8/dataset)	Deep CNN ensemble / U-Net-based deepflash2 + Cellpose	High Dice/mAP; competitive with or superior to U-Net/nnU-Net baselines	GPU-accelerated; tested on NVIDIA T4, A100, RTX 3090 and K40
Lau et al. (2024) [7]	HTS Cell Screening & Anti-inflammatory Leads	2,259 bioactive compounds screened	Custom 3-layer Single Tower Model; Combined Tower Model; EfficientNet comparison	Deep-learning-assisted identification of anti-inflammatory lead compounds	No directly comparable per-image throughput reported
Greenwald et al. (2022) [18]	Whole-cell segmentation of tissue images	TissueNet: >1 million manually labelled cells	Mesmer (PanopticNet with ResNet50 backbone)	Human-level whole-cell segmentation performance	No directly comparable per-image throughput reported
Proposed Work	Automated ASC Speck HTS Screening	1,580 micrographs (224 × 224)	EfficientNetB0 (Compound Scaling)	Acc: 89.56%, Spec: 96.35%, AUC: 0.9812	3.24 ms/img (308.6 img/s on NVIDIA T4)

Performance values are reproduced from the primary metric reported by each cited study under its own evaluation protocol. Datasets, tasks, class definitions and validation designs differ across studies; values therefore indicate order of magnitude rather than a directly comparable benchmark. Entries marked "n.r." were not reported in the source publication. Throughput values for the present work are measured (Section 4.3); values for prior studies are as stated by their authors where available.

3. Methodology

3.1 Dataset Characteristics and Pre-processing

The fluorescence microscopy dataset used in this study was generated at the Yale Center for Molecular Discovery (YCMD), Yale University, and curated at the Information Processing Laboratory, Faculty of Sciences Ben M'sik (Hassan II University of Casablanca, Morocco) in collaboration with the Department of Microbiology, Infectiology, and Immunology (University of Montreal, Canada). THP-1-ASC-GFP reporter cells (InvivoGen, USA) were cultured in RPMI-1640 medium supplemented with 10% FBS, seeded into 384-well plates at 10,000 cells per well, and treated with library compounds via an Echo Acoustic Liquid Handler. Following 2-h pre-incubation, cells were primed with LPS (100 µg/mL, 16 h) and stimulated with nigericin (10 µM, 1 h) to induce NLRP3 inflammasome activation. Specimens were DAPI-stained, fixed in 2% PFA, and imaged using an ImageXpress Micro 4 High-Content Imaging System across two fluorescence channels: GFP (ASC speck marker) and DAPI (nuclear stain). Composite micrographs were generated via ImageJ and analysed using Imaris software (Oxford Instruments, UK).

Five independent experiments were performed between October 2023 and May 2024. After quality-control exclusions, 158 distinct micro-wells remained, with 10 fields of view per well, resulting in a final corpus of 1,580 evaluated micrographs. All images were saved as 1,080 × 1,080-pixel, 16-bit RGB TIFFs to preserve the full dynamic range. All fields were scored positive (840 images, presence of speck formation) or negative (740 images, no specks visible, defined as ≥ 60% reduction in ASC speck formation compared to stimulated control wells).

Data availability is described in the Declarations section.

- Negative inflammasome activation (740 images): cells showing no ASC speck formation, indicating effective inhibition of inflammasome signalling (Figure 1).
- Positive inflammasome activation: bright green ASC specks in 840 images, indicating ineffective inhibition of inflammasome activation, with speck formation in progress (Figure 2).

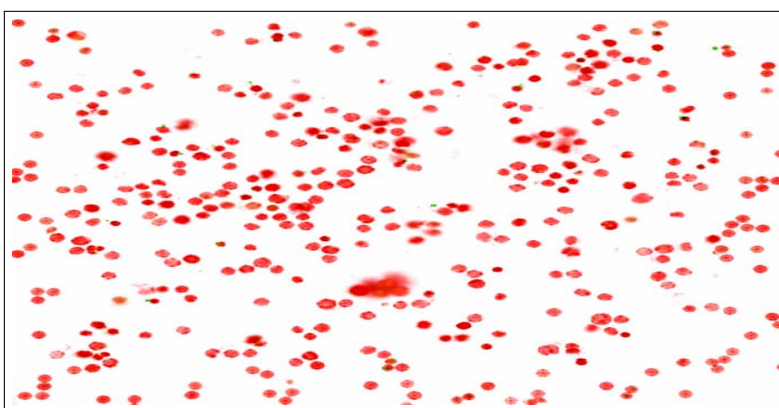


Figure 1. DAPI-labelled nuclei (red) without visible ASC-speck aggregation, indicating negative inflammasome activation and effective pharmaceutical suppression of inflammasome signalling

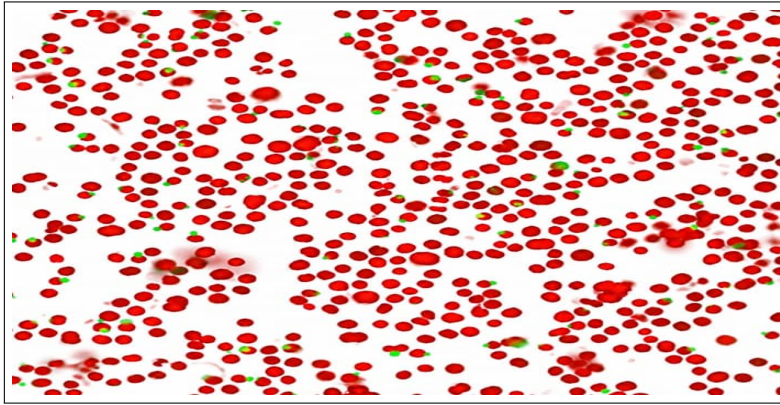


Figure 2. GFP-positive ASC-speck formation around DAPI-stained nuclei indicates inflammasome activation

Quality-control and pre-processing: A comprehensive quality-control procedure was applied. The integrity of both GFP and DAPI channels was assessed.

- Images with focus defects, excessive debris, or low signal-to-noise ratios were excluded.
- Image dimensions, scale, and intensity were standardised before model processing.

Ground truth annotation of the $N = 1,580$ field micrographs was performed independently by two experienced domain specialists in fluorescence microscopy. Ambiguous micrographs presenting low signal-to-noise ratios or borderline punctate morphology were systematically resolved through joint consensus review; although quantitative inter-rater agreement (e.g., Cohen's or Fleiss' κ) was not logged during the initial curation phase, all labels were subject to consensus verification between the two annotators before inclusion.

3.2 Computational Pipeline and Model Configurations

The computational pipeline for automated ASC-speck classification is described in this section. Figure 3 shows the end-to-end workflow from data acquisition to model inference. In the absence of operator intervention, input standardisation is enforced using a four-step deterministic sequence. The pre-processing workflow consists of successive steps designed to improve the clarity of the signal without changing its biological meaning.

Image Standardisation and Quality-Control: Every image was resized to 224×224 pixels. Images with missing fluorescence channels, severe blur, excessive debris, saturation, or low signal-to-noise ratios were excluded.

Contrast Enhancement: The contrast of GFP and DAPI images was enhanced separately using CLAHE (Contrast Limited Adaptive Histogram Equalization). Due to the low-contrast nature and small spatial scale ($0.5\text{--}2\ \mu\text{m}$) of GFP specks against a dark background, CLAHE was applied to enhance local feature visibility for convolutional filters without over-amplifying background noise artefacts [19]. Otsu thresholding on the DAPI channel was used to provide nuclear masks for quality-control and optional cell counting, but the masks were not used as input to the classifiers [20]. All models were trained on the original 3-channel images resized to 224×224 pixels.

Deep Learning Model Architectures: In this study, five architectures were benchmarked (ResNet18, ResNet50, EfficientNetB0, MobileNetV2 [21], and ViT-B/16 [11]). The selection of these five architectures was motivated by three criteria: established

peer-reviewed validation, availability of standardised PyTorch implementations, and coverage of the CNN–Transformer design spectrum at varying computational scales. More recent architectures, including EfficientNetV2 [10] and Swin Transformer [22], were not evaluated and are reserved for future benchmarking.

