scientific reports

Check for updates

OPEN

Adaptive example selection for prototype based explainable mitosis detection in digital pathology

Mita Banik[1], Ken Kreutz-Delgado[1,2], Ishan Mohanty[1], James B. Brown[1] & Nidhi Singh[1,3]✉

Understanding the decision-making process of black-box neural networks is crucial for safe use of AI in high-stakes medical tasks such as histopathology. We present Adaptive Example Selection (AES), a prototype-based explainable AI framework that improves interpretability of deep learning models for mitosis detection. AES retrieves a sparse set of supporting and contradicting real-world prototype images to locally approximate the model's confidence surface with high fidelity ($R^2 = 0.96$). The framework is integrated with a robust Faster R-CNN detector that demonstrates strong cross-tumor performance, for example achieving an F1-score of 0.84 on the Canine Cutaneous Mast Cell Tumor dataset. AES generates concise, case-specific explanations that faithfully capture local decision boundaries while linking predictions to interpretable exemplars. This enables clinicians to visualize model reasoning, assess uncertainty, and conduct contrastive analyses. Unlike prior methods focused on discrete class predictions, AES shows how similarity to mitotic and non-mitotic prototypes shapes graded confidence, enhancing transparency, trust, and practical adoption of AI-assisted mitosis detection in cancer diagnostics.

Mitosis, the process of tumor cell division, is a vital biomarker in the diagnosis and grading of many cancers. Histopathological assessment of hematoxylin and eosin (H&E)-stained tissue slides, where pathologists manually identify and count mitotic figures in tissue samples, remains the clinical gold standard[1]. However, this manual process is labor-intensive, subjective, and prone to variability due to differences in experience and interpretation among pathologists. Accurate mitosis detection is especially challenging due to their visual heterogeneity across different mitotic phases as well as the resemblance of non-mitotic cells, such as apoptotic bodies and lymphocytes, to true mitoses. Moreover, mitotic figures are often small and sparsely distributed, making them easy to overlook, especially under time constraints[2]. These factors contribute to inconsistencies and diagnostic uncertainty, limiting reliability.

Recent advances in deep learning (DL) have sparked interest in automating mitosis detection[3–5], offering improved efficiency and reduced variability. DL models analyze high-resolution digitized whole slide images (WSIs) to identify mitotic cells with performance approaching that of expert pathologists[6]. Despite these successes, DL models remain largely "black boxes" due to their highly complex, multi-layered neural networks whose complex decision-making processes are difficult to interpret[7]. This lack of transparency poses a significant barrier to clinical adoption, particularly in medical applications, where the consequences of errors can be far-reaching.

For AI to be safely and effectively adopted in medicine, models must not only be accurate but also explainable. To communicate findings responsibly, especially those derived from automation, clinicians, as ultimate decision-makers, must place as much trust in their tools as their patients place in them[8]. Explainable AI (XAI) enables clinicians to understand the rationale behind predictions, assess confidence, and identify potential errors or biases. Such transparency supports clinical decision-making and fosters trust in AI systems. Importantly, effective XAI tools should allow users to interact with models, adjusting sensitivity or focusing explanations to their diagnostic needs, and should be rigorously evaluated with end users[9,10].

In histopathology, XAI techniques typically fall into two categories: saliency maps, which highlight image regions influencing predictions, and example-based methods, which provide similar prior cases to explain

[1]Pattern Computer, Inc., Redmond, WA, United States. [2]Department of Electrical and Computer Engineering, University of California San Diego, San Diego, CA, United States. [3]Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA, United States. ✉email: nidhis@patterncomputer.com

decisions. Saliency maps (e.g., Grad-CAM[11–13]) attempt to localize the most influential regions of an image but often produce coarse, ambiguous highlights and may lack precision needed for clinical use[14]. In contrast, example-based approaches provide intuitive, concrete references that align well with how clinicians reason.[15,16].

Among example-based methods, prototype-based reasoning offers an intuitive, human-aligned framework for interpretable machine learning by grounding predictions in real-world examples.[17,18]. Technically, they clarify data embeddings and decision boundaries by anchoring new instances to representative exemplars. However, most implementations focus on classification, neglecting complex tasks like object detection that require spatial localization and confidence estimates. Few are adapted to clinical workflows or evaluated for usability and relevance, key factors for trust and adoption in real-world medical practice.

We introduce the Adaptive Example Selection (AES) mechanism, which approximates a black-box detector's local decision landscape using a truncated radial basis function (RBF) expansion of its confidence scores to generate locally faithful, example-based explanations. AES retrieves a small set of supporting and contradicting prototype images to reveal the evidence influencing each prediction. It operates as a post-hoc interpretability layer on a standard Faster R-CNN detector, providing instance-level explanations without modifying the detector architecture.

The main contributions of this work are: (1) a prototype-based explainability framework for object detection in histopathology; (2) a truncated RBF formulation for locally faithful prototype selection; (3) a contrastive retrieval strategy presenting supporting and contradicting exemplars; and (4) quantitative and qualitative evaluations demonstrating compact, case-specific explanations aligned with clinical reasoning. Figure 1 summarizes the two-stage workflow: training a high-performance mitotic figure detector and applying AES to explain and visualize each prediction at the individual-instance level. By aligning AI outputs with the way clinicians reason about evidence, AES seeks to build trust, enhance transparency, and promote practical adoption of AI-assisted mitosis detection in cancer diagnostics.

Results

We developed the Adaptive Example Selection framework for mitotic figure detection in histopathology, aiming to (i) establish a high-performance baseline detector and (ii) provide case-specific, interpretable explanations. Below, we report both predictive performance and interpretability outcomes, linking each step to clinical relevance.

Faster R-CNN mitotic figure detector

We trained a two-stage region-based convolutional neural network (Faster R-CNN) [19] to detect MFs in histopathology images. Each predicted region of interest (ROI) received a confidence score $\beta(x) \in [0, 1]$, with predictions above a decision threshold $\tau_0 = 0.969$ considered positive. This threshold was selected via cross-validation using a held-out validation set to optimize the F1-score.

Training used the MItosis DOmain Generalization++ (MIDOG++) dataset [20], which includes expert-annotated MFs across multiple human and canine tumor types with varied staining and morphology. Model performance was assessed via 5-fold cross-validation on a held-out validation set using precision, recall, and F1-score. (Table 1). F1-scores ranged from 0.57 to 0.84, matching or surpassing MIDOG++ benchmarks. Variation reflects species- and tumor-specific histological differences. These results confirm the detector as a robust but complex "black box," suitable for AES interpretability analysis.

