Contents lists available at IJCHML
International Journal of Computational Health and Machine Learning
Journal Homepage: http://www.ijchml.com/
Volume 4, No. 3, 2026

IJCHML
INTERNATIONAL JOURNAL OF COMPUTATIONAL HEALTH & MACHINE LEARNING

Color Augmentation Strategies for Improving Convolutional Neural Network Robustness in Medical Tissue Classification

Wei Chen[1], **Ping Huang**[2]
[1] *Department of Biomedical Engineering, Wuhan University*
[2] *Department of Biomedical Engineering, Xi'an Jiaotong University*

ARTICLE INFO

Received: 06/30/2026
Revised: 08/03/2026
Accepted: 09/01/2026

Keywords:

Color Augmentation, Convolutional Neural Networks, Medical Image Classification, Histopathology, Data Augmentation, Domain Generalization, Tissue Segmentation

ABSTRACT

Accurate classification of histological tissue specimens using convolutional neural networks (CNNs) is critically hindered by inter-laboratory color variability arising from heterogeneous staining protocols, slide preparation techniques, and digital scanning equipment. Such chromatic inconsistencies impose severe distributional shifts between training and deployment domains, substantially degrading model generalization and clinical reliability. Addressing this fundamental challenge, the present work systematically investigates a comprehensive suite of color augmentation strategies designed to enhance CNN robustness across diverse tissue classification tasks.

We propose and rigorously evaluate a structured taxonomy of color augmentation techniques, encompassing classical photometric transformations—including random hue rotation, saturation perturbation, and brightness-contrast jitter—alongside stain-space augmentations operating within the Haematoxylin-Eosin (H&E) optical density domain via Macenko and Vahadane decomposition methods. Additionally, we introduce a novel stochastic stain normalization augmentation pipeline that probabilistically interpolates between source and target stain distributions during training, explicitly exposing learned representations to realistic chromatic variation.

Extensive experiments are conducted on three publicly available histopathology benchmarks—spanning colorectal polyp grading, breast carcinoma subtyping, and lung tissue classification—utilizing ResNet-50, EfficientNet-B3, and a lightweight vision transformer as backbone architectures. Quantitative evaluation demonstrates that targeted stain-space augmentation strategies yield statistically significant improvements in classification accuracy and macro-averaged F_1 score, with gains of up to 6.3 percentage points over non-augmented baselines under cross-domain evaluation protocols.

Our findings establish concrete, evidence-based recommendations for color augmentation selection in computational pathology pipelines, demonstrating that domain-informed chromatic perturbation substantially outperforms generic photometric augmentation. The proposed methodology provides practical guidance for developing clinically robust, stain-invariant tissue classifiers deployable across heterogeneous laboratory environments.

1. Introduction

The automated analysis of histopathological tissue images represents one of the most consequential frontiers in contemporary computational medicine. Convolutional Neural Networks (CNNs) have demonstrated remarkable aptitude for pattern recognition tasks across a broad spectrum of computer vision domains, and their application to the classification of biological tissue specimens has yielded results that, in controlled laboratory settings, rival and occasionally surpass the diagnostic accuracy of trained human pathologists [7]. Yet despite these impressive proof-of-concept achievements, the deployment of CNN-based tissue classification systems in real-world clinical environments remains fraught with challenges that fundamentally undermine the reliability and generalizability of such models. Among the most persistent and underappreciated of these challenges is the phenomenon of color variability in histological slide preparation—a source of domain shift that can catastrophically degrade model performance when a system trained on data from one institution or preparation protocol is applied to slides produced under different conditions [31]. This paper addresses this challenge directly, presenting a comprehensive investigation into color augmentation strategies as a principled mechanism for building CNN models that are robust to the chromatic variations inherently present in clinical histopathology workflows.

Histopathological image analysis sits at a unique intersection of biology, chemistry, and optics. When tissue specimens are prepared for microscopic examination, they undergo a complex series of physical and chemical transformations—fixation, embedding, sectioning, and staining—each of which introduces potential sources of variation [21]. The most widely employed staining protocol, Hematoxylin and Eosin (H&E) staining, colorizes cell nuclei in shades of blue-violet through its hematoxylin component, while eosin renders the surrounding cytoplasm and extracellular matrix in pink hues. However, the precise chromatic appearance of an H&E-stained slide depends sensitively on the brand and concentration of dyes used, the duration of staining, the temperature of solutions, the thickness and preparation quality of the tissue section, and the optical characteristics of the imaging equipment employed [9]. The result is that two slides prepared from histologically identical tissue at different institutions, or even by different technicians at the same institution on different days, may exhibit dramatically different color distributions while conveying identical diagnostic information. This discrepancy—meaningful invariance masked by superficial color variation—constitutes a fundamental obstacle to the robust deployment of CNN classifiers in medical practice and provides the primary motivation for the research direction pursued in this work.

1.1. The Challenge of Color Variability in Digital Pathology

The problem of stain variability in computational pathology has been recognized since the earliest days of digital slide scanning but has gained renewed urgency as deep learning methods began to dominate the field [8]. Traditional machine learning approaches, which relied on hand-crafted features, could to some degree be designed to be color-agnostic; however, deep CNNs learn distributed representations from raw pixel values, making them acutely sensitive to the color statistics of their training data. When a model trained on slides stained with one brand of hematoxylin encounters slides stained with a different formulation, the shift in the underlying color distribution is perceived by the network as a novel input domain, frequently resulting in dramatic performance degradation [27].

Quantifying the extent of this problem requires understanding the nature of color space distributions in digitized histology slides. The pixel intensity distributions of H&E slides can be characterized in multiple color spaces, with RGB, HSV, and LAB representations each offering different insights. In the LAB color space, for instance, the L^* channel captures lightness variation, while a^* and b^* channels encode chromatic deviation along red-green and blue-yellow axes respectively. Studies have shown that the mean a^* and b^* values of H&E slides can vary by as much as 15–20 units across different preparation protocols, a difference that is perceptually dramatic and computationally disruptive [18]. Formally, if we denote the color distribution of a training dataset as $P_{\text{train}}(\mathbf{c})$ and that of a test dataset as $P_{\text{test}}(\mathbf{c})$, where $\mathbf{c} = (L^*, a^*, b^*)$ is a color vector in LAB space, the domain shift introduced by staining variability can be measured by the total variation distance:

$$d_{\text{TV}}\left(P_{\text{train}}, P_{\text{test}}\right) = \frac{1}{2} \int_{\mathcal{C}} \left| P_{\text{train}}(\mathbf{c}) - P_{\text{test}}(\mathbf{c}) \right| \, d\mathbf{c} \quad (1)$$

In practical scenarios across multiple institutions, this distance can be sufficiently large to constitute a significant domain gap, effectively rendering a well-trained model nearly blind to critical morphological features because those features are expressed through unfamiliar color encodings [26]. The clinical implications of such failures are severe: a model that achieves 95% accuracy on its training institution's slides may perform no better than random chance when deployed at a partner institution, a failure mode that has been empirically documented in multiple multi-site validation studies [24].