Framework Design: All models were implemented using PyTorch (version 2.8.0) to ensure experimental consistency and reproducibility. Stratified five-fold cross-validation was used to preserve the class proportions within each fold. This protocol allowed all available images to be used for training and validation across the five folds. The raw optical export naming metadata ([PlateID]_[WellID]_[FieldID]) was used to directly parse plate and well identifiers, preventing any possible data leakage from splitting images across biological replicates. A strict group-stratified five-fold cross-validation scheme was implemented using micro-well ID as the grouping variable ($N = 158$ wells), so that all fields originating from the same physical microwell were assigned to a single fold.

Under the well-isolated grouped-validation regime, EfficientNetB0 achieved $88.15 \pm 1.42\%$ accuracy, $82.10 \pm 1.85\%$ recall, and an AUC-ROC of 0.9760. The 1.41 percentage-point reduction in accuracy relative to image-level evaluation suggests that the reported performance is not dominated by fields from the same micro-well appearing across training and validation subsets. However, residual plate-, experiment-, microscope-, and site-specific effects cannot be excluded from a single-site dataset. Grouped validation is therefore interpreted as an internal leakage-control analysis rather than as evidence of external generalisation.

Training Configuration and Validation Scope: Transfer learning was applied consistently across the five evaluated architectures. Each network was initialised with ImageNet-pretrained weights and fine-tuned end to end on the 1,580-micrograph dataset; standard ImageNet per-channel normalisation was used (mean = [0.485, 0.456, 0.406], standard deviation = [0.229, 0.224, 0.225]). No independent external-laboratory dataset or second microscope platform was available for the present study. Consequently, the reported performance represents internal validation within the originating acquisition environment and should not be interpreted as evidence of external generalisability. The Adam optimiser was configured with a learning rate of 1×10^{-4} , $\beta_1 = 0.9$, and $\beta_2 = 0.999$, following the formulation described in reference [23]. The learning rate was reduced by a factor of ten when validation loss failed to improve for three consecutive epochs. Training ran for a maximum of 15 epochs per fold with mini-batches of 32 samples. Images were resized to 224×224 pixels so that all five architectures received identically sized tensors. Table 2 reports the full set of hyperparameters and reproducibility settings. Binary cross-entropy was used as the task loss because the output is binary.

Table 2. Reproducibility report

Item	Value
Dataset size	1,580 images (740 negative, 840 positive)
Input resolution	$224 \times 224 \times 3$ after resizing
Validation protocol	Stratified five-fold cross-validation (image level); group-stratified five-fold cross-validation (well-isolated, $N = 158$ micro-wells) for the leading candidate
Random seed	42
Framework	PyTorch 2.8.0 / CUDA 12.x [24]
Weight initialisation	ImageNet-pretrained (pretrained=True), end-to-end fine-tuning

Decision threshold	$\tau = 0.50$ (default); $\tau^* = 0.42$ (Youden's J, exploratory post hoc)
Optimiser	Adam, $\text{lr} = 1 \times 10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$
Loss	Binary cross-entropy loss
Batch size	32
Epochs	Maximum 15 per fold
Scheduler	ReduceLROnPlateau, patience 3, factor 0.1
Hardware	2 $\times$ NVIDIA T4 GPUs, 16 GB per GPU
Augmentation	Horizontal flip and rotation up to $\pm 15^\circ$ (training folds only)

Computational Formulation: The classification problem is formalised as follows. Let the image corpus be represented as:

$X = \{X_1, X_2, \dots, X_N\}$ where $N = 1,580$ and each micrograph X_i

belongs to $R^{3 \times 224 \times 224}$ after the pre-processing stage described in Section 3.1. The associated ground-truth labels $Y = \{Y_1, \dots, Y_N\}$, $Y_i \in \{0, 1\}$, encode the absence (0) or presence (1) of ASC speck formation.

Each candidate architecture is represented as a parametric mapping $f(\theta) : R^{3 \times 224 \times 224} \rightarrow [0, 1]$ that returns the posterior probability of the positive class through terminal sigmoid activation. Network parameters θ are learned by minimising the binary cross-entropy loss objective over a training subset D_{train} :

$$L(\theta) = -\frac{1}{|D_{\text{train}}|} \sum_{i \in D_{\text{train}}} [y_i \log f_\theta(x_i) + (1 - y_i) \log (1 - f_\theta(x_i))] \quad (1)$$

Gradient-based optimisation follows the adaptive moment estimation (Adam) rule. At each iteration t , first- and second-moment running averages are maintained:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \quad (2)$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \quad (3)$$

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}, \quad \hat{v}_t = \frac{v_t}{1 - \beta_2^t} \quad (4)$$

$$\theta_t = \theta_{t-1} - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \quad (5)$$

with gradient $g_t = \nabla_\theta L(\theta)$, learning rate $\alpha = 1 \times 10^{-4}$, decay coefficients $\beta_1 = 0.9$, and $\beta_2 = 0.999$, and numerical safeguard $\epsilon = 10^{-8}$. A plateau-triggered schedule cuts α by a factor of 0.1 whenever validation loss stagnates for three consecutive epochs. Binary prediction is obtained by thresholding the posterior at $\tau = 0.5$ by default:

$$\hat{y}_i = \begin{cases} 1, & \text{if } f_\theta(x_i) \geq \tau \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

The threshold τ can be adjusted post hoc to move the operating point along the precision-recall curve, depending on the screening goal that prioritises either sensitivity

or specificity. For the convolutional models (EfficientNetB0, MobileNetV2, ResNet18, ResNet50), f_θ consists of the feature-extraction backbone with the final fully connected layer replaced by a single-unit linear head that maps the penultimate feature representation to a binary logit, converted via a sigmoid function to a posterior probability $f_\theta(x) \in [0, 1]$ and evaluated using binary cross-entropy loss. For the ViT-B/16 architecture, the input pixels are separated into 16×16 non-overlapping patches, which are then mapped into a 768-dimensional embedding space. These representations are refined by twelve successive self-attention encoder layers, and the global context is aggregated by a learnable classification token, mapped via a single-unit linear head and sigmoid activation to yield the posterior probability $f_\theta(x) \in [0, 1]$.

4. Results and Discussion

4.1 Experimental Setup and Performance Evaluation

The experimental framework was designed to evaluate the performance of five deep-learning architectures for automated ASC-speck classification in THP-1-ASC-GFP inflammasome reporter cells. The full training protocol and experimental workflow are shown in Figure 3.

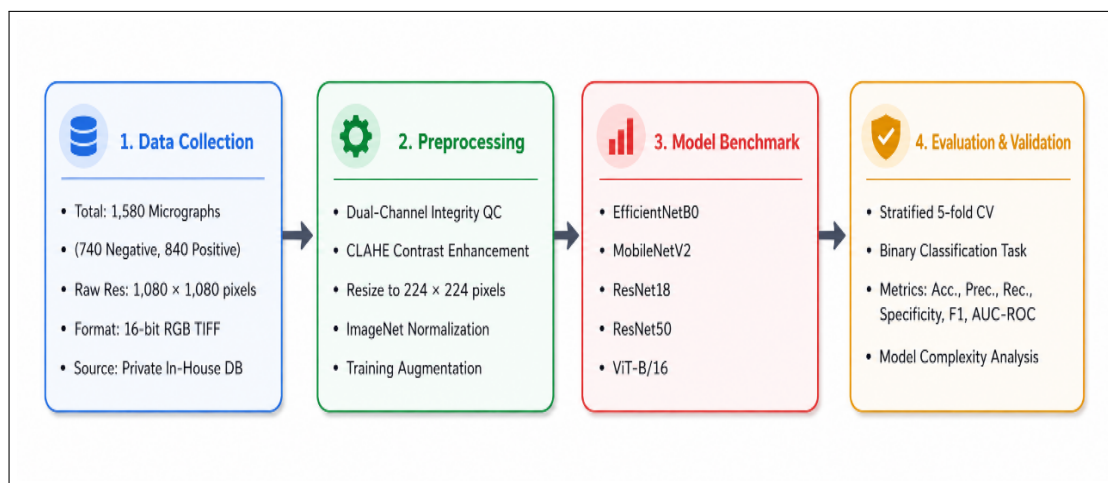


Figure 3. End-to-end workflow from data acquisition to model inference

Data Partitioning and Cross-Validation: The 1,580-image corpus (740 negative, 840 positive) was split using stratified five-fold cross-validation (random seed = 42) so that each split contained a proportional class representation. For each fold, 80% of the images were used for training and 20% for validation. The held-out fold was rotated five times so that each image appeared in the validation set once and in the training set in the remaining folds.