Adaptive example selection

AES generates case-specific, interpretable explanations by retrieving a small set of prototypical images that support (Positive Decision Figures, PDFs) or contradict (Negative Decision Figures, NDFs) each prediction (Fig. 2). Input includes the detector's bounding boxes x, feature embeddings, and confidence scores $\beta(x)$. Output is a compact set of prototypes approximating the local decision boundary.

To emphasize predictions near the decision threshold, scores are transformed using a shifted Box–Cox (SBC) function:

$$f(x) = \mathrm{SBC}(\beta(x); \tau_0) = \mathrm{BC}(\beta(x)) - \mathrm{BC}(\tau_0), \tag{1}$$

where τ_0 separates positive and negative predictions. AES models $f(x)$ with a sparse Radial Basis Function (RBF) expansion:

$$\widehat{f}(x) = \sum_{y_i \in \mathcal{D}} c_i \, \rho^{(\rho_0)}_{k_{y_i}, y_i}(x), \tag{2}$$

with $\mathcal{D}$ the set of prototype images, c_i their weights, and k_{y_i} the concentration parameter of the RBF centered at y_i. RBFs are truncated below $\rho_0 = 0.1$ to ensure locality (see Supplemental Information, Section S2).

AES learns prototype weights $\boldsymbol{c}$ and concentrations $\boldsymbol{k}$ by alternately minimizing: (i) a fidelity and sparsity loss $\mathcal{L}_1$, and (ii) an overlap reduction loss $\mathcal{L}_2$ to enforce label-consistent and diverse prototypes. Further details on the loss functions are provided in the Methods section. Early stopping is applied to $\mathcal{L}_1$ to prevent overfitting. Once trained, AES provides both a global approximation and a localized explanation set for each prediction:

$$\mathrm{AES}(x) = \left\{ i_{y_i} \;\middle|\; |c_{y_i} \rho^{(\rho_0)}_{k_{y_i}, y_i}(x)| \geq \mu_2(x), \;\; y_i \in \mathcal{D} \right\}, \tag{3}$$

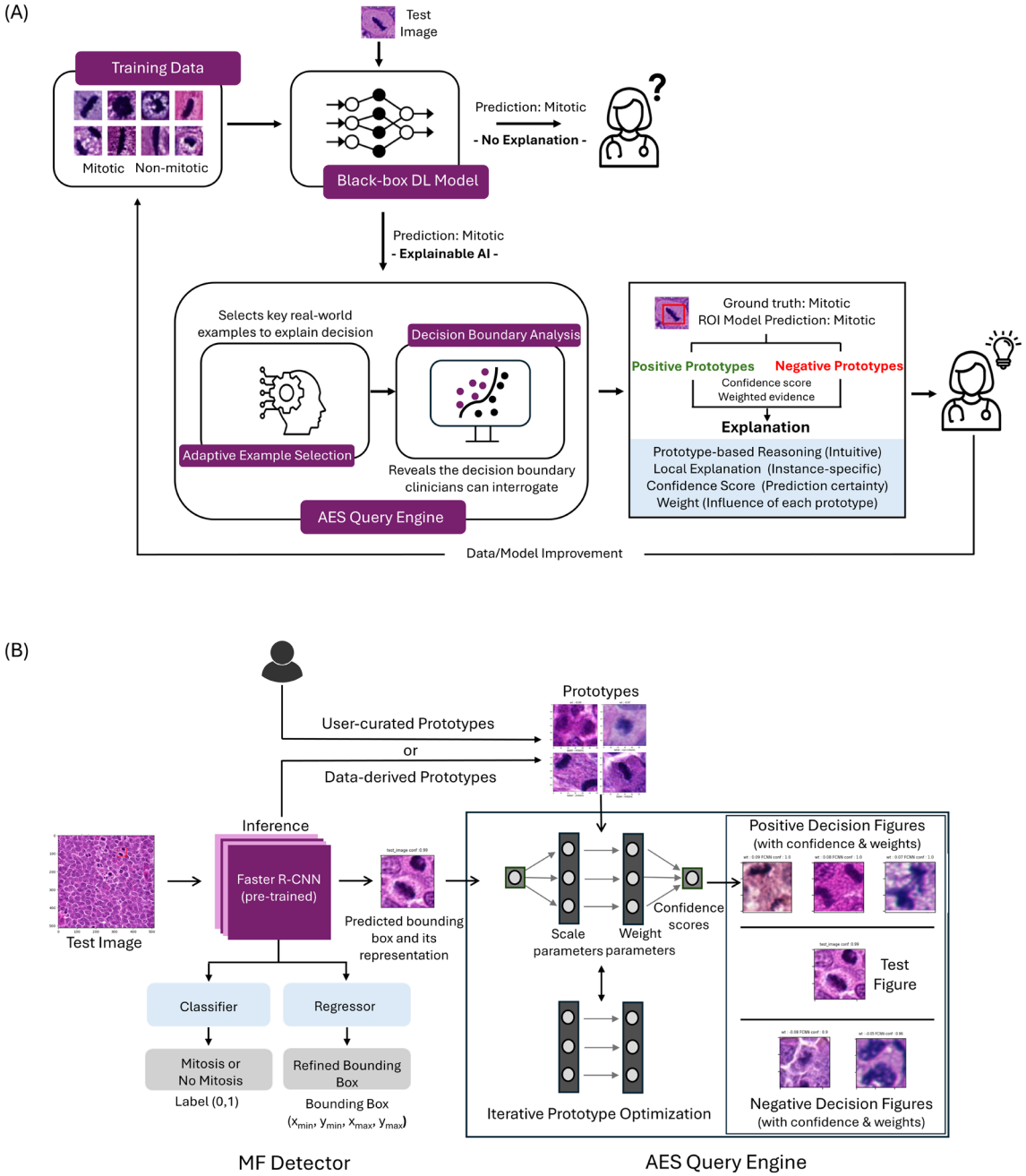

Fig. 1. (**A**) Conceptual illustration showing that the need for explainability grows with the clinical importance of AI-assisted diagnostic applications. (**B**) Overview of the XAI-driven mitosis detection workflow. A Faster R-CNN detector identifies candidate mitotic cells by generating localized bounding boxes in test images. The Adaptive Example Selection (AES) module then retrieves visually similar, mitotic and non-mitotic examples from the training data. These prototypes provide case-based explanations, enabling clinicians to interpret predictions and evaluate decision confidence in a transparent, contrastive manner.

with $\mu_2(x)$ chosen to explain a fixed proportion (e.g., 90%) of the predicted score.

