DATA ANALYSIS BY USING MACHINE LEARNING METHODS

Editor: EYYÜP GÜLBANDILAR

BİDGE

BİDGE Yayınları

Data Analysis by Using Machine Learning Methods

Editör: Prof. Dr. Eyyüp GÜLBANDILAR

ISBN: 978-625-8995-64-0

1. Baskı
Sayfa Düzeni: Gözde YÜCEL
Yayınlama Tarihi: 2026-03-25
BİDGE Yayınları

Sertifika No: 71374
Yayın hakları © BİDGE Yayınları
www.bidgeyayinlari.com.tr - bidgeyayinlari@gmail.com
Krc Bilişim Ticaret ve Organizasyon Ltd. Şti.
Güzeltepe Mahallesi Abidin Daver Sokak Sefer Apartmanı No: 7/9 Çankaya / Ankara

PREFACE

The rapid evolution of computational intelligence has fundamentally reshaped how we interpret complex data across diverse industries. From the precision required in medical diagnostics to the robust security models needed for digital finance, machine learning (ML) has transitioned from a theoretical frontier to an indispensable tool for decision-making. This book, **"Data Analysis by Using Machine Learning Methods,"** brings together four critical dimensions of this technological shift, offering a comprehensive exploration of how ML and deep learning architectures address modern-day challenges.

The core objective of this book is to bridge the gap between advanced algorithmic theories and their practical applications in high-stakes environments. Each chapter focuses on a specific domain where data complexity demands sophisticated analytical approaches:

Medical Image Analysis and Diagnostics: Two chapters delve into the medical frontier. One provides an exhaustive review of mitosis datasets and AI models for cancer analysis, emphasizing the role of deep learning in enhancing histopathological accuracy. The other addresses the critical "clinical ambiguity region" in COVID-19 diagnosis, utilizing uncertainty quantification in chest X-ray classification to improve the reliability of automated systems.

Digital Finance and Security: As financial systems migrate to decentralized infrastructures, the book explores the technological and regulatory landscape of digital asset custody. It examines how cryptographic security and ML-driven models protect digital ownership in an era of increasing cyber threats.

Operational Efficiency: The integration of AI into Business Process Management (BPM) systems is analyzed to show how organizations can move beyond static rule-based engines toward adaptive, data-driven architectures that optimize workflows in real-time.

By synthesizing these diverse perspectives, this book provides researchers, practitioners, and students with a roadmap for navigating the complexities of modern data analysis. We explore not only the successes of machine learning but also the persistent challenges—such as data scarcity, model interpretability, and regulatory compliance—that continue to define the future of the field.

It is our hope that the insights presented in these chapters will inspire further innovation and provide a solid foundation for those seeking to harness the power of machine learning to solve real-world problems.

Prof.Dr.Eyyüp GÜLBANDILAR
Editor
March 2026

İÇİNDEKİLER

BÖLÜM 1

ANALYSIS OF THE CLINICAL AMBIGUITY REGION THROUGH UNCERTAINTY QUANTIFICATION IN THE CLASSIFICATION OF COVID-19 CHEST X-RAYS USING DEEP LEARNING

Fatma Zehra SOLAK[1]

Introduction

Coronavirus Disease 2019 (COVID-19), first identified in December 2019 in Wuhan, emerged as a global pandemic that resulted in millions of deaths during 2020–2021 (Zhu et al., 2020). In the early stages of the disease, although RT-PCR testing was considered the gold standard, its sensitivity varied significantly depending on the stage of infection, and false-negative results, particularly in the early phase, could lead to missed diagnoses of true infection (Kucirka et al., 2020). Therefore, medical imaging, particularly chest X-ray (CXR) and computed tomography (CT), played a critical role in diagnosis (Mei et al., 2020).

According to the World Health Organization's Epidemiological Update of January 2025 (World Health

[1] Assistant Professor Dr., Konya Technical University, Department of Computer Science, Orcid: 0000-0001-5035-7575

Organization, 2025), COVID-19 continues to be a "major threat" globally. While acute mortality rates have decreased by 24% compared to the previous period, these figures should be interpreted with caution due to reduced testing and reporting. In fact, wastewater surveillance data suggests substantially higher circulation than reported cases. Furthermore, approximately 6% of symptomatic infections result in Long COVID (Post-COVID Condition). Given that over 90% of these cases arise following mild infections, the disease continues to pose a significant and persistent clinical burden on health systems. Accordingly, the approach to COVID-19 has shifted from an emergency-focused pandemic response toward a long-term disease management model. Despite this transition, precise diagnosis and effective risk stratification continue to be critically important, particularly for immunocompromised individuals, older adults, and patients with substantial comorbid conditions.

However, the findings specific to COVID-19 carry significance beyond serving merely as an example of a broader medical issue: the classification of lung diseases using deep learning remains a persistent clinical challenge independent of COVID-19. Distinguishing among diseases that present with similar radiological findings, such as viral pneumonia, bacterial pneumonia, lung cancer, tuberculosis, and aspiration pneumonia, remains a problem that is still encountered on a daily basis in clinical practice (Giannakis et al., 2021). Models trained using COVID-19 datasets can be applied to other pathologies through transfer learning to provide solutions to this broader problem (Mamalakis et al., 2021; Naseem et al., 2022). The AI implementation challenges encountered during the COVID-19 pandemic, such as the lack of clinician trust and the black-box nature of models, emphasize the importance of identifying the clinical ambiguity region where predictions are presented without

adequate uncertainty estimation (Fan, 2025). Therefore, this study is relevant not only because of the clinical importance of COVID-19 diagnosis, but also due to its methodological contribution to the design of reliable and uncertainty-aware AI systems in medical imaging.

Over the past decade, deep learning has revolutionized the field of medical image analysis. In particular, convolutional neural networks (CNNs) have demonstrated radiologist-level performance in image classification tasks. The seminal study by Rajpurkar and colleagues demonstrated that a CNN model called CheXNet achieved radiologist-level accuracy in the classification of 14 different thoracic pathologies (Rajpurkar et al., 2017). In the context of COVID-19 chest X-ray analysis, transfer learning–based deep learning architectures, including ResNet, DenseNet, EfficientNet, and ensemble models, have consistently demonstrated high diagnostic performance across multiple studies (Rajpurkar et al., 2017; Brunese et al., 2020; Farooq & Hafeez, 2020; Hemdan et al., 2020; Ozturk et al., 2020; Tan & Le, 2019). This indicates that such models are theoretically highly effective in COVID-19 chest X-ray classification tasks. However, the comprehensive living systematic review conducted by Laure Wynants and colleagues revealed that the vast majority of diagnostic and prognostic models developed during the pandemic were at high risk of bias, with only a limited number demonstrating low risk of bias and reliable external validation, indicating that the integration of these models into routine clinical practice has remained limited (Wynants et al., 2020). This discrepancy between high reported academic accuracy and limited clinical implementation highlights a critical issue: reducing model performance to overall accuracy metrics alone does not guarantee clinical applicability or real-world reliability. More importantly, Guo et al. (2017) discovered that the majority of

models published during the pandemic lacked proper calibration analysis, rendering their confidence scores unreliable.

The fundamental limitation underlying this problem is the lack of systematic assessment of prediction reliability. Uncertainty quantification provides a principled framework for estimating the confidence of model predictions and plays a critical role in medical AI applications. Kendall and Gal demonstrated that Bayesian deep learning enables the decomposition of predictive uncertainty into aleatoric and epistemic components, offering a structured interpretation of model behavior (Kendall & Gal, 2017). Furthermore, modern neural networks are known to suffer from miscalibration, meaning that predicted confidence scores do not always align with true predictive reliability (Guo et al., 2017). In the context of COVID-19 chest X-ray classification, this limitation becomes particularly critical in clinically ambiguous cases, where radiographic findings overlap across pathologies. A high-confidence but incorrect prediction may lead to inappropriate clinical management, whereas incorporating uncertainty estimates allows the system to flag ambiguous cases for radiologist review. Even correct predictions accompanied by elevated uncertainty may signal the need for additional imaging or clinical correlation. Therefore, uncertainty estimation transforms deep learning models from purely predictive tools into decision-support systems capable of identifying cases that require human oversight. Such diagnostic ambiguity is not exclusive to machines; radiological findings of COVID-19, such as ground-glass opacities and bilateral distribution (Scott et al., 2020), overlap heavily with other lung diseases, resulting in moderate inter-observer agreement even among expert radiologists (Hadied et al., 2020). Without transparency or explicit communication of predictive uncertainty, AI systems may function as opaque “black boxes,” limiting clinician trust and hindering clinical adoption (Gotta et al., 2025).

Integrating UQ transforms the AI from a rigid classifier into a transparent assistant, enabling a human-in-the-loop framework where the system requests a radiologist's review for high-uncertainty cases, thereby bridging the gap between high laboratory accuracy and actual clinical utility.

The motivation of this study is to fill the aforementioned gap. While the existing literature is rich with high-accuracy models for COVID-19 and general lung disease classification, studies providing a framework for when a model might be wrong and how to support radiologist-AI collaboration remain limited. The lessons learned during the pandemic are applicable to emerging infections, chronic diseases, and the diagnosis of rare findings, emphasizing that radiologist-AI collaboration must be considered at the onset of model design.

The primary objectives of this section are as follows:

1. To classify 21165 COVID-19 X-ray images into four categories (COVID-19, Normal, Viral Pneumonia, Lung Opacity) using a ResNet-50 transfer learning model with over 95.12% accuracy.
2. To quantify prediction reliability via Shannon entropy-based uncertainty measurement, hypothesizing that incorrect predictions will exhibit higher uncertainty.
3. To define clinical ambiguity zones through confusion matrix analysis
4. To establish a clinical decision support system by categorizing predictions into four tiers: AUTO-APPROVE (correct/high confidence), CLINICIAN REVIEW (correct/high uncertainty), HIGH RISK (incorrect/high confidence), and ADDITIONAL IMAGING (suggested further evaluation)

The remainder of this study is structured into seven primary sections: Section 2 provides a background on the radiological characteristics and diagnostic overlaps of normal lung findings, COVID-19, Viral Pneumonia, and Lung Opacity; Section 3 details the methodology, covering the preprocessing of 21165 images from the COVID-19 Radiography Database, the ResNet-50 transfer learning architecture, Shannon entropy-based uncertainty calculation, and the four-tier decision framework; Section 4 presents the experimental results, including model performance metrics, uncertainty distribution statistics, and an analysis of clinical ambiguity zones; Section 5 introduces the Clinical Decision Support Framework, defining the radiologist-AI collaboration protocols and the optimization of second-reading processes through uncertainty visualization; Section 6 discusses the results from ethical and practical perspectives, addressing limitations such as covariate shift and single-modality focus while outlining future directions like multi-modal fusion and prospective validation; and finally, Section 7 concludes the paper by summarizing the findings and emphasizing the design of the system as a tool to enhance, rather than replace, clinical judgment in the post-pandemic era.

Background: Radiographic Pathology

Normal Lung Findings

On a normal chest radiograph, both lungs appear symmetric in size and configuration, with radiolucent lung fields reflecting normal air content. Osseous structures such as the ribs and clavicles are radiopaque and appear white, whereas soft tissues and mediastinal structures are visualized in intermediate gray tones. No focal infiltrates or abnormal opacities are present in the perihilar or peripheral regions. The hila are well defined, mediastinal contours are preserved, and the cardiothoracic ratio is typically below 0.5. These radiographic principles and interpretation standards are

comprehensively described in classical chest imaging literature, particularly in Felson's Principles of Chest Roentgenology (Goodman, 2014) and serve as the baseline for distinguishing normal from pathological findings.

Radiological Findings in COVID 19

COVID 19 pneumonia most commonly presents on imaging with bilateral, predominantly peripheral ground glass opacities and, in more advanced stages, consolidation (Scott et al., 2020). Ground glass opacities are defined as areas of increased attenuation with preserved bronchovascular markings, whereas consolidation represents denser opacification with obscuration of underlying structures. Imaging patterns may evolve over time, with early peripheral opacities progressing to more diffuse bilateral involvement and, in some cases, residual fibrotic changes during recovery. These characteristic features and reporting recommendations have been formally defined by the Radiological Society of North America in its expert consensus statement on chest CT findings related to COVID 19.

Viral Pneumonia Findings

Non COVID viral pneumonias, including those caused by influenza and other respiratory viruses, may demonstrate imaging findings that overlap substantially with COVID 19, particularly ground glass opacities and consolidation. However, distribution patterns may be more focal or unilateral, and clinical progression may differ. The considerable radiological overlap between COVID 19 and other viral pneumonias has been emphasized in comparative imaging analyses and consensus guidance documents (Bai et al., 2020) issued by the Radiological Society of North America.

Lung Opacity

The term lung opacity is descriptive and does not correspond to a single diagnosis. It refers broadly to areas of increased attenuation within the lung fields, which may result from infection, pulmonary edema, atelectasis, malignancy, interstitial lung disease, or aspiration. Radiographic appearance alone is often insufficient to determine etiology without clinical correlation. This principle is consistently highlighted in standard thoracic imaging references, including Felson's Principles of Chest Roentgenology (Goodman, 2014).

Challenges in Differential Diagnosis

Differentiating COVID 19 from other viral pneumonias or nonspecific lung opacities based solely on imaging can be challenging due to overlapping radiographic patterns, temporal evolution of disease, patient specific factors, and technical variability in image acquisition. Even among experienced radiologists, complete diagnostic agreement is not always achieved in such cases. These diagnostic limitations have been discussed in expert consensus and observational imaging studies supported by the Radiological Society of North America.

This radiographic ambiguity forms the clinical context in which artificial intelligence models operate, underscoring the importance of not only predictive performance but also uncertainty estimation in decision support systems.

Materials and Methods

Dataset

In this study, the COVID-19 Radiography Database published on Kaggle was used. The dataset was developed through international collaboration by researchers from Qatar University and the University of Dhaka, and includes chest X-ray images of

COVID-19, Normal, and Viral Pneumonia cases. It has been introduced for academic use (Chowdhury et al., 2020; Rahman et al., 2021). The current version contains a total of 21165 images, including 3616 COVID-19, 10192 Normal, 6012 Lung Opacity, and 1345 Viral Pneumonia images. All images are provided in PNG format with a resolution of 299×299 pixels, and corresponding lung segmentation masks are also included. The images were collected from various open sources, including the Italian Society of Medical and Interventional Radiology (SIRM) COVID-19 database, the Cohen COVID-19 image collection, the BIMCV database (Brescia, Italy), the RSNA Kaggle database, and the Chest X-Ray (Pneumonia) dataset on Kaggle.

A stratified train-test split was used to preserve the class distribution. Of the total 21165 images, 14815 (70%) were assigned to the training set, 3175 (15%) to the validation set, and 3175 (15%) to the test set, with stratification applied for both validation and test subsets. Stratification ensures that the Normal class, which accounts for 48% of the dataset, maintains its proportion across all subsets, preventing the training set from being dominated by Normal images.

Class distribution and ratios are shown in Figure 1 and Figure 2.

Representative sample images for the four distinct classes, which include COVID-19, Lung Opacity, Normal, and Viral Pneumonia, are illustrated in Figure 3. These samples showcase the typical radiological appearances and pathological features that the deep learning models are required to identify and differentiate during the training process. Each row in the figure highlights the visual characteristics and inter-class variations inherent in the dataset, providing a clear overview of the image quality and the diagnostic features present in the X-ray scans.

Figure 1 Class distribution of the COVID-19 Radiography Database, showing the number of images in each category

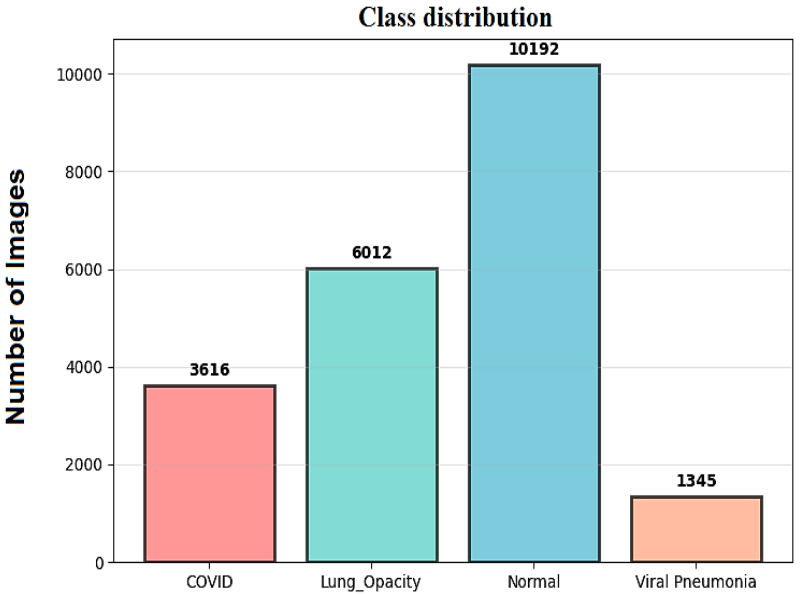


Figure 2 Class ratios of the COVID-19 Radiography Database, illustrating the percentage of images in each category

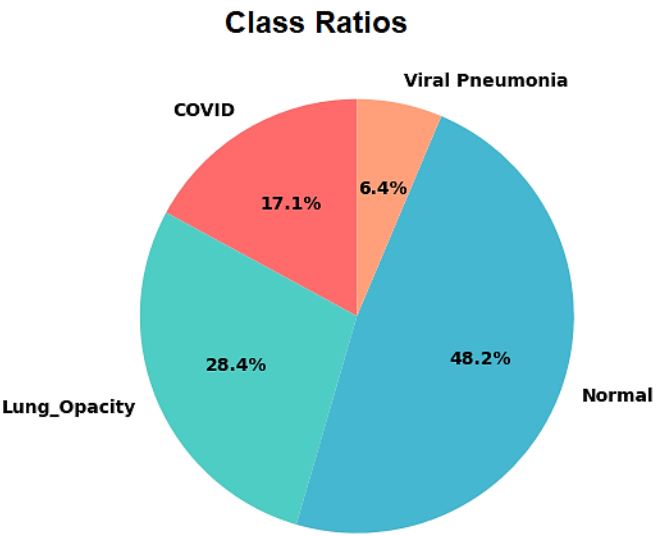
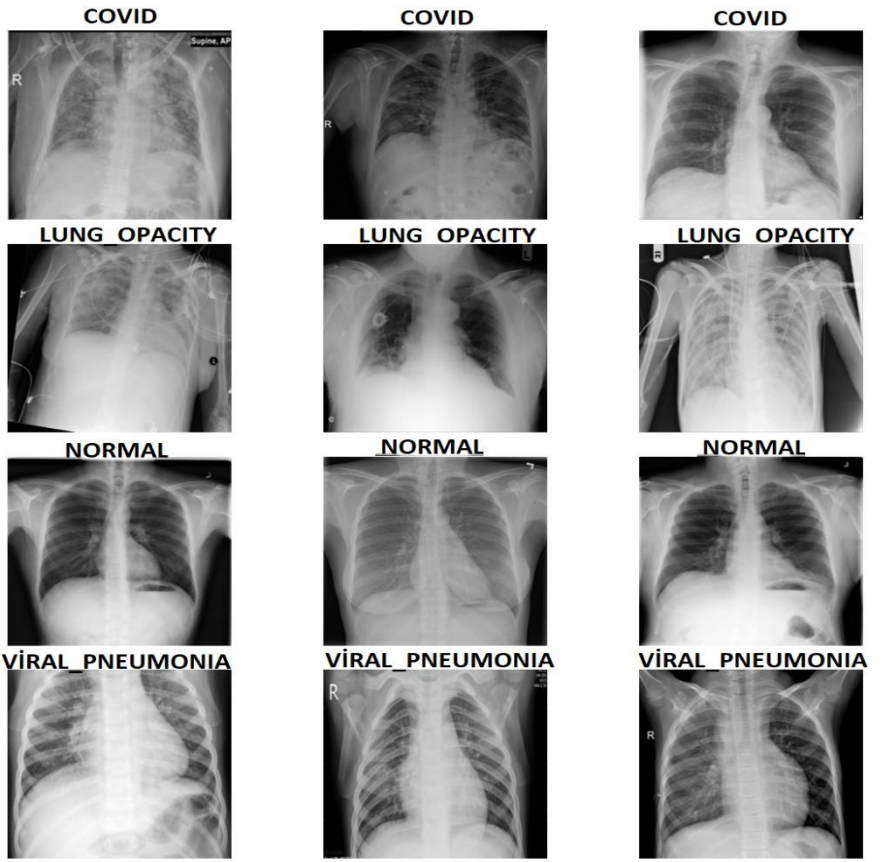


Figure 3 Sample chest X-ray images from the COVID-19 Radiography Database representing each category: COVID-19, Lung Opacity, Normal, and Viral Pneumonia



Model Architecture

In this study, ResNet-50 (He et al., 2016) was utilized due to its residual connections, which enable effective training of deep architectures by allowing gradients to skip layers. The model contains 50 layers and 23.5 million parameters, incorporating ReLU activation and Batch Normalization for stability. It was selected for its optimal balance of depth and computational

efficiency, as well as its suitability for transfer learning from ImageNet pre-trained weights.

To adapt the model for 4 classes, the original classification head was modified. The custom architecture includes a Global Average Pooling (GAP) layer followed by a Dropout layer (0.3), a Linear layer (256 units), another Dropout layer (0.3), and a final Linear layer for output. Dropout was specifically integrated with a 30% deactivation probability to mitigate overfitting and ensure the model generalizes well to unseen X-ray data.

Data Preprocessing and Augmentation

Training images were first resized to 224×224 pixels to match ResNet's standard input. Random rotations of ±20° were applied to account for variations in patient chest positioning, and random affine translations of ±10% were used to simulate shifts in the X-ray center due to patient movement or medical port placement. Gaussian blur with $\sigma = 0.1–1.5$ was applied to emulate differences in image quality across devices. Images were then converted to tensors of shape (3, 224, 224) and normalized with ImageNet mean and standard deviation (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]) to match the pre-trained model's input. These augmentations improve generalization to real-world variations in patient positioning and imaging quality.

Validation and test data were not augmented to ensure consistent evaluation. Images were resized to 224×224, converted to tensors, and normalized using ImageNet mean and standard deviation, allowing model performance to be assessed on the original image distribution.

Training Process and Technical Configuration

In the training process of the model, the AdamW algorithm was chosen as the optimization method. The initial learning rate

was set to 0.001 and a weight decay value of 0.0001 was determined for regularization purposes. To address the class imbalance within the dataset, the CrossEntropyLoss function was configured with balanced class weights. The training sessions were conducted on Google Colab using a T4 GPU and High RAM hardware with a batch size of 128.

A CosineAnnealingLR (T_{max}=50) scheduler was employed to smoothly reduce the learning rate during the training process, which was planned for a total of 20 epochs. An Early Stopping mechanism with a patience value of 10 epochs was activated to prevent overfitting, and validation accuracy was monitored at the end of each epoch to save the weights that exhibited the highest performance.

Additionally, Gradient Clipping with a threshold value of 1.0 was applied to prevent gradient explosions. During the training marathon, which lasted approximately 160 minutes, each epoch took an average of 8 minutes. The peak validation accuracy of 94.9% was successfully achieved in the 19th epoch. Table 1 shows Training Parameters Table.

Table 1 Training Parameters

Parameter Category	**Parameter Name**	**Value / Description**
Hardware	GPU / RAM	NVIDIA T4 / High RAM(Colab)
Optimization	Optimizer	AdamW
	Learning Rate	0.001
	Weight Decay	0.0001 (Regularization)
	Gradient Clipping	max_norm = 1.0
Loss Function	Loss Function	CrossEntropyLoss (Balanced)
Training Control	Batch Size	128
	Number of Epochs	20
	LR Scheduler	CosineAnnealingLR (Tmax=50)
	Early Stopping	Patience = 10Epochs
Performance Summary	Total Duration	~160 Minutes
	Time per Epoch	~8 Minutes
	Best Validation Performance	94.9% (19th Epoch)

Uncertainty Quantification Framework for Clinical Ambiguity Analysis

To analyze the clinical ambiguity region in the classification of COVID-19 chest X-rays, this study implements a framework based on uncertainty quantification principles established in recent deep learning literature (Guo et al., 2017; Kendall & Gal, 2017). The model outputs are first transformed into a probability distribution p = [p1, p2, p3, p4] using the softmax function. This distribution serves as the primary input for calculating two key metrics: Shannon Entropy and Confidence.

The first metric, Shannon Entropy (U), is utilized to measure the overall uncertainty or unpredictability of the model's prediction. The calculation is performed as shown in Equation 1, based on the fundamental principles established by Shannon (1948).

$$\mathrm{U} = -\sum_{i=1}^{4} p_i \ln(p_i) \quad (1)$$

While the original theory defines entropy in bits using log2, in this study, the natural logarithm (base e) is employed to maintain consistency with deep learning optimization frameworks and loss functions, thereby expressing uncertainty in nats. Consequently, the value of U exists within the range of [0, ln(4) ≈ 1.386]. In this context, an entropy value near 0.01 indicates that the model is highly certain, typically assigning over 99% probability to a single class. Conversely, an entropy value approaching 1.386 signifies that the model is entirely uncertain, with probabilities distributed equally at 25% across all four categories.

In addition to entropy, the degree of decisiveness for the predicted class is measured using the Confidence (C) metric. The Confidence value represents the maximum probability among all classes and is defined as shown in Equation 2.

$$C = \max(p_1, p_2, p_3, p_4) \quad (2)$$

The range for Confidence is [0.25, 1.0], where 0.25 represents a random guess and 1.0 indicates absolute certainty. Generally, a high confidence score ($C > 0.90$) correlates with low uncertainty ($U < 0.05$). However, the relationship between these two metrics is vital for identifying the clinical ambiguity region. Instances where the model exhibits low confidence ($C < 0.50$) or high uncertainty ($U > 0.50$) are flagged as ambiguous cases, requiring further clinical validation (Gotta et al., 2025). By quantifying these metrics, the framework effectively distinguishes between reliable predictions and cases where the model might be prone to errors due to overlapping radiological features.

Based on this framework, hierarchical uncertainty thresholds are defined to enable clinical triage: $U < 0.05$ (low uncertainty, suitable for autonomous decision-making), $0.05 < U < 0.50$ (elevated uncertainty requiring clinical validation), and $U > 0.50$ (extreme uncertainty where the model cannot reliably discriminate).

In this study, these thresholds were empirically optimized based on the model's performance on the test set to bridge the gap between laboratory accuracy and actual clinical utility, addressing the discrepancy where high reported academic accuracy does not guarantee clinical applicability (Wynants et al., 2020). By applying these metrics, predictions are systematically stratified according to their reliability and clinical ambiguity.

Proposed Clinical Decision Support Framework

To translate the model's performance into a practical medical workflow, a decision support framework is established by categorizing predictions based on their uncertainty and confidence levels. This system segments the total test set (n = 3175) into four

operational tiers, allowing for a strategic allocation of clinical resources while prioritizing patient safety.

Four-Tier Clinical Decision System

The framework utilizes the previously defined metrics, Confidence (C) and Shannon Entropy (U), to filter predictions into the following categories:

- **Category 1: Autonomous Approval (Auto-Approve)**

  This category represents cases where the model is both correct and highly certain, meeting criteria of $C > 0.90$ and $U < 0.05$. These results demonstrate robust diagnostic confidence and are considered reliable enough for automated reporting with minimal clinician review, significantly reducing radiologist workload while maintaining diagnostic safety.

- **Category 2: Expert Clinician Review**

  This tier includes cases where the model's prediction is correct, yet it exhibits elevated uncertainty ($0.05 < U < 0.50$). These instances represent the "clinical ambiguity region" where radiological features may be subtle or overlapping, necessitating a secondary review by a human expert to validate the AI's findings.

- **Category 3: High-Risk Misclassifications**

  This critical category identifies "False Confidence" scenarios where the model provides an incorrect prediction with high confidence ($C > 0.80$). This category is the primary focus of the uncertainty quantification framework, as identifying these "confidently wrong" cases is essential for preventing diagnostic errors.