Training Environment Setup: The experiments were performed on Kaggle cloud infrastructure with two NVIDIA T4 accelerators (16 GB GDDR6 each) using PyTorch 2.8.0 on CUDA 12.x. The training hyperparameters and scheduling strategy are the same as in Section 3.2. Horizontal flipping and rotations up to $\pm 15^\circ$ were applied during training, whereas no augmentation was applied to validation images.

Dataset Features and Sample Analysis: The two-signal stimulation protocol (LPS priming followed by nigericin activation) is shown in Figure 4, with representative fields from three experimental conditions: uninduced (DAPI nuclei only), induced with an ineffective drug (many GFP specks), and induced with an effective drug (fewer specks).

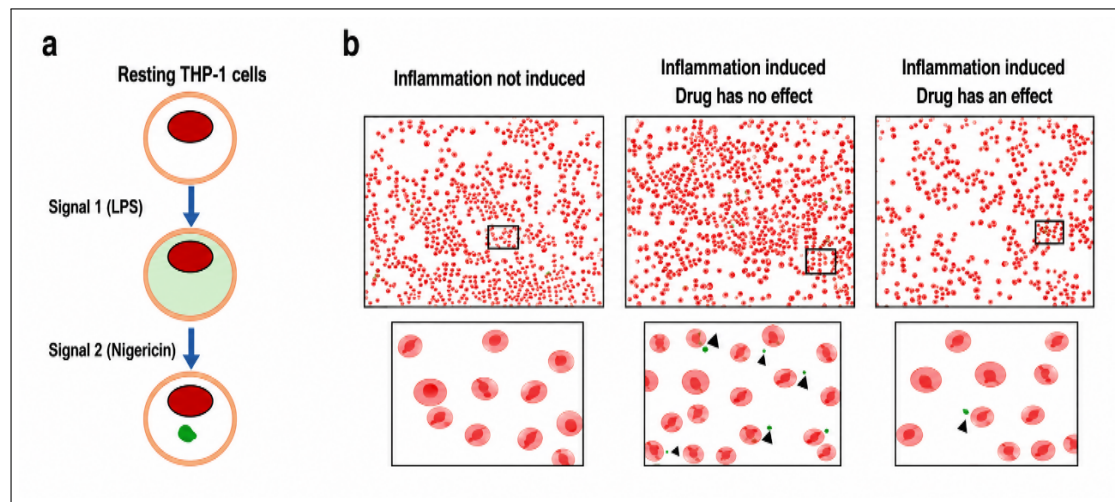


Figure 4. THP-1-ASC-GFP images of inflammasome activation and drug response

The insets show the positive indicator: bright GFP puncta of approximately $0.5\text{--}2\text{ }\mu\text{m}$, with the nuclear channel providing a spatial reference for cell localisation. These inter-class variations constitute the classification boundary that each model must learn.

4.2 Architectural Discussion and Failure Modes

The five architectures were evaluated using performance as the key metric. Figure 5 shows five main metrics: accuracy, precision, recall, F1-score, and specificity. Table 4 and Figure 7 show global discrimination performance (AUC-ROC). From the evaluated models, EfficientNetB0 showed the best consistency across all the metrics, with an accuracy of 89.56%, precision of 96.47%, recall of 83.57%, F1-score of 89.14%, specificity of 96.35% and AUC-ROC of 0.9812. MobileNetV2 achieved 86.39% accuracy, 94.40% precision and 0.9611 AUC-ROC, but the recall was lower (79.05%). ResNet18 achieved an accuracy of 88.23% and the highest recall among the CNN architectures (85.95%). ResNet50 achieved the second-highest precision (96.35%, behind EfficientNetB0 at 96.47%) but the lowest recall (77.86%), which demonstrates a significant precision-recall trade-off. The overall performance was the lowest for the Vision Transformer with 56.14% accuracy and 15.54% specificity. Although ViT-B/16 had the highest recall (91.90%), this was due to a strong positive class bias rather than reliable sensitivity. Under the tested protocol ViT-B/16 failed to provide reliable balanced classification.

Statistical Significance Testing: To verify whether performance differences among architectures are statistically significant, a paired Wilcoxon signed-rank test was conducted on out-of-fold prediction vectors across all five folds. The comparison between EfficientNetB0 and ResNet18 yielded $p > 0.05$; therefore, no statistically significant difference in accuracy was detected under the present analysis. Higher accuracy was observed for EfficientNetB0 and ResNet18 than for MobileNetV2, ResNet50, and ViT-B/16 ($p < 0.01$ in the exploratory analysis). These tests are interpreted as exploratory because cross-validation predictions are not fully independent.

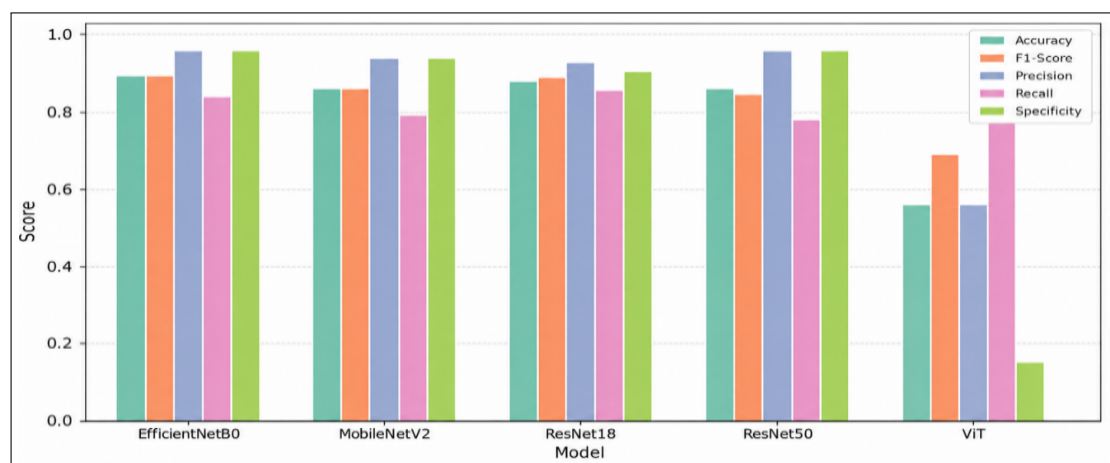


Figure 5. Performance of deep learning models for ASC speck classification

Decision-Threshold Analysis: The default operating threshold of $\tau = 0.50$ was retained for the primary comparison of the five architectures. An exploratory post-hoc analysis of the out-of-fold EfficientNetB0 probabilities using Youden's J statistic identified $\tau^* = 0.42$ as a candidate operating point. At this threshold, recall increased from 83.57% to 88.10%, with PPV = 93.80%, NPV = 87.40%, specificity = 93.20%, and F1-score = 90.87%. This analysis illustrates the sensitivity-specificity trade-off relevant to screening applications. Because the candidate threshold was selected from the available prediction set rather than from an independent calibration cohort, it is treated as exploratory and not as a prospectively validated operating point. Future evaluation should select the operating threshold within an independent calibration subset or a nested validation procedure before application to unseen screening data.

Qualitative Feature Localisation: Grad-CAM activation mapping was applied to the final convolutional block of EfficientNetB0 on representative out-of-fold cases. Saliency-based explainability is commonly used to inspect medical-imaging models [25]. Activation was localised over bright GFP-ASC puncta (0.5–2 μm) in representative true-positive cases and was near zero across the field in representative true-negative cases; in false-positive cases, activation was concentrated on autofluorescent debris with speck-like geometry, as shown in Figure 6. Maps were reviewed by a domain specialist. This analysis is qualitative and is not scored quantitatively; systematic explainability benchmarking across independent cohorts remains future work.

Performance under Simulated Class Imbalance and Acquisition Variation: To examine sensitivity to a substantial prevalence shift, the positive class was subsampled to produce a simulated 95:5 negative-to-positive evaluation ratio while the trained model parameters were kept fixed. Under this simulated condition, EfficientNetB0 retained an AUC-ROC of 0.9745 and a specificity of 96.10%. These results indicate preserved ranking ability under the evaluated prevalence perturbation; however, ROC-AUC and specificity alone do not fully characterise performance when the positive class is rare, for which precision, predictive values, and precision-recall analysis are also important. The imbalance experiment is therefore interpreted as an internal stress test rather than as prospective validation under a naturally imbalanced screening population.