Quantitative evaluation

AES was evaluated on sparsity (dictionary size $|\mathcal{D}|$), approximation fidelity (R^2), and interpretability (relevance scores, RS). The mean and median values of $|\mathrm{AES}(x)|$ across the test set were reported as Mean-RS and Med-RS, respectively, representing the average and median number of AES prototypes required to explain 90% of the detector's predicted score for each test bounding box. Lower RS values indicate that fewer dictionary images are needed to explain model outputs, potentially reducing the cognitive load for clinicians by allowing them to focus on fewer, more meaningful examples. RBF fit quality was further assessed using the condition number $\kappa(\mathcal{D})$ of the RBF matrix, where lower values indicate better feature-space coverage and numerical stability. Table 2

Tumor Type	Precision	Recall	F1-score	F1-score†
Human Breast Cancer	0.70 ± 0.03	0.81 ± 0.02	0.75 ± 0.01	0.71 ± 0.02
Canine Lung Cancer	0.66 ± 0.03	0.63 ± 0.02	0.64 ± 0.02	0.68 ± 0.02
Canine Lymphosarcoma	0.79 ± 0.01	0.73 ± 0.01	0.760 ± 0.004	0.68 ± 0.01
Canine Cutaneous Mast Cell Tumor	0.89 ± 0.02	0.80 ± 0.02	0.84 ± 0.01	0.82 ± 0.01
Human Neuroendocrine Tumor	0.50 ± 0.04	0.67 ± 0.04	0.57 ± 0.02	0.59 ± 0.01
Canine Soft Tissue Sarcoma	0.68 ± 0.04	0.74 ± 0.01	0.71 ± 0.03	0.69 ± 0.01
Human Melanoma	0.75 ± 0.03	0.80 ± 0.01	0.77 ± 0.01	0.81 ± 0.01

Table 1. Performance of the MF detector across tumor types. †Benchmark F1-scores from[20] are included for comparison. See Supplementary Information, Table S1, for a confusion matrix.

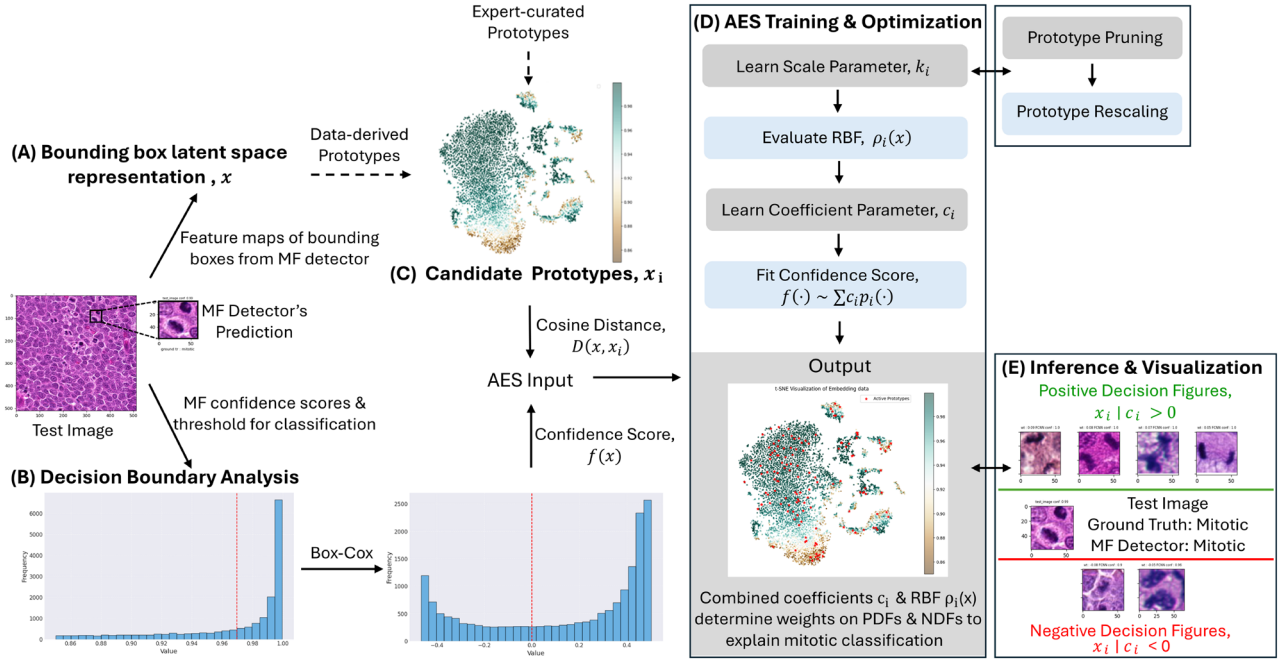

Fig. 2. The AES Query Engine for interpretability and decision boundary analysis: (**A**) Feature maps and confidence scores for Mitotic Figures (MF) classification are extracted for each predicted bounding box in a test image. (**B**) Confidence scores are normalized using a variance-stabilizing power transformation, improving the detector's behavior near the decision boundary. (**C**) Candidate prototypes are selected from either the training set or an expert-curated reference database. (**D**) During AES training and optimization, redundant prototypes are pruned, and radial basis functions (RBFs) are scaled, yielding a sparse, diverse, and informative dictionary of active prototypes, see Figure S1 in the Supplementary Information. (**E**) At inference, for each bounding box, the trained AES assigns signed coefficients that quantify the influence of nearby prototypes, Positive Decision Figures (PDFs; mitotic) and Negative Decision Figures (NDFs; non-mitotic), providing visual explanations of mitotic predictions.

summarizes AES simulation results across different configurations, with values reported as means averaged over 10 runs with different random seeds. The optimal setup (Row 7) achieved $R^2 = 0.96$ with a compact median of 10 prototypes per case, yielding a highly interpretable global dictionary of ∼190 images. See Section S3 in the Supplementary Information for details on standard errors and hyperparameters.