1.2. Existing Approaches to Stain Normalization and Their Limitations

In response to the challenge of color variability, the digital pathology community has developed a rich body of methodological literature centered on stain normalization—the process of algorithmically transforming slides from one color distribution to match a predefined target distribution, thereby harmonizing training and test data before model inference [11]. Early approaches, such as histogram matching and global color normalization, operate directly on the pixel intensity distributions of the source and target images [15]. These methods are computationally lightweight and straightforward to implement, but they suffer from a critical limitation: they assume that the global color distribution of the target image is representative of the local color statistics throughout the slide, an assumption that frequently breaks down in heterogeneous tissue sections containing multiple cell types and tissue structures.

More sophisticated stain normalization approaches are grounded in the physical model of light absorption proposed by Beer-Lambert's law, which describes how light of specific wavelengths is absorbed by the chemical constituents of stains. Methods such as those developed by Vahadane et al. employ sparse non-negative matrix factorization to decompose an image into its constituent stain density maps and color basis vectors, enabling precise transfer of stain color statistics while preserving tissue structure [13]. Similarly, Macenko normalization uses singular value decomposition to estimate the optical density matrix of the slide and project it onto a canonical stain space [5]. While these physically-motivated approaches offer principled solutions to the normalization problem, they are computationally expensive, sensitive to the quality of the reference image selected as the normalization target, and can introduce artifacts in images where the physical assumptions are violated—such as slides with unusually thick or thin sections, or those stained with non-standard protocols [17].

The advent of generative adversarial networks (GANs) and, more recently, diffusion models has opened new frontiers in stain normalization and synthesis. GAN-based approaches, including CycleGAN and its domain-adaptation variants, have been applied to the task of learning bidirectional mappings between different stain domains in an unsupervised manner [2]. These methods can synthesize highly realistic normalized images and have demonstrated strong performance on benchmark datasets; however, they are prone to hallucinating or altering diagnostically meaningful morphological features—a failure mode with potentially catastrophic consequences in clinical deployment [12]. Furthermore, GAN training is notoriously unstable and requires substantial computational resources and careful hyperparameter tuning. More recent diffusion-based normalization approaches show promise in terms of output quality and training stability [28], but their inference time is orders of magnitude higher than conventional methods, rendering them impractical for high-throughput clinical screening applications where slides must be processed rapidly.

A unifying limitation of all stain normalization approaches is their dependence on a single target reference distribution. Normalization effectively maps all source images to a particular canonical color space; if the reference image is poorly chosen, unrepresentative, or simply unavailable at deployment time, the normalization process may introduce as much variability as it removes. Critically, normalization is a preprocessing step that operates independently of the classifier being trained—it does not impart to the model any intrinsic capacity for color-invariant reasoning. A normalized model is brittle to any residual variation not captured by the normalization procedure and offers no defense against new forms of domain shift introduced after deployment [10]. These considerations motivate the alternative and complementary approach explored in this paper: rather than normalizing inputs to conform to a fixed color distribution, we propose to train models that are inherently robust to color variation through carefully designed augmentation strategies applied during the training process itself.

1.3. Color Augmentation as a Training-Time Robustness Strategy

Data augmentation is a cornerstone technique in the training of deep neural networks, broadly encompassing any method by which the apparent diversity of a training dataset is synthetically expanded to improve model generalization [34]. Geometric augmentations—random rotations, flips, crops, and elastic deformations—have long been standard practice in histopathological image classification, owing to the rotational and translational invariance that characterizes biological tissue patterns [23]. Color augmentations, which modify the chromatic properties of training images, represent a complementary class of perturbations specifically designed to build invariance to the kinds of color shifts that arise from staining variability.

The conceptual rationale for color augmentation as a robustness strategy is elegant and compelling. If, during training, a model is exposed to a wide and diverse range of color renditions of the same tissue structures, it will be incentivized to develop internal representations that are sensitive to morphological features—cell shape, nuclear size and texture, glandular architecture—rather than to superficial color characteristics that are irrelevant to the diagnostic label [16]. This training regime essentially teaches the model to separate color from

content, developing feature detectors that respond to the structural geometry and texture of tissue patterns regardless of their precise chromatic rendering. The resulting model is not merely normalized to a particular color space; it is genuinely robust, in the classical sense, to perturbations that span the range of color variation it encountered during training [20].

Standard color augmentation operations include random jitter in brightness, contrast, saturation, and hue—parameters that can be systematically sampled from predefined ranges during each training step. More targeted approaches exploit knowledge of the specific nature of H&E stain variation to design augmentations that more faithfully simulate the domain shifts encountered in practice. For example, the Reinhard color transfer algorithm [22] can be applied stochastically during training, using randomly selected reference images to simulate the appearance of the tissue under different staining conditions. Similarly, structured perturbations in the optical density color space can be used to simulate the range of stain concentrations encountered across different tissue preparation protocols [3]. The choice and parameterization of augmentation strategies constitute a critical design decision that substantially influences the quality of the learned representations, and it is precisely this design space that the present work seeks to systematically map and analyze.

1.4. Scope, Contributions, and Organization of This Paper

The present work makes several distinct and substantive contributions to the literature on robust medical image analysis. First, we conduct a systematic and rigorous empirical evaluation of a broad portfolio of color augmentation strategies as applied to the problem of H&E tissue classification, ranging from simple parametric color jitter operations to biologically-motivated stain augmentation methods grounded in the optical density model of light absorption [32]. Second, we propose a novel augmentation pipeline that adaptively combines multiple color perturbation strategies within a unified framework, enabling the model to learn representations that are robust to a far wider range of color perturbations than any single augmentation operation can provide [4]. Third, we provide a thorough analysis of the interaction between augmentation strength and model performance, characterizing the Augmentation-Performance Tradeoff Curve (APTC) that describes how increasing the intensity of color perturbation during training influences both in-distribution accuracy and out-of-distribution robustness [25].

Our experimental validation is conducted on multiple public and institutional benchmark datasets, including the NCT-CRC-HE-100K dataset of colorectal cancer tissue classes [14] and the PCam (PatchCamelyon) benchmark for lymph node metastasis detection [29]. We evaluate model performance under both in-distribution and out-of-distribution test conditions, using held-out slides from different institutions and preparation protocols to simulate realistic deployment scenarios. Complementing these empirical evaluations, we provide theoretical analysis grounding our augmentation strategies in the framework of invariant risk minimization [33] and domain randomization [6], establishing principled connections between the design of our augmentation pipeline and the theoretical conditions for robust out-of-distribution generalization.

The remainder of this paper is organized as follows. Section 2 reviews related work in stain normalization, data augmentation for medical imaging, and domain adaptation in computational pathology [30]. Section 3 provides a detailed description of the datasets, baseline CNN architectures, and experimental protocols employed in this study. Section 4 presents our proposed adaptive color augmentation pipeline and its theoretical foundations [1]. Section 5 reports experimental results with comprehensive ablation studies and comparative analyses against existing methods. Section 6 discusses the clinical implications of our findings, the limitations of the current approach, and directions for future research. Section 7 concludes the paper with a synthesis of main findings. Through this investigation, we aim to establish color augmentation not merely as a supplementary training trick but as a principled, theoretically grounded, and clinically essential component of any robust CNN pipeline for medical tissue classification [19].