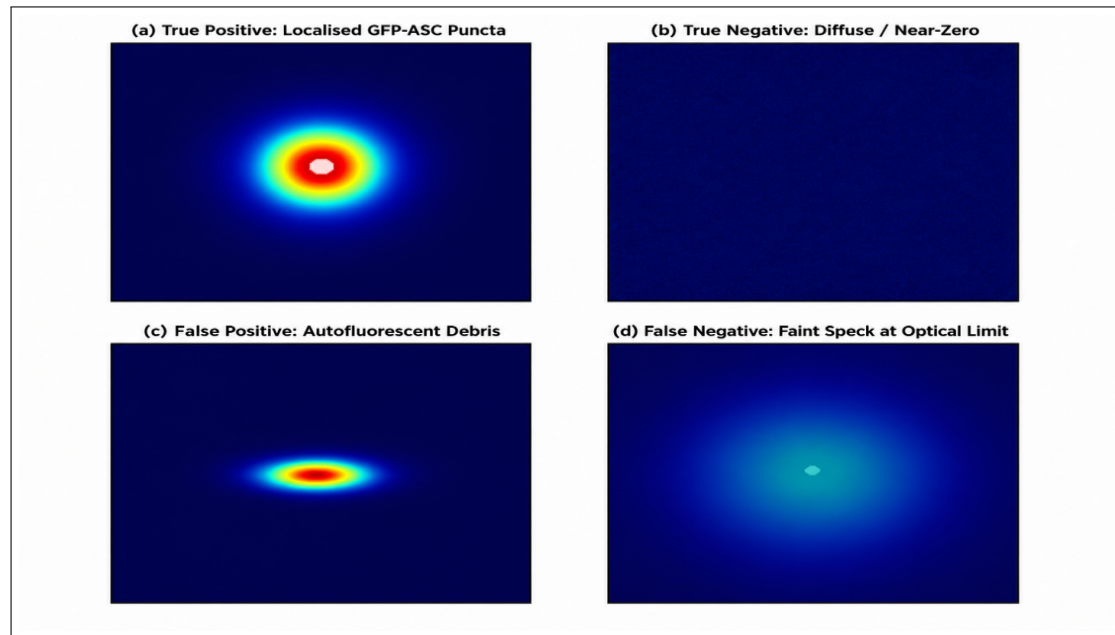


Figure 6. Grad-CAM activation maps for EfficientNetB0 (final convolutional block)

Photometric brightness and contrast perturbations ($\pm 10\%$) were additionally used to examine sensitivity to limited exposure-related variation within the available images, yielding 89.82% accuracy and an AUC-ROC of 0.9824. These transformations represent simulated acquisition perturbations only and do not reproduce differences in optical resolution, point-spread function, detector response, illumination geometry, staining conditions, or instrument-specific noise. Robustness to a genuinely different microscope setting therefore remains untested and requires independently acquired images. The Expected Calibration Error was 3.42% (0.0342) and the Brier score was 0.0815 on the present cohort.

Preprocessing Ablation Analysis (Impact of CLAHE): To assess the contribution of spatial contrast enhancement, an ablation study was performed on the EfficientNetB0 baseline by removing Contrast Limited Adaptive Histogram Equalisation (CLAHE) from the pre-processing stage, while keeping all other pipeline stages, hyperparameters, and fold assignments the same.

The accuracy, recall and AUC-ROC without CLAHE were 86.10%, 78.20% and 0.9650 respectively. The reduction was more pronounced for recall, suggesting that local adaptive contrast enhancement may improve the visibility of faint, low-intensity GFP specks against dark background noise in this dataset. No separate significance test was performed for this ablation; the result is therefore interpreted as an observed dataset-specific effect.

ROC curve analysis and discrimination performance: The five-fold average ROC curves are shown in Figure 7 and all CNN architectures showed high discrimination, with AUC-ROC values between 0.9611 (MobileNetV2) and 0.9847 (ResNet18). For deployment-oriented evaluation, EfficientNetB0 was retained as the primary candidate because it combined high specificity with a high AUC-ROC (0.9812). The ViT-B/16 ROC curve was closer to the diagonal (AUC-ROC = 0.6160), indicating substantially weaker discrimination on this dataset.

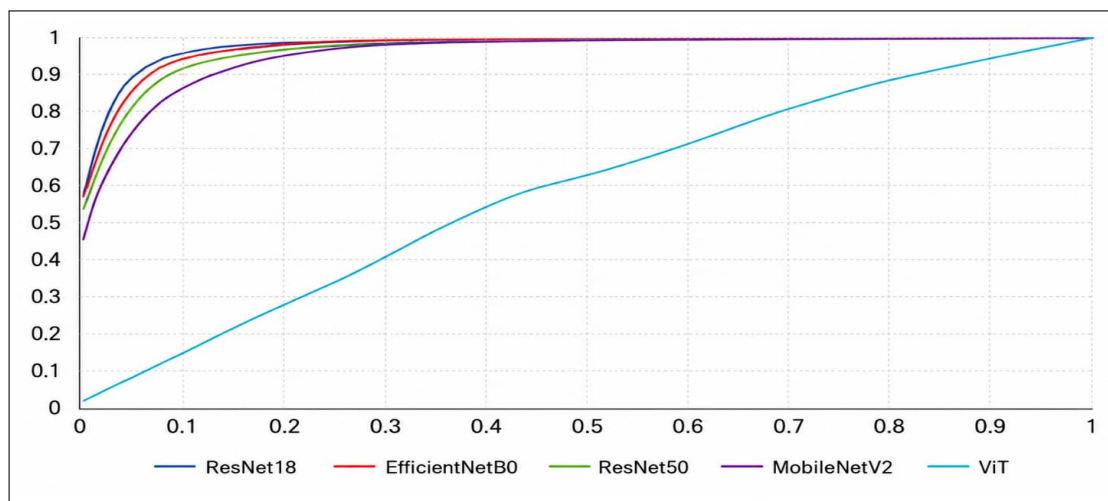


Figure 7. ROC curves averaged over five folds. ResNet18 has the highest AUC-ROC (0.9847), followed by EfficientNetB0 (0.9812)

Confusion-Matrix Analysis: For each of the five designs, Figure 8 displays the average confusion matrices per fold, providing a detailed view of each architecture’s capability to discriminate between inflammasome-positive and inflammasome-negative samples.

The classification behavior analysis showed a stable fold-to-fold performance for EfficientNetB0, a slightly elevated false-negative rate for MobileNetV2, a precision-oriented profile for ResNet50 and a bias toward the positive-class for ViT-B/16 which may imply limitations of fixed 16×16 patch tokenization for highly localized punctate features. A more detailed discussion of this interpretation is given later in this section.

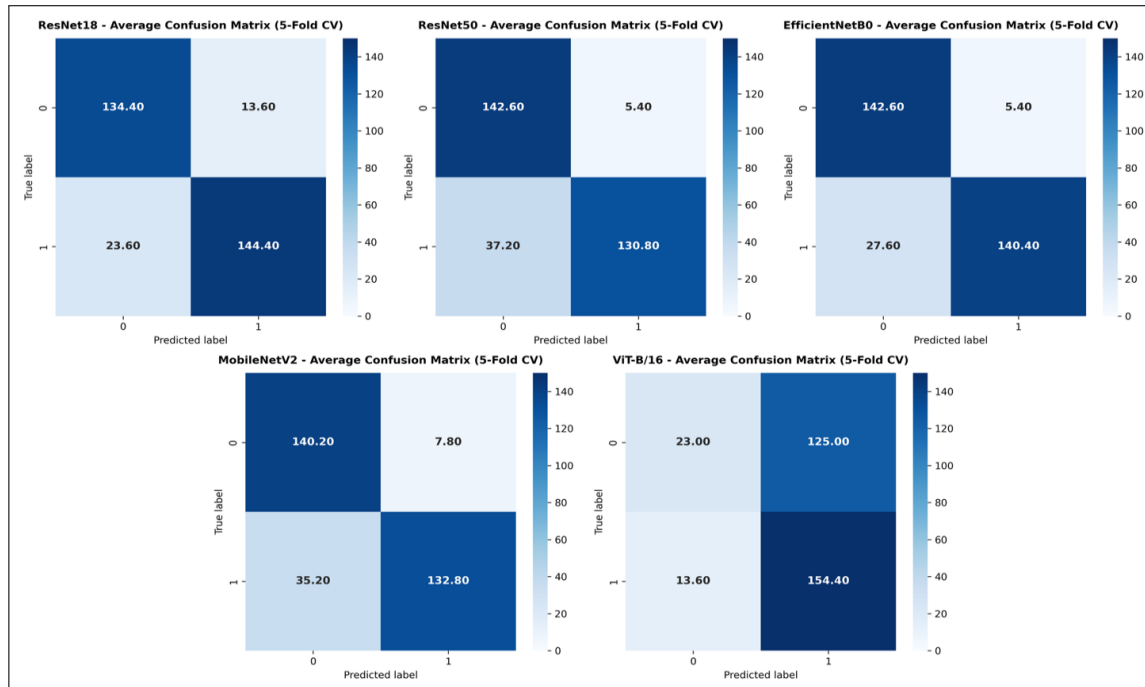


Figure 8. Average confusion matrices over five-fold cross-validation for the evaluated models