Interpretability outcomes

Beyond quantitative performance, AES interpretability was evaluated across three diagnostic scenarios: correct predictions, false positives, and false negatives (Fig. 3). Prototypes are labeled as supporting (PDFs, green) or contradicting (NDFs, red) with associated influence weights and confidence scores, providing contrastive, localized explanations.

a. Correct Prediction - High-confidence Mitotic Classification: AES consistently retrieved strongly weighted mitotic prototypes (PDFs) and no NDFs for high-confidence mitotic classifications, reflecting unambiguous

Row	Case Study	Parameter Values	R^2	Mean-RS	Med-RS	$\lvert\mathcal{D}\rvert$	$\log_{10}(\kappa(\mathcal{D}))$
1	No trim; $\mathcal{L}_2$ off	constant k, $\gamma_1 = \gamma_2 = 0$	0.97	606.95	146.10	1647.50	5.70
2	No trim; $\mathcal{L}_2$ off	constant k, non-zero γ_1, γ_2	0.95	253.94	238.90	1576.20	4.93
3	No trim; $\mathcal{L}_2$ off	variable $\boldsymbol{k}$, non-zero γ_1, γ_2	0.96	268.85	245.90	620.90	4.46
4	Trimmed; $\mathcal{L}_2$ off	constant k, $\gamma_1 = \gamma_2 = 0$	0.96	621.50	791.00	1619.10	4.16
5	Trimmed; $\mathcal{L}_2$ off	constant k, non-zero γ_1, γ_2	0.94	182.03	210.60	1844.50	4.36
6	Trimmed; $\mathcal{L}_2$ off	variable $\boldsymbol{k}$, non-zero γ_1, γ_2	0.96	21.17	8.50	269.40	2.97
7	Trimmed; $\mathcal{L}_2$ on	variable $\boldsymbol{k}$, non-zero γ_1, γ_2	0.96	14.81	10.00	190.20	2.87

Table 2. AES performance under different configurations. "Trimmed" indicates $\rho_0 = 0.1$; "No trim" sets $\rho_0 = 0$. Variable $\boldsymbol{k}$ allows each RBF concentration to be learned independently; constant k applies a shared value. Reported values are means averaged over 10 runs with different random seeds; therefore, dictionary size $|\mathcal{D}|$ may appear non-integer. Mean RS and Med RS denote the average and median number of AES prototype images required to explain 90% of a prediction's score; lower values indicate fewer exemplars and reduced interpretive burden. Row 7 balances high fidelity with the smallest prototype dictionary.

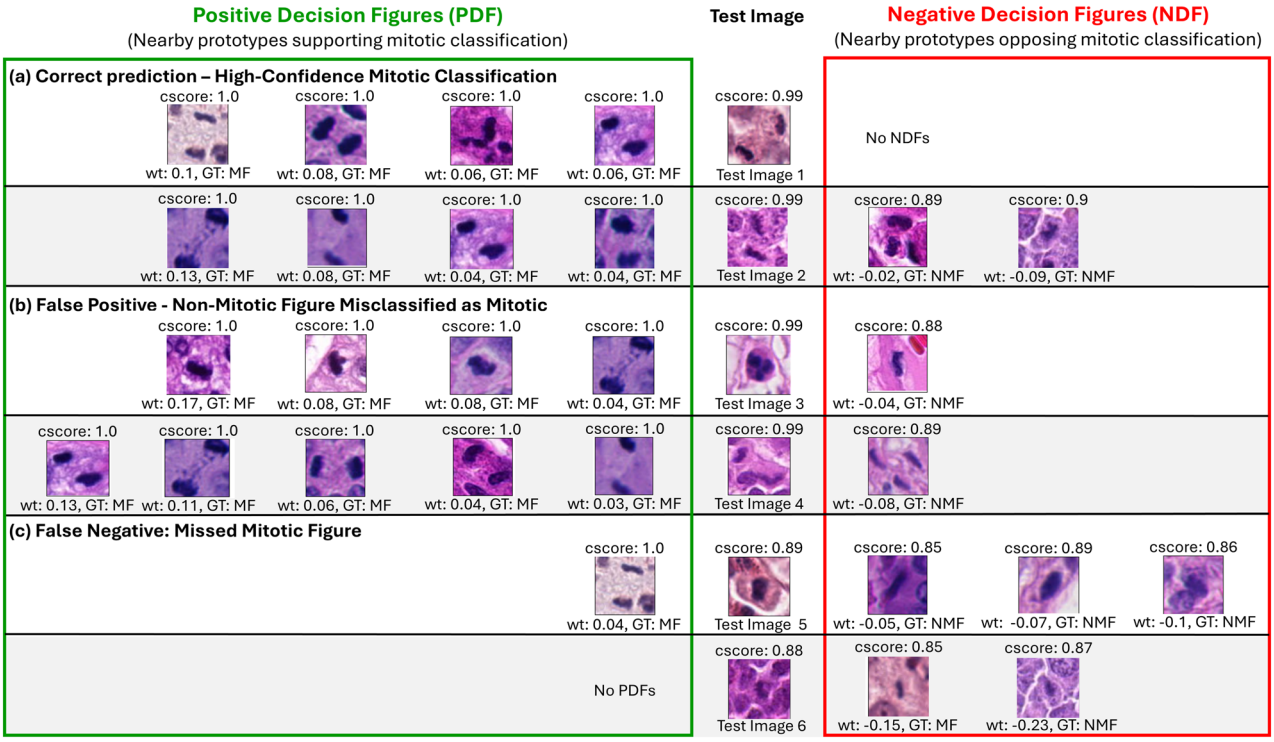

Fig. 3. The AES framework illustrates model interpretability under varying conditions. The global detection threshold is set to $\tau_0 = 0.969$, while Positive Decision Figures (PDFs) and Negative Decision Figures (NDFs) are defined as $\text{PDF} = \text{PDF}\epsilon(\mathcal{I})$ and $\text{NDF} = \text{NDF}\epsilon(\mathcal{I})$, respectively, using a local threshold of $\tau_\ell = 0.85$. For each predicted bounding box, AES retrieves nearby PDFs and NDFs, which serve as visual, example-based explanations of the model's decision. These prototypes, drawn from training data or expert-curated examples, highlight the features most influential in the detector's classification of mitotic versus non-mitotic figures. Abbreviations: cscore: confidence score; wt: weight; GT: ground truth; MF: Mitotic figure; NMF: Non-mitotic figure.

reasoning (Fig. 3a). Borderline cases included a few weak NDFs (weights -0.02 to -0.09) but mitotic PDFs predominated, showing subtle uncertainty while maintaining interpretability.

b. False Positive - Non-mitotic Figure Misclassified as Mitotic: In misclassified non-mitotic figures, AES retrieved multiple mitotic PDFs with strong positive influence (weights up to 0.17) and only weak NDFs (e.g., -0.04; Fig. 3b). Visual similarity in chromatin and intensity patterns explains why the detector was misled.

c. False Negative - Missed Mitotic Figure: For missed mitotic cases, AES retrieved either a single weak PDF with several influential NDFs or, in extreme cases, only NDFs. Occasionally, an NDF carried a mislabeled

mitotic ground truth, highlighting underrepresented or ambiguous morphologies in the training data. Contradictory prototypes can shift the decision toward non-mitotic classification despite mitotic features.