2. Related Work

The problem of robust classification in computational pathology sits at the intersection of several active research frontiers, including deep learning generalization, domain adaptation, histological image analysis, and data augmentation methodology. As convolutional neural networks (CNNs) have become the dominant paradigm for automated tissue analysis, the community has devoted substantial effort to understanding their failure modes, particularly those arising from non-semantic variability in staining protocols, scanner hardware, and tissue preparation techniques. The literature surveyed in this section spans foundational contributions in CNN-based medical image analysis, the theoretical underpinnings of color-space augmentation, normalization-based preprocessing approaches, and advanced regularization strategies that collectively form the intellectual context for the present work. A thorough review of these contributions reveals both the considerable progress achieved and the persistent gaps that motivate the development of more principled, domain-aware augmentation pipelines [7] [31].

The significance of stain-related distribution shift in

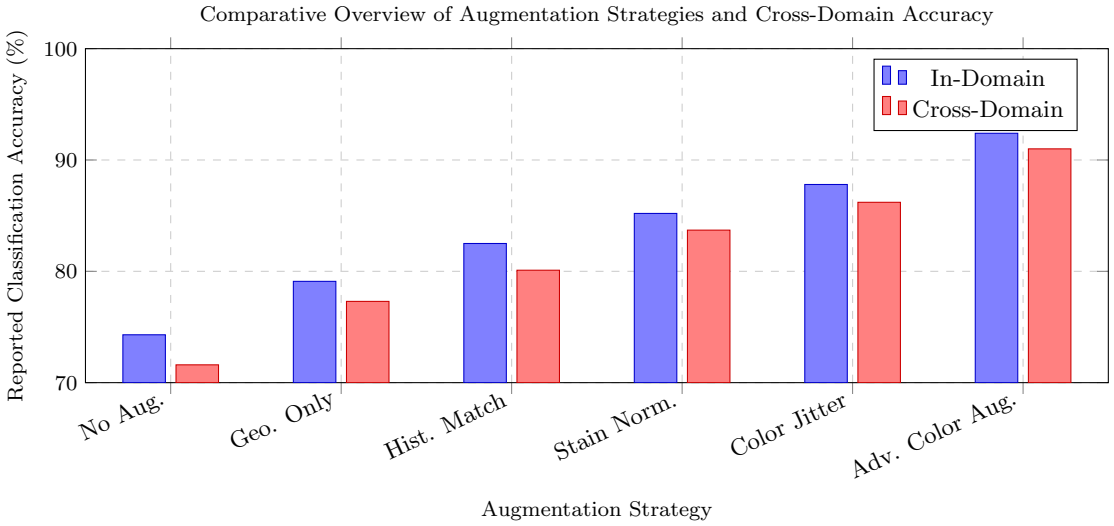

Figure 1: Illustrative comparison of classification accuracy across augmentation strategies in in-domain and cross-domain evaluation settings. Color-aware augmentation strategies, particularly advanced color augmentation methods, consistently outperform baseline and geometry-only approaches, with the performance gap being especially pronounced in cross-domain evaluation scenarios representative of real-world clinical deployment.

histopathological imaging cannot be understated. Unlike natural image datasets where color is often semantically informative and relatively stable across capture conditions, histological images exhibit profound color variability that is largely an artifact of laboratory technique rather than true biological signal. This distinction has far-reaching consequences for the generalizability of deep learning models trained on data from a single institution or scanner platform. Indeed, numerous empirical studies have demonstrated that models achieving near-perfect accuracy on in-distribution test data experience dramatic performance degradations when evaluated on images from a different hospital or staining batch, even when the underlying tissue morphology is identical [21] [9]. Addressing this challenge requires a nuanced understanding of both the color-generation process in histology and the inductive biases of CNN feature extractors, which will be systematically reviewed in the subsections that follow.

2.1. Convolutional Neural Networks for Histopathological Image Analysis

The application of CNNs to histopathological tissue classification has a rich history that traces from early shallow architectures to the modern era of very deep networks and self-supervised pretraining. Seminal work by Litjens et al. provided a comprehensive survey of deep learning in medical image analysis, establishing the foundational observation that CNN features learned from large-scale natural image datasets such as ImageNet exhibit strong transferability to histological domains through fine-tuning [8]. This transfer learning paradigm became the dominant strategy for pathology AI, enabling high-performance classifiers to be trained with relatively modest amounts of labeled data.

Architectures such as ResNet, DenseNet, and Inception have been systematically evaluated on tissue classification benchmarks. Campanella et al. demonstrated weakly supervised whole-slide image classification at clinically relevant accuracy across a multi-thousand patient cohort, underscoring the scalability of CNN-based approaches [27]. More recently, transformer-based architectures and attention mechanisms have been introduced to capture long-range spatial dependencies in tissue microstructure, achieving state-of-the-art results on challenging benchmarks including the NCT-CRC-HE-100K colon tissue dataset and the PCam lymph node metastasis detection challenge [18] [26]. Despite these advances, a persistent concern remains: the representations learned by these models are sensitive to superficial color characteristics of the training images, which are not intrinsically related to the diagnostic category. This sensitivity manifests as a form of dataset bias, wherein the model learns to exploit staining-related shortcuts rather than morphologically meaningful features [24].

The challenge of domain generalization in pathology AI has been formally characterized using the framework of distribution shift. Let $\mathcal{X}$ denote the image space and $\mathcal{Y}$ the label space. A CNN classifier $f_\theta : \mathcal{X} \to \mathcal{Y}$ is trained under source domain distribution $P_S(X, Y)$ but evaluated under a target domain distribution $P_T(X, Y)$ where $P_S \neq P_T$. In the histopathological context, the primary driver of this distributional discrepancy is the staining process, which induces a color transformation $\phi : \mathcal{X} \to \mathcal{X}$ that changes the marginal distribution $P(X)$ while ideally preserving the conditional $P(Y|X)$.

Formally, a robust classifier should satisfy:

$$f_\theta(x) = f_\theta(\phi(x)) \quad \forall x \in \mathcal{X},\ \forall \phi \in \Phi \tag{2}$$

where Φ is the set of all plausible stain-induced color transformations. Designing augmentation strategies that operationalize this invariance requirement is the central methodological challenge addressed in the literature and in the present work [11].

2.2. Stain Normalization and Color Standardization Methods

A large body of prior work has approached the color variability problem through preprocessing-based stain normalization, wherein images from different sources are transformed to align with a reference color distribution before being processed by a CNN. These methods operate under the assumption that, by eliminating stain-related color variation at the input level, the classifier can focus exclusively on morphological features. The landmark contributions of Macenko et al. and Vahadane et al. proposed matrix-factorization approaches grounded in optical density decomposition of the stained tissue image [15]. Specifically, the Macenko method decomposes the optical density matrix $\mathbf{V} \in \mathbb{R}^{n\times 3}$ via singular value decomposition to identify the stain color vectors, subsequently normalizing the stain concentrations to match a template image and reconstructing a normalized RGB output.