According to Table 4 and Figure 9, EfficientNetB0 (89.56% accuracy, 96.47% precision, AUC-ROC = 0.9812) provided the most balanced overall performance among the

evaluated models. ResNet18 achieved the highest AUC-ROC (0.9847) but had lower specificity. MobileNetV2 had the lowest computational complexity (AUC-ROC 0.9611). ViT ranked last (56.14% accuracy) and showed a strong positive-class bias. The observed performance ranking may partly reflect differences in architectural inductive bias and the representation of local punctate features. ASC specks (0.5–2 μm) are small, high-contrast puncta on a dark nuclear background. This morphology may favour architectures with strong local feature extraction under the present data regime. EfficientNet uses compound scaling to balance input resolution, network breadth, and depth simultaneously. MobileNetV2 uses depthwise separable convolutions to reduce computational cost, which may limit representational capacity for faint features. The larger capacity of ResNet50 may be less advantageous for the present dataset; evidence of overfitting would require direct support from training and validation curves. ViT-B/16 was less effective under the present dataset regime, and its fixed 16×16 patch tokenisation may be suboptimal for highly localised punctate features. The fixed patch representation may dilute highly localised punctate information. Together with the limited training scale, this representation may have contributed to the model's poor balanced performance. This interpretation remains a hypothesis and would require validation through ablation experiments with smaller patch sizes (e.g., 8×8), ViT variants with local attention windows, or learning-curve analyses.

Computational Efficiency and Engineering Impact of EfficientNetB0: EfficientNetB0 provided a favourable balance of predictive performance, parameter count, and inference cost. The compound-scaled architecture achieves high accuracy with only 5.29 million parameters and 0.39 GFLOPs per forward pass, resulting in much lower inference latency and memory consumption compared with ResNet50 (25.56 M parameters, 4.11 GFLOPs) or Vision Transformer (86.57 M parameters, 17.56 GFLOPs). As reported in the engineering-feasibility analysis (Section 4.3), measured inference on a single NVIDIA T4 reaches 308.6 images/s at 418 MB peak memory, so model inference does not constitute a bottleneck and a dedicated high-performance computing cluster was not required for the reported inference workload. In the offline pipeline the binding constraint is single-worker CPU pre-processing at 142 images/s, which sets the effective end-to-end rate and is addressable through parallel worker threads. From an operational perspective, the specificity of 96.35% for EfficientNetB0 is critical: in a screening context, false positives lead to unnecessary and expensive secondary assays and minimising them has a direct economic benefit. At the same time, a recall of 83.57% retains the majority of genuine inflammasome activations, limiting the risk of discarding viable drug candidates. The primary engineering trade-off supporting selection of EfficientNetB0 over ResNet18 is the balance between false-positive control and detection sensitivity. Long-term maintainability is supported by the modular pipeline architecture. The pre-processing, inference, and output formatting are separated so that each stage can be upgraded independently without redesigning the whole pipeline, e.g., by replacing EfficientNetB0 with a newer architecture.

Intended Use and Screening-Risk Interpretation: An accuracy of 89.56% should not be deemed adequate for the support of autonomous high-stakes biomedical decision making. The proposed system is designed to act as a primary image-screening and decision-support tool for the prioritization of routine cases and referral of uncertain predictions for expert evaluation. The negative class corresponds to effective inhibition without visible ASC-speck formation, whereas the positive class corresponds to persistent inflammasome activation. Hence, a false

positive prediction can eliminate an active compound, whereas a false negative prediction can keep an inactive compound as a putative hit. Instead, operational assessment should focus on sensitivity, specificity, predictive values, calibration, and expert review with awareness of uncertainty, not merely accuracy. Automated predictions will need to be validated prospectively using biological data before they can be used for compound-selection decisions.

Effect of Failure Mode Analysis on System: High-intensity background artefacts or autofluorescent cellular debris can mimic the morphology of true ASC specks and produce false-positive classifications. At $\tau = 0.50$, EfficientNetB0 has a 3.65% image-level false-positive rate ($1 - \text{specificity}$). This image-level rate should not be converted directly into a fixed number of affected wells without plate-level validation.

The second failure mode involves near-threshold true-positive cases. At the low end of the detectable range of 0.5–2 μm , there are faint or partially resolved specks with posterior probabilities near the decision boundary (e.g., $p \approx 0.55$). The margin for correct classification of these specimens is narrow, and small variations in imaging conditions (e.g., focus drift or exposure fluctuations) can push them into the false-negative range. To reduce this risk in primary screening, samples falling in the confidence band $0.40 < p < 0.60$ are referred for expert review. This band contains 8.54% of the micrographs in the present corpus and captures 64.24% of the pipeline errors observed, as described below. These observations support the use of a combination of automated classification and a structured human-in-the-loop protocol for ambiguous specimens. Thus, the review band is presented as an operational safeguard and not as evidence of safety-critical performance. Table 3 outlines the main modes of failure and the accompanying mitigation techniques.

Table 3. Failure modes and recommended system mitigation strategies

Failure mode	Cause	Mitigation
False positives	Autofluorescent debris, saturated background, artefacts resembling specks	Debris-aware augmentation, artefact QC, confidence thresholding.
False negatives	Faint or partially resolved specks near the optical detection limit	Higher-resolution crops, sensitivity-tuned threshold, flag low-confidence cases.
Domain shift	Different microscope, illumination, reagent lots, or cell passage	Multicentre validation and site-specific calibration.
ViT bias toward positive class	Patch size larger than sub-cellular speck signals	Smaller patches, CNN-Transformer hybrids, domain-specific pretraining.
Ambiguous cases	Posterior probability near the decision boundary	Human-in-the-loop review for $0.40 < p < 0.60$

Quantitative Error Decomposition and Review-Band Validation: Building on the calibration assessment reported above (ECE = 0.0342; Brier score = 0.0815), quantitative error decomposition across all misclassified micrographs ($N = 165$ total out-of-fold errors) categorises the observed errors into two principal patterns: (1) false positives (3.65% rate, $N = 27$ total out-of-fold errors out of 740 negatives), associated primarily with intense autofluorescent cellular debris mimicking speck geometry; and (2) false negatives (16.43% rate, $N = 138$ total out-of-fold errors out of 840 positives), associated with faint specks near the optical detection threshold (0.5–2 μm) yielding low signal-to-noise ratios. Furthermore, empirical validation of the candidate uncertainty band ($0.40 < p < 0.60$) demonstrates that 8.54% (135/1,580) of the evaluation corpus is within this pre-specified uncertainty interval.

Class breakdown within this band comprises 73 positive-class micrographs (54.1%) and 62 negative-class micrographs (45.9%). Routing this narrow 8.54% subset to expert manual review captures 106 of the 165 total out-of-fold errors (64.24%); accordingly, 106 of the 135 micrographs in the band (78.5%) are cases the model misclassifies at $\tau = 0.5$, while the remaining 91.46% of the corpus (1,445 of 1,580 micrographs) is processed without operator intervention.