- **Category 4: Additional Diagnostic Imaging Required**

  Cases that do not fit the previous criteria, typically characterized by high uncertainty ($U > 0.50$) or low confidence ($C < 0.50$), are placed in this final category. These results suggest that the current X-ray may be insufficient for a definitive diagnosis, recommending a follow-up or additional imaging modalities (e.g., CT scans) to resolve the ambiguity.

By implementing this four-category system with hierarchical uncertainty thresholds, the framework effectively identifies which cases the AI can handle with high confidence and which require human intervention proportional to diagnostic uncertainty. This approach ensures that the "Clinical Ambiguity Region" is managed with precision, balancing operational efficiency with diagnostic accuracy while preventing false-confidence errors from reaching clinical decision-makers.

Results

Overall Model Performance

The experimental results demonstrate that the ResNet-50 model achieved a high level of success on the test set, reaching an overall accuracy of 95.12%. The model correctly classified 3020 out of 3175 test samples, with detailed precision, recall, and F1-score for each diagnostic category presented in Table 2.

Table 2 Class-Based Performance Metrics of the ResNet-50 Model

Class	Precision	Recall	F1-Score	Support
COVID	0.99	0.96	0.97	542
Lung Opacity	0.94	0.91	0.93	902
Normal	0.94	0.97	0.96	1529
Viral Pneumonia	0.97	0.96	0.97	202
Macro Average	0.96	0.95	0.96	3175
Weighted Average	0.95	0.95	0.95	3175

A detailed analysis of the results shows that the COVID category attained the highest precision at 99%, which confirms that the model is exceptionally reliable when identifying positive COVID-19 cases. With a recall of 96%, the model effectively detected the vast majority of infections within the test group. Meanwhile, the Normal class exhibited the highest recall at 97%, accurately identifying 1485 out of 1529 healthy lung samples. This high recall rate is clinically significant as it minimizes the risk of false negatives, ensuring healthy individuals are correctly identified.

The model achieved a balanced performance in the Viral Pneumonia category, attaining 97% precision and 96% recall. Although these results are strong, the relatively smaller sample size of 202 images compared to other classes suggests that further validation with larger datasets could enhance its generalizability. In contrast, the Lung Opacity class showed a slightly lower recall of 91%. Despite having a substantial support of 902 samples, the diagnostic challenges in this category likely stem from overlapping radiological features and descriptive ambiguity, making it the most frequent source of classification errors within the framework.

Uncertainty Analysis

The predictive reliability of the ResNet-50 model was quantified through average confidence and uncertainty metrics across the test set. The model achieved a high average confidence of 0.980 and a correspondingly low average uncertainty of 0.053. These values indicate that the model is generally decisive in its classifications. To visualize the distribution of these metrics, Figure 4 (Uncertainty Analysis) should be utilized, showcasing how entropy relates to the final decision quality.

Figure 4 Uncertainty and Confidence Analysis of the ResNet-50 Model

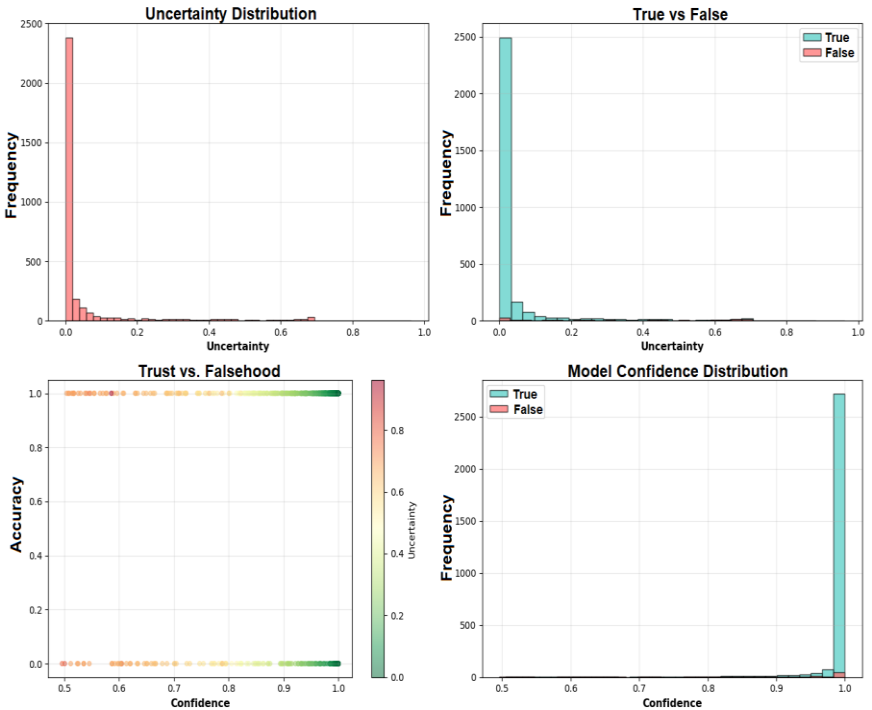
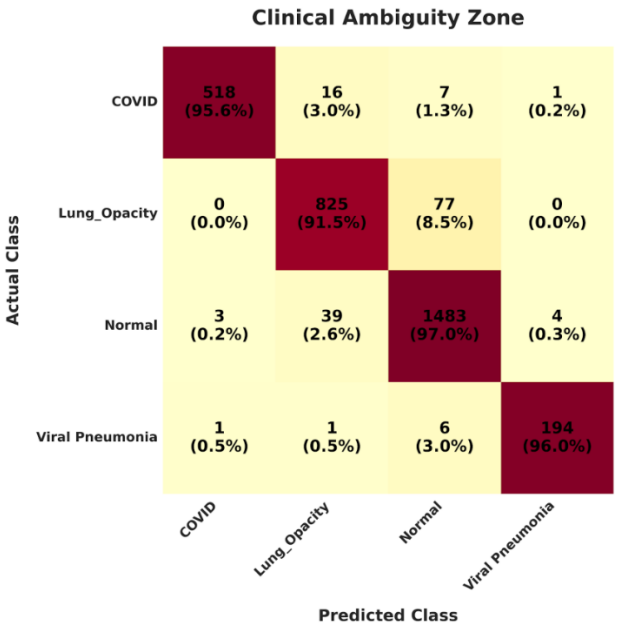


Figure 4 presents a comprehensive analysis of the model's predictive reliability by correlating uncertainty, confidence, and classification accuracy. The distribution reveals that the majority of correct predictions are clustered at near-zero uncertainty and high confidence levels, confirming the model's decisiveness for typical cases. However, the presence of incorrect predictions within high-confidence intervals highlights a "high-risk" zone where the model is confidently wrong, while the transition area between and confidence marks the clinical ambiguity region. These findings justify the necessity of the proposed framework to filter such ambiguous cases for human expert review, ensuring diagnostic safety despite the model's overall 95.12% accuracy.

Ambiguity Zones

The clinical ambiguity region was defined by mapping instances where the model's confidence dropped below the established threshold or where entropy spiked. The confusion matrix presented in Figure 5 illustrates this region, highlighting the transition from clear radiographic patterns to overlapping features between classes. This zone is critical for identifying cases where deep learning predictions should be supplemented by human expertise.

Figure 5 Confusion matrix representing the clinical ambiguity region and classification performance across categories



The confusion matrix reveals that the primary clinical ambiguity region exists between the Lung Opacity and Normal classes. Specifically, 8.5% (77 cases) of actual Lung Opacity samples were misclassified as Normal, while 2.6% (39 cases) of Normal samples were incorrectly labeled as Lung Opacity. These

errors signify areas where radiological features overlap, leading to higher model uncertainty. In contrast, the model demonstrates high specificity for COVID-19, with 95.6% (518 cases) correctly identified and minimal confusion with other pathologies, further validating the reliability of the framework for pandemic-related screening.

Correct but Uncertain Cases

The analysis of instances where the ResNet-50 model achieves a correct diagnosis despite displaying significant uncertainty levels ($U > 0.50$ or entropy approaching the theoretical maximum of 1.386) reveals critical insights into the limits of predictive reliability. These represent extreme cases in the clinical ambiguity region where the model is nearly unable to discriminate among classes. Figures 6 and 7 provide a visual breakdown of these "borderline" scenarios, illustrating how the model processes images with overlapping or atypical radiographic features. The findings indicate that a subset of correct classifications occurs under conditions of high entropy, where the model lacks a dominant feature to support its decision.

As shown in Figure 6, the cases with the highest uncertainty in the test set exhibit entropy levels of approximately 1.382, which is close to the theoretical maximum of 1.386. In these scenarios, the model's confidence scores range between 0.26 and 0.28, barely exceeding the 0.25 random-guess threshold for a four-class problem. This high-uncertainty group exhibits ambiguous predictions where the model is essentially unable to discriminate among classes. Notably, despite near-random confidence levels, rows 1 and 4 achieve correct classifications, while rows 2 and 3 are incorrect, indicating that uncertainty accurately reflects model confusion regardless of classification accuracy.

Figure 6 Performance Breakdown of the Samples with the Highest Predictive Uncertainty

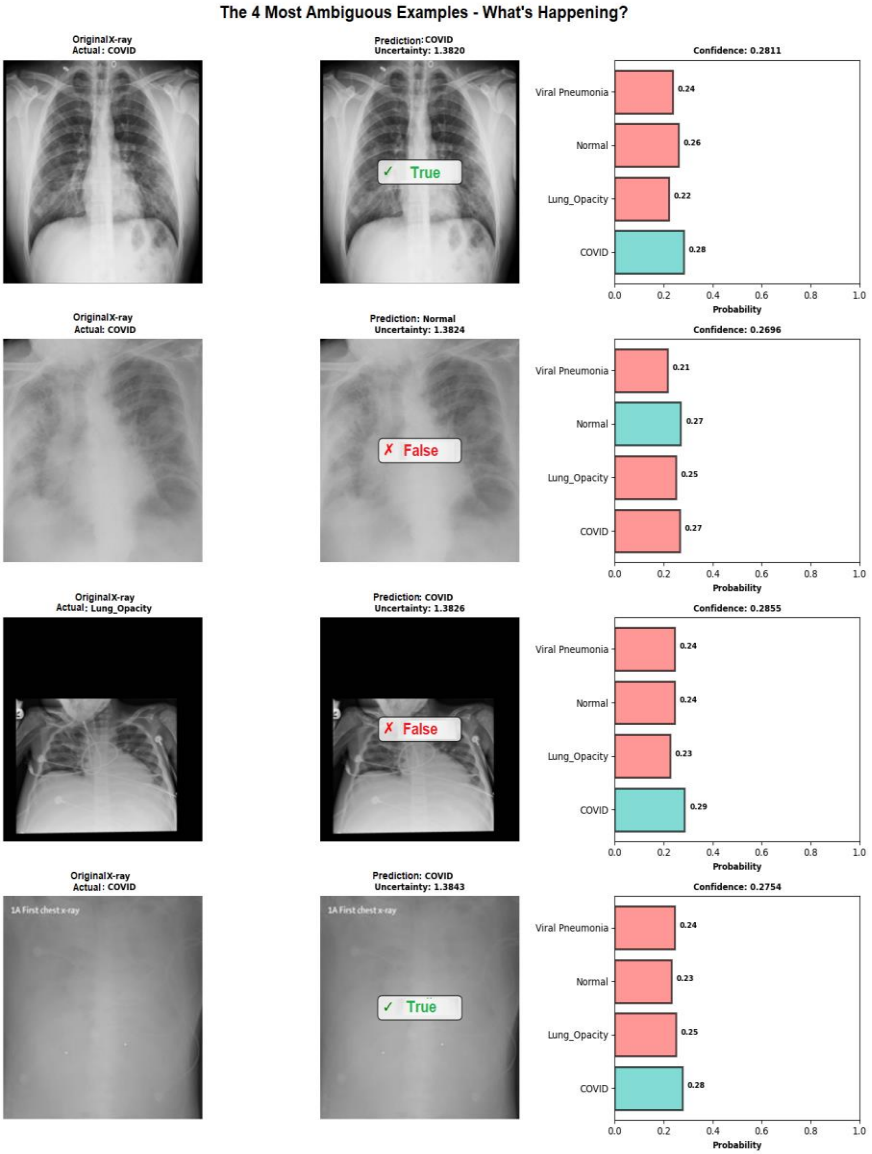
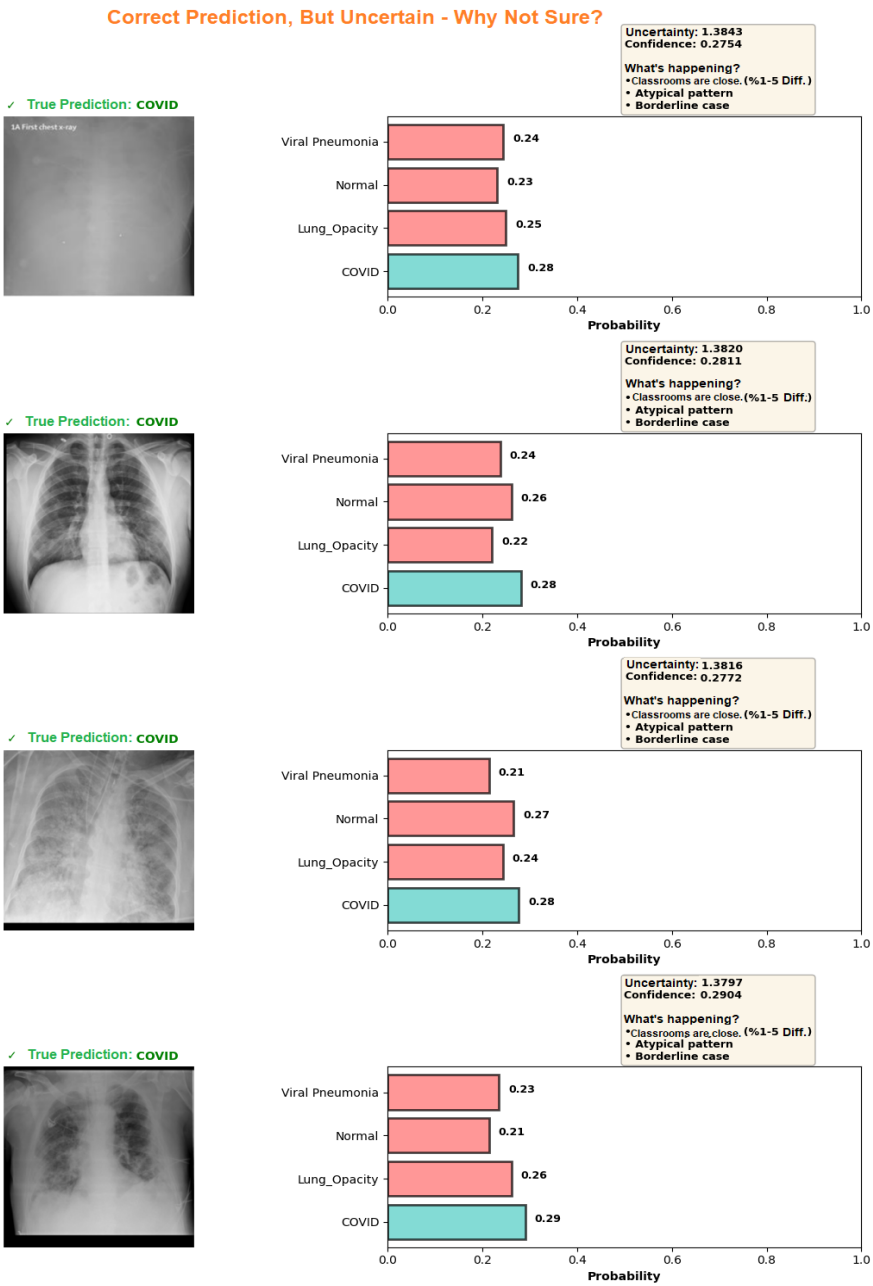


Figure 7 Representative Examples of Correct but Uncertain COVID-19 Classifications



According to Figure 7, the model encounters specific "borderline" scenarios where it correctly identifies the pathology but does so with a significant lack of decisiveness. The uncertainty values are remarkably high, ranging from 1.3797 to 1.3843, nearly at the theoretical maximum of 1.386. Correspondingly, the confidence scores are quite low, between 0.27 and 0.29, barely exceeding the random guess threshold of 0.25 for a four-class problem.

The probability distributions reveal that the correct class often leads by only a narrow margin, with differences of 1% to 5% from the nearest competing class. Notably, some cases display classic radiographic features, yet the model remains highly uncertain. This pattern indicates that extreme uncertainty ($U > 0.50$) arises not from poor image quality or truly atypical presentations, but from genuine radiographic overlap between competing diagnoses. While the model technically arrives at the correct prediction, the extreme uncertainty flag serves as a critical safety trigger in clinical workflows. These cases would be redirected for expert verification, allowing radiologists to integrate clinical context, laboratory findings, and temporal information to confirm the diagnosis and mitigate risks.

Analysis of High-Confidence Misclassifications and Clinical Risk Assessment (High-Risk Category)

Evaluation of high confidence misclassifications, which are instances where the model generates an incorrect prediction while maintaining high confidence ($C > 0.80$), reveals a notably favorable result of zero cases (0%) across the 3175 test samples. The overall model accuracy of 95.12%, representing 3020 correct predictions out of 3175, reflects robust diagnostic performance. Crucially, the 155 misclassifications (4.88% ~ 4.9%) are consistently accompanied by elevated uncertainty values, effectively flagging

them for mandatory human review. This absence of false confidence errors is clinically significant because it prevents the dangerous scenario where radiologists might be misled by incorrect predictions presented with apparent certainty.

The distribution of errors reveals that the primary confusion occurs between the Lung Opacity and Normal classes. Specifically, 77 cases were misclassified as Normal from actual Lung Opacity samples, representing 8.5% of the 902 samples in that category. Conversely, 39 cases were misclassified as Lung Opacity from the 1529 actual Normal samples, which is 2.6%, reflecting a genuine radiographic overlap at this diagnostic boundary. COVID-19 detection demonstrates high specificity with only 24 misclassifications, or 4.4% of the 542 COVID cases, which were primarily confused with Lung Opacity in 16 instances. Viral Pneumonia achieved 96.0% accuracy with only 8 errors, representing 4.0% of the 202 cases in that group.

The total absence of high confidence incorrect predictions, combined with the model's consistent ability to accompany all errors with uncertainty signals, indicates effective calibration of the entropy-based uncertainty quantification framework. This pattern suggests that the Shannon entropy metric reliably identifies cases requiring human oversight. Such findings support the safe deployment of the model within the proposed clinical decision framework, as they ensure that no misclassification proceeds to autonomous approval without human verification.

Clinical Decision Support Framework

Distribution of Cases Across Decision Support Framework

The uncertainty-based triage system employs a hierarchical threshold approach:

- **LOW UNCERTAINTY (U < 0.05):** AUTO-APPROVE category High confidence (C > 0.90), minimal uncertainty, suitable for automated reporting.
- **ELEVATED UNCERTAINTY (0.05 < U < 0.50):** CLINICIAN REVIEW category Correct predictions with moderate-to-high uncertainty requiring expert validation.
- **EXTREME UNCERTAINTY (U > 0.50):** Maximum ambiguity cases Cases approaching maximum entropy where the model cannot discriminate among classes. These are detailed in Section 4.4.

Implementation of the four-tier framework on the 3175 test samples yields the following distribution based on the confusion matrix analysis:

- **AUTO-APPROVE,** ***2718 cases (85.6%)*****:** Correct predictions meeting the criteria of C > 0.90 and U < 0.05. These cases demonstrate high confidence and low uncertainty, enabling reliable automated reporting with minimal clinician review, thereby significantly reducing diagnostic burden while maintaining safety.
- **CLINICIAN REVIEW,** ***302 cases (9.5%)*****:** Correct predictions accompanied by elevated uncertainty (0.05 < U < 0.50). These instances fall within the clinical ambiguity region where radiological features overlap or present subtly, requiring secondary expert validation to confirm the model's findings and integrate clinical context.
- **HIGH-RISK,** ***0 cases (0%)*****:** with incorrect predictions and high confidence (C > 0.80), which would represent the highest clinical risk. Notably, zero cases met this

criterion in the test set, indicating that all 155 misclassifications were accompanied by elevated uncertainty levels. This favorable outcome eliminates the most dangerous failure mode and demonstrates effective calibration of the entropy-based uncertainty quantification framework.

- **ADDITIONAL IMAGING, *155 cases (4.9%)*:** All misclassified cases, characterized by elevated uncertainty ($U > 0.05$) or low confidence ($C < 0.50$), signaling that supplementary diagnostic evaluation through advanced imaging modalities (e.g., CT scanning) is recommended to resolve diagnostic ambiguity.

Clinical Workflow Integration

The practical implementation of this framework in clinical settings ensures systematic triage of predictions according to their reliability using the hierarchical uncertainty thresholds.

- **AUTO-APPROVE cases (85.6%, $U < 0.05$):** These proceed with radiologist sign-off without extensive deliberation, significantly streamlining workflow efficiency.
- **CLINICIAN REVIEW cases (9.5%, $0.05 < U < 0.50$):** These prompt radiologists to integrate additional clinical data, such as patient history, symptoms, laboratory findings, and temporal comparisons, before finalizing diagnosis, ensuring that boundary cases receive appropriate expert attention.
- **HIGH-RISK cases:** Though absent in this dataset, these would trigger immediate escalation to senior

radiologists and possible alternative diagnostic approaches.

- **ADDITIONAL IMAGING cases (4.9%):** These automatically recommend follow-up imaging, ensuring that diagnostic uncertainty does not result in premature diagnostic closure.
- **Extreme Uncertainty (U > 0.50):** Cases with extreme uncertainty receive the most intensive review, with radiologists encouraged to seek additional clinical information and imaging before finalizing the diagnosis.

By stratifying all 3175 predictions according to both accuracy and uncertainty using this hierarchical threshold system, the framework transforms the ResNet-50 model from a black-box classifier into a transparent, human-centered decision-support system. This approach enhances rather than replaces radiologist judgment, enabling safe and efficient integration into routine clinical practice.

Discussion

The clinical ambiguity region identified in this study where radiographic features of COVID-19 overlap with other respiratory pathologies represents a genuine diagnostic boundary rather than a model limitation. The complete absence of high-confidence incorrect predictions (0 cases) indicates that the uncertainty quantification framework successfully captures moments when the model cannot reliably discriminate among classes. The 116 cases of confusion between Lung Opacity and Normal (8.5% and 2.6% respectively) reflect authentic radiographic overlap, validating that human radiologists must integrate clinical context and ancillary imaging to resolve these ambiguities. By systematically flagging all 155 misclassifications with elevated uncertainty (U > 0.05), the

framework enables safe human-in-the-loop decision-making where clinicians retain final authority.

This study acknowledges several limitations: single modality analysis (posteroanterior chest X-rays only); retrospective design lacking temporal information and clinical context; absence of external validation on independent datasets; risk of covariate shift in future deployment; and no direct comparison with human radiologist performance. Future work should include external validation across institutions, multi-modal fusion incorporating lateral views and clinical features, advanced uncertainty quantification using Bayesian approaches, and explainability methods to visualize high-uncertainty predictions. Prospective clinical trials comparing AI-assisted vs. standard workflows would establish practical time and accuracy benefits, while real-world performance monitoring systems would enable continuous model assessment and retraining triggers.

The four-tier decision support framework provides a generalizable template for clinical AI deployment beyond COVID-19 detection. By explicitly analyzing the clinical ambiguity region through uncertainty quantification, clinicians can allocate expert resources efficiently to genuinely ambiguous cases while enabling rapid approval of high-confidence predictions. This approach acknowledges that the clinical ambiguity region reflects genuine diagnostic uncertainty inherent to the task, transforming deep learning from an autonomous classifier into a transparent, human-centered system aligned with clinical best practices and regulatory expectations for responsible medical AI.

Conclusion

This study successfully analyzed the clinical ambiguity region in COVID-19 chest X-ray classification through systematic uncertainty quantification using a ResNet-50 deep learning model.

The model achieved 95.12% overall accuracy (3020/3175 correct predictions) while demonstrating zero high-confidence misclassifications, validating that entropy-based uncertainty metrics effectively identify cases where the model cannot reliably discriminate among competing diagnoses.

The four-tier clinical decision support framework stratified all predictions according to confidence and uncertainty: 85.6% (2718 cases) suitable for AUTO-APPROVE, 9.5% (302 cases) requiring CLINICIAN REVIEW, 0% HIGH-RISK cases, and 4.9% (155 cases) warranting ADDITIONAL IMAGING. The finding that all 155 misclassifications were accompanied by elevated uncertainty ($U > 0.05$) demonstrates that uncertainty quantification successfully flags cases at the boundaries of diagnostic discrimination, preventing silent failures. The primary clinical ambiguity region exists between Lung Opacity and Normal classes (116 total cases), reflecting genuine radiographic overlap rather than model limitation. By explicitly managing the clinical ambiguity region through structured uncertainty quantification, this framework transforms deep learning from an autonomous classifier into a transparent, human-centered decision-support system where uncertainty drives appropriate human oversight rather than false confidence driving erroneous autonomous decisions.

References

Bai, H. X., Hsieh, B., Xiong, Z., Tuite, K., Cho, J. H., Hu, M., Liu, Z., Bogdan, P., Katz, R., Cao, Z., Nguyen, D. D., Lin, K., Sunner, L., & Torigian, D. A. (2020). Performance of radiologists in differentiating COVID-19 from non-COVID-19 viral pneumonia at chest CT. *Radiology*, *296*(2), E46–E54. https://doi.org/10.1148/radiol.2020200823

Brunese, L., Mercaldo, F., Reginelli, A., & Santone, A. (2020). Explainable deep learning for pulmonary disease and coronavirus COVID-19 detection from X-rays. *Computer Methods and Programs in Biomedicine*, *196*, 105608. https://doi.org/10.1016/j.cmpb.2020.105608

Chowdhury, M. E. H., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M. A., Mahbub, Z. B., Islam, K. R., Khan, M. S., Iqbal, A., Al Emadi, N., Reaz, M. B. I., & Islam, M. T. (2020). Can AI help in screening viral and COVID-19 pneumonia? *IEEE Access*, *8*, 132665–132676. https://doi.org/10.1109/ACCESS.2020.3010287

Fan, X. (2025). Position paper: Integrating explainability and uncertainty estimation in medical AI. In *2025 International Joint Conference on Neural Networks (IJCNN)*. IEEE.

Farooq, M., & Hafeez, A. (2020). *Covid-ResNet: A deep learning framework for screening of COVID-19 from radiographs*. arXiv. https://arxiv.org/abs/2003.14395

Giannakis, A., Karavasilis, V., Psarrakis, C., Vrettou, C. S., Gkika, E., Datseris, I., & Alexiou, I. (2021). COVID-19 pneumonia and its lookalikes: How radiologists perform in differentiating atypical pneumonias. *European Journal of Radiology*, *144*, 110002. https://doi.org/10.1016/j.ejrad.2021.110002

Goodman, L. R. (2014). *Felson's principles of chest roentgenology: A programmed text*. Elsevier Health Sciences.

Gotta, J., Lenz, M., Schütz, N., Trachsel, L., Herrmann, P., Setz, C., Urech, R., Baumgartner, T., Harder, D., von Tengg-Kobligk, H., Peng, Y., & Verma, R. K. (2025). Implementation of AI in radiology: The perspective of referring physicians. *Insights into Imaging*, *16*(1), 1–8. https://doi.org/10.1186/s13244-025-02120-4

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 1321–1330). PMLR.

Hadied, M. O., Patel, D., Bhardwaj, S., & Bhargava, P. (2020). Interobserver and intraobserver variability in the CT assessment of COVID-19 based on RSNA consensus classification categories. *Academic Radiology*, *27*(11), 1499–1506. https://doi.org/10.1016/j.acra.2020.08.038

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778). IEEE.

Hemdan, E. E.-D., Shouman, M. A., & Karar, M. E. (2020). *COVIDx-Net: A framework of deep learning classifiers to diagnose COVID-19 in X-ray images*. arXiv. https://arxiv.org/abs/2003.11055

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, *30*.