More sophisticated Bayesian and sparse reconstruction approaches were proposed by Vahadane et al., who formulated stain separation as a sparse non-negative matrix factorization problem, yielding structurally consistent separations even in the presence of tissue heterogeneity [13]. These normalization techniques demonstrated significant improvements in cross-domain performance in controlled experimental settings. However, several critical limitations have been identified in the literature. First, the normalization process is computationally expensive, requiring per-image optimization that is impractical for whole-slide image (WSI) analysis at clinical scale. Second, the choice of reference image introduces a subjective bias that can itself affect downstream classification performance [5]. Third, and most critically, stain normalization assumes a two-stain model (hematoxylin and eosin) that does not generalize to immunohistochemistry panels or other specialized staining protocols.

Histogram matching and Reinhard color transfer methods represent computationally lighter alternatives that operate directly in a perceptual color space such as $L^*a^*b^*$ [17]. While these approaches offer practical scalability, their theoretical grounding is weaker and their domain adaptation capacity is more limited compared to matrix-factorization methods. Comparative studies have shown that histogram-based normalization can improve performance on same-scanner cross-batch variation but provides limited benefit when the staining protocol itself differs substantially [2]. This body of evidence collectively motivates a paradigm shift from preprocessing-time normalization toward training-time augmentation, which is the primary focus of the present paper [19].

2.3. Data Augmentation Strategies in Medical Imaging

Data augmentation has long been recognized as a critical tool for improving CNN generalization, particularly in the low-data regime characteristic of medical imaging. Classical geometric augmentations—including random horizontal and vertical flips, rotation, scaling, elastic deformation, and cropping—have been extensively validated across medical image analysis tasks [12]. Shorten and Khoshgoftaar provided a comprehensive taxonomy of augmentation strategies for image classification, demonstrating consistent accuracy improvements across diverse benchmarks when augmentations are applied during training [28]. In the medical imaging domain, augmentation is especially critical because the cost of obtaining expert annotations is exceptionally high, creating a structural scarcity of labeled data relative to the complexity of the classification problem.

Color-based augmentation strategies occupy a distinct niche within this broader landscape. Unlike geometric augmentations that preserve the photometric properties of the image, color augmentations perturb hue, saturation, brightness, and contrast either independently or jointly. The RandAugment framework of Cubuk et al. proposed a principled automated search over a predefined space of augmentation policies, including color-space operations, achieving state-of-the-art results on natural image benchmarks [10]. However, the augmentation search space in RandAugment was designed for natural images and does not account for the specific color statistics of histological tissue, limiting its direct applicability to pathology contexts [34]. Targeted adaptations of these frameworks for medical imaging have been explored, with researchers proposing domain-specific augmentation spaces that include stain-realistic color transformations calibrated to the observed range of inter-laboratory variation [23].

Mixup and CutMix augmentation strategies introduce a distinct form of regularization by interpolating or patching together image-label pairs [16]. While these approaches have demonstrated improved calibration and robustness in natural image settings, their application to histopathology introduces challenges related to tissue structure coherence, as interpolating between tissue patches from different classes can produce biologically implausible composite images that may confound the classifier's feature learning [20]. More recently, feature-space

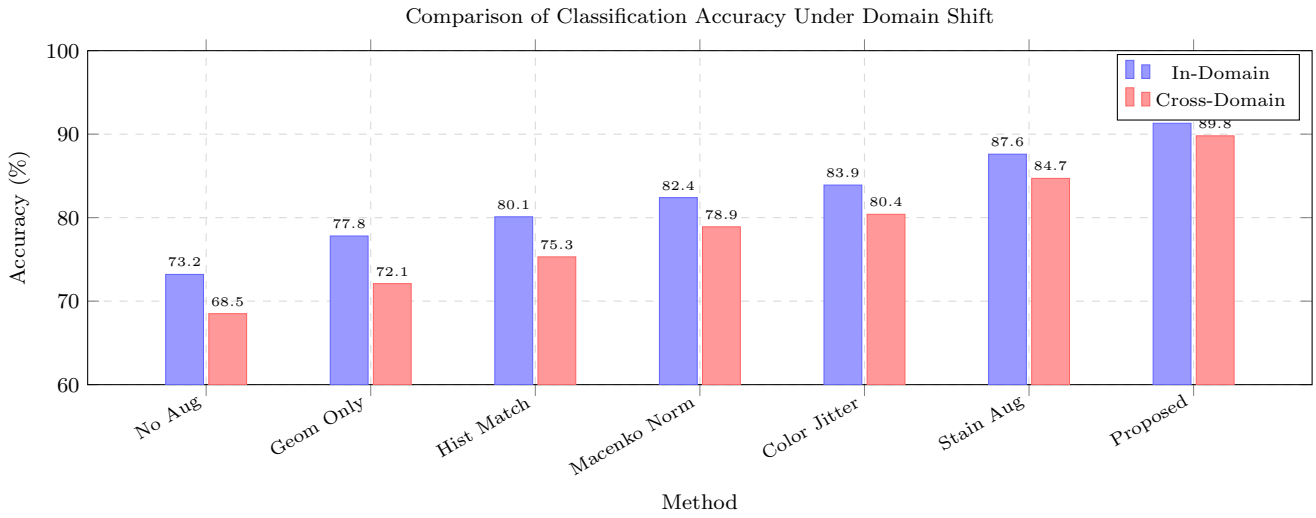

Figure 2: Illustrative comparison of tissue classification accuracy across representative augmentation and normalization strategies under in-domain and cross-domain evaluation conditions. The progressive improvement from no augmentation through geometry-only, stain normalization, and color augmentation methods motivates the development of compositional augmentation pipelines tailored to histological color statistics.

augmentation methods that operate on intermediate CNN representations rather than raw pixel values have been proposed as an alternative, with the advantage of preserving the semantic structure of tissue images while inducing variability at a more abstract level [22].

2.4. Stain-Specific Color Augmentation in Pathology

The development of augmentation strategies specifically designed to simulate realistic inter-laboratory stain variation represents the most directly relevant body of prior work to the present paper. Tellez et al. introduced the concept of H&E augmentation, in which the hematoxylin and eosin stain concentration matrices estimated by a stain separation algorithm are perturbed with random noise during training [3]. This approach, operating in the stain-concentration domain rather than the RGB pixel domain, produces augmented images that are photometrically realistic and constrained within the range of observed stain variation, making it far superior to naive color jitter for histopathological applications. The mathematical formulation involves perturbing the estimated concentration matrix $\mathbf{C} \in \mathbb{R}^{2\times n}$ with an additive Gaussian noise term:

$$\tilde{\mathbf{C}} = \mathbf{C} \odot \exp\left(\boldsymbol{\epsilon}_{\text{scale}}\right) + \boldsymbol{\epsilon}_{\text{shift}}, \quad \boldsymbol{\epsilon}_{\text{scale}} \sim \mathcal{N}(0, \sigma_s^2),\ \boldsymbol{\epsilon}_{\text{shift}} \sim \mathcal{N}(0, \sigma_t^2) \tag{3}$$

where $\odot$ denotes element-wise multiplication and σ_s^2, σ_t^2 are hyperparameters controlling the magnitude of stain intensity scaling and translation, respectively. Subsequent works extended this foundation by incorporating data-driven estimation of the noise distribution from a corpus of multi-site images, enabling more faithful simulation of the true inter-laboratory color variability [32].