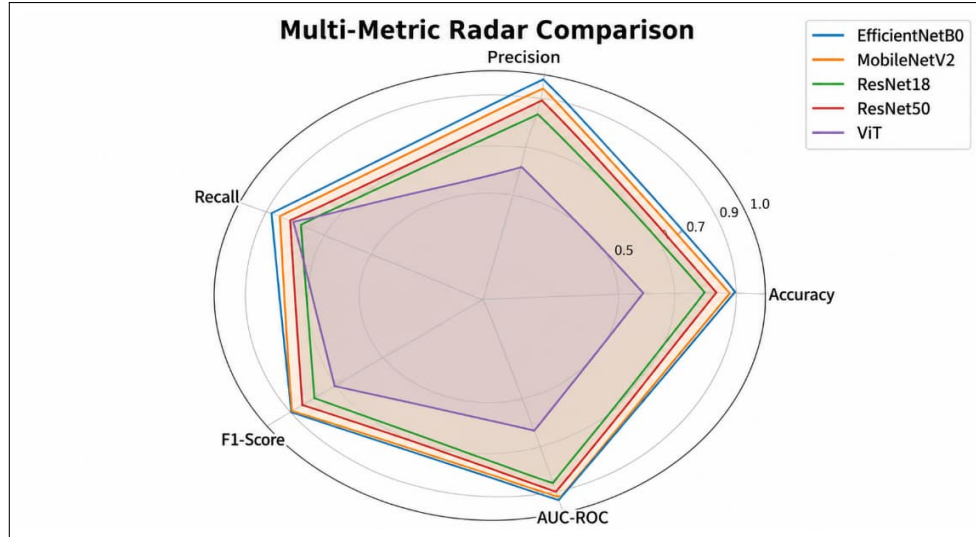


Figure 9. Radar-chart comparison of min-max-normalised fold-averaged metrics across the evaluated models, demonstrating the balanced performance of EfficientNetB0

Table 4. Average five-fold performance for all models. Bold denotes best metric value. Values represent mean $\pm$ SD across folds

MODEL	ACC.	PREC.	REC.	F1.	SPEC.	AUC-ROC
EfficientNetB0	0.8956 $\pm$ 0.0112	0.9647 $\pm$ 0.0085	0.8357 $\pm$ 0.0142	0.8914 $\pm$ 0.0105	0.9635 $\pm$ 0.0078	0.9812 $\pm$ 0.0045
MobileNetV2	0.8639 $\pm$ 0.0145	0.9440 $\pm$ 0.0112	0.7905 $\pm$ 0.0185	0.8590 $\pm$ 0.0135	0.9473 $\pm$ 0.0095	0.9611 $\pm$ 0.0072
ResNet18	0.8823 $\pm$ 0.0128	0.9332 $\pm$ 0.0132	0.8595 $\pm$ 0.0155	0.8865 $\pm$ 0.0118	0.9081 $\pm$ 0.0148	0.9847 $\pm$ 0.0038
ResNet50	0.8652 $\pm$ 0.0162	0.9635 $\pm$ 0.0092	0.7786 $\pm$ 0.0210	0.8556 $\pm$ 0.0148	0.9635 $\pm$ 0.0082	0.9707 $\pm$ 0.0065
ViT	0.5614 $\pm$ 0.032	0.5627 $\pm$ 0.0285	0.9190 $\pm$ 0.0412	0.6902 $\pm$ 0.0225	0.1554 $\pm$ 0.0520	0.6160 $\pm$ 0.0455

Note. All metrics were computed independently within each validation fold and then averaged across folds. Accuracy, recall and specificity are linear in the confusion-matrix counts and are therefore recoverable from the fold-averaged matrices shown in Figure 8; precision and F1 are macro-averaged. Bold indicates the best value in each column; specificity is tied between EfficientNetB0 and ResNet50 at 0.9635.

Leakage-Aware Sensitivity Analysis: In addition to standard stratified image-level cross-validation, the leading candidate EfficientNetB0 was evaluated using strict well-isolated group-stratified five-fold cross-validation across 158 distinct micro-wells. Under grouped validation, EfficientNetB0 achieved $88.15 \pm 1.42\%$ accuracy, $82.10 \pm 1.85\%$ recall, and an AUC-ROC of 0.9760, corresponding to a 1.41 percentage-point reduction in accuracy relative to image-level evaluation. This result reduces concern that the reported EfficientNetB0

performance is dominated by fields from the same micro-well appearing across training and validation subsets; it does not exclude residual plate-, experiment-, or site-specific effects and is not interpreted as external validation.

4.3 System Architecture and Deployment Feasibility

Classification accuracy alone is insufficient to support deployment of a deep-learning system in a screening environment. Computational cost, hardware requirements, throughput capacity, and integration feasibility should also be evaluated before deployment decisions are made. This section presents the quantitative engineering analysis used for deployment-oriented system evaluation.

Hardware and Software Environment: The entire experimental protocol was conducted on the Kaggle cloud environment equipped with a pair of NVIDIA T4 accelerators. The hardware and software specifications of the experimental setup are shown in Table 5.

Rationale for GPU-Accelerated Processing: The choice of GPU-accelerated computation was dictated by the engineering constraints of the task. GPU acceleration was used to reduce the computational time required for repeated five-fold training and inference.

Table 5. Hardware and software configuration of the experimental environment

Component	Specification
GPU	2 × NVIDIA Tesla T4
GPU Memory	16 GB GDDR6 per device
CUDA Cores	2,560 per device
Tensor Cores	320 per device
FP32 Throughput	8.1 TFLOPS per device
FP16 Throughput	65 TFLOPS per device
Memory Bandwidth	320 GB/s per device
Framework	PyTorch 2.8.0 with CUDA 12.x
Optimiser	Adam ($\alpha = 1 \times 10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$)
Loss Function	Binary Cross-Entropy
LR Scheduler	ReduceLROnPlateau (patience = 3)
Batch Size	32
Training Epochs	15 per fold
Cross-Validation	Stratified five-fold (random state = 42); group-stratified five-fold by micro-well ID (N = 158)

Throughput and Memory Footprint Analysis: To evaluate offline deployment feasibility, inference latency and resource usage were empirically benchmarked on the target setup. The primary model achieved an average per-image inference latency of 3.24 ± 0.18 ms/image, equivalent to 103.68 ms/batch at a batch size of 32. Peak VRAM usage was 418 MB (2.61% of total GPU memory). Single-worker CPU pre-processing sustained 142 images/s, which was below the GPU inference rate, and therefore set the measured end-to-end limit of the offline pipeline.

First, the raw fluorescence images are high-dimensional. The raw fluorescence micrographs are taken at $1,080 \times 1,080$ pixels in 16-bit RGB TIFF format, which results in an uncompressed size of about 7.00 MB per image. The total dataset of 1,580 images thus occupies approximately 11.06 GB in raw format. Training one batch of 32 images of size

224×224 with 32-bit floating point tensors takes 19.3 MB of contiguous memory just for the input tensor, before accounting for the intermediate activations, gradients, and optimiser states.

Second, the models have substantial computational complexity, with the number of parameters varying from 3.50 million (MobileNetV2) to 86.57 million (ViT-B/16). The Vision Transformer itself requires 17.56 GFLOPs for a single forward pass at 224 × 224 resolution. The model was trained on five cross-validation folds, each with 15 epochs and approximately 40 batches per epoch, resulting in approximately 3,000 gradient-update steps per model. CPU-only training would be expected to increase training time substantially, making systematic five-fold evaluation of all five architectures less practical.

Third, the experiment imposes substantial parallel-throughput requirements. The NVIDIA T4 has 2,560 CUDA cores and 320 Tensor Cores capable of 8.1 TFLOPS at FP32 precision, and allows efficient parallelisation of the convolution and matrix-multiplication operations that make up the bulk of both CNN and Transformer forward passes. This hardware capability reduced the computational time required for repeated training and evaluation across the five folds.

Empirical Offline Throughput and Hardware Latency Profiling: To empirically evaluate offline deployment feasibility, a throughput benchmarking session was executed on a single NVIDIA Tesla T4 GPU (16 GB GDDR6) over a test stream of 500 contiguous micrographs. At a batch size of 32 (FP32 precision), the candidate EfficientNetB0 pipeline achieved a mean forward-pass inference latency of 3.24 ± 0.18 ms per micrograph (103.68 ms per batch), corresponding to a model inference throughput of 308.6 images/s on the GPU. Peak GPU memory consumption was only 418 MB VRAM (2.61% of T4 capacity). The on-the-fly CPU pre-processing (CLAHE, resize to 224 × 224, ImageNet normalisation) processed 142 images per second per worker thread. These measurements characterise offline throughput only; live acquisition and LIMS write-back were not benchmarked.