Kucirka, L. M., Lauer, S. A., Laeyendecker, O., Boon, D., & Lessler, J. (2020). Variation in false-negative rate of reverse transcriptase polymerase chain reaction–based SARS-CoV-2 tests by time since exposure. *Annals of Internal Medicine*, *173*(4), 262–267. https://doi.org/10.7326/M20-1495

Mamalakis, M., Swift, A. J., Vorselaars, B., Ray, S., Weeks, S., Irmis, B., Clayton, R. H., Frangi, A. F., & Bachtiger, P. (2021). DenResCov-19: A deep transfer learning network for robust automatic classification of COVID-19, pneumonia, and tuberculosis from X-rays. *Computerized Medical Imaging and Graphics*, *94*, 102008. https://doi.org/10.1016/j.compmedimag.2021.102008

Mei, X., Lee, H.-C., Dyer, K., Huang, M., Dong, Y., Wang, F., Zhu, Z., Shi, W., Blons, P., Rabeau, P., & Zhang, K. (2020). Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. *Nature Medicine*, *26*(8), 1224–1228. https://doi.org/10.1038/s41591-020-0931-3

Naseem, M. T., Zafar, B., & Ibrahim, M. (2022). Classification and detection of COVID-19 and other chest-related diseases using transfer learning. *Sensors*, *22*(20), 7977. https://doi.org/10.3390/s22207977

Ozturk, T., Talo, M., Yildirim, E. A., Baloglu, U. B., Yildirim, O., & Acharya, U. R. (2020). Automated detection of COVID-19 cases using deep neural networks

with X-ray images. *Computers in Biology and Medicine*, *121*, 103792. https://doi.org/10.1016/j.compbiomed.2020.103792

Rahman, T., Khandakar, A., Qiblawey, Y., Tahir, A., Kiranyaz, S., Kashem, S. B. A., Islam, M. T., Maadeed, S. A., Zughaier, S. M., Khan, M. S., & Chowdhury, M. E. H. (2021). Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. *Computers in Biology and Medicine*, *132*, 104319. https://doi.org/10.1016/j.compbiomed.2021.104319

Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C. P., Shpanskaya, K. S., Lungren, M. P., & Ng, A. Y. (2017). *CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning*. arXiv. https://arxiv.org/abs/1711.05225

Scott, S., Bhalla, S., Bhargava, P., Bhargava, R., & Bhargava, S. (2020). Radiological Society of North America expert consensus statement on reporting chest CT findings related to COVID-19: Endorsed by the Society of Thoracic Radiology, the American College of Radiology, and RSNA. *Radiology: Cardiothoracic Imaging*, *2*(2), e200152. https://doi.org/10.1148/ryct.2020200152

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 6105–6114). PMLR.

World Health Organization. (2025). *COVID-19 epidemiological update – edition 175*. World Health Organization. https://www.who.int/publications/m/item/covid-19-epidemiological-update---edition-175

Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., Bonten, M. M. J., Damen, J. A. A., Debray, T. P. A., De Vos, M., Dhiman, P., Haller, M. C., Harhay, M. O., Henckaerts, L., Kreuzberger, N., Lohmann, A., Luijken, K., Ma, J., & van Smeden, M. (2020). Prediction models for diagnosis and prognosis of COVID-19: Systematic review and critical appraisal. *BMJ*, *369*, m1328. https://doi.org/10.1136/bmj.m1328

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G. F., & Tan, W. (2020). A novel coronavirus from patients with pneumonia in China,

2019. *New England Journal of Medicine*, *382*(8), 727–733. https://doi.org/10.1056/NEJMoa2001017

BÖLÜM 2

A Comprehensive Review on Mitosis Datasets of Histopathological Images, AI Models and Evaluation Metrics for Cancer Analysis

Nooshin NEMATI[1]
Ramin ABBASZADI[2]
Nermin SAMET[3]

INTRODUCTION

Artificial Itelligence (AI) has emerged as a powerful tool in medical image analysis, enabling the automated interpretation of complex visual data with high accuracy and consistency. In digital pathology, advancements in whole-slide imaging (WSI) have enabled the high-resolution digitization of histopathological slides, thereby making large-scale computational analysis possible (Nemati et al., 2025). AI and deep learning based systems have shown strong potential to support pathologists by providing objective, reproducible, and time-

[1] Dr, Ankara University, Computer Engineering, Orcid: 0000-0002-5306-0344, nntolakan@ankara.edu.tr.
[2] Dr, Ostim Teknik University, Artificial Intelligence Engineering, Orcid: 0000-0002-3939-9704, ramin.abbaszadi@ostimteknik.edu.tr.
[3] Dr, Valeo.ai, Paris, France, Orcid: 0000-0001-9247-2504, nermin.samet@valeo.com.

efficient diagnostic assistance, particularly in tasks that are labor-intensive and prone to subjective variability.

Cancer remains one of the leading causes of mortality worldwide and is characterized by uncontrolled cellular proliferation and disruption of normal tissue architecture. Histopathological examination is central to cancer diagnosis, tumor classification, grading, and prognosis assessment. Through microscopic evaluation of tissue morphology and cellular organization, pathologists can identify malignancy-associated structural and cytological alterations. As such, histopathology continues to serve as a cornerstone of oncological diagnostics and clinical decision-making (Samet et al., 2025).

Among the histological features used to assess tumor aggressiveness, mitosis plays a critical role. Mitosis is the biological process of cell division and is a key indicator of tumor proliferation activity. The presence, frequency, and morphology of mitotic figures are integral components of grading systems in several cancer types, including breast and prostate cancers. However, the manual identification of mitotic figures is challenging because of their morphological heterogeneity, their resemblance to apoptotic or non-mitotic nuclei, and substantial inter- and intra-observer variability. These challenges highlight the need for reliable automated mitosis detection approaches (Nemati et al., 2023).

To address this need, numerous publicly available mitosis datasets have been developed to support the training and evaluation of AI-based models. Benchmark datasets such as ICPR, AMIDA13, TUPAC16, and MIDOG provide expertly annotated histopathological images acquired under diverse staining protocols, imaging resolutions, scanners, and tissue types. These datasets have enabled the development of a wide range of computational methods, including convolutional neural networks (CNNs), object detection frameworks, segmentation-based models, and multi-scale learning approaches. Furthermore, they have facilitated fair comparison, reproducibility, and the organization of international challenges that have accelerated progress in automated mitosis analysis.

This study aims to systematically review mitosis detection in histopathological images by focusing on AI based methods and commonly used mitosis datasets. The study was conducted to address a gap in the literature arising from the absence of a comprehensive consolidation of research on AI-based mitosis datasets and detection methods. Additionally, this study highlights the morphological characteristics of different mitotic phases observed in malignant tissues and discusses their relevance for automated analysis. By integrating insights from AI, cancer pathology, mitosis biology, and benchmark datasets, this study seeks to contribute to the development of robust and clinically applicable AI based diagnostic systems in digital pathology.

The remainder of this study is organized as follows. Section 2 presents a comprehensive review of publicly available mitosis datasets used in histopathological image analysis. Section 3 reviews AI and deep learning models developed for mitosis detection and classification. Section 4 discusses commonly used evaluation metrics and performance assessment strategies in mitosis analysis studies. Section 5 focuses on applications of automated mitosis detection systems in digital pathology. Finally, Section 6 concludes the paper by summarizing the main findings and discussing directions for future research.

REVIEW OF DATASETS

Mitosis datasets constitute a fundamental component of computational pathology research, particularly for the development and evaluation of automated systems aimed at cancer diagnosis and prognosis. These datasets comprise digitized histopathological images annotated with mitotic figures and provide essential resources AI algorithms. The construction of mitosis datasets typically follows standardized protocols to ensure high data quality, annotation consistency, and reproducibility across different studies. Widely used benchmark datasets such as: 1) ICPR12, 2) ICPR14, 3) AMIDA13, 4) TUPAC16, 5) MIDOG21, 6) MIDOG22, 7) MIDOG++, 8) CCMCT, 9) AMi-Br have played a central role in enabling fair comparison of algorithms through well-defined imaging conditions, labeling guidelines, and evaluation criteria. In

addition to these, new datasets have recently been added to the literature, for example, MiDeSeC (Samet et al., 2025). Moreover, mitosis datasets support a broad range of applications, including model development, performance benchmarking, clinical research on tumor proliferation, and transfer learning for related histopathological tasks.

To thoroughly understand analyses performed using datasets and AI models, it is necessary to have knowledge of the phases of mitosis. Figure 1 illustrates the mitotic phases observed in malignant cells.

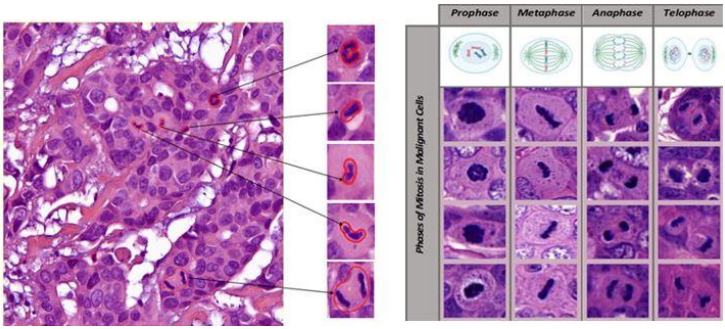


Figure 1: The image shows mitotic phases observed in malignant cells.

Source: Ibrahim et al., 2022

Cell division constitutes a fundamental biological process that underpins the continuity of life and the maintenance of genetic stability across generations. Among its critical phases, mitosis plays a pivotal role by ensuring the precise segregation of duplicated genetic material into two genetically identical daughter cells (Figure 1). Mitosis is a fundamental cellular process comprising four main phases. During prophase, chromatin condenses into distinct chromosomes and the nuclear envelope begins to disintegrate. In metaphase, chromosomes align along the cell's equatorial plane. During anaphase, sister chromatids separate and migrate toward opposite poles of the cell. Finally, in telophase, chromosomes decondense, nuclear membranes re-form around each chromosome set, and cytokinesis completes cell division. As artificial intelligence continues to advance, the importance of large-scale, diverse, and standardized mitosis datasets is expected to increase, with future efforts focusing on multi-center data, multi-stain imaging, and

adherence to fair data principles to enhance transparency and reproducibility in medical AI research. In the following section, a detailed review and comparative analysis of the major publicly available mitosis datasets are presented.

ICPR12 Dataset

As part of the MICO project, a mitosis-detection challenge was organized for ICPR12 using H&E breast cancer slides. Detecting mitoses is difficult because they are small, vary greatly in shape, and can be confused with other structures. To make the task more comprehensive, images were collected from two different slide scanners and a multispectral microscope. Five selected slides produced 50 high-power fields at 40× magnification, and a pathologist manually annotated all mitotic figures, resulting in about 320 mitoses. Out of 129 registered teams, 17 submitted results, and the best achieved a recall of 0.70, a precision of 0.89, and an F-measure of 0.78. Although the results are promising, the dataset is too limited to fully assess the reliability of the proposed algorithms. Since mitotic count is an important factor in breast cancer grading but suffers from poor inter-observer agreement, automated tools could improve consistency and reduce workload. A key goal of this challenge was to provide a dataset to stimulate research in mitosis detection, with future plans to expand the database using more slides, different cancer types, and annotations from multiple pathologists. The ICPR12 dataset consists of 5 breast cancer biopsy slides containing 50 H&E-stained images of 2084 × 2084 pixels selected by pathologists. In the ICPR12 dataset, 35 images containing 226 mitoses are used for training, and 15 images containing 101 mitoses are used for evaluation. All mitotic pixels were annotated by pathologists. This dataset is considered a strong dataset that provides an example of ground truth mitotic cells for scanners in Figure 2 (Ludovic et al., 2013).

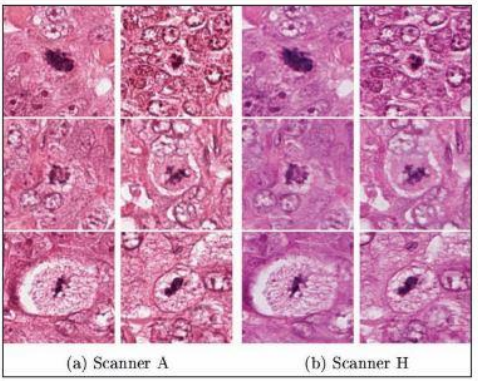


Figure 2: Sample of ground truth mitosis for scanners.
Source: Ludovic et al., 2013

AMIDA13 Dataset

The AMIDA13 challenge was organized as part of the MICCAI Grand Challenge to advance research in automatic mitosis detection in breast cancer histopathology. The authors provided a large, carefully annotated dataset of H&E-stained whole-slide images to evaluate and compare algorithmic performance under standardized conditions. This challenge highlighted the difficulty of detecting mitosis due to their varied appearance and the presence of many visually similar structures. By benchmarking numerous state-of-the-art methods, the study demonstrated both the potential and the current limitations of automated mitosis detection, emphasizing the need for robust algorithms that can generalize across diverse tissue samples. The proliferative activity of breast tumors, typically assessed by counting mitotic figures in H&E sections, is a key prognostic indicator. However, manual mitosis counting is time-consuming, subjective, and often inconsistent across pathologists. With the increasing use of WSI in pathology, automated analysis has emerged as a promising alternative. The AMIDA13 challenge evaluated this potential by providing a dataset of 12 training and 11 testing cases, containing over a thousand mitotic figures annotated by multiple experts. Eleven different algorithms were tested, and their performance compared. Notably, the best-performing method achieved an error rate similar to the variability observed among human pathologists, demonstrating encouraging progress toward

reliable automated mitosis detection. The AMIDA13 initiative played a key role in stimulating further work in computational pathology and establishing a foundation for future challenges in digital histology. The AMIDA13 dataset includes 606 HPF images of 2000×2000 pixels. The test dataset consists of 295 HPFs obtained from 11 subjects. In total, this dataset contains 1083 mitoses — 550 in the training set and 533 in the test set. Since pathologists annotated only the central pixels of the mitoses, this dataset shows limited performance. Separation into high-power fields is shown in Figure 3.

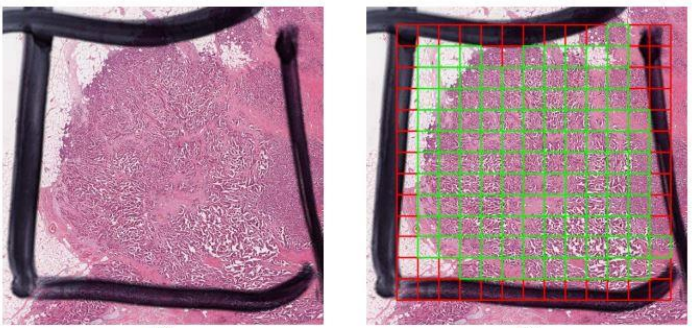


Figure 3: Separation into high-power fields (HPFs). (A) An example slide showing the annotation area marked in black. (B) Each rectangle in the grid represents one HPF. HPFs intersecting the black marker boundaries (shown in red) are excluded from the dataset.

Source: Veta et al., 2015

ICPR14 Dataset

The ICPR14 dataset is a widely used benchmark for mitosis detection in histopathological breast cancer images. It contains a total of 1696 high-resolution tissue images, of which 1200 are labeled and 496 are unlabeled. The labeled portion includes both mitotic and non-mitotic nuclei, enabling researchers to train and evaluate detection models using a balanced set of positive and negative examples. A notable characteristic of the dataset is the way annotations were generated: pathologists provided labels by marking only the central pixel of each mitotic figure. This means the dataset does not include precise boundaries or segmentation masks for

mitoses; instead, it provides approximate locations. As a result, the ICPR14 dataset is often considered to have limited annotation quality, and this constraint poses additional challenges for model training and evaluation. To overcome these limitations, many studies incorporate preprocessing strategies, region-expansion techniques, or advanced data augmentation methods to better capture the mitotic regions. Despite its annotation limitations, the ICPR14 dataset has served as a foundational resource in the field of mitosis detection. It played a significant role in the evolution of early deep-learning-based approaches and continues to be used as a reference dataset due to its historical importance and the challenge it presents for robust algorithm development (Roux et al., 2014). Figure 4 shows two examples of Aperio scanner images from the ICPR14 dataset.

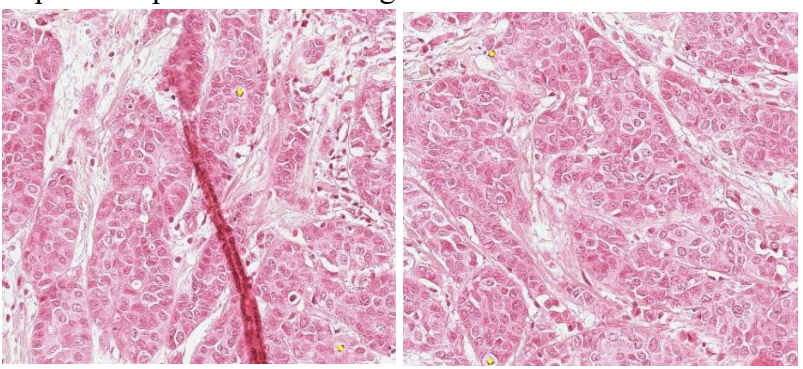


Figure 4: Two samples of the Aperio scanner images of the ICPR14 dataset(mitosis marked with yellow circles).

Source: (Roux et al., 2014)

TUPAC16 Dataset

The TUPAC16 dataset was created using a large collection of 500 H&E WSIs obtained from routine breast cancer diagnostics, with the goal of supporting automated tumor proliferation assessment. To construct the proliferation ground truth, each slide was assigned two complementary types of proliferation indicators: molecular proliferation scores derived from gene-expression profiling, and visual proliferation scores provided by experienced pathologists, including the clinically used mitotic score from the Nottingham grading system. In addition to these slide-level scores, the authors created a dedicated mitosis detection training set by selecting 73

WSIs from the full cohort; for each of these slides, a pathologist chose 1–5 representative tissue regions of 512×512 µm, which were then examined in detail by expert annotators who manually labeled every visible mitotic figure, producing high-quality region-level ground truth for model training. To ensure unbiased and reliable evaluation, the challenge organizers also prepared independent test sets, including a large internal WSI test collection and a separate mitosis detection test set annotated by two independent pathologists, who marked mitoses within predefined fields of view. Together, this multi-level annotation strategy—combining molecular data, expert grading, and meticulously labeled mitotic regions—resulted in a comprehensive dataset designed to benchmark both mitosis detection and tumor proliferation prediction algorithms under realistic clinical conditions (Veta et al., 2014). Figure 5 shows two low-magnification whole-slide image examples, and Figure 6 shows examples from the mitosis detection task.

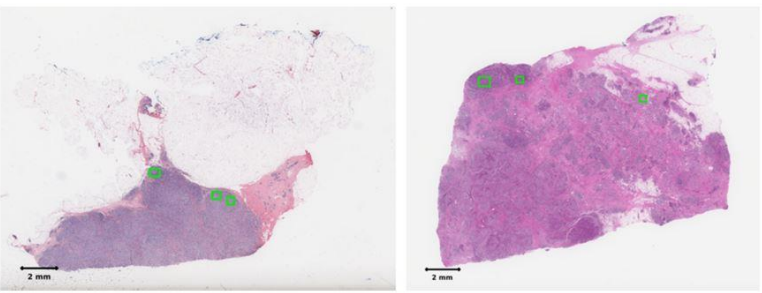


Figure 5: Samples of two low-magnification WSI from the auxiliary region-of-interest (ROI) dataset, each annotated by a pathology resident with three ROIs (green rectangular boxes). These ROIs indicate areas where a pathologist would typically perform mitosis counting.

Source: Veta et al., 2019

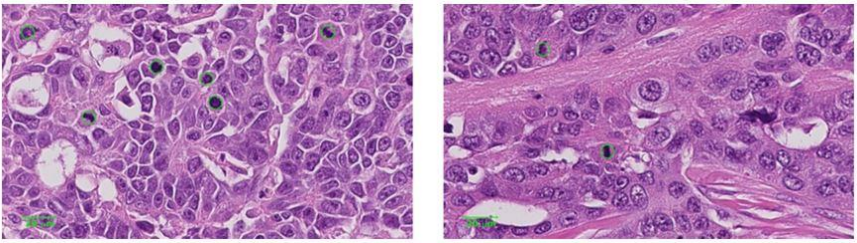


Figure 6: Illustrated are examples from the auxiliary mitosis detection dataset in which mitosis are annotated with green circles, reflecting agreement by a minimum of two pathologists.

Source: Veta et al., 2019

MIDOG21 Dataset

The MIDOG21 dataset, released for the 2021 MIDOG Challenge, consists of a large-scale collection of 280 breast cancer WSIs obtained from UMC Utrecht in the Netherlands for mitosis detection. The goal is to develop robust algorithms to reduce inconsistencies in WSIs. The tissues were scanned at 40× magnification, producing WSIs of approximately 8000×8000 pixels (depending on the scanner). The WSIs are divided into training and test datasets with no overlapping cases. The training dataset consists of human breast cancer tissue samples scanned using four different WSI: Hamamatsu XR NanoZoomer 2.0, Hamamatsu S360 (0.5 NA), Aperio ScanScope CS2, and Leica GT450. For each scanner, WSI from 50 distinct breast cancer cases were acquired. A trained pathologist selected a 2 mm^2 region from each slide—corresponding to roughly 10 high-power fields following the Elston and Ellis grading criteria—and these cropped areas were provided as TIFF files to facilitate processing. Mitotic figures were annotated through a well-established multi-expert blind annotation pipeline designed to capture all mitotic figures, and additional annotations were included for hard examples or non-mitotic look-alikes. Annotations are available only for scanners 1 to 3; scanner 4 serves as additional reference data for methods such as unsupervised domain adaptation. The training set contains 1,721 mitotic figures and 2,714 hard examples. The test set includes images prepared in the same manner but from different tumor cases and incorporates two scanners from the training phase, along with two undisclosed scanners, totaling 80

cases (20 per scanner). A preliminary test set is also provided for self-evaluation within submitted Docker containers, consisting of 5 cases per scanner (20 cases in total) and accessible only for a limited time. Figure 7 shows examples from mitosis detection in the MIDOG21 dataset.

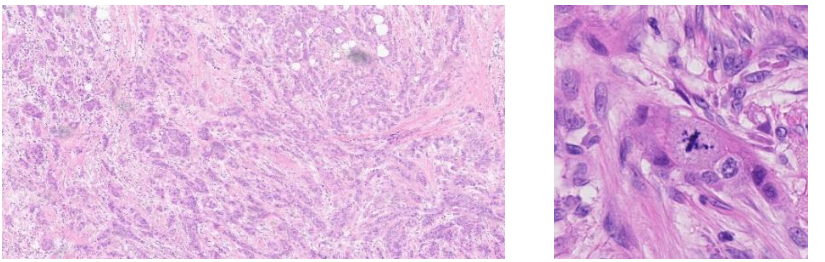


Figure 7: Samples from the mitosis detection MIDOG21 dataset
Source: Aubreville et al., 2021

MIDOG22 Dataset

The MIDOG22 dataset was developed within the scope of the Mitosis Domain Generalization Challenge 2022 to evaluate the robustness of mitotic figure detection algorithms under substantial domain shifts commonly encountered in real-world histopathological imaging. The dataset is specifically designed to study domain generalization across variations in tumor type, species, laboratory processing, and whole-slide imaging scanners. It consists of regions of interest extracted from hematoxylin and eosin–stained whole-slide images, where each region corresponds to an area of 2 mm^2 selected by expert pathologists as the tumor region with the highest expected mitotic activity in accordance with established diagnostic guidelines.

The training portion of the dataset comprises 405 tumor cases, corresponding to 405 individual patients or animals, and includes a total of 9,501 annotated mitotic figures. These cases are distributed across six distinct tumor domains, each defined by a unique combination of tumor type, species, laboratory of origin, and scanning device. The training domains include human breast carcinoma, canine lung carcinoma, canine lymphoma, canine cutaneous mast cell tumor, human pancreatic and gastrointestinal neuroendocrine tumors, and human melanoma. Among these, the

melanoma domain is provided without annotations and serves exclusively as an additional source of data diversity for unsupervised or semi-supervised domain generalization approaches.

To ensure biological and visual diversity, the dataset incorporates samples from multiple species, including human, canine, and feline tumors, and covers a wide range of tumor morphologies. These morphologies are broadly categorized into aggregated cell patterns, round cell morphology, and spindle cell morphology. Such diversity introduces significant variability in cellular appearance and tissue architecture, requiring algorithms to generalize beyond simple appearance-based cues.

The evaluation framework includes two independent test sets designed to rigorously assess generalization performance. A preliminary test set containing cases from four unseen tumor domains is used for technical validation of submitted algorithms, while the final test set consists of 100 cases drawn from ten completely independent tumor domains. The final test set introduces additional domain shifts, including previously unseen species, novel tumor morphologies, and different scanning devices, ensuring that algorithm performance reflects true domain generalization rather than adaptation to familiar data characteristics.

Ground truth annotations for mitotic figures were established using a blinded majority vote of three expert pathologists, supported by a machine-learning–assisted candidate detection process to reduce the likelihood of missed mitotic figures. All annotators had more than five years of experience in mitotic figure identification. In addition to the conventional hematoxylin and eosin–based ground truth, an alternative reference standard was created for the final test set using immunohistochemistry staining for phospho-histone H3. This PHH3-assisted annotation provides a more objective and biologically grounded reference, enabling a more comprehensive evaluation of algorithmic performance.(Aubreville et al., 2024). Figure 8 shows an overview of test set domains with representative examples.

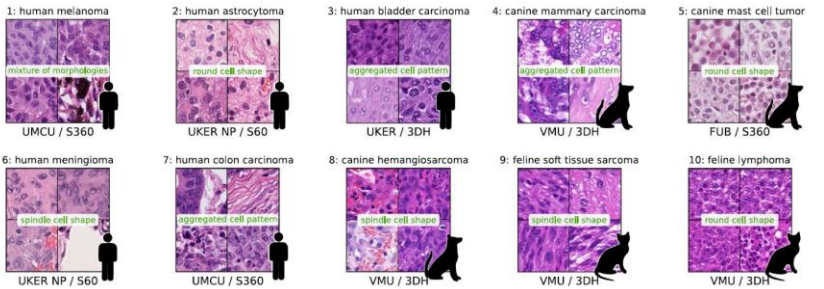


Figure 8: Overview of test set domains with representative 256 × 256 px image patches from randomly selected samples. Captions indicate tissue origin and scanner type. Tumor types are grouped based on tissue morphology into aggregated, round, and spindle cell patterns.

Source: Aubreville et al., 2024

MIDOG++ Dataset

The MIDOG++ (Mitosis Domain Generalization) dataset was used for mitotic figure detection in histopathological images. MIDOG++ is a large-scale, publicly available, multi-domain dataset developed to address the problem of domain shift in deep learning based mitosis detection. It extends the datasets released for the MIDOG21 and MIDOG22 challenges and is designed to support research on robust and generalizable mitotic figure detection across heterogeneous data sources. The dataset comprises 503 histological cases, each represented by a single region of interest (ROI) with a fixed area of 2 mm^2, which approximates the standard diagnostic practice of evaluating 10 high-power fields at 400× magnification in routine pathological assessment. Across all cases, a total of 11,937 mitotic figures are annotated, accompanied by a large number of annotated non-mitotic structures (hard negatives) that closely resemble mitoses in morphology and are therefore essential for training discriminative detection models.

The dataset includes samples from seven different tumor types derived from both human and canine specimens, introducing substantial biological diversity. The human tumor types comprise breast carcinoma, neuroendocrine tumors, and cutaneous melanoma, while the canine tumor types include lung carcinoma,

lymphosarcoma, cutaneous mast cell tumor, and (sub)cutaneous soft tissue sarcoma. These tumor entities exhibit marked differences in cellular architecture, mitotic density, cell size, and tissue organization. For instance, lymphosarcoma samples typically present with very high mitotic activity and smaller cell sizes, whereas neuroendocrine tumors often show sparse mitotic figures. Such variability reflects real-world diagnostic conditions and poses significant challenges for automated mitotic figure detection systems.