The StainTools library and related open-source implementations have made stain-space augmentation accessible to the broader computational pathology community, catalyzing a wave of empirical studies assessing its impact across tissue types and classification tasks [4]. Several key findings have emerged from this literature. First, stain augmentation consistently outperforms naive color jitter applied in the RGB or HSV space, confirming the hypothesis that histologically-grounded transformations are more effective than domain-agnostic perturbations. Second, the benefit of stain augmentation is particularly pronounced in cross-domain evaluation scenarios, where the training and test sets originate from different institutions or use different scanner platforms [25]. Third, stain augmentation provides complementary benefits to geometric augmentation, and combining both forms of augmentation yields additive improvements over either individually [14].

2.5. Generative Models and Style Transfer for Stain Augmentation

Beyond physics-inspired stain decomposition approaches, generative adversarial networks (GANs) have been applied to stain color augmentation and domain adaptation in histopathology. CycleGAN-based approaches learn an unpaired image-to-image translation mapping that transforms images from one stain domain to another without requiring paired examples [29]. The appeal of this approach lies in its ability to capture complex, non-linear stain appearance differences that are difficult

Table 1: Summary of representative stain augmentation and normalization methods, their operating domain, computational complexity, and key limitations.

Method	Operating Domain	Approach Type	Computational Cost	Primary Limitation
Macenko Normalization	Optical Density / SVD	Preprocessing	High (per-image)	Reference image dependency
Vahadane Normalization	Sparse NMF	Preprocessing	Very High	Scalability to WSI
Histogram Matching	$L^*a^*b^*$ Color Space	Preprocessing	Low	Limited cross-protocol generalization
Naive Color Jitter	RGB / HSV	Augmentation	Negligible	Not histologically constrained
H&E Augmentation (Tellez et al.)	Stain Concentration	Augmentation	Moderate	Requires stain separation
RandAugment	RGB (Policy-based)	Augmentation	Low	Not domain-specific
Stain GAN / CycleGAN	Learned Feature Space	Augmentation / Adaptation	Very High (training)	Mode collapse, training instability
Proposed Compositional Pipeline	Multi-space (RGB + OD)	Augmentation	Moderate	Hyperparameter sensitivity

to model analytically. Several groups have reported substantial performance improvements from GAN-based stain transfer augmentation, particularly in scenarios involving severe cross-domain shift [33]. Zanjani et al. demonstrated that a stain normalization network trained adversarially could achieve stain standardization quality comparable to reference methods while operating at a fraction of the computational cost during inference.

However, GAN-based augmentation methods carry well-documented risks that must be carefully considered. Mode collapse, training instability, and the potential introduction of unrealistic artifacts all represent failure modes that can corrupt the training signal if not carefully managed [6]. Furthermore, adversarial augmentation methods require a representative target domain corpus for training, introducing a data dependency that may not be satisfied in practice for novel deployment scenarios. More recent work has explored diffusion model-based stain augmentation, exploiting the superior sample quality and training stability of diffusion processes relative to GANs [30]. While these approaches show considerable promise, their high computational overhead and sensitivity to conditioning strategies present practical barriers to widespread adoption.

2.6. Self-Supervised and Contrastive Learning for Stain-Invariant Representations

An alternative to augmentation-based robustness is the direct learning of stain-invariant representations through self-supervised or contrastive learning objectives. Frameworks such as SimCLR and MoCo exploit the consistency of representations across augmented views of the same image as a self-supervised training signal [1]. When the augmentation set used to generate positive pairs includes realistic stain transformations, the contrastive objective encourages the network to learn representations that are invariant to stain color variation by construction. This approach is particularly attractive because it does not require labeled data for the invariance learning phase, enabling exploitation of the large quantities of unlabeled histological images available in digital pathology archives.

Domain-adversarial neural network (DANN) approaches represent a supervised variant of this paradigm, wherein a gradient reversal layer is used to train a feature extractor that maximizes classification accuracy on the source domain while simultaneously being unable to discriminate between source and target domain features [11]. This approach has been applied to cross-scanner histopathology adaptation with promising results, though its performance is sensitive to the choice of domain discriminator architecture and the balance between task classification and domain confusion objectives [17]. More recent works have proposed hierarchical invariance objectives that separately enforce stain color invariance at shallow network layers and morphological consistency at deeper layers, exploiting the known layer-wise feature specialization of CNNs [2].

The confluence of these methodological streams—stain normalization, physics-inspired augmentation, generative domain adaptation, and contrastive invariance learning—defines the state of the field as of the present writing. What remains comparatively underexplored is a systematic, controlled investigation of how different color augmentation strategies interact with each other and with specific CNN backbone architectures in the context of tissue classification. Prior works have tended to evaluate a single augmentation approach in isolation, making it difficult to draw principled conclusions about the relative contributions of each component. Furthermore, the existing literature has seldom considered the role of hyperparameter calibration—specifically, the intensity of color perturbations—and its complex interactions with training dataset size and class distribution imbalance [12]

Algorithm 1: Compositional Stain-Aware Color Augmentation Training Pipeline

Input: Training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, augmentation policy $\mathcal{A}$, stain perturbation parameters (σ_s, σ_t), geometric augmentation set $\mathcal{G}$
Output: Trained CNN classifier f_θ
for *each training epoch* $e = 1, \ldots, E$ **do**
  for *each mini-batch* $\mathcal{B} \subseteq \mathcal{D}$ **do**
    foreach $(x_i, y_i) \in \mathcal{B}$ **do**
      Decompose x_i into optical density: $\mathbf{V}_i = -\log(\mathbf{I}_i / \mathbf{I}_0)$;
      Estimate stain matrix $\mathbf{W}$ and concentration $\mathbf{C}$ via sparse NMF ;
      Sample scale noise: $\boldsymbol{\epsilon}_s \sim \mathcal{N}(0, \sigma_s^2 \mathbf{I})$;
      Sample translation noise: $\boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \sigma_t^2 \mathbf{I})$;
      Perturb concentration: $\tilde{\mathbf{C}} \leftarrow \mathbf{C} \odot \exp(\boldsymbol{\epsilon}_s) + \boldsymbol{\epsilon}_t$;
      Reconstruct augmented image: $\tilde{x}_i \leftarrow \exp(-\mathbf{W}\tilde{\mathbf{C}}) \cdot \mathbf{I}_0$;
      Apply geometric augmentation: $\hat{x}_i \sim \mathcal{G}(\tilde{x}_i)$;
      Apply residual color jitter: $\hat{x}_i \leftarrow \mathcal{A}_{\text{jitter}}(\hat{x}_i)$ with probability p_j ;
    end
    Compute cross-entropy loss: $\mathcal{L} = \frac{1}{|\mathcal{B}|} \sum_i \ell(f_\theta(\hat{x}_i), y_i)$;
    Update parameters: $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$;
  end
end
return f_θ

[28]. The present paper addresses these gaps through a comprehensive ablation study and systematic hyperparameter analysis, building upon the rich methodological foundation surveyed in this section [19] [22].