AES explanations extend beyond abstract performance metrics: they expose the evidence supporting or contradicting each decision, reveal systematic error patterns, and highlight clinically relevant ambiguities at decision boundaries.

Overall, (i) the MF detector achieves robust cross-species performance suitable for clinical deployment, and (ii) AES produces compact, case-specific prototype explanations that approximate the detector's decision boundary and provide interpretable morphological evidence, enhancing transparency and trustworthiness in AI-assisted cancer diagnostics.

Discussion

Deep learning has shown promise in automating mitotic figure detection, yet its clinical adoption remains limited by the opacity of model decisions. In high-stakes domains such as pathology, where treatment decisions hinge on image interpretation, transparency is a requirement. It is essential for trust, error mitigation, and safe deployment. We address this challenge with Adaptive Example Selection, a prototype-based explainability framework that complements a high-performance Faster R-CNN MF detector with concise, human-interpretable, and case-specific explanations.

AES is grounded in the way pathologists reason through ambiguous findings: comparing a case to a mental library of known examples, weighing supporting and contradicting evidence before making a call. To emulate this reasoning process, AES retrieves a small set of real, expert-annotated *prototypes* from a reference library. These prototypes are divided into (a) Positive Decision Features (PDFs), which represent supporting evidence, and (b) Negative Decision Features (NDFs), which contradict the current prediction. This dual set allows clinicians to understand not only why a prediction was favored, but also why alternatives were rejected, enabling contrastive reasoning at the decision boundary.

Technically, AES operates as an interpretable query engine layered on top of a trained detector. The detector predicts bounding boxes and confidence scores for MF candidates. A calibrated threshold converts scores into binary mitotic/non-mitotic calls. For each candidate region, AES fits a local regression model that approximates the detector's confidence landscape in that neighborhood of feature space. This is achieved with a truncated RBF expansion, chosen for its ability to represent smooth decision surfaces while enforcing locality. From the global library, AES selects a sparse subset of prototypes whose weighted influence best explains each prediction. Selection is guided by: (a) Visual similarity in learned feature space. (b) Magnitude and sign of influence on the local decision. (c) Adaptive concentration parameters controlling the spatial extent of each prototype's influence. The retrieved PDFs and NDFs, along with their influence weights and similarity rankings, are presented to the clinician. This provides an interpretable, human-scale explanation, typically a median of 10 prototypes per case, without overwhelming with irrelevant examples.

Compared with prior prototype-based methods [21–23], AES advances interpretability in three ways. First it introduces (*contrastive reasoning*): by retrieving both supporting (PDFs) and contradicting (NDFs) prototypes, AES provides counterfactual evidence that clarifies not only *why* a prediction was made, but also *why* an alternative was rejected. Second, it ensures (*local fidelity*): the truncated RBF approximation closely matches the detector's decision surface ($R^2 = 0.96$) while constraining prototype influence, keeping explanations focused, and clinically meaningful. Third, it promotes (*sparsity and stability*): adaptive scaling and regularization yield compact, well-conditioned prototype sets, avoiding the overly dense or unstable explanations seen in prior RBF-based approaches. Together, these advances address long-standing limitations in prototype-based interpretability, making AES both technically faithful and cognitively aligned with how pathologists reason through ambiguous cases [9,24].

We evaluated AES through ablation studies focused on sparsity, fidelity, and stability. Without regularization, the model achieved high predictive accuracy ($R^2 = 0.97$) but required more than 1,600 prototypes, resulting in explanations that were dense and difficult to interpret. Introducing regularization and adaptive concentration parameters dramatically reduced dictionary size while maintaining accuracy. In the best configuration, AES required only about 270 prototypes, with a median of fewer than 10 retrieved per case, while preserving high fidelity ($R^2 = 0.96$). A further refinement with an $\mathcal{L}_2$ placement loss pruned the set to 190 prototypes without sacrificing accuracy, concentrating explanations on the most relevant exemplars. Although our implementation used trimmed Gaussian kernels for simplicity, smoother alternatives such as Wendland kernels[25] are compatible. Collectively, these results show that AES provides concise, stable, and clinically interpretable explanations suitable for real-time decision support.

Beyond technical accuracy, clinical adoption depends on whether explanations meaningfully support decision-making. AES addresses this need by providing four complementary explanation types: (1) Prototype-based reasoning, which uses real, labeled exemplars (supporting & contradicting) that mirror how pathologists compare cases to known references; (2) Local explanation (instance-specific context), where prototypes are selected locally for each case, ensuring that the explanation reflects the reasoning for that decision rather than a global average; (3) Confidence scores, through which AES integrates model confidence, allowing clinicians to gauge when predictions are reliable versus uncertain; and (4) Influence weight, which quantifies each prototype's contribution (positive or negative) to reveal why a decision was made and which exemplars shaped it most. Applied across three recurring diagnostic scenarios (Fig. 3), these explanation types allow pathologists to confirm robust predictions when supporting evidence is consistent, scrutinize errors by identifying misleading visual similarities, and detect blind spots where atypical or underrepresented morphologies drive misclassification.

Our AES-based interpretability analysis provides clinically relevant insights into the MF detector's decision-making. First, in high-confidence correct predictions, the absence of contradictory NDFs strengthens trust

in the model, as all retrieved prototypes reinforce the mitotic classification. For pathologists, this consistency is reassuring, since the explanatory evidence aligns closely with the ground truth. Second, borderline correct predictions and false positives highlight the importance of interpretability in guiding expert oversight. When AES surfaces both mitotic and non-mitotic prototypes, the mixture signals regions of uncertainty where human review may be critical. In particular, confident false positives arose from visual similarities in chromatin morphology, underscoring how domain-specific ambiguities can mislead automated detectors. Third, false negatives often reflected either subtle or atypical mitotic morphologies, or mislabeled prototypes within the training set. These findings point toward two practical interventions: (i) curating more diverse training data to capture rare mitotic variants, and (ii) auditing prototype libraries to minimize mislabeled exemplars. Both steps would improve model robustness while maintaining transparency for clinical users. In this way, AES extends interpretability from "what the model predicted" to "why the model predicted it," thereby improving transparency, supporting expert override when needed, and laying the foundation for trust in AI-assisted histopathology.