All tissue specimens were processed using standard formalin-fixed paraffin-embedded procedures and stained with H&E. WSI were digitized at 40× magnification using five different whole slide scanners from multiple manufacturers, resulting in spatial resolutions ranging between approximately 0.23 and 0.25 μm per pixel. The specimens originated from several pathology laboratories, each employing its own routine protocols for tissue preparation, staining, and scanning. Consequently, the dataset exhibits pronounced variability in color characteristics, contrast, sharpness, and imaging artifacts. Together with differences in scanner hardware, laboratory workflows, species, and tumor types, this diversity introduces multiple sources of domain shift that are known to negatively affect the generalization performance of deep learning models trained on homogeneous datasets.

The annotation process was conducted using a rigorous multi-stage protocol to ensure high label quality. Initially, an experienced pathologist exhaustively screened each ROI and annotated all visible mitotic figures as well as morphologically similar non-mitotic structures, referred to as hard negatives. To further reduce the likelihood of missed mitotic figures, a deep learning–based detection model was trained on preliminary annotations and applied to identify additional candidate regions. All annotated candidates were subsequently reviewed in a multi-expert consensus process involving three expert pathologists. Annotations with agreement between the first two experts were directly accepted, while disagreements were resolved by a third expert. This consensus-driven strategy was employed to mitigate the well-documented inter-observer variability associated with mitotic figure identification and

to improve the reliability and consistency of the ground truth labels. Figure 9 shows mitotic figure candidates from all domains summarized.

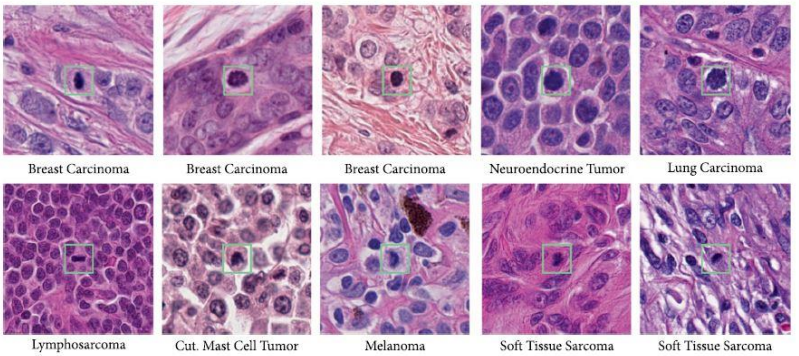


Figure 9: Mitotic figure candidates from all domains summarized

Source: Aubreville et al., 2023

CCMCT Dataset

A large-scale and comprehensive dataset designed for the assessment of mitotic figures in WSIs of canine cutaneous mast cell tumors (CCMCT) consists of 32 high-resolution H&E-stained WSIs, carefully selected to represent a broad spectrum of tumor biology, ranging from low-grade to high-grade cases with varying mitotic activity. Each slide has been fully and exhaustively annotated, making the dataset unique compared to previous resources that typically focus only on limited regions of interest. The primary focus of the annotations is on mitotic figures, which are a critical prognostic indicator in many tumor types, but the dataset also includes detailed labels for non-mitotic neoplastic mast cells, eosinophilic granulocytes, and mitotic figure look-alikes, thereby capturing the morphological complexity encountered in real diagnostic settings.

The annotation process was performed by two expert veterinary pathologists using a blinded, consensus-based workflow to minimize inter-observer variability and improve label reliability. To further enhance completeness and reduce the risk of missing rare mitotic events, algorithm-assisted annotation strategies based on deep

learning were incorporated. This resulted in three complementary dataset variants: a manually expert-labeled dataset (MEL), a hard-example augmented expert-labeled dataset (HEAEL), and an object-detection augmented expert-labeled dataset (ODAEL). In total, the dataset comprises 262,481 cell annotations, of which 44,880 correspond to mitotic figures, making it the largest dataset of its kind in terms of both annotated tumor area and number of mitoses. Figure 10 shows sample image from CCMCT dataset.

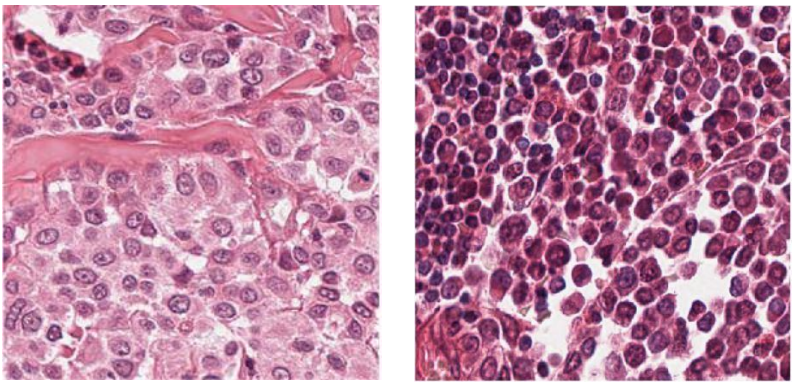


Figure 10: Sample images from CCMCT dataset

Source: Bertram et al., 2019

AMi-Br Dataset

The AMi-Br (Atypical Mitotic Figures in Breast Cancer) dataset represents the first publicly available collection specifically dedicated to distinguishing normal (typical) and atypical mitotic figures (MFs) in histological sections of human breast cancer. This dataset addresses a critical need in computational pathology: while mitotic figure density is a well-established a prognostic indicator of tumor proliferation and aggressiveness in breast cancer and other malignancies, recent studies highlight that the proportion of atypical mitotic figures (AMFs), characterized by abnormal morphologies such as polar asymmetry, multipolar spindles, lagging chromosomes, or bridging chromosomes, may serve as an independent prognostic indicator. These AMFs reflect underlying genetic mutations in cell cycle regulation, leading to chromosomal instability (aneuploidy) and promoting tumor progression.

The dataset was created by combining and re-annotating mitotic figures from two of the most prominent public mitosis detection challenges: the TUPAC16 dataset (with improved alternative labels) and the MIDOG 2021 dataset. These sources cover diverse scanning devices (6 different whole-slide scanners) and pathology centers, ensuring broad domain variability. In total, the AMi-Br dataset includes 3,720 mitotic figures extracted from 223 tumor cases: 832 atypical MFs (0.224) and 2,888 normal MFs (0.776).

Annotation was performed rigorously: three expert pathologists independently classified each 128×128 pixel patch (centered on the original MF location) as normal or atypical in a blinded manner. The final labels were determined by majority vote. Normal MFs follow classic mitotic phases (prometaphase, metaphase, ring-shaped metaphase, ana-/telophase), while atypical ones include categories like polar asymmetry (bipolar or tri-/multipolar) and chromosome segregation errors (e.g., lagging or bridging chromosomes). Additional subclassifications into finer morphological categories are available from two experts for deeper analysis. A main CSV file (AMI-BR.csv) containing metadata, coordinates, individual expert labels, and the majority-vote atypical flag. Cropped 128×128 pixel patches of the annotated MFs in the patches/ folder. Original images and Jupyter notebooks demonstrating baseline classification experiments.

Baseline results using Monte Carlo cross-validation and class-imbalance mitigation strategies yielded average balanced accuracies of up to 0.806 (patch-level split) and 0.713 (patient-level split), confirming the dataset's usability while highlighting challenges such as morphological overlap between classes and severe class imbalance. By enabling research into automated AMF detection and classification, often framed as a two-step pipeline (mitosis detection followed by atypical/normal patch classification) AMi-Br paves the way for more reproducible, efficient, and prognostically valuable computational tools in breast cancer histopathology. It has already inspired follow-up benchmarks and extensions in related works on domain generalization and foundation model adaptation for atypical mitosis classification. Researchers can download the patches directly from the repository and combine them with the original MIDOG

2021 and TUPAC images for full experimentation (Bertram et al., 2025). Figure 11 shows the Mitotic and Atypical Figure Dataset in Breast Cancer.

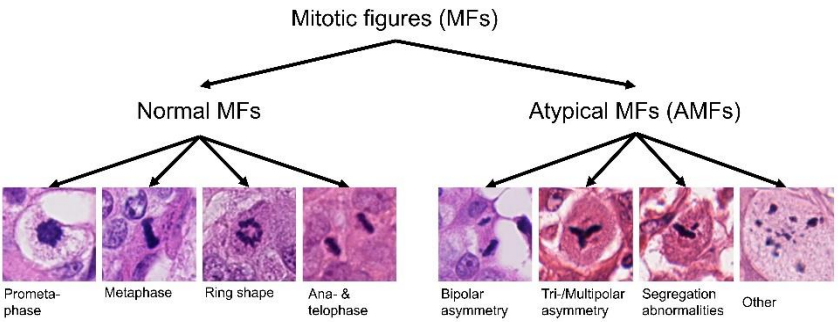


Figure 11: Mitotic and Atypical Figure AMi-Br Dataset

Source: Bertram et al., 2025

MiDeSeC Dataset

This dataset was curated using H&E-stained breast tissue slides from 25 anonymous cases from the Ankara University Faculty of Medicine, Department of Pathology archive. The patients underwent surgery (excisional biopsy, lumpectomy, or mastectomy) with a diagnosis of invasive breast carcinoma. The selected glass slides were scanned using a 3D HISTECH Panoramic Scanner P250 flash 3 to obtain WSIs. A pathologist annotated the tumor areas on each slide, excluding stromal components and benign breast tissue from this study. Only the most representative invasive tumor areas with good fixation, high cellularity, and no artifacts or necrosis were selected. Typical and atypical mitotic figures within these tumor areas were identified and marked. In the next stage, 1024×1024 pixel regions were cropped from these selected tumor fields and extracted in RGB 8 bit format from the TIFF images. The QuPath program was used to annotate these extracted patches, and the pathologist manually marked typical and atypical mitotic figures. Finally, the annotated images and their corresponding coordinate files (csv) were prepared for use in deep learning algorithms. Two thirds of the images were reserved for training, with the remaining third for testing. Test and training sets were randomly selected. Figure 12 shows samples of MiDeSeC dataset images (Samet et al., 2025).

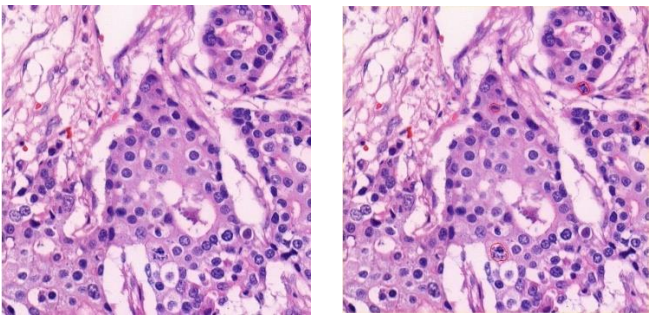


Figure 12: The Sample Images of MiDeSeC Dataset

Source: Samet et al., 2025

General Assessment

The field of mitosis detection in computational pathology has undergone significant evolution from 2012 to the present. Early datasets were typically small-scale, focused exclusively on human breast cancer, and limited to a few hundred mitotic figures across a modest number of high-power fields (HPFs). Annotations in these initial collections were often either complete cell delineations or, more commonly, only the central pixel of each mitosis, resulting in relatively restricted quality and limited capacity for training robust models. While these early efforts produced promising results, the generalization ability of algorithms developed on such datasets remained severely constrained when applied to real-world clinical scenarios.

Over time, datasets have grown dramatically in both scale and diversity. Modern collections now contain thousands of annotated mitotic figures, incorporate images acquired from multiple whole-slide scanners, and reflect realistic variations in staining protocols, laboratory procedures, color profiles, contrast, and imaging artifacts—collectively known as domain shift. Recent datasets have expanded far beyond human breast cancer to include samples from different species (such as canine), various organ sites (lung, lymphoma, mast cell tumors, neuroendocrine tumors, melanoma, etc.), and a wide range of tumor morphologies (aggregated cell patterns, round cell, spindle cell). This biological and visual heterogeneity closely mirrors the complexity encountered in actual diagnostic practice. Annotation quality has also improved

substantially. Early single-pathologist or single-center labeling has largely been replaced by rigorous multi-expert protocols involving blinded majority voting among several experienced pathologists, frequently supported by machine-learning-assisted candidate detection to minimize missed events. Many contemporary datasets also include extensive labeling of morphologically similar non-mitotic structures (so-called “hard negatives”), which are essential for training discriminative models. Some provide exhaustive annotations across entire whole-slide images, while others focus on standardized 2 mm^2 tumor regions with the highest expected mitotic activity, consistent with routine diagnostic guidelines.

Furthermore, the most recent advancements have introduced datasets specifically dedicated to the detection and classification of atypical (abnormal morphology) mitotic figures. These atypical mitoses, characterized by features such as multipolar spindles, lagging chromosomes, or bridging chromosomes, are increasingly recognized as independent prognostic markers reflecting chromosomal instability and tumor aggressiveness.

In summary, the mitosis detection field has moved decisively away from small, homogeneous, and annotation-limited datasets toward large-scale, highly heterogeneous, multi-domain collections with superior annotation quality and challenging negative examples. Today, the performance of state-of-the-art models is evaluated primarily based on their robustness across diverse scanners, species, tumor types, laboratories, and morphological variations. The field continues to advance rapidly, not only by addressing technical generalization challenges but also by enabling the discovery and automated assessment of new prognostic indicators, such as the proportion of atypical mitotic figures, with substantial potential to enhance reproducibility and clinical value in digital histopathology. Each dataset is compared based on its content, type, number of mitoses, and difficulty level, as presented in Table 1.

Table 1: Mitosis datasets in the current literature

Data Set	Year	Species	Sample	Mitos	Content	Features
ICPR12	2012	Human breast	50 images	327	Small	Beginner level
AMIDA13	2013	Human breast	23 HPF	1083	H&E	Medium resolution
ICPR14	2014	Human breast	1.696 images	-	Mitosis + Atypical	Atypical distinction is challenging
TUPAC16	2016	Human breast	500 WSI	1552	Includes spread score	Score prediction can be made
MIDOG21	2021	Human breast	280 WSI	-	From 4 different scanners	Domain generalization
MIDOG22	2022	Human, dog, cat	520 case	-	6 tumor types, 5 scanners	Multiple species and variations
MIDOG++	2023	Multi-species	503 sample	11937	Multi-area labeled ROIs	Quality test dataset
CCMCT	2022	Dog mast	21	13907	Dog skin texture	WSI
AMi-Br	2025	Human breast	-	3720	832 atypical, 2,888 typical mitoses	Atypical distinction can be made
MiDeSeC	2025	Human breast	50 image 1024× 1024 pixels	500	H&E	Comprehensive

Source: Samet et al, 2025

The datasets have evolved significantly from 2012 to 2025 in terms of both scale and diversity; they have progressed from small and introductory datasets such as ICPR12 to advanced datasets like

MIDOG++ and CCMCT, which include multiple species, multiple scanners, and thousands of annotations. Although human breast tissue data remain predominant, MIDOG22 and MIDOG++ introduce multi-species content, providing opportunities for generalization studies. The distinction of atypical mitoses is available only in certain datasets and stands out as a critical feature for clinical accuracy and model performance. Moreover, multi-scanner and multi-domain annotations enable modern deep learning models to learn variations from different data sources, while smaller datasets still serve well for methodological development and initial experimentation. Because MiDeSeC focuses specifically on breast cancer and provides detailed annotations for detection, segmentation, and classification, it fills a gap in existing medical image resources and helps develop more robust AI models for cancer grading and diagnostic support.

REVIEW OF AI MODELS

Mitosis analysis in computational pathology comprises three closely related yet methodologically distinct tasks, namely detection, segmentation, and classification. Deep learning based approaches developed for these tasks are largely shaped by the choice of model architecture and processing pipeline.

Detection

In particular, the small size, rarity, and high morphological variability of mitotic figures make the detection stage especially challenging and critical. Within this context, mitosis detection methods are commonly categorized into single-stage and two-stage object detection models. Single-stage models aim to directly predict the location and class of mitotic figures using a single network in an end-to-end manner. Representative examples include YOLO based architectures, RetinaNet, and SSD, which perform detection in a single forward pass. Owing to their high inference speed and computational efficiency, these models are often favored in real-time applications and studies focusing on clinical deployment.

YOLO Based Architectures: YOLO (You Only Look Once) architectures constitute a class of single-stage object detection

models that perform object localization and classification in a unified, end-to-end framework. Unlike two-stage detectors, YOLO models process the entire image in a single forward pass, directly predicting bounding boxes and class probabilities from global feature maps. This design enables high inference speed and computational efficiency, making YOLO-based approaches particularly attractive for large-scale histopathological image analysis and real-time clinical applications. Over successive versions, from early implementations to more recent variants such as YOLOv5, YOLOv7, and YOLOv8, substantial improvements have been introduced in backbone networks, feature pyramid structures, loss functions, and training strategies. These enhancements have significantly improved the detection of small and densely distributed objects, which is critical for mitosis detection.

In computational pathology, YOLO based models have been increasingly adopted for mitosis detection due to their ability to efficiently scan whole-slide images and identify candidate mitotic figures with low latency. However, the small size and high morphological variability of mitotic figures present challenges for standard YOLO configurations. To address these limitations, recent studies have incorporated multi-scale feature extraction, anchor-free detection heads, attention mechanisms, and customized loss functions to improve sensitivity to small mitotic objects while controlling false positives. Despite generally offering slightly lower localization precision compared to two-stage detectors, YOLO-based architectures provide a favorable trade-off between accuracy and speed, making them well suited as front-end detectors in multi-stage or hybrid mitosis analysis pipelines. Figur 13 shows generic architecture of single stage object detectors.

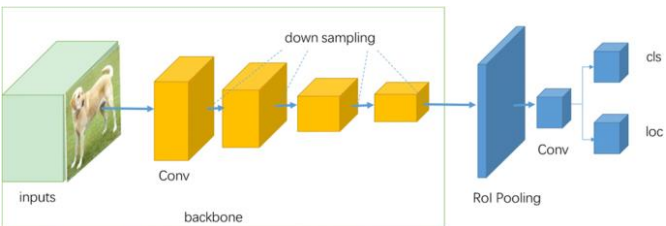


Figure 13: Generic architecture of single stage object detectors

Source: Diwan et al., 2023

RetinaNet Architecture: RetinaNet is a single-stage object detection architecture designed to address the performance gap between one-stage and two-stage detectors, particularly in scenarios characterized by extreme class imbalance. The core contribution of RetinaNet is the introduction of the focal loss function, which down-weights easy negative samples and focuses training on hard, informative examples. This property is especially relevant for mitosis detection, where mitotic figures constitute a very small fraction of the overall image content.

The RetinaNet architecture is composed of three primary components: a backbone network, a feature pyramid network (FPN), and task-specific subnetworks. The backbone is typically a deep convolutional neural network such as ResNet, which extracts hierarchical feature representations from the input image. On top of the backbone, an FPN is constructed to generate multi-scale feature maps, enabling effective detection of objects at different sizes.

RetinaNet employs two lightweight, fully convolutional subnetworks that operate on each level of the feature pyramid. The classification subnetwork predicts the probability of object presence for each anchor, while the regression subnetwork estimates bounding box offsets relative to predefined anchor boxes. The use of anchors at multiple scales and aspect ratios allows RetinaNet to robustly localize objects with varying shapes and sizes. During training, the focal loss is applied to the classification branch, effectively mitigating the dominance of background samples and improving sensitivity to rare positive instances. Figure 14 shows RetinaNet architecture.

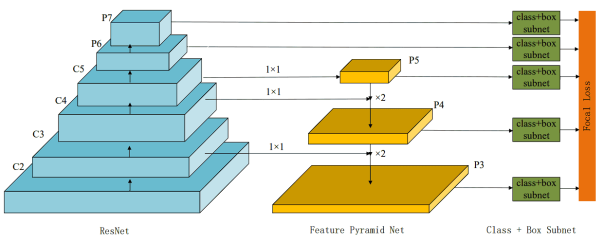


Figure 14: RetinaNet Architecture

Source: Tian et al., 2020

SSD Architecture: The Single Shot MultiBox Detector (SSD) is a single-stage object detection architecture designed to perform object localization and classification in a single forward pass through a deep convolutional neural network. By eliminating the region proposal stage used in two-stage detectors, SSD achieves high inference speed while maintaining competitive detection accuracy. This efficiency makes SSD particularly suitable for large-scale image analysis and applications requiring fast processing, such as whole-slide image scanning in computational pathology.

The SSD architecture consists of a base convolutional network, typically a pre-trained classification model such as VGG-16, ResNet, or MobileNet, which serves as the backbone for feature extraction. On top of this backbone, SSD introduces a set of additional convolutional layers that progressively decrease in spatial resolution. These layers enable the network to generate multi-scale feature maps, allowing objects of different sizes to be detected at multiple resolutions. This design is especially important for detecting small structures, such as mitotic figures, within high-resolution histopathological images. SSD predicts a fixed set of default bounding boxes (also referred to as anchor boxes) with different scales and aspect ratios at each location of the feature maps. For each default box, the network simultaneously outputs class confidence scores and bounding box offsets. The predictions from all feature maps are then combined, and non-maximum suppression is applied to remove redundant detections. By distributing detection tasks across multiple feature layers, SSD effectively balances the detection of small and large objects within a unified framework. Figure 15 shows SSD architcture.

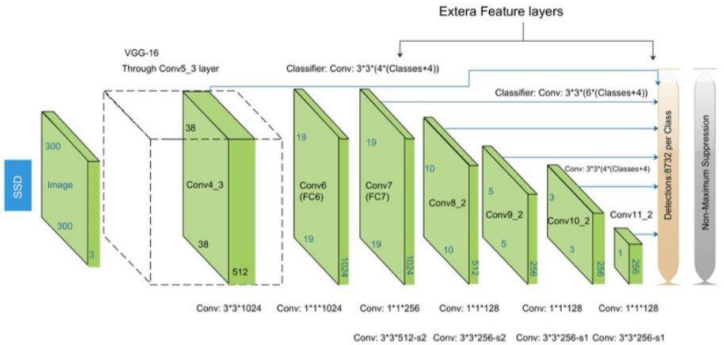


Figure 15: SSD architecture

Source: Bahaghighat et al., 2020

In contrast, two-stage models decompose the detection process into sequential steps, where candidate regions are first proposed and subsequently refined and classified. Faster R-CNN is a canonical example of this paradigm, offering high detection accuracy for small objects. Mask R-CNN further extends this framework by jointly performing detection and pixel-level segmentation, making it particularly attractive for mitosis analysis tasks that benefit from both localization and precise morphological delineation. While two-stage models generally provide higher detection accuracy for small and complex structures such as mitotic figures, single-stage models offer advantages in terms of speed and computational efficiency.

Mask R-CNN: Mask R-CNN is a two-stage object detection and instance segmentation architecture that extends Faster R-CNN by incorporating an additional branch for pixel-level mask prediction. This design enables Mask R-CNN to simultaneously perform object detection, classification, and segmentation, making it particularly suitable for tasks that require both accurate localization and precise morphological delineation, such as mitosis analysis in histopathological images.

The architecture of Mask R-CNN consists of a backbone network, a region proposal network (RPN), and multiple task-specific heads. The backbone, typically a deep convolutional neural network such as ResNet or ResNeXt, is used in conjunction with a Feature Pyramid Network (FPN) to extract multi-scale feature representations from the input image. The FPN enhances the

detection of objects at different scales, which is essential for identifying small structures like mitotic figures.

In the first stage, the RPN generates candidate object regions by predicting objectness scores and bounding box proposals over the feature maps. These region proposals are then refined and aligned with the feature maps using the RoIAlign operation, which replaces the quantization-prone RoIPooling used in Faster R-CNN. RoI align preserves exact spatial correspondence between the input features and the proposed regions, significantly improving localization accuracy and segmentation quality.

In the second stage, the aligned region features are fed into parallel task-specific branches. The classification head predicts the object class, the bounding box regression head refines the object location, and the mask head, implemented as a small fully convolutional network, produces a binary segmentation mask for each detected instance. Importantly, the mask prediction is performed independently of classification, allowing Mask R-CNN to achieve high-quality instance segmentation without compromising detection performance (Figure 16).

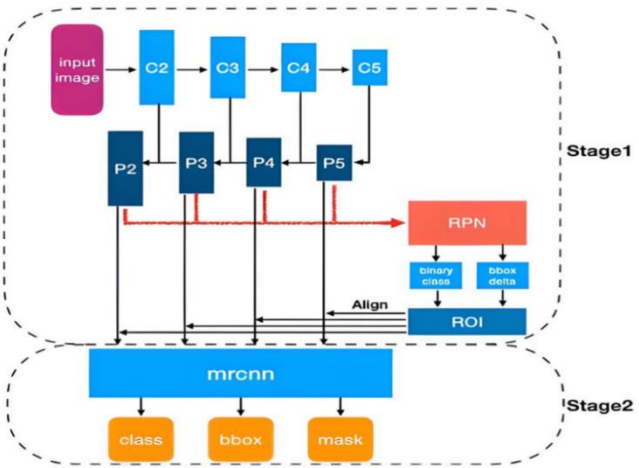


Figure 16: Mask R-CNN architecture

Source: He et al., 2017

Faster R-CNN Architecture: Faster R-CNN is a two-stage object detection architecture that significantly improves both detection accuracy and computational efficiency compared to R-CNN based methods. By integrating region proposal generation directly into the network, Faster R-CNN enables end-to-end training and has become a foundational model for high-precision object detection tasks, including mitosis detection in histopathological images. The architecture consists of three principal components: a backbone network, a Region Proposal Network (RPN), and a detection head. The backbone, commonly implemented using deep convolutional neural networks such as ResNet or VGG, extracts hierarchical feature representations from the input image. These feature maps serve as shared inputs for both region proposal generation and object detection, reducing redundant computations and improving efficiency. The RPN is a fully convolutional network that slides over the backbone feature maps to generate a set of candidate object regions. At each spatial location, the RPN predicts objectness scores and bounding box coordinates relative to a predefined set of anchor boxes with multiple scales and aspect ratios. This design allows Faster R-CNN to effectively propose regions corresponding to objects of varying sizes, which is particularly important for detecting small structures such as mitotic figures.

In the second stage, the region proposals produced by the RPN are refined and classified. Each proposed region is mapped onto the feature maps using a region pooling operation, typically RoIPooling or RoIAlign, to produce fixed-size feature representations. These features are then processed by fully connected layers that perform object classification and bounding box regression, yielding the final detection results. The use of a two-stage pipeline enables precise localization and classification, making Faster R-CNN especially effective for complex detection tasks (Figure 17).

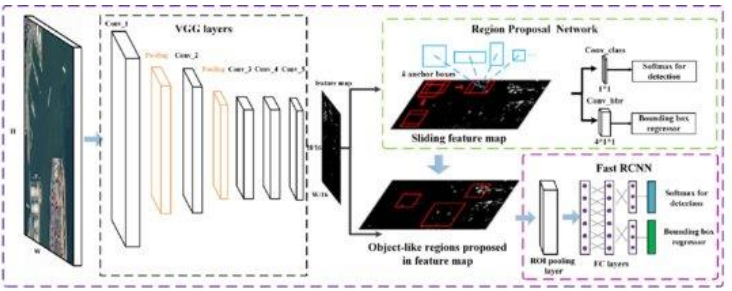


Figure 17: The architecture of Faster R-CNN

Source: Ren et al., 2016

Segmentation

Mitosis segmentation focuses on delineating mitotic figures at the pixel level, which enhances biological interpretability and supports downstream analysis. Among segmentation approaches, U-Net has emerged as the most widely adopted architecture, owing to its effectiveness even with limited annotated data. Variants such as Attention U-Net and U-Net++ have been proposed to better capture small mitotic regions and suppress background noise. In addition, models such as DeepLabv3 and DeepLabv3+ leverage multi-scale contextual information and have demonstrated strong performance in complex histopathological environments. Two-stage detection frameworks like Mask R-CNN are also frequently employed for mitosis segmentation, as they integrate region-based detection with instance-level mask prediction.

U-Net Architecture: The U-Net architecture is a deep learning model based on convolutional neural networks, originally developed for biomedical image segmentation. Its U-shaped structure consists of an encoder and a decoder (Figure 18). In the encoder part, successive convolution and pooling layers reduce the spatial dimensions of the image while extracting high-level semantic features. The decoder part then restores the spatial resolution through upsampling operations to produce a detailed segmentation map. A key feature of U-Net is the use of skip connections between corresponding encoder and decoder layers, which help preserve spatial information and improve localization accuracy at the pixel level (Ronneberger et al., 2015).