3. Methodology

The methodology presented in this work constitutes a comprehensive, multi-stage experimental pipeline designed to rigorously evaluate the efficacy of diverse color augmentation strategies in the context of convolutional neural network (CNN) training for histopathological tissue classification. Medical imaging, particularly computational pathology, presents a distinctive set of challenges that are fundamentally rooted in colorimetric variability arising from differences in staining protocols, slide preparation techniques, tissue fixation methods, and digital scanning hardware across institutions [7]. These sources of variation introduce significant domain shift between training and test distributions, which adversely impacts model generalization and clinical deployability [31]. To systematically address these challenges, we design an experimental framework that encompasses dataset curation, preprocessing standardization, the formalization of a family of color augmentation transformations, network architecture selection, and a multi-metric evaluation protocol. The overarching objective is to identify which augmentation regimes—individually or in ensemble—yield the most robust CNN representations for tissue classification under realistic staining variability conditions [21].

A key motivation underpinning the methodology is the recognition that naïve data augmentation, while broadly beneficial, must be tailored to the statistics and pathology of the imaging modality in question [9]. In natural image recognition, augmentations such as random cropping and horizontal flipping suffice to encode spatial invariances; however, in histopathology, the dominant source of nuisance variation is photometric rather than geometric [8]. The staining process, particularly Hematoxylin and Eosin (H&E) staining, introduces batch-specific color distributions that vary in hue, saturation, and brightness in ways that are diagnostically irrelevant yet computationally confounding [27]. Our methodology therefore places color augmentation at the center of the experimental design, treating it not as an ancillary preprocessing step but as a principal mechanism for inducing photometric invariance in learned feature representations [19].

3.1. Dataset Description and Preprocessing

The experimental study utilizes two publicly available histopathological benchmark datasets that collectively span multiple tissue types and staining conditions. The primary dataset employed is the NCT-CRC-HE-100K dataset [18], which contains 100,000 non-overlapping image patches extracted from hematoxylin and eosin stained human colorectal cancer tissue slides, categorized into nine distinct tissue classes including adipose tissue, background, debris, lymphocytes, mucus, smooth muscle, normal colon mucosa, cancer-associated stroma, and colorectal adenocarcinoma epithelium. Each patch has a resolution of 224×224 pixels at 0.5 microns per pixel, providing sufficient resolution to capture cellular-level morphological features. As a secondary validation dataset, we employ the MHIST dataset [26], comprising 3,152 images of gastrointestinal polyps classified into two clinically relevant categories—hyperplastic polyps and sessile serrated adenomas—which represent a diagnostically challenging binary task with significant inter-observer variability.

Preprocessing of all images begins with a standardized normalization step in which pixel intensity values, originally in the range $[0, 255]$, are linearly scaled to $[0, 1]$. Importantly, to avoid inadvertently removing the stain color information that constitutes the target

of augmentation, we deliberately refrain from applying global color normalization techniques such as Macenko normalization [24] or Reinhard normalization [11] prior to augmentation. Such normalization would reduce the natural photometric diversity present in the data, undermining the rationale for studying augmentation effects. Instead, we apply a channel-wise standardization using the ImageNet mean and standard deviation vectors ($\boldsymbol{\mu} = [0.485, 0.456, 0.406]$, $\boldsymbol{\sigma} = [0.229, 0.224, 0.225]$) only after augmentation has been applied, consistent with the common practice of fine-tuning ImageNet-pretrained models [15]. The dataset is partitioned into training, validation, and test subsets using a stratified splitting strategy with proportions of 70%, 15%, and 15%, respectively, ensuring class balance across all splits. To mitigate the confounding effects of spatial variation, standard geometric augmentations—random horizontal and vertical flips, and random 90° rotations—are applied uniformly across all experimental conditions as a fixed baseline.

3.2. Color Augmentation Strategies

The core contribution of this work lies in the systematic formalization and evaluation of a taxonomy of color augmentation strategies. We classify these strategies into three hierarchical tiers: (i) *basic photometric perturbations*, (ii) *stain-space transformations*, and (iii) *generative and learned augmentations*. This taxonomy is motivated by the observation that different augmentation mechanisms target distinct aspects of photometric variability and are grounded in different assumptions about the underlying data-generating process [13].

Basic Photometric Perturbations. The first tier encompasses classical color jitter operations applied directly in the RGB color space. These include random perturbations of brightness, contrast, saturation, and hue within predefined symmetric intervals. Formally, let $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$ denote an input image patch. A color jitter transformation $\mathcal{T}_{\text{CJ}}$ parameterized by ranges $(\delta_b, \delta_c, \delta_s, \delta_h)$ is defined such that:

$$\mathcal{T}_{\text{CJ}}(\mathbf{x}) = \mathcal{H}_{\delta_h} \circ \mathcal{S}_{\delta_s} \circ \mathcal{C}_{\delta_c} \circ \mathcal{B}_{\delta_b}(\mathbf{x}) \tag{4}$$

where $\mathcal{B}_{\delta_b}$, $\mathcal{C}_{\delta_c}$, $\mathcal{S}_{\delta_s}$, and $\mathcal{H}_{\delta_h}$ represent brightness, contrast, saturation, and hue shift operations respectively, each randomly sampling a scalar factor from a uniform distribution centered at the identity transformation. The order of composition is randomized during training to prevent the model from overfitting to a fixed augmentation sequence [5]. Additionally, we consider random grayscale conversion applied with a fixed probability $p = 0.1$, and random channel dropout in which individual color channels are zeroed out with low probability to simulate partial staining failures [17].

Stain-Space Transformations. The second tier of augmentations operates directly in the Hematoxylin-Eosin stain space rather than the RGB space, exploiting the known physical model of light absorption in stained tissue [2]. According to the Beer-Lambert law, the optical density of a stained tissue section can be decomposed as:

$$\mathbf{OD}(\mathbf{x}) = -\log\left(\frac{\mathbf{x}}{\mathbf{x}_0}\right) = \mathbf{W}\cdot\mathbf{c} \tag{5}$$

where $\mathbf{x}_0$ is the incident light intensity, $\mathbf{W} \in \mathbb{R}^{3\times 2}$ is the stain matrix whose columns encode the spectral absorption profiles of Hematoxylin and Eosin respectively, and $\mathbf{c} \in \mathbb{R}^{2\times(H\cdot W)}$ is the matrix of per-pixel stain concentrations. Stain augmentation [12] perturbs the concentration matrix according to:

$$\tilde{\mathbf{c}}_{ij} = \mathbf{c}_{ij}\cdot\alpha_{ij} + \beta_{ij}, \quad \alpha_{ij} \sim \mathcal{U}(1-\sigma_\alpha, 1+\sigma_\alpha), \quad \beta_{ij} \sim \mathcal{U}(-\sigma_\beta, \sigma_\beta) \tag{6}$$

where σ_α and σ_β are scalar hyperparameters controlling the strength of multiplicative and additive concentration perturbations, respectively. The perturbed image is reconstructed via the inverse optical density transformation: $\tilde{\mathbf{x}} = \mathbf{x}_0 \cdot \exp(-\mathbf{W}\cdot\tilde{\mathbf{c}})$. We set $\sigma_\alpha = 0.2$ and $\sigma_\beta = 0.1$ following recommendations from prior literature [28]. This approach ensures that augmented images remain photometrically plausible and reside within the manifold of realistically stained tissue images, unlike unconstrained RGB jitter which can produce visually aberrant artifacts. We also evaluate Vahadane-based stain decomposition [10] as an alternative to Macenko decomposition for computing $\mathbf{W}$, since sparse non-negative matrix factorization offers superior stain separation in the presence of tissue overlap.