Furthermore, AES' interactive, user-centric design promotes clinician–AI collaboration by revealing real-world examples that expose biases, highlight failure modes, and inform model refinement through retraining or calibration. Human-in-the-loop features allow for context-specific threshold tuning, incorporation of expert-curated prototypes, and interface customization to diverse clinical workflows, thereby strengthening trust, reliability, and equity in diagnostic decision-making.

Importantly, AES is more than an explanatory add-on. By linking ambiguous findings to curated exemplars, it functions as a clinical decision support tool. It can aid tumor grading, prognosis estimation, and training of junior pathologists. Its modular design suggests applicability beyond MF detection, including tumor subtyping, organ classification, and rare disease identification.

AES relies on a prototype library derived from the training data, and the relevance of retrieved examples depends on how well this library captures the visual diversity of mitotic and non-mitotic structures. Given the known heterogeneity of mitotic figures across tumor types, staining protocols, and imaging conditions, rare or atypical morphologies may be underrepresented, leading to less informative explanations in some cases. In addition, AES functions as a post-hoc explanation layer and does not correct errors made by the underlying detector; it explains the detector's behavior, whether correct or incorrect. Finally, domain shifts related to scanner hardware or staining variations may affect prototype similarity and would require updating or adapting the prototype library for optimal performance.

Future work will focus on scaling AES to whole-slide images, integrating it into digital pathology viewers, and evaluating usability through prospective clinician studies. By combining technical rigor with cognitive alignment, AES offers a pathway toward transparent, context-aware and trustworthy AI adoption in pathology, enabling safe, effective clinician–AI collaboration.

Methods

We developed a two-stage pipeline for interpretable MF detection: (a) A high-performance Faster R-CNN MF detector trained on the multi-species, multi-tumor MIDOG++ dataset; (b) An AES module that retrieves a small set of real-world prototypes, supporting as well as contradicting the detector's output, to approximate its local decision boundary.

Black-box faster R-CNN MF detector

Dataset

We trained and evaluated our detector on the publicly available MIDOG++ dataset (https://github.com/DeepMicroscopy/MIDOGpp), the largest multi-domain MF dataset at the time of writing. A multi-domain dataset ensures robustness across tumor types and laboratories, increasing the likelihood of reproducibility in real-world pathology workflows. The dataset contains high-resolution ROIs from 503 histological specimens, with 11,937 expert-annotated mitotic figures across seven tumor types (human and canine): breast carcinoma, lung carcinoma, lymphosarcoma, neuroendocrine tumor, cutaneous mast cell tumor, cutaneous melanoma, and subcutaneous soft tissue sarcoma. Images were acquired using five whole-slide scanners across four pathology laboratories.

Following the dataset authors' recommendations, 111 images were reserved as an independent test set. Of the remaining 392 images, 22 images without ground-truth mitotic figures were excluded from training to reduce class imbalance and computational overhead, leaving 370 images for training and validation. Image-level 5-fold cross-validation was then performed on these 370 images, with an 80%/20% split for training and validation in each fold.

Data preprocessing

Original images measure up to 7215×5412 pixels, whereas MFs are typically $\sim 50 \times 50$ pixels. To avoid loss of detail from global downsampling, we generated 512×512 pixel patches using a sliding window crop with 20% overlap, following [20]. Patches were generated separately for training, validation, and test sets, ensuring that no patches from the same image appear in multiple sets, thereby preventing potential bias in evaluation. This increased the MF-to-patch area ratio from ∼0.7% to ∼10%, improving detection sensitivity. To enhance domain generalization, we applied: (a) Geometric augmentation: random horizontal/vertical flips (50% probability each) and random rotations up to $\pm 45^{\circ}$ (50% probability). (b) Photometric augmentation: brightness shifts (± 0.2) and contrast shifts ($\pm 20\%$, 20% probability), Gaussian noise (variance 10–50, 30% probability), and saturation shifts ($\pm 20\%$, 20% probability). (c) Stain augmentation: Random Stain Normalization and Augmentation (RandStainNA) [26] with four style templates in pink/purple H&E tones. Templates were selected using *k*-means clustering on color histograms to ensure coverage of common staining modes. Preserving fine nuclear detail while simulating real-world staining variability improves the detector's reliability for routine diagnostic slides.

Faster R-CNN: training and hyperparameters
We employed Faster R-CNN with a ResNet-50 [27] backbone pretrained on MS-COCO [28], combined with a FPN for small-object detection. We chose Faster R-CNN over RetinaNet [29] (used in [20]) due to its interpretability and robust performance in histopathology detection tasks. Patches containing at least one MF were used to reduce computation overhead. Training ran for 50 epochs (batch size 2), using stochastic gradient descent (learning rate = 0.001, momentum = 0.9). Model selection was based on mean average precision (mAP) across IoU thresholds 0.5–0.95 (step 0.05). Post-training, class-confidence and bounding-box area thresholds were tuned on the validation set to maximize F1-score.

The detector outputs a confidence $\beta(x) \in [0, 1]$ that bounding box x contains an MF:

$$\beta(x) = \text{confidence}(i_x \doteq \text{MF}).$$

Positive predictions at decision threshold τ_0 form the set:

$$\mathcal{B}^+_{\tau_0} = \{x \mid \beta(x) \geq \tau_0\}$$

For integration with AES, we extracted `fc7`-layer features for each predicted bounding box. The optimal threshold $\tau_0 = 0.969$ maximised F-1 score across tumor types. It is important to note that high thresholds reduce false positives, important for pathologists who must avoid over-calling mitoses.

Evaluation metrics
We report Precision, Recall, and F1 score to evaluate the detector's performance:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{F1} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}}. \tag{4}$$

where TP, FP, and FN denote true positives, false positives, and false negatives. A detection was considered correct if the Intersection over Union (IoU) between predicted and ground-truth bounding boxes was ≥ 0.5, balancing sensitivity for small mitotic figures while aligning with common object detection benchmarks. All mAP and F1-scores in Table 1 were computed using this threshold. Balancing sensitivity and specificity ensures pathologists see true mitoses without being overwhelmed by spurious detections.

AES query engine

We developed an AES query engine to interpret and visualize the confidence scores $\beta(x)$ generated by a trained MF detector by retrieving representative *active prototypes* from the training set i.e., examples that either support or contradict predictions. This provides pathologists with visual "second opinions," enhancing interpretability and trust. Although all prototypes are drawn from the detector's training data in this study, the framework can also accommodate examples selected by domain experts.

Throughout this section, $\boldsymbol{\mathcal{B}}$ denotes the set of all candidate bounding boxes, and r_x the feature embedding of a region x obtained from the detector's penultimate layer.