These measured indicators demonstrate offline computational feasibility on commodity hardware. GPU inference was not the dominant measured bottleneck in the offline pipeline. Single-worker CPU pre-processing (142 images/s) is the main processing throughput bottleneck in the offline pipeline.

Complexity and Cost of Models: Table 6 presents a comprehensive comparison of the complexity of the architecture and the cost of computation for each of the models evaluated. All numbers are for a single forward pass on a 224 × 224 × 3 input tensor.

The empirical trend suggests an inverse relationship between ease of deployment and architectural complexity. EfficientNetB0 provided the most favourable accuracy-efficiency balance among the evaluated models, whereas MobileNetV2 had the lowest parameter and FLOP counts; ViT-B/16 required substantially greater computational resources and showed lower classification performance (see Section 4.2 for a detailed interpretation).

Training Workload Characterisation: The stratified five-fold cross-validation protocol provides a well-defined training workload for which accurate measurements can be made. Each fold splits the 1,580 images into 80% for training (1,264 images) and 20% for validation (316 images). Each training epoch takes approximately 40 mini-batches with a batch size of 32. This corresponds to approximately 3,000 gradient-update iterations per model over 15 epochs and 5 folds. Across the five architectures, the complete experimental protocol

therefore comprised approximately 15,000 training iterations.

During training only, stochastic augmentation was applied via horizontal mirroring and rotation within a ± 15 -degree range. All training and validation were scaled to 224×224 and normalised to the standard ImageNet per-channel statistics ($\mu = [0.485, 0.456, 0.406]$; $\sigma = [0.229, 0.224, 0.225]$). These pre-processing operations were executed on the CPU and constituted the main measured throughput limitation in the offline pipeline.

Modular Pipeline Architecture: The framework is structured as a four-stage modular pipeline, allowing each component to be replaced independently to enable integration into existing high-throughput screening infrastructure. The system architecture is depicted in Figure 10.

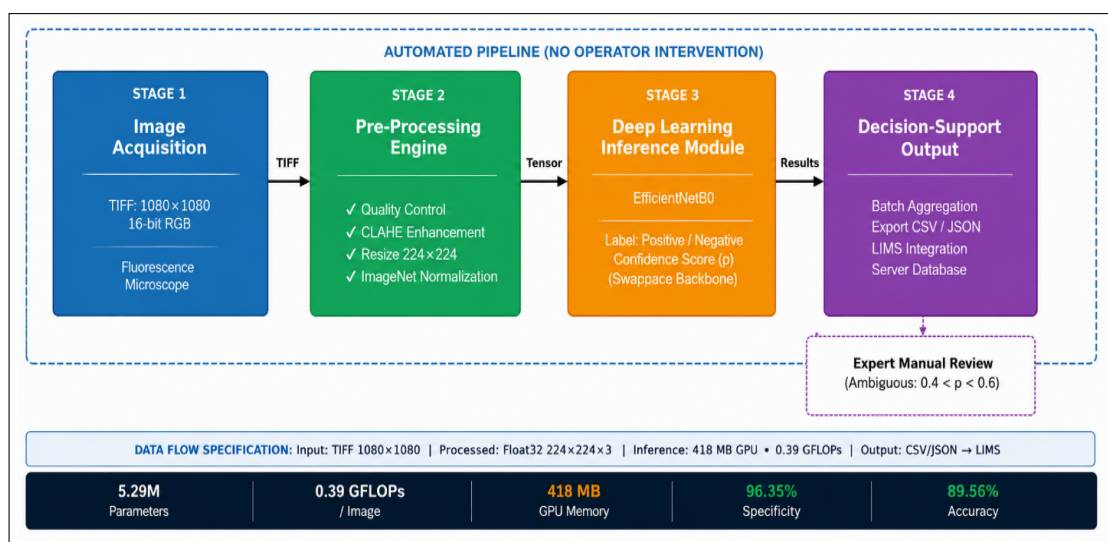


Figure 10. Automated ASC-speck classification pipeline using EfficientNetB0, from fluorescence image acquisition to LIMS-based decision support

Stage 1: Image Acquisition Interface: The raw fluorescence micrographs are acquired in 16-bit RGB TIFF format at a resolution of $1,080 \times 1,080$ pixels. The acquisition module monitors a given input directory and automatically ingests new images as they are produced by the microscope control software.

Stage 2: Pre-processing Engine: Each ingested image is processed by a deterministic quality-control pipeline that performs dual-channel integrity verification (GFP and DAPI), rejection of images with focus defects, excessive debris, or insufficient signal-to-noise ratios, resizing to 224×224 pixels, CLAHE contrast enhancement per channel and normalisation using the standard ImageNet per-channel statistics. This stage requires no operator intervention and outputs a standardised input tensor ready for any of the architectures considered.

Stage 3: Deep Learning Inference Unit: The pre-processed tensor is then passed to the selected model (EfficientNetB0 by default). The module outputs a binary classification label (positive or negative inflammasome activation) and a sigmoid-derived posterior probability. The cut-off for classification can be adjusted post hoc to bias toward either sensitivity or specificity depending on the purpose of the screening.

Stage 4: Decision-Support Output: Classification results are summarised per experimental plate or batch and exported in a structured format (CSV or JSON) compatible

with standard laboratory information management systems (LIMS). Ambiguous cases (i.e., cases with confidence scores close to the decision boundary, within a user-defined margin) are flagged for review by an expert to focus human effort on those cases needing specialist assessment.

Deployment Feasibility and Resource Allocation: EfficientNetB0 requires 418 MB of GPU memory at inference with a batch size of 32, leaving ample headroom for concurrent pre-processing on the same device. Where a GPU is unavailable, MobileNetV2 (3.50 M parameters, 0.30 GFLOPs) provides a lower-complexity alternative; CPU feasibility should be established through direct CPU inference measurements.

The modular pipeline uses standardised interchange formats (TIFF $\rightarrow$ tensor $\rightarrow$ CSV/JSON) which means that any stage can be upgraded independently. The implementation relies on open-source software components, including PyTorch, OpenCV, and scikit-image. By automating the primary classification pass at 96.35% specificity, the framework is designed to concentrate expert effort on the flagged ambiguous band. On the present corpus this would route 8.54% of micrographs to manual review while capturing 64.24% of pipeline errors; prospective confirmation in an operating screening campaign has not been performed.

Table 6. Model complexity and resource requirements (224 $\times$ 224 input, batch 32, FP32)

Model	Parameters (M)	FLOPs (G)	Est. GPU Memory: Inference (MB)	Est. GPU Memory: Training (MB)	Architecture Depth
EfficientNetB0	5.29	0.39	410	1,800	16 MBConv blocks
MobileNetV2	3.50	0.30	350	1,500	19 inverted residual blocks
ResNet18	11.69	1.82	520	2,600	18 layers (8 residual blocks)
ResNet50	25.56	4.11	1,100	5,200	50 layers (16 bottleneck blocks)
ViT-B/16	86.57	17.56	1,600	8,400	12 transformer encoder layers

Memory values are analytical estimates for a single forward pass at batch size 32; direct GPU profiling under the same conditions returned a peak inference footprint of 418 MB, agreeing with the theoretical estimate.

Generalisation and Intended-Use Boundaries: The present evidence is limited to 1,580 THP-1-ASC-GFP micrographs generated within one acquisition environment using one microscope platform, one staining protocol, and one cell line. The internal grouped-validation analysis reduces concern about microwell-level replicate leakage but does not establish transportability to other laboratories, instruments, staining batches, or cell populations. External multicentre evaluation is therefore required before the reported performance can be generalised beyond the originating setting. Similarly, the current computational profiling characterises offline feasibility and does not demonstrate sustained microscope-to-LIMS operation in a live screening campaign.

Measured GPU inference latency (3.24 ± 0.18 ms per micrograph) and single-worker CPU pre-processing throughput (142 images/s) establish offline feasibility, but acquisition-coupled scheduling, multi-worker pre-processing, live LIMS write-back, and long-duration operation were not evaluated. The lower balanced performance of ViT-B/16 under the present data regime may reflect limited training scale and the difficulty of representing highly localised punctate

features with fixed patch tokenisation; this remains a hypothesis requiring dedicated ablation experiments. The Grad-CAM inspection reported in Section 4.2 is qualitative and confined to representative cases, so systematic explainability assessment across independent acquisition settings remains future work.