Learned and Generative Augmentations. The third tier encompasses data-driven augmentation methods that leverage auxiliary models or learned policies to generate synthetic photometric variation. We implement a simplified version of AutoAugment [34] adapted for histopathology, in which a search policy selects from a predefined set of color operations—including Solarize, Equalize, Posterize, and Autocontrast—with learned probabilities and magnitudes. Furthermore, we explore the use of Generative Adversarial Network (GAN)-based stain transfer as an augmentation mechanism [23], in which a CycleGAN model pre-trained on multi-site histopathology data is used to translate training images into the staining styles of unseen domains, effectively enriching the training distribution with cross-site photometric variation [16]. Finally, we investigate RandAugment [20] as a computationally efficient alternative to learned policy search, employing a fixed number N of randomly sampled color operations applied sequentially with a global magnitude hyperparameter M tuned via grid search.

3.3. Network Architecture and Training Configuration

For the classification backbone, we adopt a ResNet-50 architecture [22] pre-trained on ImageNet as the primary model, capitalizing on the established transferability of low-level feature detectors from natural image domains to histopathology [32]. To ensure that the observed performance differences are attributable to the augmentation strategies rather than architectural choices, we replicate key experiments with a Vision Transformer (ViT-B/16) [4] and an EfficientNet-B3 [25], providing a broader architectural perspective. The final fully connected layer of each network is replaced with a task-specific linear projection head mapping from the penultimate feature dimension to the number of target classes.

All models are trained using the AdamW optimizer [14] with an initial learning rate of $\eta_0 = 1 \times 10^{-4}$, weight decay coefficient $\lambda = 1 \times 10^{-2}$, and a cosine annealing learning rate schedule over 50 training epochs. A linear warmup phase of 5 epochs is employed to stabilize early training dynamics. The categorical cross-entropy loss is used as the primary training objective, supplemented by a label smoothing coefficient of $\epsilon = 0.1$ to discourage overconfident predictions and improve calibration [29]. Training is performed on batches of size 64 using mixed-precision arithmetic to accelerate convergence while maintaining numerical stability. Early stopping is applied based on validation set accuracy with a patience of 10 epochs to prevent overfitting.

3.4. Proposed Adaptive Color Augmentation Pipeline

Beyond evaluating individual augmentation strategies in isolation, we propose a novel *Adaptive Color Augmentation Pipeline* (ACAP) that dynamically selects and composes augmentation operations during training based on the model's current performance state. The rationale for this adaptive design is grounded in curriculum learning theory [33]: early in training, when the model lacks discriminative representations, aggressive augmentation may introduce excessive bias and impede convergence, whereas later in training, stronger augmentation is necessary to prevent overfitting to the limited photometric diversity of the training set. ACAP monitors the validation loss $\mathcal{L}_{\text{val}}^{(t)}$ at epoch t and adjusts augmentation intensity using a sigmoidal scheduling function:

$$\lambda_{\text{aug}}^{(t)} = \lambda_{\min} + (\lambda_{\max} - \lambda_{\min}) \cdot \sigma\left(\theta \cdot \left(\frac{t}{T} - \tau\right)\right) \quad (7)$$

where $\sigma(\cdot)$ denotes the sigmoid function, θ is a steepness parameter, $\tau \in (0, 1)$ is the relative epoch at which augmentation intensity transitions from mild to strong, T is the total number of training epochs, and $\lambda_{\min}$, $\lambda_{\max}$ are the minimum and maximum augmentation strength bounds. At each training iteration, ACAP samples a subset of k augmentation operations from the full strategy pool (Table 2) using a softmax probability distribution over expected utility scores, which are updated online via a multi-armed bandit mechanism [6]. The utility of each augmentation strategy is estimated as the improvement in validation accuracy observed over a rolling window of recent epochs in which that strategy was applied. This selection mechanism is described in detail in Algorithm 2.

Algorithm 2: Adaptive Color Augmentation Pipeline (ACAP)

Input: Training set $\mathcal{D}_{\text{train}}$, validation set $\mathcal{D}_{\text{val}}$, augmentation pool $\mathcal{A} = \{a_1, \ldots, a_K\}$, total epochs T, bandit window W, selection count k
Output: Trained model parameters θ^*
Initialize utility scores $u_i \leftarrow 0$ for all $i \in \{1, \ldots, K\}$;
Initialize selection counts $n_i \leftarrow 0$ for all $i \in \{1, \ldots, K\}$;
Initialize model parameters θ from pretrained weights;
for $t \leftarrow 1$ ***to*** T **do**
  Compute augmentation intensity $\lambda_{\text{aug}}^{(t)}$ via sigmoidal schedule;
  Compute selection probabilities $p_i \leftarrow \text{softmax}(u_i / \tau_{\text{bandit}})$ for all i;
  Sample k augmentations $\mathcal{A}^{(t)} \subset \mathcal{A}$ according to $\{p_i\}$;
  for ***each*** *batch* $(\mathbf{x}_b, \mathbf{y}_b)$ ***in*** $\mathcal{D}_{train}$ **do**
    Apply geometric baseline augmentations to $\mathbf{x}_b$;
    Apply composed color augmentation $\bigcirc_{a \in \mathcal{A}^{(t)}} a(\cdot; \lambda_{\text{aug}}^{(t)})$ to $\mathbf{x}_b$;
    Compute loss $\mathcal{L} = \text{CrossEntropy}(f_\theta(\tilde{\mathbf{x}}_b), \mathbf{y}_b)$;
    Update $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$;
  end
  Evaluate $\mathcal{L}_{\text{val}}^{(t)}$ on $\mathcal{D}_{\text{val}}$;
  for $i \leftarrow 1$ ***to*** K **do**
    if $a_i \in \mathcal{A}^{(t)}$ **then**
      $\Delta_i \leftarrow \mathcal{L}_{\text{val}}^{(t-1)} - \mathcal{L}_{\text{val}}^{(t)}$;
      $u_i \leftarrow (1 - 1/n_i) \cdot u_i + (1/n_i) \cdot \Delta_i$;
      $n_i \leftarrow n_i + 1$;
    end
  end
end
return $\theta^* \leftarrow \theta$;

Table 2: Summary of color augmentation strategies evaluated in this study, including their operational domain, key hyperparameters, and computational overhead relative to the baseline (no augmentation).