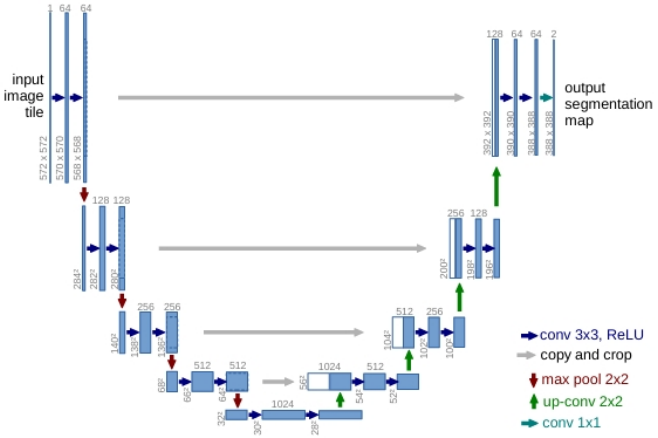


Figure 18: U-Net architecture. Blue boxes show feature maps, white boxes copied maps, arrows indicate operations.

Source: Ronneberger et al., 2015

U-Net++ Architecture: U-Net++ is an advanced variant of the U-Net architecture designed to improve segmentation accuracy by reducing the semantic gap between the encoder and decoder feature maps. Unlike the original U-Net, U-Net++ introduces nested and dense skip connections, which connect encoder and decoder sub-networks through a series of intermediate convolution layers. These redesigned skip pathways allow feature maps at different depths to be more semantically aligned before fusion, leading to better feature representation. U-Net++ also supports deep supervision, where segmentation outputs are generated at multiple decoder depths during training, improving gradient flow and convergence. Owing to these enhancements, U-Net++ achieves higher accuracy and robustness, especially in complex medical image segmentation tasks where precise boundary delineation is critical (Zhou et al., 2018). Figure19 shows (a) UNet++ consists of an encoder and decoder.

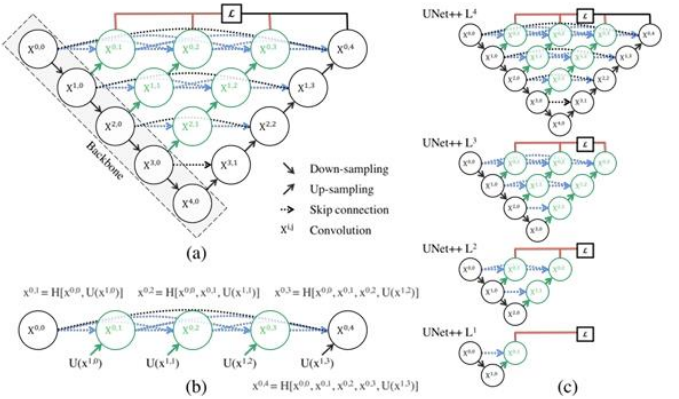


Figure 19: (a) UNet++ consists of an encoder and decoder that are connected through a series of nested dense convolutional blocks. (b) Detailed analysis of the first skip pathway of UNet++. (c) UNet++ can be pruned at inference time, if trained with deep supervision.

Source: Zhou et al., 2018

Classification

Mitosis classification is typically performed on candidate regions obtained from detection or segmentation stages and aims to distinguish mitotic from non-mitotic cells or to identify mitotic subphases. Convolutional neural network–based architectures, including ResNet and EfficientNet, are the most commonly used models for this task. ResNet benefits from residual connections that facilitate the training of deep networks and is widely used for both binary and multi-class mitosis classification. EfficientNet, on the other hand, provides an effective balance between accuracy and computational cost. More recently, transformer-based models such as Vision Transformer and Swin Transformer have been incorporated into mitosis classification pipelines, often in combination with CNN backbones, to further enhance representation learning and classification performance.

ResNet Architecture: The ResNet (Residual Network) architecture is a deep convolutional neural network introduced to solve the degradation problem that arises as network depth increases. Its key

innovation is the use of residual blocks, which include shortcut (skip) connections that allow the input of a block to be added directly to its output. Instead of learning a direct mapping, the network learns a residual function, making optimization easier and improving gradient flow during backpropagation. ResNet architectures are composed of stacked residual blocks with convolution, batch normalization, and activation layers, and they can be built at great depths, such as ResNet-50, ResNet-101, and ResNet-152. Due to their stability and strong performance, ResNet models are widely used in image classification, object detection, and many other computers vision applications (He et al., 2016). Figure 20 shows a building block.

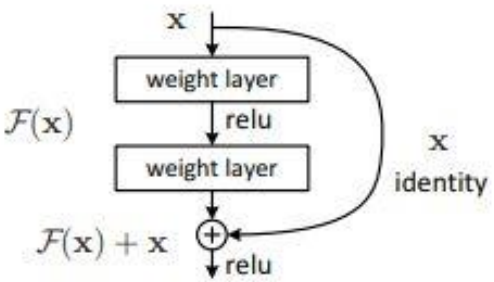


Figure 20: Residual learning: a building block

Source: He et al., 2016

EfficientNet Architecture*:* The EfficientNet architecture is a family of convolutional neural networks designed to achieve high accuracy with significantly fewer parameters and computational cost. Its main contribution is the compound scaling method, which uniformly scales network depth, width, and input resolution using a fixed set of scaling coefficients, rather than scaling these dimensions independently. EfficientNet models are built on a mobile-friendly baseline network (EfficientNet-B0) that uses MBConv (Mobile Inverted Bottleneck Convolution) blocks with depthwise separable convolutions and squeeze-and-excitation modules to enhance feature representation. By balancing efficiency and performance, EfficientNet architectures deliver state-of-the-art results on image classification tasks and are widely adopted in applications where

computational resources are limited (Tan & Le., 2019). Figure 21 shows EfficientNet architecture.

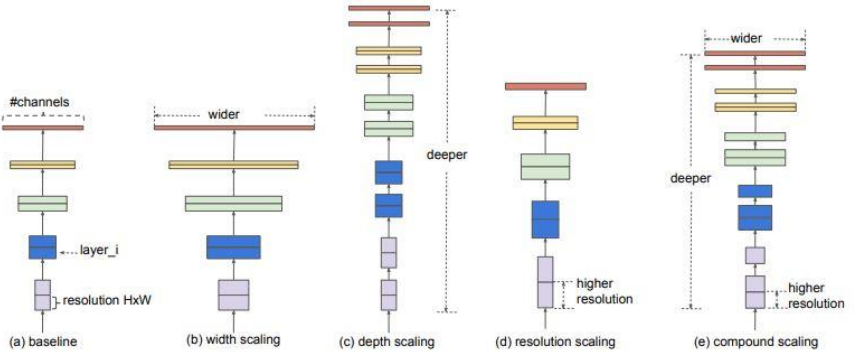


Figure 21: Model Scaling. (a) is a baseline network example; (b)-(d) are conventional scaling that only increases one dimension of network width, depth, or resolution. (e) is our proposed compound scaling method that uniformly scales all three dimensions with a fixed ratio.

Source: Tan & Le., 2019

Vision Transformer Architecture: The Vision Transformer (ViT) architecture applies the transformer model, originally developed for natural language processing, to image recognition tasks by treating an image as a sequence of patches. An input image is divided into fixed-size patches, which are flattened and linearly projected into embedding vectors. Positional embeddings are added to retain spatial information, and the resulting sequence is processed by a stack of transformer encoder layers consisting of multi-head self-attention and feed-forward networks. Instead of convolutional operations, ViT relies on global self-attention to capture long-range dependencies across the image. A special classification token or pooled representation is used for final prediction. Vision Transformers have demonstrated strong performance on large-scale image classification tasks, particularly when trained on large datasets or combined with pretraining strategies (Dosovitskiy et al., 2020). Figure 22 shows Vision Transformer architecture.

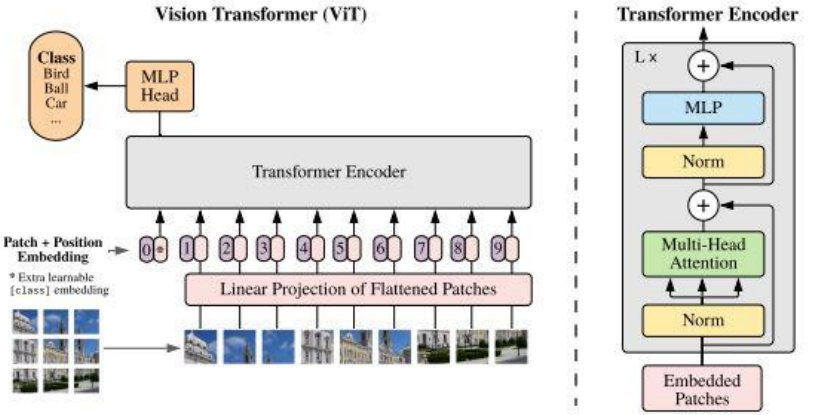


Figure 22: Model overview. The image is split into patches, embedded, combined with positional embeddings, and processed by a Transformer encoder with a classification token.

Source: Dosovitskiy et al., 2020

Swin Transformer Architecture: The Swin Transformer architecture is a hierarchical vision transformer designed to efficiently model visual data by computing self-attention within shifted local windows rather than across the entire image. The input image is first divided into non-overlapping patches, which are embedded and processed through multiple stages with increasing feature dimensions and decreasing spatial resolution. Within each stage, self-attention is computed inside fixed-size windows, and the window positions are periodically shifted to enable cross-window information exchange. This design significantly reduces computational complexity while maintaining strong modeling capacity. As a result, Swin Transformer achieves high performance and scalability across various vision tasks such as image classification, object detection, and semantic segmentation. Figure 23 shows Swin Transformer (Swin-T) architecture.

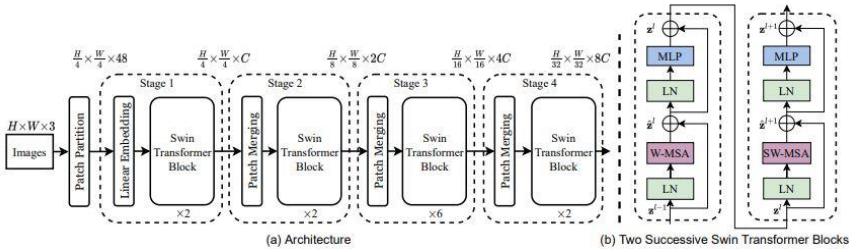


Figure 23: (a) Swin Transformer (Swin-T) architecture; (b) two consecutive Swin Transformer blocks. W-MSA and SW-MSA denote regular and shifted window-based multi-head self-attention, respectively.

Source: Liu et al., 2021

General Assessment

The field of automated mitosis detection, segmentation, and classification in histopathological images has advanced significantly over the past decade, driven by deep learning architectures and increasingly diverse, multi-domain benchmark datasets. Early methods performed reasonably on small, homogeneous datasets (e.g., ICPR12, AMIDA13) but struggled with generalization due to scanner differences, staining variations, and tumor morphology changes. Today, single-stage detectors — particularly recent YOLO versions (YOLOv5 to YOLOv8) — are widely adopted for their strong speed–accuracy balance and ability to process large whole-slide images efficiently. Two-stage models (Faster R-CNN, Mask R-CNN) continue to provide superior localization precision and remain preferred when precise bounding boxes or instance segmentation are required. For segmentation, the U-Net family (including U-Net++, Attention U-Net, and transformer-enhanced variants) dominates due to its effectiveness even with limited annotated data. In classification tasks (mitosis vs. non-mitosis, typical vs. atypical, or sub-phase identification), models such as EfficientNet, ConvNeXt, and vision transformers (Swin Transformer, ViT) are increasingly common, especially when leveraging large-scale pretraining.

State-of-the-art performance is typically achieved by combining strong domain generalization techniques (stain normalization, adversarial training, contrastive learning), ensemble strategies, hard-

negative mining, multi-task learning, and semi-supervised approaches. Despite these advances, key challenges persist: severe class imbalance, high morphological variability, hard-negative confusion, and domain shifts across scanners, laboratories, species, and tumor types.

REVIEW OF EVALUATION METRICS

In computational pathology, mitosis analysis plays a crucial role in assessing tumor proliferation and grading. Automated systems for mitosis detection, segmentation, and classification have become increasingly important for improving diagnostic accuracy and reproducibility. To evaluate such systems, a variety of quantitative performance metrics are used. These metrics allow researchers to objectively compare algorithmic results against ground-truth annotations, measuring both detection accuracy and morphological consistency.

The following tables summarize the most commonly used evaluation metrics in the literature for each stage of mitosis analysis. Table 2 shows evaluation metrics for mitosis detection, Table 3 shows evaluation metrics for mitosis segmentation, and Table 4 represents the evaluation metrics for mitosis classification. Each metric is described in terms of its test category, formula, interpretation, and unit of measurement.

Table 2. Evaluation Metrics for Mitosis Detection

Category	Metric	Formula and Explanation	Unit
Detection	Precision	Precision=$\frac{TP}{TP+FP}$ → Measures how many of the detected mitoses are true mitoses.	[0–1] %
	Recall	Recall=$\frac{TP}{TP+FN}$ → Measures how many of the actual mitoses were correctly detected.	[0–1] %

	F1-Score	$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \rightarrow$ Harmonic mean of Precision and Recall.	[0–1] %
	Average Precision (AP)	Area under the Precision–Recall curve, often computed with IoU or distance thresholds.	[0–1] %
	Average Precision (AP)	Average of AP values over multiple classes or thresholds.	[0–1] %
	Localization Accuracy	Measures how close predicted mitosis locations are to ground truth (e.g., IoU > 0.5 or within 8 µm).	µm or [0–1]

Table 3: Evaluation Metrics for Mitosis Segmentation

Category	Metric	Formula and Explanation	Unit
Segmentation	Dice Coefficient	$Dice = \frac{2TP}{2TP+FP+FN} \rightarrow$ Quantifies overlap between predicted and ground-truth regions.	[0–1] %
	Intersection over Union (IoU)	$IoU = \frac{TP}{TP+FP+FN} \rightarrow$ Ratio of intersection to union between prediction and ground truth.	[0–1] %
	Pixel Accuracy	$Acc = \frac{TP+TN}{TP+TN+FP+FN} \rightarrow$ Percentage of correctly classified pixels.	[0–1] %
	Hausdorff Distance (HD)	HD(A,B)=max(h(A,B),h(B,A)) → Measures the maximum boundary distance between prediction and ground truth (lower is better).	µm or pixels
	Boundary F1 (BFScore)	F1-like score computed only on boundary pixels; evaluates boundary accuracy.	[0–1] %

Table 4: Evaluation Metrics for Mitosis Classification

Category	Metric	Formula and Explanation	Unit
Classification	Accuracy	$Acc = \frac{TP+TN}{TP+TN+FP+FN} \rightarrow$ Overall proportion of correctly classified samples.	[0–1] %
	Precision	Precision= $\frac{TP}{TP+FP} \rightarrow$ Fraction of correctly identified positive samples among all predicted	[0–1] %
	Recall	Recall=$\frac{TP}{TP+FN} \rightarrow$ Fraction of correctly identified positive samples among all predicted	[0–1] %
	F1-Score	$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}} \rightarrow$ Fraction of true positives detected among all actual positives.	[0–1] %
	ROC-AUC	Area under the ROC curve; represents overall discriminative ability.	[0–1] %
	Cohen's Kappa	$\kappa = \frac{p_0 - p_e}{1 - p_e} \rightarrow$ Corrected measure of agreement beyond random chance.	[0–1] %

General Assessment

Mitosis image datasets represent one of the most essential components enabling the advancement of artificial intelligence–based analysis in computational pathology. As discussed in the previous subsections, these datasets play a pivotal role not only in model training and performance evaluation but also in establishing standards and ensuring comparability among different research studies. With the widespread adoption of digital pathology, datasets such as ICPR, AMIDA, TUPAC, and MIDOG have significantly enhanced the generalization capability of algorithms by incorporating variations in resolution, staining methods, and species

diversity. Studies utilizing these datasets have provided a solid benchmark for developing deep learning–based models. In particular, experiments conducted with architectures such as YOLO, Faster R-CNN, UNet, and Mask R-CNN demonstrate that model performance is strongly dependent on dataset quality, scale, and annotation accuracy. Evaluation metrics, including precision, recall, F1-score, IoU, and Dice coefficients, have enabled objective and reproducible measurement of detection, segmentation, and classification performance. The standardization of these metrics has facilitated fair comparisons between studies and promoted transparency in algorithmic evaluation.

APPLICATIONS

In the previous subsections, datasets reported in the literature related to mitosis, machine and deep learning models, and evaluation metrics were defined. In this subsection, applications related to mitosis that have been conducted for cancer analysis in the literature will be discussed and evaluated.

Accurate detection of mitotic figures in whole-slide histopathological images remains challenging due to their rarity, morphological variability, and differences in tissue preparation and staining. The MIDOG competition series has established standardized benchmarks that foster the development of generalizable deep learning models for mitosis detection. Recent works have evaluated YOLOv5 and YOLOv8 one-stage detectors on datasets such as MIDOG++, CMC, and CCMCT. Training strategies incorporating stain-invariant color perturbations and texture-preserving augmentations have proven effective in enhancing model robustness. While YOLOv5 tends to achieve higher precision and YOLOv8 better recall, ensemble approaches combining both architectures have improved sensitivity without notable precision loss. These findings underscore the potential of modern ensemble-based object detection frameworks to advance automated mitosis detection in digital pathology (Kelam et al. 2025).

Mitosis detection is essential for breast cancer grading, and the limitations of manual counting highlight the need for automated methods. To address this, Wang et al. (2025) proposed a two-stage

cascaded network named FoCasNet that was for mitosis detection. The first stage detects potential candidates, while the second stage refines these results through classification. Incorporating attention mechanisms, normalization techniques, and a hybrid anchor-branch classification subnet, the method enhances detection accuracy. FoCasNet achieved the highest reported F1-score (0.888) on the ICPR12 dataset and 0.563 on the newly released GZMH dataset, outperforming existing state-of-the-art approaches. These results demonstrate the effectiveness and generalizability of the proposed framework for automated mitosis detection. Figure 24 shows the overall architecture of the FoCasNet.

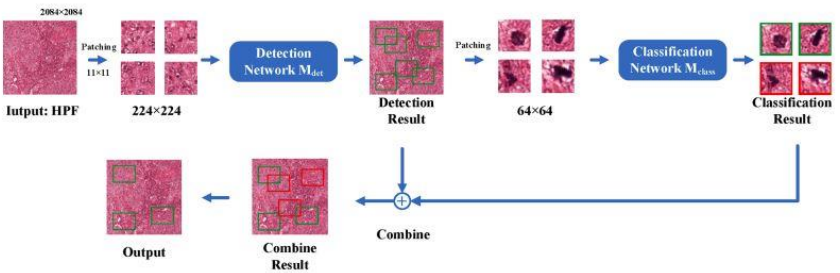


Figure 24: The overall architecture of the FoCasNet

Source: Wang et al., 2024

Manual mitosis counting is time-consuming and subjective, motivating AI-based detection, though domain shifts and class imbalance remain major challenges. To address these challenges, mitosis detection was formulated as pixel-level segmentation using a UNet-based teacher-student model with domain generalization modules, including contrastive representation learning and domain-adversarial training. Pseudo-masks are generated for annotated mitoses, hard negatives, and normal nuclei, improving feature discrimination and robustness. For atypical mitosis classification, a multi-scale CNN classifier leverages segmentation features within a multi-task learning framework. This approach was evaluated on MIDOG++, TUPAC16, AMi-Br, MIDOG25, and Octopath datasets, achieving an F1 score of 0.766 for detection and a balanced accuracy of 0.841 for classification, demonstrating the effectiveness of integrating segmentation and classification in a unified framework for robust mitosis analysis (Choe et al. 2025). Figure 25 illustrates

the overall workflow of the proposed semi-supervised framework. The model is built on a UNet backbone enhanced with attention, contrastive learning, and domain-adversarial components to improve domain generalization. A frozen teacher network generates online pseudo-masks that guide the student decoder through a semi-supervised loss. Additionally, a multi-scale classification head analyzes the encoder feature maps to perform atypical mitotic figure classification (Track 2 only).

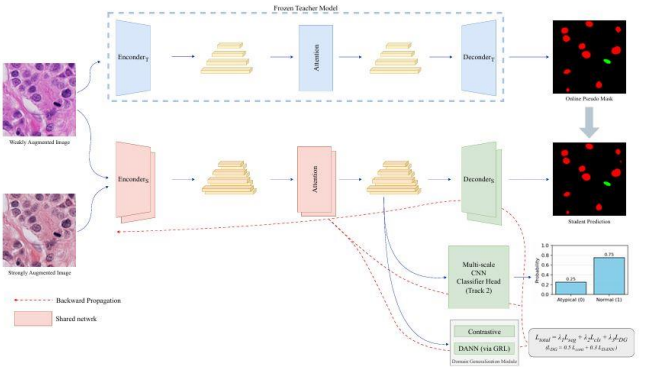


Figure 25: Overview of the proposed semi-supervised framework with UNet backbone, attention, contrastive, and domain-adversarial modules, supported by a teacher-student setup and multi-scale classification head (Track 2).

Source: Choe et al., 2025

Histopathological image analysis leverages deep learning to extract insights from H&E images, with nuclei informing diagnosis and mitosis critical for cancer grading. Nemati et al. (2025) proposed CompSegNet for nuclei segmentation and a hybrid object detection–fuzzy-classification approach for mitosis detection, validated on the MiDeSeC dataset. On MiDeSeC and ICPR12, YOLOv8 + FRF achieved the highest F1-scores (0.882 and 0.913, respectively), outperforming standard YOLOv8 and other hybrids. These results highlight that combining advanced object detection with classifier-based refinement enhances precision, recall, and overall detection accuracy in mitosis analysis. Figure 26 shows the overall structure of the mitosis detection methodology, illustrating the feature extraction backbone, the segmentation or detection modules, and the

classification components used to identify mitotic figures across multi-scale image representations.

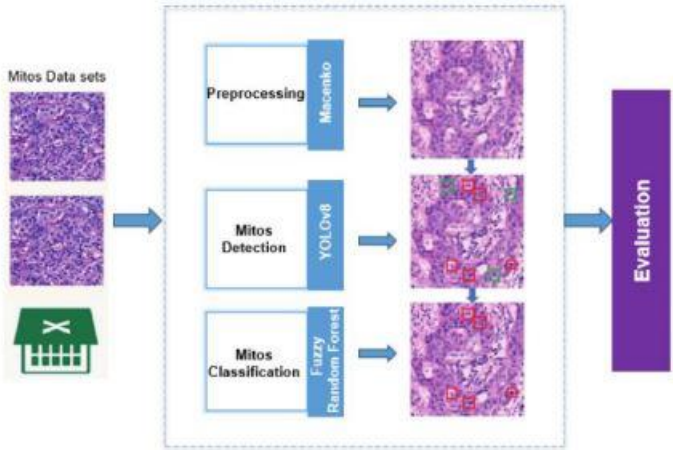


Figure 26: The overall structure of the mitosis detection methodology

Source: Nemati et al., 2025

Accurate mitosis detection is critical for cancer diagnosis but remains challenging due to class imbalance and complex morphological variations. To address this, Alhassan and Altmami (2025) developed a Customized Deep Learning (CDL) model, integrating transfer learning, skip connections, and a hybrid Jellyfish Search–Walrus Optimization mechanism to enhance feature extraction, localization, and model momentum. The CDL model was evaluated on multiple public datasets, including Mitosis WSI CCMCT, Mitosis-AIC, Mitosis Detection, and Mitosis and Non-Mitosis, achieving an F1-score of 0.994 and an accuracy of 0.988. These results demonstrate the model's effectiveness for robust mitotic figure detection, supporting pathologists in cancer diagnosis and prognosis. Future directions include fusion methodologies, real-time efficiency, and extension to broader histopathological analyses. Figure 27 shows a Block diagram of the proposed mitosis detection technique, illustrating the main processing stages, including feature extraction, candidate region generation, and the classification module used to identify mitotic figures accurately.

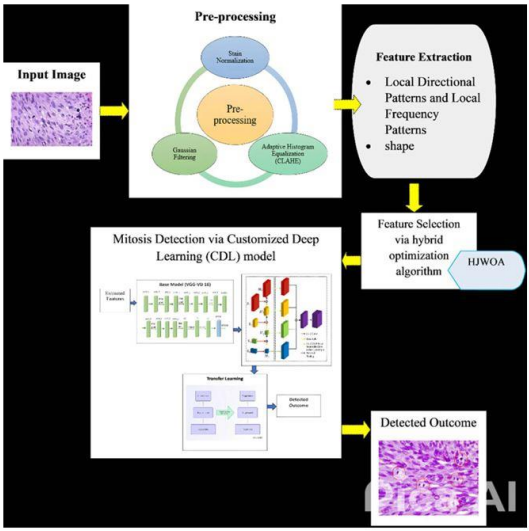


Figure 27: Block diagram of the proposed mitosis detection technique

Source: Alhassan & Altmami, 2025

Ivan et al. (2025) introduced a subphase-labeled mitotic dataset to enhance AI based cell division analysis in digital pathology. Extending MIDOG++, the dataset includes five mitotic stages (prophase, prometaphase, metaphase, anaphase, telophase) plus atypical mitotic figures, with detailed segmentation masks for precise localization. A new LUNG-MITO dataset from lung adenocarcinoma samples further improves domain diversity. Their ConvNeXt–Mask R-CNN with EfficientNet refinement achieved superior F1-scores (0.794 on stMIDOG++ and 0.761 on LUNG-MITO), setting a new benchmark for phase-specific mitosis detection and segmentation. Figure 28 shows that the refinement is performed hierarchically, i.e., first, a mitotic vs. imposter differentiation is performed, and in the second step, the final subphase classification is done on the mitotic cells. An initial decision from Mask R-CNN is only overwritten if its confidence is lower than that of the EfficientNet.

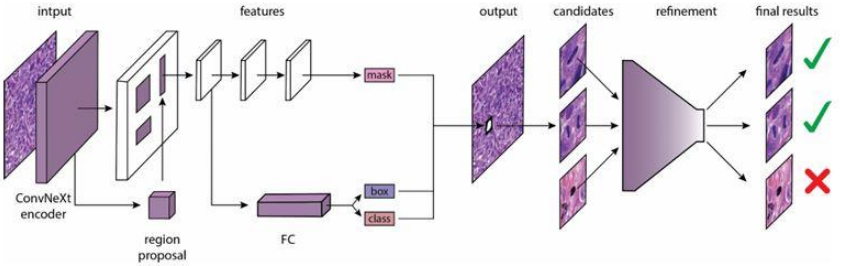


Fig 28: Overview of the proposed two-stage method. First, propose mitotic candidates with a ConvNeXt encoder-based Mask R-CNN. The candidates are reclassified with EfficientNet in the second step.

Source: Ivan et al., 2025

Traditionally, pathologists manually count mitoses in H&E biopsy sections, a complex and time-consuming task. Identifying mitotic cells is challenging due to limited datasets and visual similarities between mitotic and non-mitotic cells. Computer-assisted mitosis detection simplifies this process by automatically selecting, detecting, and labeling mitotic cells. While conventional methods rely on handcrafted image processing criteria, deep neural networks enable automatic feature extraction from histopathology images. Shihabuddin and Beevi (2024) treat mitosis detection as an object detection problem using multiple neural networks. At the tissue level, mitoses were analyzed with pre-trained Faster R-CNN on raw images. Experiments were conducted on the ICPR14 and TUPAC16 datasets, with results compared to other methods in the literature. Figure 29 shows the block diagram of the proposed mitosis detection technique.