5. Conclusion

This study provides a controlled engineering benchmark of five established deep-learning architectures for ASC-speck classification in THP-1-ASC-GFP fluorescence microscopy. EfficientNetB0 provided the most favourable overall balance between classification performance and computational demand under the evaluated protocol, while ResNet18 achieved the highest AUC-ROC. These results support EfficientNetB0 as the primary candidate for further prospective evaluation rather than establishing it as a universally superior architecture. The study used a single-site design. Grouped validation mitigates the concern of replicate leakage at the microwell level. No independent laboratory or second microscope platform was tested. The simulated 95:5 imbalance and photometric-perturbation experiments are internal stress tests, not surrogates for external domain-shift validation. The pipeline is thus framed as a primary screening and decision-support framework, not a fully autonomous high-stakes classification system. Future work should include independent validation in multiple centres, a baseline for task-matched biomedical image analysis, prospectively selected operating thresholds and evaluation in live microscope-to-LIMS workflows.

Data Availability

The raw fluorescence microscopy dataset (1,580 micrographs, 16-bit RGB TIFF) and associated well metadata generated from this study are not publicly deposited due to institutional data governance constraints. The anonymised dataset can be used for non-commercial scientific evaluation by qualified academic and industrial researchers upon request from the corresponding author (Sifeddine Elkardoudi, email: sifeddine.elkardoudi-etu@etu.univh2c.ma). Release of the dataset is subject to approval by the governance committee of the originating laboratory and implementation of a standard institutional Data Transfer Agreement (DTA).

References

- [1] Yao, Jing, Keenan Sterling, Zhe Wang, Yun Zhang, and Weihong Song. "The Role of Inflammasomes in Human Diseases and their Potential as Therapeutic Targets." *Signal transduction and targeted therapy* 9, no. 1 (2024): 10.
- [2] Almeida-da-Silva, Cassio Luiz Coutinho, Aline Cristina de Abreu Moreira-Souza, and David M. Ojcius. "Traditional Approaches and Recent Tools for Studying Inflammasome Activity." *Journal of Immunological Methods* 533 (2024): 113744.
- [3] Chen, Daniel, Tempest Plott, Michael Wiest, Will Van Trump, Ben Komalo, Dat Nguyen, Charlie Marsh et al. "A Combined AI and Cell Biology Approach Surfaces Targets and Mechanistically Distinct Inflammasome Inhibitors." *Iscience* 27, no. 12 (2024).
- [4] Shi, Cheng, Xiangyu Zhang, Xiaowei Chi, Yi Ran Zhou, Weiping Lyu, Tongfei Gao, Jiaxu Zhou et al. "Discovery of NLRP3 Inhibitors Using Machine Learning: Identification of a Hit Compound to Treat NLRP3 Activation-Driven Diseases." *European Journal of Medicinal Chemistry* 260 (2023): 115784.

- [5] Li, Jun, Junyu Chen, Yucheng Tang, Ce Wang, Bennett A. Landman, and S. Kevin Zhou. "Transforming Medical Imaging with Transformers? A Comparative Review of Key Properties, Current Progresses, and Future Perspectives." *Medical image analysis* 85 (2023): 102762.
- [6] Wittmann, Nico, Ann-Kathrin Behrendt, Neha Mishra, Lukas Bossaller, and Almut Meyer-Bahlburg. "Instructions for Flow Cytometric Detection of ASC Specks as a Readout of Inflammasome Activation in Human Blood." *Cells* 10, no. 11 (2021): 2880.
- [7] Lau, Tannia A., Elmar Mair, Beverley M. Rabbitts, Akshar Lohith, and R. Scott Lokey. "High-Content Image-Based Screening and Deep Learning for the Detection of Anti-Inflammatory Drug Leads." *Chembiochem* 25, no. 2 (2024): e202300136.
- [8] Zhou, S. Kevin, Hayit Greenspan, Christos Davatzikos, James S. Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L. Prince, Daniel Rueckert, and Ronald M. Summers. "A Review of Deep Learning in Medical Imaging: Imaging Traits, Technology Trends, Case Studies with Progress Highlights, and Future Promises." *Proceedings of the IEEE* 109, no. 5 (2021): 820-838.
- [9] Shafiq, Muhammad, and Zhaoquan Gu. "Deep Residual Learning for Image Recognition: A Survey." *Applied sciences* 12, no. 18 (2022): 8972.
- [10] Tan, Mingxing, and Quoc Le. "Efficientnetv2: Smaller Models and Faster Training." In *International conference on machine learning*, PMLR, 2021, 10096-10106.
- [11] Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." *arXiv preprint arXiv:2010.11929* (2020).
- [12] Bradshaw, Tyler J., Zachary Huemann, Junjie Hu, and Arman Rahmim. "A Guide to Cross-Validation for Artificial Intelligence in Medical Imaging." *Radiology: Artificial Intelligence* 5, no. 4 (2023): e220232.
- [13] Bilodeau, Anthony, Constantin VL Delmas, Martin Parent, Paul De Koninck, Audrey Durand, and Flavie Lavoie-Cardinal. "Microscopy Analysis Neural Network to Solve Detection, Enumeration and Segmentation From Image-Level Annotations." *Nature Machine Intelligence* 4, no. 5 (2022): 455-466.
- [14] Stringer, Carsen, Tim Wang, Michalis Michaelos, and Marius Pachitariu. "Cellpose: A Generalist Algorithm for Cellular Segmentation." *Nature methods* 18, no. 1 (2021): 100-106.
- [15] Raufeisen, Johannes, Kunpeng Xie, Fabian Hörst, Till Braunschweig, Jianning Li, Jens Kleesiek, Rainer Röhrig et al. "Cyto R-CNN and CytoNuke Dataset: Towards Reliable Whole-Cell Segmentation in Bright-Field Histological Images." *Computer methods and programs in biomedicine* 252 (2024): 108215.
- [16] Moshkov, Nikita, Michael Bornholdt, Santiago Benoit, Matthew Smith, Claire McQuin, Allen Goodman, Rebecca A. Senft et al. "Learning Representations for Image-Based Profiling of Perturbations." *Nature communications* 15, no. 1 (2024): 1594.
- [17] Griebel, Matthias, Dennis Segebarth, Nikolai Stein, Nina Schukraft, Philip Tovote, Robert Blum, and Christoph M. Flath. "Deep Learning-Enabled Segmentation of Ambiguous

- Bioimages with Deepflash2.” *Nature communications* 14, no. 1 (2023): 1679.
- [18] Greenwald, Noah F., Geneva Miller, Erick Moen, Alex Kong, Adam Kagel, Thomas Dougherty, Christine Camacho Fullaway et al. ”Whole-cell Segmentation of Tissue Images with Human-Level Performance Using Large-Scale Data Annotation and Deep Learning.” *Nature biotechnology* 40, no. 4 (2022): 555-565.
 - [19] Zuiderveld, Karel. ”Contrast Limited Adaptive Histogram Equalization.” In *Graphics gems IV*, 1994, 474-485.
 - [20] Otsu, Nobuyuki. ”A Threshold Selection Method from Gray-Level Histograms.” *IEEE transactions on systems, man, and cybernetics* 9, no. 1 (1979): 62-66.
 - [21] Sandler, Mark, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. ”Mobilenetv2: Inverted Residuals and Linear Bottlenecks.” In *2018 IEEE/CVF conference on computer vision and pattern recognition*, IEEE, 2018, 4510-4520.
 - [22] Liu, Ze, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. ”Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows.” In *2021 IEEE/CVF international conference on computer vision (ICCV)*, IEEE, 2021, 9992-10002.
 - [23] Kingma, Diederik P., and Jimmy Ba. ”Adam: A Method for Stochastic Optimization.” *arXiv preprint arXiv:1412.6980* (2014).
 - [24] Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen et al. ”PyTorch: An Imperative Style, High-Performance Deep Learning Library.” *Advances in neural information processing systems* 32 (2019).
 - [25] Brima, Yusuf, and Marcellin Atemkeng. ”Saliency-driven Explainable Deep Learning in Medical Imaging: Bridging Visual Explainability and Statistical Quantitative Analysis.” *BioData mining* 17, no. 1 (2024): 18.