Pre-processing and candidate selection
Initial filtering removes bounding boxes with implausible sizes or excessive overlap. Specifically, we retain boxes with area of at least 2400 pixels (equivalent to 50×50 at the dataset resolution) and apply non-maximum suppression (NMS) to eliminate redundancy.

We define:

$$\mathcal{B}^+_{\tau_0} = \{x \mid \beta(x) > \tau_0\}, \quad \mathcal{B}_{[\tau_\ell, \tau_0]} = \{x \mid \beta(x) \in [\tau_\ell, \tau_0]\}.$$

and combine these subsets to obtain $\mathcal{B}^+_{\tau_\ell}$. This construction prioritizes borderline detections near the decision threshold τ_0, where interpretability is most clinically valuable.

Decision boundary analysis via shifted box-cox transformation
We stabilize and recenter the distribution of confidence scores using a shifted Box-Cox (SBC) transformation:

$$\text{SBC}(y; \tau_0) = \frac{y^\lambda - 1}{\lambda} - \frac{\tau_0^\lambda - 1}{\lambda}.$$

The parameter λ is estimated from data. This transformation centers the decision boundary (at which $\text{SBC}(\tau_0; \tau_0) = 0$), improving comparability across detections and ensuring prototype selection reflects clinically meaningful thresholds.

Prototype selection via truncated radial basis functions
Each bounding box x is represented by normalized detector features $\hat{r}_x = r_x / \|r_x\|$. Similarity between boxes is quantified by a radial basis function (RBF):

$$\rho_{k_{y_i}, y_i}(x) = \exp(-k_{y_i}(1 - \cos(\hat{r}_x, \hat{r}_{y_i}))),$$

which is truncated at ρ_0 to retain only influential prototypes. Here k_{y_i} controls the concentration (inverse width) of each radial basis function, and $\rho_0 \in (0, 1)$ sets the minimum cosine similarity required for a prototype to influence the prediction. The transformed confidence is then approximated as:

$$f(x) \approx \widehat{f}(x) = \sum_{y_i \in \mathrm{AES}(x)} c_{y_i}\, \rho^{(\rho_0)}_{k_{y_i}, y_i}(x).$$

where $\mathrm{AES}(x)$ is the minimal set of prototypes explaining most of the score. Clinically, this shows pathologists which specific examples influenced a decision, mirroring human diagnostic reasoning.

Optimization

The goal of AES optimization is to learn a compact prototype dictionary $\mathcal{D}$ that serves as the set of RBF centers in $\widehat{f}_{\boldsymbol{k},\boldsymbol{c}}(x)$, approximating the transformed detector confidence $f(x)$. A smaller, well-structured dictionary reduces cognitive load for clinicians while maintaining explanatory power. Here $\boldsymbol{k} = [k_{y_1}, \ldots, k_{y_{|\mathcal{D}|}}]$ and $\boldsymbol{c} = [c_{y_1}, \ldots, c_{y_{|\mathcal{D}|}}]$ denote the vectors of kernel concentrations and coefficients, respectively, and $\gamma_1, \gamma_2, \gamma_3$ are scalar regularization weights controlling sparsity and kernel localization.

Optimization is guided by four principles: (i) prototypes must be informative for explanations, i.e. $\mathrm{AES}(x) \subset \mathcal{D}$; (ii) centers $y_i \in \mathcal{D}$ should be well-separated to ensure broad coverage; (iii) kernels should remain localized through high concentrations k_{y_i} values; (iv) coefficients c_i should remain sign-consistent with $f(x)$ to preserve decision alignment.

Starting from an initial prototype set $\mathcal{D}_0$, dictionary elements are divided into ϵ-positive and ϵ-negative decision figures:

$$\mathrm{PDF}_\epsilon = \{x \in \boldsymbol{\mathcal{B}}^+_{\tau_0+\epsilon} \cap \mathcal{D}_0\} \quad \text{and} \quad \mathrm{NDF}_\epsilon = \{x \in \boldsymbol{\mathcal{B}}_{[\tau_\ell, \tau_0-\epsilon]} \cap \mathcal{D}_0\}, \tag{5}$$

which provide boundaries for prototype selection. Here $\epsilon > 0$ defines a small confidence margin around the decision boundary τ_0, used to separate strongly positive and negative prototype regions. The optimized $\mathcal{D}$ must therefore satisfy: 1. Sign alignment: $c_i \geq 0$ for $y_i \in \mathrm{PDF}_\epsilon$, and $c_i \leq 0$ for $y_i \in \mathrm{NDF}_\epsilon$. 2. Sparsity and spread: dictionary size should be much smaller than the candidate set ($|\mathcal{D}| \ll |\mathcal{D}_0|$) and well-conditioned to avoid redundancy. 3. Localization: explanations should remain compact, i.e. $\mathrm{R}^\alpha_{\mathcal{D}}(x)$ is small (see Supplementary Information, Section S4, for a formal definition), so that only a few prototypes contribute to each decision. For sharply peaked RBFs, this implies $\mathrm{sign}(c_i) = \mathrm{sign}(f(y_i))$ for all $y_i \in \mathcal{D}$.

Training alternates between two complementary objectives. The first ensures fidelity, sparsity, and localization:

$$\mathcal{L}_1(\boldsymbol{c}, \boldsymbol{k}) = \frac{1}{|\boldsymbol{\mathcal{B}}^+_{\tau_\ell}|} \sum_{x_j \in \boldsymbol{\mathcal{B}}^+_{\tau_\ell}} \left(f(x_j) - \sum_{y_i \in \mathcal{D}_0} c_i \rho^{(\rho_0)}_{k_{y_i}, y_i}(x_j) \right)^2 + \gamma_1 \|\boldsymbol{c}\|_1 + \gamma_2 \left(\|\boldsymbol{k}^{-1}\|_1 + \gamma_3 \|\boldsymbol{k}^{-1}\|_\infty \right).$$

The second promotes separation by penalizing kernel overlap:

$$\mathcal{L}_2(\boldsymbol{k}) = \frac{1}{|\mathcal{D}|^2} \sum_{x_j \in \mathcal{D}} \sum_{y_i \in \mathcal{D}} \rho^{(\rho_0)}_{k_{y_i}, y_i}(x_j).$$

Optimization alternates between $\mathcal{L}_1$ and $\mathcal{L}_2$ for 10 epochs each. Early stopping is triggered when $\mathcal{L}_1$ validation loss ceases to improve or when the R^2 score falls below 99.5% of its maximum. Training and test partitions follow the Faster R-CNN split (358 training, 111 test cases), restricted to bounding boxes with $\tau_\ell \geq 85\%$. Note that although 370 images (291 for training and 79 for validation) were used to train the Faster R-CNN detector, only 358 images produced admissible bounding boxes for AES training. The independent test set of 111 images, however, was identical for both Faster R-CNN evaluation and AES evaluation.