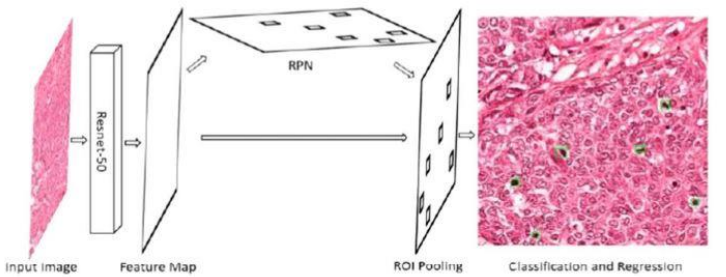


Figure 29: Mitosis detection using Faster R-CNN

Source: Shihabuddin & Beevi, 2024

Mitotic count (MC) is a critical parameter for cancer detection and grading, traditionally assessed by pathologists under high-power microscopy. Accurate identification of MC cells is challenging due to segmentation difficulties, and feature extraction often depends heavily on precise segmentation. To address this, the Coati Optimization Algorithm with Deep Learning-Driven Mitotic Nuclei Segmentation and Classification (COADL-MNSC) method is proposed. It employs median filtering for preprocessing, the Hybrid Attention Fusion U-Net (HAU-UNet) for mitotic nuclei segmentation, and a capsule network (CapsNet) for feature extraction, with hyperparameters optimized using the Coati Optimization Algorithm. Classification is performed using a bidirectional long short-term memory (BiLSTM) model. Extensive experiments on mitotic nuclei image datasets demonstrate that COADL-MNSC achieves superior performance, attaining an accuracy of 0.9889 compared to existing methods (Alshardan et al., 2024). Figure 30 shows the working process of the COADL-MNSC model.

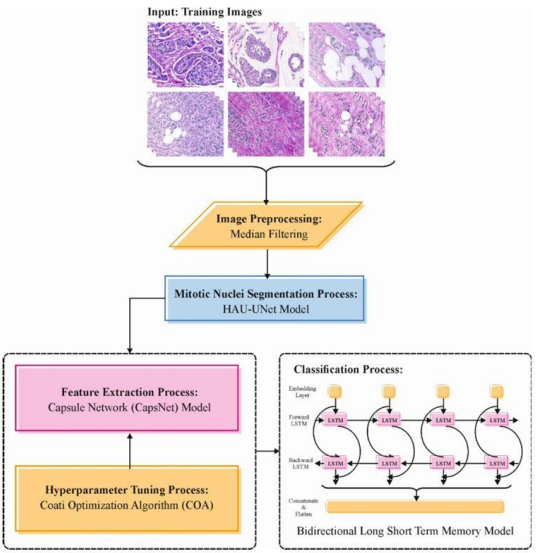


Figure 30: Workflow of the COADL-MNSC technique

Source: Alshardan et al., 2024

The main challenges include intra-class variance of mitotic cells, presence of hard negatives (non-mitotic cells resembling mitoses),

and histopathology domain shifts across datasets caused by variations in tissues, organs, scanners, and laboratories. To address these issues, Han et al. (2025) proposed the Domain Generalized Dynamic Mitosis Detector (DGDMD), which features a dynamic mitosis feature extractor based on residual structured depth-wise convolution and domain shift alignment. This extractor handles intra-class variance from differing sizes and shapes of mitotic cells, as well as hard negatives, while the domain generalization schedule aligns histopathology-mitosis domain shifts to manage variations between training and test datasets. Validate is DGDMD on the MIDOG++ dataset and other mitosis datasets, including MIDOG21, ICPR14, AMIDA13, and TUPAC16. Experiments demonstrate state-of-the-art performance in domain-generalized mitosis detection across tissues, organs, scanners, and data sources, highlighting the effectiveness of the approach. Figure 31 shows the proposed domain generalized mitosis detection system.

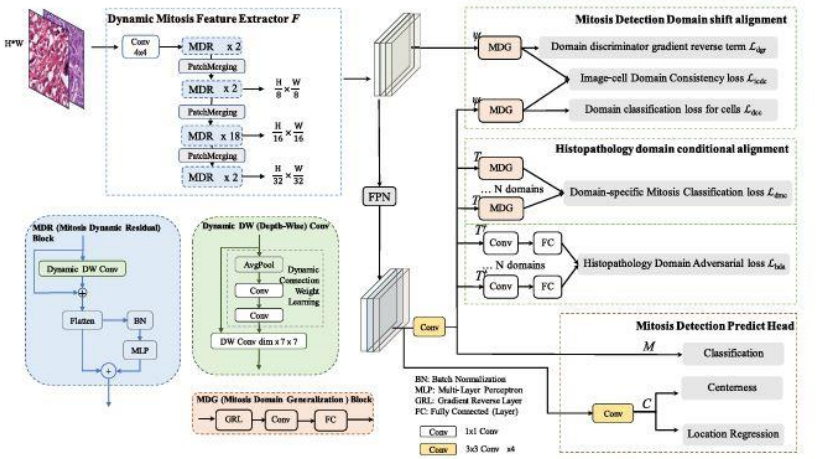


Figure 31: Overall architecture of the proposed domain generalized mitosis detection system

Source: Han et al., 2025

Salmerón et al. (2025) investigate automated mitosis detection in stained histopathological images using deep learning, with a focus on object detection models. They propose a two-stage Faster R-CNN model to effectively detect mitoses, complemented by stain

augmentation and normalization to address domain shifts in histopathology images. Experiments on the MIDOG++ dataset show that Faster R-CNN with stain techniques achieves the most accurate and reliable detection, while previous one-stage RetinaNet frameworks offer faster performance. Our results demonstrate strong F1-scores across various scenarios and tumor types, emphasizing the importance of addressing domain shifts and mitotic figure counts for robust diagnostic tools. Figure 32 shows Mitotic figures candidates from all domains.

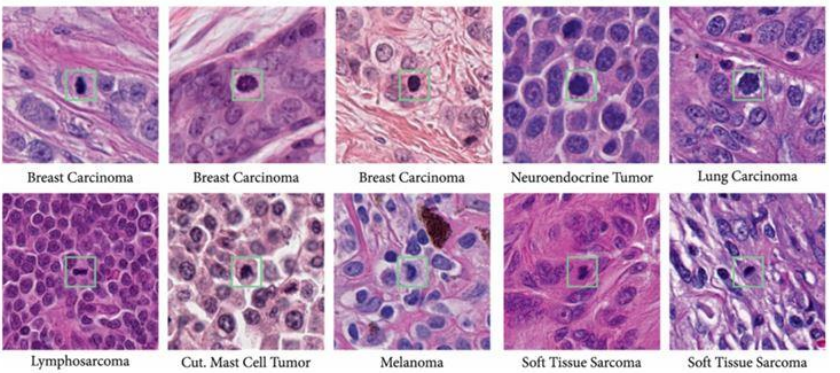


Figure 32: Mitotic figures candidates from all domains

Source: Salmerón et al., 2025

Shen et al., (2024) propose an AI approach for detecting MFs in digitized H&E-stained whole slide images (WSIs). Progress in this area has been limited by the scarcity and diversity of cancer datasets. To address this, they created the largest pan-cancer MF dataset by combining an in-house soft tissue tumour dataset (STMF) with five open-source datasets (ICPR, TUPAC, CCMCT, CMC, MIDOG++), totaling 74,620 MFs and 105,538 mitotic-like figures. Then a two-stage framework, the Optimised Mitoses Generator Network (OMG-Net), was applied, which first uses the Segment Anything Model (SAM) for automated contouring, followed by an adapted ResNet18 for MF classification. OMG-Net achieved an F1-score of 0.84 for pan-cancer MF detection, outperforming the previous MIDOG++ benchmark by 16% on breast cancer detection, demonstrating superior accuracy across various tumor types and scanner

conditions. Figure 33 shows the data preparation workflow. Figure 34 shows the OMG-Net Architecture.

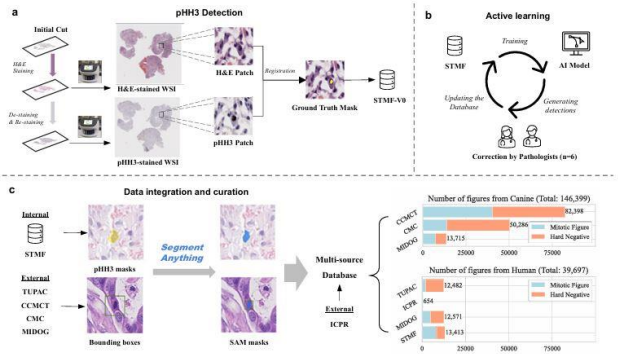


Figure 33: Data Preparation Workflow: H&E-stained WSIs were de-stained and labeled with pHH3 antibody (STMF-V0). An initial Mask R-CNN detected MFs, verified by six pathologists to refine STMF. MF masks from STMF and bounding boxes from external datasets were refined with SAM and combined with ICPR to form the final dataset.

Source: Shen et al., 2024

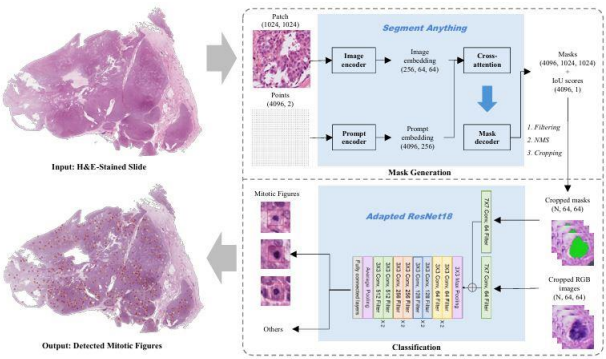


Figure 34: OMG-Net Architecture: OMG-Net has a two-step design: first, cell masks are generated from patched WSIs using SAM with a point grid prompt; second, the RGB image and binary mask of each cell are classified as mitotic figures using an adapted ResNet18.

Source: Shen et al., 2024

Accurate and efficient identification of mitotic figures (MFs) is essential for diagnosing and grading cancers such as glioblastoma (GBM), a highly aggressive brain tumor. Manual counting in WSIs is labor-intensive and prone to interobserver variability. Liu et al., (2025) Using GBM WSIs from The Cancer Genome Atlas (TCGA), the framework combines CNNs with active learning. A CNN is first trained on a small annotated set, then identifies uncertain samples from unlabeled data for expert review. Verified cases are used to iteratively retrain the model. The approach achieved 0.8175 precision, 0.8248 recall for MF detection, and 84.1% accuracy for MF subclass classification. Annotation time was reduced by nearly 50%, saving ~900 minutes across 66 WSIs. The deep active learning framework significantly improves efficiency and accuracy in MF detection and classification for GBM, reducing reliance on large annotated datasets. This method is generalizable to other medical imaging tasks, supporting broader healthcare applications. Figure 35 shows the Active Learning framework overview. Figure 36 shows the Inference pipeline overview, and Figure 37 represents a two-stage MF detection and classification.

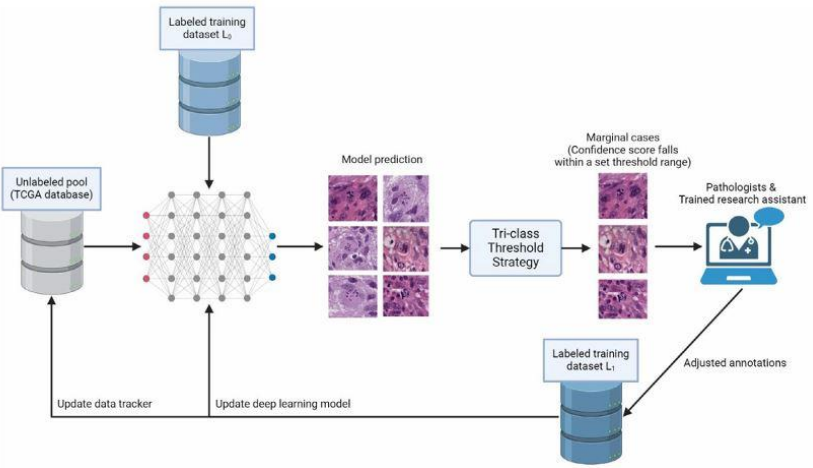


Figure 35: Active Learning Framework Overview: Images from the unlabeled pool are processed through the CNN inference pipeline. Actively selected samples are verified by pathologists and trained assistants, then used to retrain the deep learning model.

Source: Liu et al., 2025

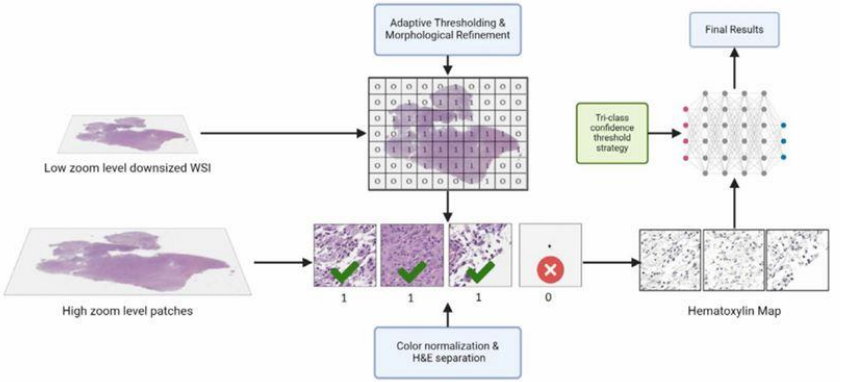


Figure 36: Inference Pipeline Overview, Each WSI is processed at low and high zoom levels. The low-zoom image generates a binary tissue map to select relevant high-zoom patches. These patches undergo H&E separation to isolate nuclei, which are then analyzed by the model to produce final detection results.

Source: Liu et al., 2025

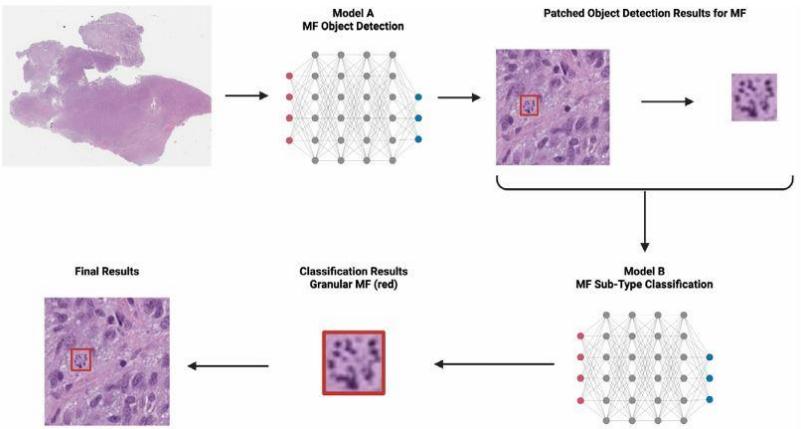


Figure 37: Two-Stage MF Detection and Classification: Stage 1 (Model A) detects candidate MFs from WSI patches. Detected bounding boxes are passed to Stage 2 (Model B) to classify each MF as normal, atypical, or granular. For illustration, only granular MFs are shown.

Source: Liu et al., 2025

CONCLUSIONS

Mitotic figure detection and analysis play an indispensable role in digital pathology for evaluating tumor proliferation, cancer grading, and prognostic assessment. Over the past decade, benchmark datasets such as ICPR, AMIDA, TUPAC, the MIDOG series, CCMCT, AMi-Br, and MiDeSeC have provided a critical foundation for the systematic development, comparison, and clinical validation of deep learning models in this domain. Despite significant progress, major challenges persist, including the rarity of mitotic figures, their high morphological variability, the abundance of visually similar hard-negative examples, and—most importantly—domain shifts arising from differences in scanners, staining protocols, laboratories, species, organs, and tumor types. While early datasets (e.g., ICPR12, AMIDA13) were small-scale and relatively homogeneous, limiting generalization performance, more recent collections such as MIDOG++, CCMCT, and AMi-Br have introduced multi-domain, multi-species, multi-scanner data along with atypical mitosis differentiation, thereby better reflecting real-world complexity and marking a substantial advancement in the field. State-of-the-art models currently achieve the highest performance by combining modern single-stage object detectors (YOLOv5, YOLOv7, YOLOv8), two-stage frameworks (Faster R-CNN, Mask R-CNN), U-Net family architectures and derivatives for segmentation, powerful classification backbones such as EfficientNet, ConvNeXt, and Swin Transformer, together with techniques including stain normalization, contrastive learning, domain-adversarial training, hard-negative mining, ensemble strategies, and semi-supervised learning. Advanced tasks such as the detection of atypical mitotic figures and the classification of mitotic subphases (from prophase to telophase) are receiving increasing attention, enabling not only quantitative proliferation indices but also additional biological insights into chromosomal instability and tumor aggressiveness. Future research priorities include the creation of large-scale, multi-center, multi-national, multi-cancer-type, and multi-species open datasets; the standardization of transparent, multi-expert consensus-based annotation protocols; prospective validation of the prognostic value of atypical mitosis ratios in large cohorts; privacy-preserving

model training through federated learning and differential privacy techniques; the development of real-time, explainable, and clinically deployable systems; and deep modeling of relationships between mitotic activity and genetic/molecular markers (Ki-67, PHH3, NGS data, etc.). In summary, mitotic figure detection and analysis remain one of the most dynamic and rapidly evolving research areas in the transition of digital pathology toward improved cancer diagnosis and personalized medicine. Continued progress will depend on enhancing dataset diversity and quality, raising annotation standards, maturing domain generalization techniques, and conducting rigorous clinical validation studies. The speed and success of this transformation will ultimately rely on sustained interdisciplinary collaboration among pathologists, computer scientists, and clinicians.

References

Alhassan, A. M., & Altmami, N. I. (2025). Mitosis detection in histopathological images using customized deep learning and hybrid optimization algorithms. *PLoS One*, *20*(7), e0327567.

Alshardan, A., Ahmad, N., Miled, A. B., Alshuhail, A., Alzahrani, Y., & Mahmud, A. (2024). Transferable deep learning with coati optimization algorithm based mitotic nuclei segmentation and classification model. *Scientific Reports*, *14*(1), 30557.

Aubreville, M., Stathonikos, N., Bertram, C. A., Klopfleisch, R., Ter Hoeve, N., Ciompi, F., & Breininger, K. (2023). Mitosis domain generalization in histopathology images—the MIDOG challenge. *Medical Image Analysis*, *84*, 102699.

Aubreville, M., Wilm, F., Stathonikos, N., Breininger, K., Donovan, T. A., Jabari, S., & Bertram, C. A. (2023). A comprehensive multi-domain dataset for mitotic figure detection. *Scientific data*, *10*(1), 484.

Bahaghighat, M., Xin, Q., Motamedi, S. A., Zanjireh, M. M., & Vacavant, A. (2020). Estimation of wind turbine angular velocity remotely found on video mining and convolutional neural network. *Applied Sciences*, *10*(10), 3544.

Bertram, C. A., Aubreville, M., Marzahl, C., Maier, A., & Klopfleisch, R. (2019). A large-scale dataset for mitotic figure assessment on whole slide images of canine cutaneous mast cell tumor. *Scientific data*, *6*(1), 274.

Bertram, C. A., Weiss, V., Donovan, T. A., Banerjee, S., Conrad, T., Ammeling, J., & Aubreville, M. (2025, March). Histologic Dataset of Normal and Atypical Mitotic Figures on Human Breast Cancer (AMi-Br). In *BVM Workshop* (pp. 113-118). Wiesbaden: Springer Fachmedien Wiesbaden.

Choe, S., Qin, X., Shafique, A., Dy, A., Done, S., Androutsos, D., & Khademi, A. (2025). Teacher-Student Model for Detecting and Classifying Mitosis in the MIDOG 2025 Challenge. *Arxiv preprint arXiv:2509.03614*.

Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

García-Salmerón, J., García, J. M., Bernabé, G., & González-Férez, P. (2025). Automated mitosis detection in stained histopathological images using Faster R-CNN and stain techniques. *Journal of Integrative Bioinformatics*, (0), 20240049.

Han, J., Wang, S., Wu, L., & Liu, W. (2025). Dynamic feature extraction and histopathology domain shift alignment for mitosis detection. *Image and Vision Computing*, *158*, 105541.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, 770-778.

Ivan, Z. Z., Hirling, D., Grexa, I., Ammeling, J., Micsik, T., Dobra, K., ... & Horvath, P. (2025). Subphase-Labeled Mitotic Dataset for AI-powered Cell Division Analysis. *bioRxiv*, 2025-07.

Kelam, N. S., Parekh, A., Bonthu, S., & Singhal, N. (2025). Ensemble YOLO Framework for Multi-Domain Mitotic Figure Detection in Histopathology Images. *Arxiv preprint arXiv:2509.02957*.

Liu, E., Lin, A., Kakodkar, P., Zhao, Y., Wang, B., Ling, C., & Zhang, Q. (2025). A deep active learning framework for mitotic figure detection with minimal manual annotation and labelling. *Histopathology*.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012-10022.

Ludovic, R., Daniel, R., Nicolas, L., Maria, K., Humayun, I., Jacques, K., & Gilles, L. N. (2013). Mitosis detection in breast cancer histological images An ICPR 2012 contest. *Journal of pathology informatics*, *4*(1), 8.

Manorost, P., Deckers, T., Bloemen, V., & Aerts, J. M. (2025). Explainable handcrafted features for mitotic event detection and classification. *Scientific Reports*, *15*(1), 7382.

Nemati, N., Hancer, E., & SAMET, R. (2025). A MITOTIC CELL DETECTION APPROACH WITH DEEPLABV3+ AND MOBILENETV2. *Appl. Comput. Math*, *24*(3), 349-363.

Nemati, N., & Samet, R. (2025). HR-YOLOv8: an innovative model to detect mitosis and identify cancer regions in histopathological images. *Neural Computing and Applications*, *37*(29), 24441-24460.

Nemati, N., Samet, R., Hancer, E., Yildirim, Z., & Akkas, E. E. (2023). A hybridized Deep learning methodology for mitosis detection and classification from histopathology images. *Journal of Machine Intelligence and Data Science (JMIDS)*, *4*(1), 35-43.

Nemati, N., Samet, R., Hancer, E., Yildirim, Z., & Traore, M. (2023). A mitosis detection and classification methodology with yolov5 and fuzzy classifiers. In *Proceedings of the 9th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS)* (Vol. 111).

Nemati, N., Hancer, E., Samet, R., Yıldırım, Z., & Traore, M. (2022). A comparative study of deep semantic segmentation architectures for mitosis detection in histopathology images. *9ROXPH*, 363.

Richardson, E., Trevizani, R., Greenbaum, J. A., Carter, H., Nielsen, M., & Peters, B. (2023). The ROC-AUC accurately assesses imbalanced datasets. *Available at SSRN 4655233*.

Rijsbergen, C. V. (1979). *Information Retrieval, Butterworths*, *2*.

Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, 234-241.

Roux, L., Racoceanu, D., Capron, F., Calvo, J., Attieh, E., Le Naour, G., & Gloaguen, A. (2014). MITOS & ATYPIA-Detection of mitosis and evaluation of nuclear atypia score in breast cancer histological images. IPAL, Agency Sci, Technol Res Inst Infocom Res. *Technol. Res. Inst. Infocom Res., Singapore, Tech. Rep*.

Samet, R., Nemati, N., Hançer, E., Sak, S., & Kırmızı, B. A. (2024, October). Histopatolojik Görüntülerde Dogru Mitoz Tespiti için Geliştirilmiş Renk Normalleştirme Yöntemi: Enhanced Stain Normalization Method for Accurate Mitosis Detection in Histopathological Images. In *2024 9th International Conference on Computer Science and Engineering (UBMK)*, 371-376.

Samet, R., Nemati, N., Hancer, E., Sak, S., & Kirmizi, B. A. (2025). An ensemble KANs method with XAI for mitosis segmentation in histopathological images: R. Samet et al. *Signal, Image and Video Processing*, *19*(11), 928.

Samet, R., Nemati, N., Hancer, E., Sak, S., Kirmizi, B. A., & Yildirim, Z. (2025). MiDeSeC: A Dataset for Mitosis Detection and Segmentation in Breast Cancer Histopathology Images. *arXiv preprint arXiv:2507.14271*.

Shen, Z., Simard, M., Brand, D., Andrei, V., Al-Khader, A., Oumlil, F., & Fekete, C. A. C. (2024). OMG-Net: A Deep Learning

Framework Deploying Segment Anything to Detect Pan-Cancer Mitotic Figures from Haematoxylin and Eosin-Stained Slides. *Arxiv preprint arXiv:2407.12773*.

Shihabuddin, A. R., & Beevi, S. (2024). Efficient mitosis detection: leveraging pre-trained faster R-CNN and cell-level classification. *Biomedical Physics & Engineering Express*, *10*(2), 025031.

Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 6105-6114.

Traore, M., Hancer, E., Samet, R., Yıldırım, Z., & Nemati, N. (2024). CompSegNet: An enhanced U-shaped architecture for nuclei segmentation in H&E histopathology images. *Biomedical Signal Processing and Control*, *97*, 106699.

Veta, M., Heng, Y. J., Stathonikos, N., Bejnordi, B. E., Beca, F., Wollmann, T., & Pluim, J. P. (2019). Predicting breast tumor proliferation from whole-slide images: the TUPAC16 challenge. *Medical image analysis*, *54*, 111-121.

Veta, M., Van Diest, P. J., Willems, S. M., Wang, H., Madabhushi, A., Cruz-Roa, A., & Pluim, J. P. (2015). Assessment of algorithms for mitosis detection in breast cancer histopathology images. *Medical image analysis*, *20*(1), 237-248.

Vieira, S. M., Kaymak, U., & Sousa, J. M. (2010, July). Cohen's kappa coefficient as a performance measure for feature selection. In *International conference on fuzzy systems* (pp. 1-8). IEEE.

Wang, H., Xu, H., Li, B., Pan, X., Zeng, L., Lan, R., & Luo, X. (2024). A novel dataset and a two-stage mitosis nuclei detection method based on hybrid anchor branch. *Biomedical Signal Processing and Control*, *87*, 105374.

Wilm, F., Marzahl, C., Breininger, K., & Aubreville, M. (2021, September). Domain adversarial RetinaNet as a reference algorithm for the MItosis DOmain generalization challenge. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 5-13). Cham: Springer International Publishing.

Yıldırım, Z., Hançer, E., Samet, R., Mali, M. T., & Nemati, N. (2022, May). Effect of color normalization on nuclei segmentation problem in h&e stained histopathology images. In *2022 30th Signal Processing and Communications Applications Conference (SIU)*, pp. 1-4.

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018, September). Unet++: A nested u-net architecture for medical image segmentation. In *International workshop on deep learning in medical image analysis* (pp. 3-11). Cham: Springer International Publishing.

BÖLÜM 3

AI-ENHANCED BPM SYSTEMS: A COMPREHENSIVE REVIEW OF TOOLS, ARCHITECTURES, AND CHALLENGES

AHMET TOPRAK[1]

Introduction

Business Process Management (BPM) [1] has long been a cornerstone of organizational efficiency, enabling enterprises to design, execute, monitor, and optimize structured workflows. Traditionally, BPM systems have relied on formal process modeling techniques and rule-based execution engines to standardize operations and ensure consistency across business functions. These systems have been widely adopted in domains such as finance, healthcare, manufacturing, and public administration, where clearly defined processes and compliance requirements are essential. However, the increasing complexity of modern business environments, coupled with the exponential growth of data, has exposed fundamental limitations in traditional BPM approaches.

Conventional BPM systems [2] are largely built upon static process models and predefined decision rules, which restrict their

[1] Dr, İstanbul Ticaret University, Computer Engineering, Orcid: 0000-0001-7046-8512

ability to adapt to dynamic and uncertain environments. In practice, business processes often involve unstructured data, evolving conditions, and context-dependent decisions that cannot be fully captured through rigid workflows. As a result, organizations frequently encounter challenges in handling exceptions, optimizing processes in real time, and leveraging the vast amounts of data generated during process execution. These limitations have created a growing demand for more intelligent, adaptive, and data-driven process management solutions.

The integration of Artificial Intelligence (AI) [3] into BPM systems represents a transformative shift from rule-based automation to intelligent process management. AI technologies, including machine learning, deep learning, natural language processing (NLP) [4], and large language models (LLMs) [5], enable BPM systems to analyze historical and real-time data, identify patterns, predict outcomes, and support decision-making processes. This evolution has led to the emergence of AI-enhanced BPM systems, where processes are not only automated but also continuously optimized through data-driven insights and adaptive mechanisms.

In parallel, the rise of complementary technologies such as Robotic Process Automation (RPA) [6], process mining, and cloud-native architectures has further expanded the capabilities of BPM systems. RPA enables the automation of repetitive tasks at the user interface level, while process mining techniques allow organizations to discover and analyze actual process flows based on event logs. Cloud-native technologies, including microservices and containerization, provide the scalability and flexibility required to deploy and manage AI-driven BPM solutions [7] in distributed environments. Together, these technologies contribute to the broader concept of hyper automation, where multiple automation tools are orchestrated to achieve end-to-end process optimization.

Another significant development in the BPM landscape is the emergence of lightweight workflow automation platforms, such as n8n and similar integration tools. Unlike traditional BPM suites, which emphasize formal process modeling and centralized orchestration, these platforms focus on event-driven workflows, API-based integration [8], and rapid automation development. Their ability to seamlessly integrate with AI services and external systems makes them particularly relevant in the context of AI-enhanced BPM, where flexibility and interoperability are critical.

Despite these advancements, the adoption of AI-enhanced BPM systems introduces new challenges related to system design, integration complexity, data quality, and governance. The incorporation of AI models into process workflows raises concerns regarding explainability, reliability, and ethical decision-making, particularly in high-stakes domains. Additionally, the integration of heterogeneous systems and technologies can lead to increased architectural complexity, making it difficult to ensure consistency, scalability, and security.

Given these developments, a comprehensive review of AI-enhanced BPM systems is both timely and necessary. This paper aims to provide a systematic analysis of the evolution, architectures, tools, and challenges associated with AI-driven process management. Specifically, the objectives of this study are to examine the transformation of BPM systems through AI integration, analyze emerging architectural paradigms, evaluate leading tools and platforms, identify key challenges and limitations, and highlight future research directions. By synthesizing insights from both academic research and industry practices, this work seeks to contribute to a deeper understanding of intelligent process management in the era of artificial intelligence.

Background and Fundamental Concepts

The concept of Business Process Management is rooted in the systematic analysis and improvement of organizational workflows. A business process can be defined as a sequence of structured activities designed to achieve a specific organizational objective, often involving multiple actors, systems, and decision points. Traditional BPM approaches emphasize formalization, standardization, and control, relying on well-defined process models and execution engines to ensure consistency and efficiency. These models are typically represented using standardized notations such as BPMN [9], which provide a visual and semantic framework for describing process logic, control flow, and interactions.

The BPM lifecycle traditionally consists of several interconnected phases, including process identification, modeling, execution, monitoring, and optimization. In the modeling phase, processes are designed based on business requirements and domain knowledge, often involving collaboration between business analysts and technical experts. Once defined, these models are deployed within workflow engines that execute process logic according to predefined rules. Monitoring mechanisms track process performance through key performance indicators (KPIs) [10], enabling organizations to identify inefficiencies and areas for improvement. Finally, optimization involves refining process models based on insights derived from monitoring and analysis.

While this lifecycle provides a structured approach to process management, it is inherently limited by its reliance on static models and deterministic rules. In dynamic business environments, processes are subject to frequent changes, uncertainties, and exceptions that cannot be fully anticipated during the design phase. This limitation has led to the incorporation of data-driven techniques, particularly those derived from artificial intelligence and machine learning, into BPM systems.

AI introduces the ability to learn from data, adapt to changing conditions, and make probabilistic decisions. Machine learning techniques enable BPM systems to analyze historical process data, identify patterns, and predict future outcomes such as process delays, bottlenecks, or failures. Unsupervised learning methods can be used to detect anomalies and uncover hidden structures within process data, while reinforcement learning approaches allow systems to optimize decision-making policies over time. These capabilities transform BPM systems from static execution engines into adaptive and intelligent systems capable of continuous improvement.

Process mining plays a crucial role in bridging the gap between traditional BPM and data-driven intelligence. By analyzing event logs generated during process execution, process mining techniques can automatically discover process models, evaluate conformance between actual and expected behavior, and identify performance bottlenecks. This data-driven perspective provides a more accurate and objective understanding of organizational processes, enabling more effective optimization strategies.

Recent advancements in natural language processing and large language models have further expanded the scope of AI in BPM. These technologies enable new forms of interaction between users and process systems, such as natural language-based process modeling, automated documentation generation, and conversational workflow management. By reducing the reliance on formal modeling expertise, these capabilities make BPM systems more accessible and flexible.

In addition to AI technologies, the integration of Robotic Process Automation has significantly enhanced the automation capabilities of BPM systems. RPA tools can automate repetitive, rule-based tasks by mimicking human interactions with software applications, thereby complementing BPM systems that focus on process orchestration. The combination of BPM, AI, and RPA forms

the foundation of intelligent automation ecosystems, where processes are not only automated but also continuously optimized and adapted based on real-time data.

Overall, the evolution of BPM toward AI-enhanced systems reflects a broader shift from static, rule-based automation to dynamic, data-driven process management. This transformation lays the foundation for the architectural and technological developments discussed in the subsequent sections.

Architectures of AI-Enhanced BPM Systems

The evolution of Business Process Management systems toward AI-enhanced paradigms has led to significant transformations in system architecture. Traditional BPM architectures were primarily designed around centralized workflow engines, rule-based decision systems, and static process models. While effective for structured and predictable workflows, these architectures lack the flexibility and scalability required to support intelligent, data-driven, and adaptive processes. The integration of artificial intelligence and related technologies has necessitated the development of more modular, distributed, and extensible architectural frameworks.

In classical BPM architectures, the system is typically organized around a core workflow engine responsible for executing process models defined using standardized notations. This engine interacts with a business rules management system (BRMS) [11], which governs decision logic, and a data layer that stores process instances and execution logs. While this architecture provides a clear separation of concerns, it is inherently limited in its ability to incorporate real-time data analytics and adaptive decision-making. The reliance on predefined rules restricts the system's capacity to respond dynamically to changing conditions, making it unsuitable for complex and unstructured environments.

AI-enhanced BPM architectures extend this traditional model by introducing additional layers dedicated to data processing, machine learning, and intelligent decision-making. One of the key components in such architectures is the integration of predictive analytics modules, which leverage historical process data to forecast outcomes such as delays, resource utilization, or process failures. These modules are often supported by machine learning pipelines that include data ingestion, preprocessing, model training, and inference stages. By embedding predictive capabilities into process execution, BPM systems can transition from reactive to proactive process management.

Another important architectural element is the incorporation of process mining and event stream processing components. Process mining tools analyze event logs to derive insights about actual process behavior, enabling continuous monitoring and optimization. When combined with real-time event streaming technologies, such as message brokers and complex event processing (CEP) engines [12], BPM systems can respond to events as they occur, facilitating near real-time decision-making. This event-driven approach represents a significant departure from traditional batch-oriented BPM systems and aligns with the requirements of modern, data-intensive applications.

The integration of large language models and natural language processing capabilities introduces a new layer of interaction and intelligence within BPM architectures. LLM-based components [13] can be used to interpret unstructured inputs, generate process documentation, assist in decision-making, and enable conversational interfaces for process control. These components are typically exposed through APIs and integrated into BPM systems as external services, allowing for flexible and scalable deployment. Their inclusion enhances the system’s ability to handle

ambiguity and human-centric tasks that are difficult to formalize using traditional modeling techniques.

Cloud-native design principles play a central role in enabling the scalability and flexibility of AI-enhanced BPM systems. Modern architectures increasingly adopt microservices-based approaches, where different functionalities—such as workflow execution, data processing, AI inference, and integration services—are implemented as independent, loosely coupled components. These components communicate through APIs and messaging systems, enabling seamless integration and horizontal scalability. Containerization technologies and orchestration platforms further facilitate deployment, resource management, and fault tolerance, making it possible to operate BPM systems in distributed and dynamic environments.

In this context, lightweight workflow automation platforms, such as n8n [14] and similar tools, have emerged as complementary components within BPM ecosystems. Unlike traditional BPM suites, which emphasize centralized control and formal modeling, these platforms adopt an event-driven and integration-centric approach. They enable rapid development of workflows through visual, node-based interfaces and provide extensive support for API integration. Their ability to connect with AI services, data sources, and external applications makes them particularly suitable for orchestrating AI-driven tasks and integrating heterogeneous systems. As a result, they are increasingly used as orchestration layers within broader AI-enhanced BPM architectures.

Security and governance considerations are also integral to architectural design. AI-enhanced BPM systems must ensure that sensitive data is protected, access to process components is controlled, and decision-making processes are auditable. This requires the implementation of authentication and authorization mechanisms, secure communication protocols, and logging systems

that capture both process execution and AI-driven decisions. Additionally, model governance frameworks are needed to manage the lifecycle of AI models, including versioning, validation, and monitoring for drift or bias.

Overall, the architecture of AI-enhanced BPM systems reflects a shift toward modularity, intelligence, and adaptability. By integrating AI components, adopting cloud-native principles, and leveraging event-driven designs, these systems are capable of supporting complex, dynamic, and data-intensive processes. This architectural transformation provides the foundation for the advanced capabilities and tools discussed in the subsequent sections.

Tools and Platforms for AI-Enhanced BPM

The rapid evolution of AI-enhanced BPM systems has been accompanied by the emergence of a diverse ecosystem of tools and platforms designed to support intelligent process management. These platforms vary significantly in terms of architectural design, target users, and capabilities, ranging from enterprise-grade BPM suites to lightweight workflow automation tools and open-source frameworks. Understanding the strengths and limitations of these platforms is essential for both researchers and practitioners seeking to design and implement effective AI-driven process solutions.

Enterprise BPM platforms have traditionally dominated the process management landscape, offering comprehensive solutions that integrate process modeling, execution, monitoring, and optimization within a unified environment. Modern enterprise platforms such as Camunda [15], Appian [16], IBM Business Automation Workflow [17], and Pegasystems [18] have increasingly incorporated AI capabilities into their offerings. These systems typically support standardized process modeling languages, such as BPMN, and provide robust workflow engines capable of handling complex, long-running processes. The integration of AI in these

platforms often includes predictive analytics, decision automation, and, in some cases, embedded machine learning models that enhance process optimization and decision-making.

Camunda, for instance, has evolved into a developer-centric platform that emphasizes process orchestration in distributed systems. Its architecture supports integration with external AI services, allowing organizations to incorporate machine learning models into process workflows. Similarly, Appian provides a low-code environment with built-in AI capabilities, enabling rapid development of intelligent applications that combine process automation with data analytics. IBM's BPM solutions integrate process mining and AI-driven insights, facilitating end-to-end process optimization. Pegasystems, on the other hand, focuses on decision automation, leveraging AI to dynamically adapt workflows based on contextual data.

In parallel, process mining platforms such as Celonis [19] and Apromore [20] have gained prominence as essential components of AI-enhanced BPM ecosystems. These platforms specialize in analyzing event logs to discover, monitor, and optimize business processes. By applying machine learning techniques to process data, they provide insights into process inefficiencies, bottlenecks, and deviations from expected behavior. While process mining tools are not full BPM systems in themselves, they play a critical role in enabling data-driven process improvement and are often integrated with BPM platforms to create closed-loop optimization systems.

Robotic Process Automation platforms, including UiPath [21] and Automation Anywhere [22], also contribute significantly to AI-enhanced BPM systems. RPA tools are designed to automate repetitive, rule-based tasks by interacting with user interfaces and legacy systems. When combined with AI technologies such as computer vision and natural language processing, RPA platforms can handle more complex and unstructured tasks. In the context of BPM,

RPA serves as an execution layer that complements process orchestration, enabling end-to-end automation across both modern and legacy systems.

In contrast to traditional enterprise platforms, lightweight workflow automation tools have emerged as flexible and developer-friendly alternatives. Platforms such as n8n [14], Zapier [23], and Apache Airflow [24] focus on integration, orchestration, and event-driven automation rather than formal process modeling. These tools provide visual interfaces for constructing workflows composed of interconnected nodes, each representing a specific function or service. Their strength lies in their ability to rapidly integrate with a wide range of APIs, databases, and external services, including AI platforms.

Among these tools, n8n is particularly notable for its open-source nature and extensibility. It enables users to design workflows that integrate AI services, such as large language models, with minimal configuration. Its event-driven architecture allows for real-time process execution, making it well-suited for dynamic and data-intensive applications. While n8n does not natively support formal process modeling standards like BPMN, its flexibility and integration capabilities make it a valuable component within AI-enhanced BPM ecosystems, particularly for orchestrating microservices and AI-driven tasks.

The distinction between enterprise BPM platforms and lightweight orchestration tools highlights a broader trend toward hybrid architectures. In many modern implementations, organizations combine multiple tools to leverage their respective strengths. For example, a BPM platform may be used for high-level process orchestration and governance, while lightweight tools handle integration and event-driven automation, and AI services provide predictive and decision-making capabilities. This layered approach enables greater flexibility and scalability, allowing

organizations to adapt their process management strategies to evolving requirements.

Despite the advancements in BPM tools and platforms, several challenges remain. Integration complexity is a significant concern, as organizations must ensure seamless communication between heterogeneous systems. Data consistency and synchronization across platforms can also be difficult to maintain, particularly in distributed environments. Additionally, the incorporation of AI introduces challenges related to model management, explainability, and reliability, which must be addressed to ensure trust and compliance.

Overall, the landscape of AI-enhanced BPM tools reflects a shift toward more flexible, modular, and intelligent systems. The convergence of BPM, process mining, RPA, and AI technologies has created a rich ecosystem of platforms that support end-to-end process automation and optimization. As these tools continue to evolve, their integration and effective utilization will play a critical role in the success of intelligent process management initiatives.

Challenges and Limitations of AI-Enhanced BPM

Despite the significant advancements brought by the integration of artificial intelligence into Business Process Management systems, the adoption of AI-enhanced BPM introduces a range of challenges and limitations that must be carefully addressed. These challenges arise from both the inherent complexities of AI technologies and the structural characteristics of BPM systems, particularly in large-scale and heterogeneous organizational environments. Understanding these limitations is essential for designing robust, reliable, and sustainable intelligent process management solutions.

One of the most critical challenges in AI-enhanced BPM systems is the issue of data quality and availability. AI models rely

heavily on historical and real-time data to generate predictions and support decision-making processes. However, process data is often incomplete, inconsistent, or noisy, particularly in organizations where data is collected from multiple systems and sources. Event logs, which are fundamental for process mining and predictive analytics, may contain missing timestamps, incorrect sequences, or fragmented traces, leading to inaccurate models and unreliable insights. Poor data quality not only reduces the effectiveness of AI components but can also propagate errors throughout the entire process lifecycle.

Another significant limitation concerns the explainability and transparency of AI-driven decisions. Traditional BPM systems are inherently interpretable, as their behavior is defined by explicit rules and process models. In contrast, many AI models, particularly those based on deep learning and large language models, operate as black boxes, making it difficult to understand how specific decisions are made. This lack of explainability poses challenges in domains where accountability, compliance, and trust are essential, such as finance, healthcare, and public administration. Organizations must therefore balance the benefits of AI-driven automation with the need for interpretable and auditable decision-making processes.

Integration complexity represents another major challenge in the implementation of AI-enhanced BPM systems. Modern organizations typically operate within complex IT ecosystems that include legacy systems, cloud-based services, and various third-party applications. Integrating AI models, process mining tools, RPA platforms, and workflow engines into a cohesive system requires careful architectural design and robust integration mechanisms. The use of heterogeneous technologies and data formats can lead to interoperability issues, increased maintenance overhead, and potential system inconsistencies. Moreover, ensuring real-time

communication and synchronization across distributed components adds further complexity.

The management of AI models within BPM systems introduces additional operational challenges. AI models require continuous monitoring, updating, and validation to ensure their accuracy and relevance over time. Issues such as model drift, where the performance of a model degrades due to changes in underlying data patterns, can significantly impact process outcomes. Organizations must establish model governance frameworks that address version control, performance evaluation, retraining strategies, and deployment pipelines. Without proper governance, AI components can become unreliable and introduce risks into critical business processes.

Ethical and legal considerations also play a crucial role in AI-enhanced BPM systems. The use of AI for decision-making raises concerns related to bias, fairness, and accountability. Biased training data can lead to discriminatory outcomes, particularly in processes involving customer interactions, credit decisions, or resource allocation. Additionally, the use of AI must comply with regulatory requirements related to data protection, transparency, and automated decision-making. Ensuring that AI-enhanced BPM systems adhere to ethical principles and legal standards is essential for maintaining trust and avoiding potential liabilities.

Scalability and performance constraints further complicate the adoption of AI-enhanced BPM. While cloud-native architectures and distributed systems provide scalability, the integration of AI models can introduce latency, particularly in real-time decision-making scenarios. Complex machine learning models may require significant computational resources, which can impact system performance and increase operational costs. Balancing performance requirements with the computational demands of AI components is therefore a critical design consideration.

Finally, organizational and cultural barriers can hinder the successful adoption of AI-enhanced BPM systems. The implementation of intelligent process management often requires changes in workflows, roles, and decision-making practices. Resistance to change, lack of technical expertise, and insufficient collaboration between business and IT teams can impede progress. Organizations must invest in training, change management, and cross-functional collaboration to fully realize the benefits of AI-enhanced BPM.

In summary, while AI-enhanced BPM systems offer significant potential for improving efficiency and decision-making, their successful implementation requires careful consideration of technical, organizational, and regulatory challenges. Addressing these limitations is essential for building reliable and sustainable intelligent process management solutions.

Future Trends and Research Directions

The continued evolution of AI-enhanced BPM systems is expected to be shaped by several emerging trends and research directions that aim to address current limitations and unlock new capabilities. As organizations increasingly seek to automate complex and dynamic processes, the integration of advanced AI technologies, novel architectural paradigms, and improved governance frameworks will play a central role in the future of process management.

One of the most prominent trends is the growing use of large language models and generative AI in BPM systems. These technologies enable new forms of interaction and automation, including natural language-based process modeling, automated generation of workflow definitions, and conversational interfaces for process execution and monitoring. By reducing the reliance on formal modeling expertise, generative AI has the potential to

democratize BPM and make it more accessible to non-technical users. Furthermore, the ability of LLMs to interpret unstructured data and generate context-aware responses can significantly enhance decision-making processes within workflows.

Another important research direction is the development of autonomous and self-optimizing processes. By leveraging reinforcement learning and continuous feedback mechanisms, BPM systems can evolve from reactive and predictive models to fully autonomous systems capable of optimizing their behavior over time. These systems can dynamically adjust process flows, resource allocation, and decision policies based on real-time data, leading to improved efficiency and adaptability. The concept of "self-healing" workflows, which automatically detect and correct anomalies or inefficiencies, represents a key milestone in this direction.

The integration of digital twins into BPM systems is also gaining attention as a promising approach for process optimization. Digital twins are virtual representations of real-world processes that can be used to simulate, analyze, and optimize process behavior under different scenarios. By combining digital twins with AI-driven analytics, organizations can evaluate the impact of potential changes before implementing them in real-world environments, thereby reducing risk and improving decision-making.

Another emerging trend is the convergence of BPM with hyper automation frameworks, where multiple automation technologies—including BPM, RPA, AI, and integration platforms—are orchestrated to achieve end-to-end process automation. This approach emphasizes the seamless integration of different tools and technologies, enabling organizations to automate complex workflows that span multiple systems and domains. Lightweight workflow platforms and API-driven orchestration tools are expected to play an increasingly important role in this ecosystem.

From a governance perspective, the development of frameworks for responsible and trustworthy AI in BPM is becoming increasingly important. These frameworks aim to ensure that AI-driven decisions are transparent, fair, and accountable, addressing concerns related to bias, explainability, and compliance. Research in this area includes the development of explainable AI techniques, model auditing tools, and standardized evaluation metrics for AI performance within process environments.

Advancements in cloud computing and edge computing are also expected to influence the future of AI-enhanced BPM. Cloud-native architectures will continue to provide scalability and flexibility, while edge computing will enable real-time processing and decision-making closer to data sources. This is particularly relevant for applications involving Internet of Things (IoT) devices, where low-latency processing is critical.

Finally, the standardization and interoperability of BPM systems represent a key area for future research. As organizations increasingly adopt heterogeneous systems and tools, the ability to integrate and coordinate these components becomes essential. Standardized interfaces, data formats, and communication protocols can facilitate interoperability, reduce integration complexity, and enable more cohesive process ecosystems.

In conclusion, the future of AI-enhanced BPM systems is characterized by increasing intelligence, autonomy, and integration. While significant challenges remain, ongoing research and technological advancements are expected to drive the development of more adaptive, efficient, and user-friendly process management solutions. The continued collaboration between academia and industry will be essential for realizing the full potential of intelligent process management in the coming years.

References

[1] Helbin, T., & Van Looy, A. (2021). Is Business Process Management (BPM) Ready for Ambidexterity? Conceptualization, Implementation Guidelines and Research Agenda. Sustainability, 13(4), 1906. https://doi.org/10.3390/su13041906

[2] Taherdoost, H., & Madanchian, M. (2024). Blockchain and Business Process Management (BPM) Synergy: A Comparative Analysis of Modeling Approaches. Information, 15(1), 9. https://doi.org/10.3390/info15010009

[3] A. Mabona, D. Van Greunen and K. Kevin, "Integration of Artificial Intelligence (AI) in Academic Libraries: A Systematic Literature Review," 2024 IST-Africa Conference (IST-Africa), Dublin, Ireland, 2024, pp. 1-9, doi: 10.23919/IST-Africa63983.2024.10569288.

[4] Y. Rajanak, R. Patil and Y. P. Singh, "Language Detection Using Natural Language Processing," 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2023, pp. 673-678, doi: 10.1109/ICACCS57279.2023.10112773.

[5] M. A. K. Raiaan et al., "A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges," in IEEE Access, vol. 12, pp. 26839-26874, 2024, doi: 10.1109/ACCESS.2024.3365742.

[6] Patrício, L., Varela, L., Silveira, Z., Felgueiras, C., & Pereira, F. (2025). A Framework for Integrating Robotic Process Automation with Artificial Intelligence Applied to Industry 5.0. Applied Sciences, 15(13), 7402. https://doi.org/10.3390/app15137402

[7] Owen, Benjamin. (2025). AI-Driven Business Process Management (BPM).

[8] K. Tair and S. Boukhedouma, "Integration of Internet of Things in BPM Lifecycle: Concepts and Comparison of Approaches," 2022 First International Conference on Big Data, IoT, Web Intelligence and Applications (BIWA), Sidi Bel Abbes, Algeria, 2022, pp. 71-76, doi: 10.1109/BIWA57631.2022.10037916.

[9] Drakopoulos, P., Malousoudis, P., Nousias, N., Tsakalidis, G., & Vergidis, K. (2026). Do LLMs Speak BPMN? An Evaluation of Their Process Modeling Capabilities Based on Quality Measures. Computation, 14(1), 10. https://doi.org/10.3390/computation14010010

[10] Bigwanto, A., Widayati, N., Wibowo, M. A., & Sari, E. M. (2024). Key Performance Indicators (KPI) to Measure Effectiveness of Lean Construction in Indonesian Project. Sustainability, 16(15), 6461. https://doi.org/10.3390/su16156461

[11] Bürkner, Paul-Christian. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. Journal of Statistical Software. 80. 10.18637/jss.v080.i01.

[12] S. Kuboi, K. Baba, S. Takano and K. Murakami, "An Evaluation of a Complex Event Processing Engine," 2014 IIAI 3rd International Conference on Advanced Applied Informatics, Kokura, Japan, 2014, pp. 190-193, doi: 10.1109/IIAI-AAI.2014.48.

[13] Khasanova Zafar kizi, M., & Suh, Y. (2025). Design and Performance Evaluation of LLM-Based RAG Pipelines for Chatbot Services in International Student Admissions. Electronics, 14(15), 3095. https://doi.org/10.3390/electronics14153095

[14] D A. Tuyishime, F. Basciani, L. Iovino, J. L. C. Izquierdo, J. Cabot and A. Pierantonio, "Bridging Workflow Automation Tools and EMF Modeling Ecosystems," 2023 ACM/IEEE International Conference on Model Driven Engineering Languages and Systems Companion (MODELS-C), Västerås,

Sweden, 2023, pp. 893-897, doi: 10.1109/MODELS-C59198.2023.00140.

[15] G. David, D. R. Zmaranda, R. -Ş. Győrödi and C. A. Győrödi, "Exploring the Impact of Workflow Engines on Business Process Management in Enterprise Applications. A case-study: Camunda," 2023 17th International Conference on Engineering of Modern Electric Systems (EMES), Oradea, Romania, 2023, pp. 1-4, doi: 10.1109/EMES58375.2023.10171706.

[16] Ashok, Shruthi. (2025). Appian Applications in Supply Chain.

[17] Represa, J. G., Larrinaga, F., Varga, P., Ochoa, W., Perez, A., Kozma, D., & Delsing, J. (2023). Investigation of Microservice-Based Workflow Management Solutions for Industrial Automation. Applied Sciences, 13(3), 1835. https://doi.org/10.3390/app13031835

[18] Ghorai, Aindrila. (2023). Enhancing PEGA Knowledge Management with AI: Transforming Customer Service. International Journal of Science and Research (IJSR). 12. 2193-2195. 10.21275/SR24615143621.

[19] M. A. R. D. Julca, Á. R. L. Cárdenas, J. Armas-Aguirre and S. A. Mayorga, "Optimization model for healthcare processes using Process Mining," 2023 18th Iberian Conference on Information Systems and Technologies (CISTI), Aveiro, Portugal, 2023, pp. 1-9, doi: 10.23919/CISTI58278.2023.10211813.

[20] Fornari, F., La Rosa, M., Polini, A., Re, B., Tiezzi, F. (2018). Checking Business Process Correctness in Apromore. In: Mendling, J., Mouratidis, H. (eds) Information Systems in the Big Data Era. CAiSE 2018. Lecture Notes in Business Information Processing, vol 317. Springer, Cham. https://doi.org/10.1007/978-3-319-92901-9_11

[21] Y. Ketkar and S. Gawade, "Effectiveness of Robotic Process Automation for data mining using UiPath," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 2021, pp. 864-867, doi: 10.1109/ICAIS50930.2021.9396024.

[22] D. Baweja, "A Comparative Analysis of Automation Anywhere, UiPath, and BluePrism," 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2023, pp. 1715-1718, doi: 10.1109/ICACITE57410.2023.10182777.

[23] S. Padwal, R. Sardesai, A. Huddar and U. P. Gurav, "Deep Learning & Zapier Automation based Medicinal Leaf Talk:Unraveling Nature’s Secrets," 2024 Asian Conference on Intelligent Technologies (ACOIT), KOLAR, India, 2024, pp. 1-6, doi: 10.1109/ACOIT62457.2024.10939936.

[24] Dylan Intorf; Dylan Storey; Kendrick van Doorn, Apache Airflow Best Practices: A practical guide to orchestrating data workflow with Apache Airflow , Packt Publishing, 2024.

BÖLÜM 4

A COMPREHENSIVE REVIEW OF DIGITAL ASSET CUSTODY: TECHNOLOGIES, SECURITY MODELS, AND REGULATORY CHALLENGES

AHMET TOPRAK[1]

Introduction

The rapid evolution of blockchain technologies and the widespread adoption of digital assets have fundamentally transformed the landscape of financial systems and asset management. Digital assets, including cryptocurrencies, tokenized securities, decentralized finance (DeFi) instrument [1], and non-fungible tokens (NFTs) [2], have introduced new paradigms of ownership, transferability, and programmability. Unlike traditional financial assets [3], which are managed through centralized intermediaries such as banks and custodians, digital assets rely on decentralized infrastructures where ownership is cryptographically enforced through private keys. Consequently, the secure storage and management of these cryptographic keys commonly referred to as digital asset custody has emerged as a critical challenge.

[1] Dr, İstanbul Ticaret University, Computer Engineering, Orcid: 0000-0001-7046-8512

In conventional financial systems [4], custodians play a central role in safeguarding assets, ensuring regulatory compliance, and facilitating transactions. However, in blockchain-based systems, the concept of custody is fundamentally different. Ownership is not determined by legal records maintained by centralized institutions but by the control of private keys associated with blockchain addresses. This shift introduces both opportunities and risks. On one hand, users gain full control over their assets without reliance on intermediaries; on the other hand, the loss or compromise of private keys results in irreversible asset loss, as blockchain transactions are immutable and typically lack recovery mechanisms.