Prototype metrics and clinical relevance

For each test box x, prototype contributions are:

$$t_i(x) = c_{y_i}\, \rho^{(\rho_0)}_{k_{y_i}, y_i}(x).$$

The locally active set $\mathrm{AES}(x)$ is the smallest set capturing fraction α of $\widehat{f}(x)$.

Mean and median relevance scores across the dataset indicate interpretability: lower values show that fewer prototypes are required for clinical explanation. This helps pathologists quickly understand model reasoning and focus on diagnostically significant patterns.

All experiments ran on an NVIDIA RTX A5000 GPU (24 GB VRAM), AMD EPYC 75F3 CPU (64 cores, 3.8 GHz), and 256 GB RAM. Software: PyTorch 2.0, CUDA 11.8, cuDNN 9.1, Python 3.8.10.

Data availability

All data utilized in this study is publicly available from the sources cited in the manuscript.

Code availability

The code for the proposed method is available at GitHub: https://github.com/mitabanik/Adaptive-Example-Selection. This facilitates reproducibility, enables validation of the results, and encourages further research and development in the field.

Received: 4 November 2025; Accepted: 11 February 2026
Published online: 18 February 2026

References

1. Gurcan, M. N. et al. Histopathological image analysis: a review. *IEEE Rev. Biomed. Eng.* **2**, 147–171 (2009).
2. Balkenhol, M. C. A. et al. Deep learning assisted mitotic counting for breast cancer. *Lab. Invest.* **99**(11), 1596–1606 (2019).
3. Veta, M., van Diest, P. J., Jiwa, M., Al-Janabi, S. & Pluim, J. P. W. Mitosis counting in breast cancer: Object-level interobserver agreement and comparison to an automatic method. *PLoS One* **11**(8), e0161286 (2016).
4. Bertram, C. A. et al. Computer-assisted mitotic count using a deep learning-based algorithm improves interobserver reproducibility and accuracy. *Vet. Pathol.* **59**(2), 211–226 (2022).
5. Aubreville, M. et al. Mitosis domain generalization in histopathology images - the MIDOG challenge. *Med. Image Anal.* **84**(102699), 102699 (2023).
6. Aubreville, M. et al. Deep learning algorithms out-perform veterinary pathologists in detecting the mitotically most active tumor region. *Sci. Rep.* **10**(1), 16447 (2020).
7. Jia, X., Ren, L. & Cai, J. Clinical implementation of AI technologies will require interpretable AI models. *Med. Phys.* **47**(1), 1–4 (2020).
8. Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R. & Yu, B. Definitions, methods, and applications in interpretable machine learning. *Proc. Natl. Acad. Sci. U. S. A.* **116**(44), 22071–22080 (2019).
9. Evans, T. et al. The explainability paradox: Challenges for xai in digital pathology. *Future Gener. Comput. Syst.* **133**, 281–296 (2022).
10. Plass, M. et al. Explainability and causability in digital pathology. *J. Pathol.: Clin. Res.* **9**(4), 251–260 (2023).
11. Selvaraju, R. R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626 (2017).
12. Ribeiro, M. T., Singh, S. & Guestrin, C. why should I trust you?: Explaining the predictions of any classifier. *CoRR*, arXiv:1602.04938 (2016).
13. Lundberg, S. M. & Lee, S-I. A unified approach to interpreting model predictions. *CoRR* arXiv:1705.07874 (2017).
14. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**(5), 206–215 (2019).
15. Koh, P. W. & Liang, P. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning* (2017).
16. Hegde, N. et al. Similar image search for histopathology: SMILY. *NPJ Digit. Med.* **2**(1), 56 (2019).
17. Chen, C. et al. This looks like that: Deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems* (Eds Wallach, H.et al.) vol. 32, page 8930–8941. (Curran Associates Inc, 2019).
18. Li, O., Liu, H., Chen, C. & Rudin, C. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. ArXiv, arXiv:1710.04806 (2017).
19. Ren, S., He, K., Girshick, R. B. & Sun, J. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, arXiv:1506.01497 (2015).
20. Aubreville, M. et al. A comprehensive multi-domain dataset for mitotic figure detection. *Sci. Data* **10**(1), 484 (2023).
21. Bien, J. & Tibshirani, R. Prototype selection for interpretable classification. *Ann. Appl. Stat.* **5**(4), 2403–2424 (2011).
22. Arık, S. Ö. & Pfister, T. Protoattend: Attention-based prototypical learning. *J. Mach. Learn. Res.* **21**(211), 1–35 (2020).
23. Snell, J., Swersky, K. & Zemel, R. S. Prototypical networks for few-shot learning. In *Adv. Neural Inf. Process. Syst.* **30**, 4077-4087 (2017).
24. Chen, V. et al. Applying interpretable machine learning in computational biology-pitfalls, recommendations and opportunities for new developments. *Nat. Methods* **21**(8), 1454–1461 (2024).
25. Wendland, H. Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Adv. Comput. Math.* **4**(1), 389–396 (1995).
26. Shen, Y. & Ke, J. Randstainna: Learning stain-agnostic features from histology slides by bridging stain augmentation and normalization. arXiv preprint arXiv:2206.12694 (2022).
27. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *CoRR*, arXiv:1512.03385 (2015).
28. Lin, T.-Y. et al. Microsoft coco: Common objects in context. arXiv preprint (2015).
29. Lin, T.-Y, Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, p 2980–2988 (2017).

Acknowledgements

We would like to thank Laura Oiye for her invaluable assistance in proofreading the manuscript.

Author contributions

MB: Conceptualization (equal); Investigation (lead); Methodology (lead); Writing - Original Draft (equal), Writing - Review, and Editing. KKD: Investigation (supporting); Methodology (supporting); Writing - Original Draft (equal), Writing-Review, and Editing. IM: Methodology (supporting); Data Curation; Writing - Original Draft (supporting), Writing - Review, and Editing. BB: Writing - Review and Editing. NS: Supervision (lead); Conceptualization (equal); Methodology (supporting); Writing - Original Draft (equal), Writing - Review, and Editing.

Funding

This work did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-40283-2.

Correspondence and requests for materials should be addressed to N.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.