The increasing institutional adoption of digital assets has further elevated the importance of robust custody solutions. Financial institutions, asset managers, hedge funds, and exchanges require secure, scalable, and compliant custody infrastructures capable of handling large volumes of digital assets. This demand has led to the development of sophisticated custody technologies that combine cryptographic techniques, hardware-based security mechanisms, and distributed trust models. Solutions such as Hardware Security Modules (HSMs), Multi-Party Computation (MPC), and multi-signature schemes have been widely adopted to mitigate risks associated with key management and unauthorized access [5].

In addition to technological advancements, the digital asset custody landscape is shaped by evolving threat models and increasingly complex attack vectors. Cyberattacks targeting exchanges, custodial services, and individual wallets have demonstrated the vulnerabilities inherent in both centralized and decentralized custody approaches. Threats such as phishing, malware, insider attacks, and protocol-level exploits highlight the necessity of layered security architectures and continuous

monitoring systems [6]. Moreover, operational risks, including key mismanagement and system failures, remain significant concerns.

Regulatory developments also play a crucial role in shaping digital asset custody practices [7]. Governments and financial authorities worldwide are actively developing frameworks to regulate digital asset service providers, including custodians. However, regulatory approaches vary significantly across jurisdictions, leading to challenges in compliance, interoperability, and cross-border operations. Issues such as asset classification, custody licensing, capital requirements, and liability frameworks remain areas of ongoing debate and development.

Given the rapid growth of digital assets and the critical importance of secure custody, a comprehensive and systematic review of digital asset custody is both timely and necessary. This paper aims to analyze the technological foundations, security models, and regulatory frameworks that underpin digital asset custody systems. Specifically, the objectives of this study are as follows:

- To examine the evolution and fundamental concepts of digital asset custody
- To classify and analyze different custody models, including custodial and non-custodial approaches
- To evaluate key technologies used in secure key management, such as HSM, MPC, and multi-signature schemes
- To identify and assess major security threats and vulnerabilities in custody infrastructures
- To analyze global regulatory developments and their implications for custody providers

- To highlight current challenges and propose future research directions

By synthesizing insights from both academic research and industry practices, this paper provides a comprehensive perspective on the design, implementation, and governance of digital asset custody systems in an increasingly decentralized financial ecosystem.

Background and Fundamental Concepts

The emergence of digital assets has introduced a fundamental shift in how value is created, stored, and transferred within modern financial and computational systems. Digital assets broadly encompass cryptocurrencies, tokenized financial instruments, decentralized finance (DeFi) assets, and non-fungible tokens (NFTs), all of which are typically recorded and managed on distributed ledger technologies, most notably blockchain systems [8]. Unlike traditional financial assets, which are maintained through centralized institutions and legal ownership records, digital assets are inherently tied to cryptographic mechanisms that define and enforce ownership in a decentralized environment. This paradigm eliminates the need for trusted intermediaries while simultaneously introducing new challenges related to security and asset management.

At the core of digital asset systems lies the concept of blockchain-based ownership, which is fundamentally different from conventional ownership models. In traditional systems, ownership is established through legal frameworks and enforced by centralized authorities such as banks or custodians. In contrast, blockchain systems rely on a decentralized ledger that records all transactions in an immutable and transparent manner. Each transaction updates the state of ownership, and once validated through consensus mechanisms, it becomes permanently embedded in the ledger. This

immutability ensures data integrity and trust among participants who may not have prior relationships. However, it also introduces a critical limitation: transactions cannot be reversed, and ownership cannot be reclaimed once lost.

The concept of ownership in digital asset systems is intrinsically linked to cryptographic key management. Specifically, control over digital assets is determined by possession of a private key, which is part of an asymmetric cryptographic key pair [9]. The public key functions as an address that can be shared openly and used to receive assets, while the private key remains confidential and is used to authorize transactions. This asymmetric relationship ensures that only the holder of the private key can initiate transfers, while others can verify the authenticity of transactions using the corresponding public key. As a result, private key security becomes synonymous with asset security. The loss of a private key results in permanent inaccessibility of the associated assets, whereas unauthorized access to a private key enables complete control by malicious actors. This absence of recovery mechanisms distinguishes digital asset systems from traditional financial infrastructures and significantly elevates the importance of secure custody solutions.

To facilitate the management of cryptographic keys, digital wallets have been developed as essential components of the digital asset ecosystem. Despite their name, wallets do not store assets directly; instead, they store the cryptographic keys that provide access to assets recorded on the blockchain. Wallets can be implemented in various forms, including software-based applications and hardware devices, each offering different trade-offs between usability and security. Software wallets, typically deployed on personal computers or mobile devices, provide convenience and ease of access but are inherently exposed to online threats such as malware and phishing attacks. In contrast, hardware wallets store

private keys in isolated physical devices, significantly reducing exposure to external attacks but requiring additional operational considerations.

Another important distinction in digital wallet [10] design relates to their connectivity. Wallets are often categorized as hot, cold, or warm depending on their exposure to network connectivity. Hot wallets, which remain connected to the internet, enable real-time transaction processing and are commonly used in environments requiring high liquidity, such as cryptocurrency exchanges. However, their continuous connectivity makes them more vulnerable to cyberattacks. Cold wallets, on the other hand, operate offline and are therefore significantly more secure against online threats, although they introduce latency in transaction execution and operational complexity. Warm wallets attempt to balance these trade-offs by maintaining limited or controlled connectivity, thereby providing a compromise between accessibility and security. These distinctions play a critical role in the design of custody architectures, particularly in institutional contexts where both security and operational efficiency are essential.

In addition to wallet types, digital asset custody can be broadly classified into custodial and non-custodial models [11], each reflecting a different approach to key management and control. In custodial models, private keys are managed by third-party entities such as exchanges, financial institutions, or specialized custody providers. These entities assume responsibility for securing the assets, implementing institutional-grade security measures, and ensuring regulatory compliance. While custodial solutions offer advantages such as ease of use, professional management, and potential recovery mechanisms, they also introduce counterparty risk and reduce user autonomy. Conversely, non-custodial models place full responsibility for key management on the user, eliminating reliance on intermediaries and enhancing privacy and control.

However, this approach requires a high level of technical competence and exposes users to risks associated with key loss, mismanagement, and operational errors.

The increasing adoption of digital assets, particularly by institutional investors, has amplified the importance of robust and scalable custody solutions. High-profile security breaches, exchange failures, and incidents of private key loss have demonstrated the significant risks associated with inadequate custody practices. As a result, digital asset custody has evolved from simple key storage mechanisms to complex systems incorporating layered security controls, advanced cryptographic techniques, and comprehensive operational frameworks. These developments reflect the growing recognition that secure custody is not merely a technical challenge but a multidimensional problem encompassing technology, security, governance, and regulation.

Digital Asset Custody Models and Architectures

The design of digital asset custody systems has evolved significantly in response to increasing security requirements, institutional adoption, and the growing complexity of blockchain ecosystems. At a fundamental level, custody architectures are concerned with how private keys are generated, stored, accessed, and used to authorize transactions. However, in practice, modern custody systems extend far beyond simple key storage, incorporating multi-layered security controls, distributed trust mechanisms, and operational workflows that ensure both security and efficiency. These architectures must balance competing requirements such as accessibility, performance, resilience, and regulatory compliance [12].

One of the most widely adopted approaches in digital asset custody is the use of tiered storage architectures, commonly referred to as hot, warm, and cold storage models. Rather than relying on a

single storage mechanism, institutional custody providers distribute assets across multiple layers based on their usage patterns and risk profiles. Hot storage is typically used for assets that require frequent access, such as those involved in active trading or liquidity provisioning. These wallets remain connected to the internet, enabling real-time transaction execution but also exposing them to higher security risks. In contrast, cold storage systems operate in offline or air-gapped environments, significantly reducing their attack surface. These systems are commonly used to store the majority of institutional holdings, where security takes precedence over immediacy. Warm storage serves as an intermediary layer, allowing limited and controlled access to assets while maintaining stronger security guarantees than hot wallets. By combining these storage models, custody providers can achieve a balance between operational flexibility and robust security [13].

Beyond storage strategies, custody architectures can also be classified based on how trust and control are distributed within the system. Centralized custody models, often employed by exchanges and financial institutions, rely on a single entity to manage private keys and enforce security policies. While this approach simplifies management and enables streamlined operations, it introduces a single point of failure and increases the risk of insider threats and large-scale breaches [14]. Decentralized and distributed custody models attempt to mitigate these risks by distributing key management responsibilities across multiple entities or systems. These approaches reduce reliance on any single trusted party and enhance resilience against both external and internal threats.

A critical component of modern custody architectures is the use of advanced cryptographic techniques to secure private keys and control transaction authorization. Multi-signature schemes represent one of the earliest and most widely used approaches in this context. In a multi-signature system, multiple private keys are associated with

a single address, and a predefined subset of these keys is required to authorize a transaction. This mechanism introduces redundancy and reduces the risk of single-key compromise, as an attacker would need to gain control over multiple keys to execute unauthorized transactions. Multi-signature schemes are particularly useful in organizational settings, where approval workflows can be enforced through distributed signing authority.

More recently, Multi-Party Computation (MPC) [15] has emerged as a powerful alternative to traditional key management approaches. Unlike multi-signature systems, which rely on multiple complete keys, MPC divides a private key into multiple cryptographic shares that are distributed across different parties or systems. These shares are never combined to reconstruct the full key; instead, cryptographic operations such as transaction signing are performed collaboratively through secure protocols. This approach significantly enhances security by ensuring that the private key is never fully exposed, even during transaction execution. Additionally, MPC enables flexible policy enforcement and dynamic access control, making it particularly suitable for institutional custody environments.

Another foundational technology in custody architectures is the Hardware Security Module (HSM) [16], which provides a tamper-resistant environment for generating and storing cryptographic keys. HSMs are widely used in traditional financial systems and have been adapted for digital asset custody to provide high levels of physical and logical security. These devices ensure that private keys never leave the secure hardware boundary and that cryptographic operations are performed in a controlled environment. While HSM-based systems offer strong security guarantees, they can be limited in flexibility and scalability, particularly when compared to distributed approaches such as MPC.

In addition to cryptographic mechanisms, modern custody architectures incorporate comprehensive operational workflows to manage asset movement and access control. Transaction approval processes often involve multiple layers of verification, including automated checks, policy enforcement rules, and manual approvals for high-value transfers. Access to key management systems is typically restricted through role-based access control (RBAC) [17], multi-factor authentication, and secure audit logging. These operational controls are essential for mitigating insider threats and ensuring compliance with regulatory requirements.

Another emerging aspect of custody architecture is the integration of real-time monitoring and anomaly detection systems. By leveraging machine learning and behavioral analytics, custody platforms can identify unusual transaction patterns, unauthorized access attempts, and potential security breaches. These systems enable proactive risk management and contribute to the overall resilience of custody infrastructures.

Ultimately, digital asset custody architectures represent a convergence of cryptographic innovation, system design, and operational governance. The increasing sophistication of threats, combined with the growing value of digital assets, necessitates the continuous evolution of these architectures. As institutional adoption accelerates, the demand for scalable, secure, and compliant custody solutions will continue to drive innovation in this domain.

Key Technologies in Digital Asset Custody

The security and reliability of digital asset custody systems are fundamentally dependent on the cryptographic and hardware-based technologies used to protect private keys and authorize transactions. As the value of digital assets has increased and threat actors have become more sophisticated, custody solutions have evolved from simple key storage mechanisms to complex, multi-

layered security infrastructures. Among the most prominent technologies in this domain are multi-signature schemes, Hardware Security Modules (HSMs), and Multi-Party Computation (MPC), each offering distinct advantages and trade-offs in terms of security, flexibility, and operational complexity.

Multi-signature (multisig) [18] schemes represent one of the earliest approaches to enhancing key security in blockchain systems. In a multisig configuration, a single blockchain address is associated with multiple private keys, and a predefined threshold of these keys must be used collectively to authorize a transaction. For example, in a “2-of-3” scheme, any two out of three key holders must sign a transaction before it can be executed. This model reduces reliance on a single point of failure and provides resilience against key compromise, as an attacker would need to gain access to multiple independent keys. Multisig also enables governance structures within organizations by distributing signing authority across different roles or departments. However, multisig implementations often require blockchain-level support and may introduce operational overhead, particularly in environments where rapid transaction execution is required. Additionally, key management complexity increases as the number of participants grows, and improper configuration can lead to accessibility issues.

Hardware Security Modules (HSMs) provide a different approach by focusing on the secure generation, storage, and use of cryptographic keys within tamper-resistant hardware devices. Originally developed for traditional financial systems, HSMs have been adapted for digital asset custody to offer strong protection against both physical and logical attacks. These devices are designed to ensure that private keys never leave the secure hardware boundary and that all cryptographic operations, such as transaction signing, are performed internally. HSMs typically include features such as secure key storage, access control enforcement, audit logging, and

resistance to side-channel attacks. Their use is particularly prevalent in institutional custody environments where regulatory compliance and auditability are critical. However, HSM-based systems are inherently centralized and may introduce scalability limitations. Furthermore, they can be costly to deploy and maintain, and their integration into decentralized blockchain environments may require additional architectural considerations.

Multi-Party Computation (MPC) [19] has emerged as a highly advanced and flexible cryptographic approach for digital asset custody. Unlike multisig schemes, which rely on multiple complete keys, MPC distributes a single private key into multiple cryptographic shares that are held by different parties or systems. These shares are generated in such a way that no single party can reconstruct the full key independently. Instead, cryptographic operations are performed collaboratively through secure protocols, enabling transaction signing without ever reconstructing the private key in a single location. This approach significantly enhances security by eliminating single points of failure and reducing the risk of key exposure during both storage and usage.

One of the key advantages of MPC is its flexibility in enforcing dynamic access control policies. For example, the threshold required for transaction approval can be adjusted without changing the underlying key structure, allowing organizations to implement adaptive governance models. Additionally, MPC can be integrated into cloud-native architectures, enabling distributed key management across geographically separated systems. This makes it particularly suitable for modern custody platforms that require scalability, resilience, and high availability. However, MPC systems are computationally more complex than traditional approaches and may introduce latency in transaction processing. The implementation of secure MPC protocols also requires careful

design to prevent vulnerabilities arising from protocol misconfiguration or implementation flaws.

In addition to these primary technologies, digital asset custody systems often incorporate complementary security mechanisms to enhance overall protection. Encryption techniques are used to secure data at rest and in transit, while secure enclaves and trusted execution environments (TEEs) [20] provide isolated environments for sensitive computations. Key sharding, secret sharing schemes such as Shamir's Secret Sharing, and hierarchical deterministic (HD) wallet structures are also commonly employed to improve key management and recovery capabilities. These mechanisms contribute to a layered security model, where multiple independent controls work together to reduce the likelihood and impact of security breaches.

When comparing multisig, HSM, and MPC approaches, it becomes evident that each technology addresses different aspects of the custody problem. Multisig provides a relatively simple and transparent method for distributing trust but may lack flexibility and scalability. HSMs offer strong hardware-based security and are well-suited for regulated environments, yet they introduce centralization and operational constraints. MPC, on the other hand, represents a more sophisticated and adaptable solution, combining strong security guarantees with distributed control, albeit at the cost of increased complexity. As a result, many modern custody platforms adopt hybrid architectures that integrate multiple technologies to leverage their respective strengths.

The selection of appropriate custody technologies ultimately depends on the specific requirements of the use case, including security level, transaction volume, regulatory constraints, and operational considerations. As digital asset ecosystems continue to evolve, the ongoing development and refinement of these

technologies will play a crucial role in ensuring the secure and efficient management of digital assets.

Threat Landscape and Security Vulnerabilities

The rapid growth of digital assets and their increasing financial value have made digital asset custody systems a primary target for a wide range of cyber threats and operational risks. Unlike traditional financial systems, where multiple layers of institutional protection and recovery mechanisms exist, digital asset ecosystems operate under a fundamentally different paradigm in which transactions are irreversible, and ownership is solely determined by cryptographic key control. As a result, any compromise of custody infrastructure can lead to immediate and permanent asset loss. This unique characteristic significantly amplifies the impact of security vulnerabilities and necessitates a comprehensive understanding of the evolving threat landscape.

One of the most prevalent categories of threats targeting digital asset custody systems is external cyberattacks. These attacks often exploit vulnerabilities in software, network infrastructure, or user interfaces to gain unauthorized access to private keys or transaction authorization mechanisms. Exchange hacks represent a prominent example, where attackers infiltrate centralized custody platforms and extract large volumes of digital assets. Such incidents are frequently facilitated by weaknesses in hot wallet management, insufficient access controls, or unpatched system vulnerabilities. In many cases, attackers leverage sophisticated techniques, including zero-day exploits and advanced persistent threats (APTs) [21], to bypass traditional security defenses.

In addition to external threats, insider attacks pose a significant risk to custody systems, particularly in custodial environments where key management responsibilities are concentrated within organizations. Malicious insiders, or employees

with privileged access, may intentionally misuse their authority to extract sensitive information or initiate unauthorized transactions. Even in the absence of malicious intent, human error and operational mismanagement can lead to severe security incidents, such as accidental key exposure or improper configuration of access controls. These risks highlight the importance of implementing strict governance frameworks, role-based access controls, and multi-layered approval mechanisms within custody infrastructures.

Phishing and social engineering attacks also represent a major threat vector, particularly for non-custodial users and retail participants. In these scenarios, attackers manipulate users into revealing private keys, seed phrases, or authentication credentials through deceptive communication channels. Malware-based attacks further exacerbate this risk by targeting endpoint devices, capturing keystrokes, or injecting malicious code into wallet applications. These attacks are particularly effective against software-based wallets and highlight the vulnerabilities associated with user-managed key storage.

Another critical category of threats involves protocol-level vulnerabilities and smart contract exploits. In decentralized finance (DeFi) ecosystems, custody is often embedded within smart contracts that manage asset flows programmatically. Flaws in smart contract logic, such as reentrancy vulnerabilities, integer overflows, or improper access control, can be exploited to drain funds or manipulate system behavior. Unlike traditional custody systems, where security controls are implemented at the infrastructure level, smart contract vulnerabilities are often immutable once deployed, making them particularly dangerous.

Key management failures represent an additional and often overlooked source of risk. These failures may arise from inadequate backup procedures, improper key generation practices, or insufficient redundancy mechanisms. For example, the loss of a

single private key in a non-custodial system can render assets permanently inaccessible, while improper implementation of multi-signature or MPC schemes can lead to deadlocks or unauthorized access. These risks underscore the importance of robust key lifecycle management, including secure generation, storage, rotation, and recovery procedures.

Furthermore, emerging threats such as supply chain attacks and vulnerabilities in third-party dependencies are becoming increasingly relevant. Custody systems often rely on external libraries, APIs, and infrastructure providers, creating potential attack surfaces that can be exploited indirectly. Compromised software updates or malicious dependencies can introduce backdoors into otherwise secure systems, highlighting the need for rigorous software validation and dependency management practices.

To address these diverse threats, modern custody solutions adopt a layered security approach, combining cryptographic safeguards, hardware-based protections, operational controls, and real-time monitoring systems. However, as attackers continue to evolve their techniques, the security of digital asset custody systems remains an ongoing challenge that requires continuous adaptation and innovation.

Regulatory Challenges and Future Research Directions

The regulatory landscape surrounding digital asset custody is rapidly evolving, reflecting the growing importance of digital assets within the global financial system. Governments and regulatory authorities are increasingly recognizing the need to establish frameworks that ensure the security, transparency, and integrity of digital asset services, including custody providers. However, the decentralized and borderless nature of blockchain technologies presents significant challenges for regulatory harmonization and enforcement.

One of the primary challenges in regulating digital asset custody lies in the classification of digital assets themselves. Different jurisdictions adopt varying definitions, categorizing digital assets as commodities, securities, currencies, or entirely new asset classes. This lack of consensus complicates the development of standardized regulatory frameworks and creates uncertainty for custody providers operating across multiple regions. The classification of assets directly influences licensing requirements, compliance obligations, and the legal responsibilities of custodians.

Another critical regulatory issue concerns the licensing and supervision of custody providers. In traditional financial systems, custodians are subject to strict regulatory oversight, including capital requirements, risk management standards, and audit obligations. Extending these requirements to digital asset custodians presents unique challenges due to the technological complexity and diversity of custody models. Regulators must account for differences between custodial and non-custodial systems, as well as the use of advanced cryptographic techniques such as MPC and multi-signature schemes. Ensuring that these systems meet regulatory standards without stifling innovation remains a delicate balance.

Data protection and privacy regulations also play a significant role in shaping digital asset custody practices. Custody providers often handle sensitive user information, including identity data and transaction histories, which must be protected in accordance with legal frameworks such as data protection laws. At the same time, the transparency of blockchain systems can conflict with privacy requirements, creating tension between regulatory compliance and the inherent properties of distributed ledgers. Addressing this tension requires the development of privacy-enhancing technologies and regulatory approaches that accommodate the unique characteristics of blockchain systems.

Cross-border regulatory inconsistencies further complicate the custody landscape. Digital asset transactions are inherently global, yet regulatory frameworks are typically defined at the national or regional level. This discrepancy creates challenges for custody providers seeking to operate internationally, as they must navigate a complex web of overlapping and sometimes conflicting regulations. Issues such as jurisdictional authority, asset custody standards, and legal liability remain areas of ongoing debate.

In addition to regulatory challenges, several open research directions are emerging in the field of digital asset custody. One promising area is the integration of advanced cryptographic techniques, such as threshold signatures and post-quantum cryptography, to enhance the long-term security of custody systems. As quantum computing technologies advance, existing cryptographic schemes may become vulnerable, necessitating the development of quantum-resistant solutions.

Another important research direction involves the application of artificial intelligence and machine learning to custody systems. AI-driven monitoring and anomaly detection can enhance the ability to identify suspicious activities and prevent security breaches in real time. Additionally, intelligent automation can improve operational efficiency and reduce the risk of human error in key management processes.

The concept of decentralized custody is also gaining traction, with emerging solutions aiming to eliminate reliance on centralized intermediaries while maintaining high levels of security and usability. These approaches often leverage distributed trust models, combining elements of MPC, blockchain-based governance, and decentralized identity systems. However, achieving scalability, usability, and regulatory compliance in such systems remains an open challenge.

Finally, the development of standardized frameworks and interoperability protocols represents a key area for future research. As the digital asset ecosystem continues to expand, the ability of different custody systems to interact seamlessly will become increasingly important. Standardization efforts can facilitate integration, improve security practices, and support regulatory compliance across the industry.

In conclusion, digital asset custody sits at the intersection of technology, security, and regulation, presenting a complex and rapidly evolving research domain. While significant progress has been made in developing secure and scalable custody solutions, ongoing challenges related to security threats, regulatory uncertainty, and technological innovation continue to shape the future of this field. Addressing these challenges will require collaborative efforts between academia, industry, and regulatory bodies to ensure the safe and sustainable growth of digital asset ecosystems.

References

[1] Gramlich, V., Guggenberger, T., Principato, M. et al. A multivocal literature review of decentralized finance: Current knowledge and future research avenues. Electron Markets 33, 11 (2023). https://doi.org/10.1007/s12525-023-00637-4

[2] S. I. Saifullah, S. Islam, M. S. Ferdous and F. Chowdhury, "Non-Fungible Token (NFT): Analyzing Marketplaces and Non-User Perspectives," 2022 25th International Conference on Computer and Information Technology (ICCIT), Cox's Bazar, Bangladesh, 2022, pp. 1044-1051, doi: 10.1109/ICCIT57492.2022.10054719.

[3] Tuckett, D. (2011). The Special Characteristics of Financial Assets. In: Minding the Markets. Palgrave Macmillan, London. https://doi.org/10.1057/9780230307827_1

[4] Khurram, A., Iqbal, A. & Pappas, V. Systemic risk: new evidence from alternative financial systems. Rev Quant Finan Acc 66, 731–755 (2026). https://doi.org/10.1007/s11156-025-01413-5

[5] K. Gai, D. Wang, J. Yu, L. Zhu and W. Meng, "A Scheme of Robust Privacy-Preserving Multi-Party Computation via Public Verification," in IEEE Transactions on Dependable and Secure Computing, vol. 22, no. 5, pp. 4896-4910, Sept.-Oct. 2025, doi: 10.1109/TDSC.2025.3555284.

[6] Zhang, S., Qu, G., Zhang, Z. et al. Efficient and secure multi-party computation protocol supporting deep learning. Cybersecurity 8, 46 (2025). https://doi.org/10.1186/s42400-024-00343-4

[7] Donoghue, Seamus. (2024). Custody in the age of digital assets : The path to building market infrastructure fit for a tokenised

economy. Journal of Securities Operations & Custody. 16. 10.69554/LHMW1512.

[8] Kukman, T., & Gričar, S. (2025). Blockchain for Quality: Advancing Security, Efficiency, and Transparency in Financial Systems. FinTech, 4(1), 7. https://doi.org/10.3390/fintech4010007

[9] Cheng, Y., Liu, Y., Zhang, Z., & Li, Y. (2023). An Asymmetric Encryption-Based Key Distribution Method for Wireless Sensor Networks. Sensors, 23(14), 6460. https://doi.org/10.3390/s23146460

[10] Babel, M., Willburger, L., Lautenschlager, J. et al. Self-sovereign identity and digital wallets. Electron Markets 35, 28 (2025). https://doi.org/10.1007/s12525-025-00772-0

[11] Erinle, Y., Feng, Y., Xu, J., Vadgama, N., & Tasca, P. (2025). Shared-Custodial Wallet for Multi-Party Crypto-Asset Management. Future Internet, 17(1), 7. https://doi.org/10.3390/fi17010007

[12] Arshadi, N., & Dombrowski, T. (2026). Applications and Management of Blockchain Technologies in Financial Services. Journal of Risk and Financial Management, 19(3), 224. https://doi.org/10.3390/jrfm19030224

[13] Jiang, S. (2025). Big Data Sharing: A Comprehensive Survey. Data, 10(11), 182. https://doi.org/10.3390/data10110182

[14] Diana, L., Dini, P., & Paolini, D. (2025). Overview on Intrusion Detection Systems for Computers Networking Security. Computers, 14(3), 87. https://doi.org/10.3390/computers14030087

[15] Borges, R., Ferreira, B., Antunes, C. M., Maximiano, M., Gomes, R., Távora, V., Dias, M., Bezerra, R. C., & Domingues, P. (2026). Using Secure Multi-Party Computation to Create Clinical Trial Cohorts. Journal of Cybersecurity and Privacy, 6(1), 2. https://doi.org/10.3390/jcp6010002

[16] V. A. Kokovin, A. A. Evsikov, A. N. Sytin, V. V. Skvortsov and S. U. Uvaysov, "Development and Research of a Hardware Security Module to Control and Protect Access to Industrial Equipment," 2024 International Seminar on Electron Devices Design and Production (SED), Sochi, Russian Federation, 2024, pp. 1-5, doi: 10.1109/SED63331.2024.10741050.

[17] Singh, Jaibir & Rani, Suman & Kumar, Vipin. (2024). Role-Based Access Control (RBAC) Enabled Secure and Efficient Data Processing Framework for IoT Networks. International Journal of Communication Networks and Information Security (IJCNIS). 10.17762/ijcnis.v16i2.6697.

[18] Zhang, P., Ge, F., Tang, Z., & Xie, W. (2025). Achieving High Efficiency in Schnorr-Based Multi-Signature Applications in Blockchain. Electronics, 14(9), 1883. https://doi.org/10.3390/electronics14091883

[19] Gamiz, I., Regueiro, C., Lage, O. et al. Challenges and future research directions in secure multi-party computation for resource-constrained devices and large-scale computations. Int. J. Inf. Secur. 24, 27 (2025). https://doi.org/10.1007/s10207-024-00939-4

[20] M. Sabt, M. Achemlal and A. Bouabdallah, "Trusted Execution Environment: What It is, and What It is Not," 2015 IEEE Trustcom/BigDataSE/ISPA, Helsinki, Finland, 2015, pp. 57-64, doi: 10.1109/Trustcom.2015.357.

[21] Erdodi, L., Abraham, D., & Houmb, S. H. (2025). Improving Detectability of Advanced Persistent Threats (APT) by Use of APT Group Digital Fingerprints. Information, 16(9), 811. https://doi.org/10.3390/info16090811


BİDGE