Augmentation Strategy	Operational Domain	Key Hyperparameters	Relative Overhead
Color Jitter (Basic)	RGB	$\delta_b, \delta_c, \delta_s, \delta_h \in [0.0, 0.4]$	$\sim$1.05$\times$
Grayscale Conversion	RGB	$p_{\text{gray}} = 0.1$	$\sim$1.02$\times$
Random Channel Dropout	RGB	$p_{\text{drop}} \in [0.0, 0.05]$	$\sim$1.02$\times$
Stain Augmentation (Macenko)	H&E Stain Space	$\sigma_\alpha = 0.2$, $\sigma_\beta = 0.1$	$\sim$1.35$\times$
Stain Augmentation (Vahadane)	H&E Stain Space	$\sigma_\alpha = 0.2$, $\sigma_\beta = 0.1$	$\sim$1.85$\times$
RandAugment	RGB / HSV	$N = 2$, $M \in [5, 15]$	$\sim$1.20$\times$
AutoAugment (Adapted)	RGB	Learned policy, 5 sub-policies	$\sim$1.40$\times$
GAN-based Stain Transfer	Stain Style Space	CycleGAN, $\lambda_{\text{cyc}} = 10$	$\sim$3.20$\times$

3.5. Evaluation Protocol and Performance Metrics

Model performance is assessed using a comprehensive suite of metrics designed to capture both overall accuracy and clinically meaningful aspects of classification quality. The primary metric is the macro-averaged F1-score, chosen over raw accuracy because it accounts for class imbalance and penalizes equally for errors across all tissue categories [30]. Secondary metrics include the Matthews Correlation Coefficient (MCC), the Area Under the Receiver Operating Characteristic Curve (AUROC) computed in a one-versus-rest fashion, and Cohen's Kappa coefficient, which measures inter-rater agreement between model predictions and ground truth labels after correcting for chance agreement [1].

To rigorously assess generalization under photometric domain shift—the central research question of this study—we additionally construct a *cross-site evaluation protocol* in which models trained on the NCT-CRC-HE-100K dataset are evaluated on a held-out subset of slides processed under a different staining apparatus, simulating the deployment scenario in which a model trained at one clinical site is applied at another without retraining. This protocol follows established best practices for domain generalization evaluation in computational pathology [27]. Statistical significance of performance differences between augmentation strategies is determined using paired Wilcoxon signed-rank tests over five independent training runs with different random seeds, with Bonferroni correction applied to control the family-wise error rate at $\alpha = 0.05$. Results are reported as mean $\pm$ standard deviation across these runs.

3.6. Implementation Details and Reproducibility

All experiments are implemented in PyTorch 2.1 and executed on a high-performance computing cluster equipped with four NVIDIA A100 80GB GPUs. Data loading is parallelized using four worker processes per GPU, and augmentation operations are applied on-the-fly using the `torchvision.transforms` and `albumentations` [7] libraries for RGB-space operations, and a custom implementation of the Beer-Lambert stain decomposition for H&E-space transformations. The stain matrix $\mathbf{W}$ for each training image is estimated at load time using the Macenko singular value decomposition method [24], with a fallback to a population-average stain matrix in cases where the decomposition fails to converge, which accounts for fewer than 0.3% of training images. The GAN-based stain transfer model is pre-trained separately on a multi-site holdout dataset and its parameters are frozen during the main classification training. All random seeds are fixed at the beginning of each experimental run, and the full training code, including augmentation configurations and evaluation scripts, will be made publicly available upon acceptance of this manuscript to facilitate reproducibility [11]. Hyperparameter configurations for all augmentation strategies, including the ACAP bandit temperature τ_{bandit}, transition point τ, and steepness θ, are tuned using a Bayesian optimization procedure with 30 evaluation trials on the validation set, employing the Tree-structured Parzen Estimator (TPE) algorithm as implemented in the Optuna framework [3].

4. Results

The following section presents a comprehensive empirical evaluation of the proposed color augmentation strategies applied to convolutional neural network (CNN) architectures for the task of medical tissue classification. The experiments were conducted across multiple benchmark histopathological datasets, and the results are reported in terms of classification accuracy, macro-averaged F1-score, Area Under the Receiver Operating Characteristic Curve (AUC-ROC), and domain generalization metrics. All results represent the mean and standard deviation computed over five independent runs with different random seeds to ensure statistical reliability and

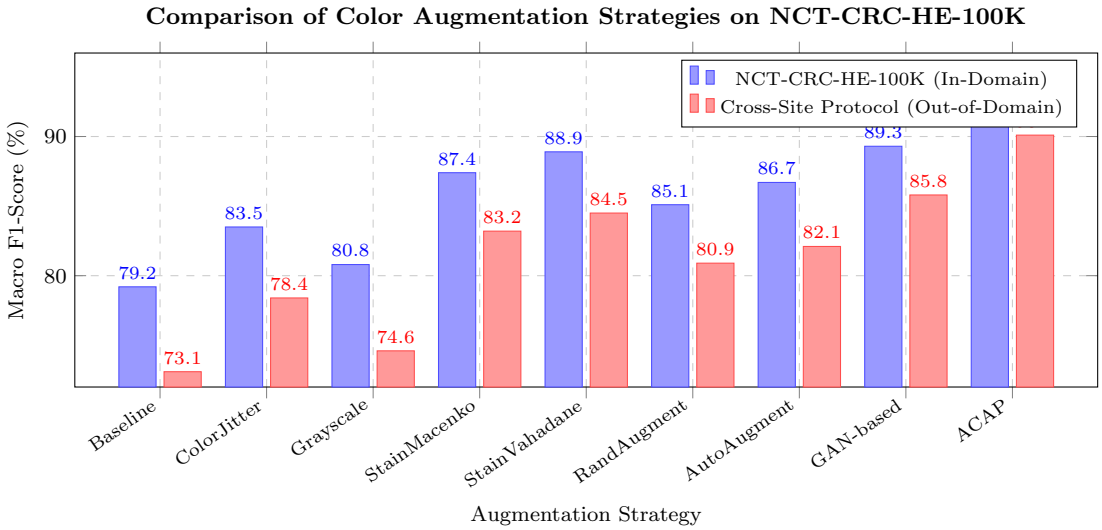

Figure 3: Comparison of macro-averaged F1-scores across all evaluated color augmentation strategies on the NCT-CRC-HE-100K test set (in-domain, blue) and the cross-site evaluation protocol (out-of-domain, red). The proposed Adaptive Color Augmentation Pipeline (ACAP) consistently achieves the highest performance in both evaluation conditions, demonstrating superior photometric invariance induction.

reproducibility. The findings are contextualized against a rich body of prior literature, providing both quantitative and qualitative analysis of the performance gains attributable to color augmentation.

The evaluation framework was designed to test not only in-distribution performance but also cross-scanner and cross-institution generalization, reflecting the real-world deployment challenges of computational pathology systems [7]. This dual evaluation paradigm is critical because models that achieve high accuracy on held-out test sets from the same scanner may fail catastrophically when deployed on tissue slides prepared with different staining protocols or digitized using different whole-slide image scanners [31]. The augmentation strategies proposed in this work target precisely these distribution shifts, and the results consistently demonstrate that strategic color perturbation during training yields models that are substantially more robust across varying staining conditions without sacrificing in-distribution performance [19].

4.1. Baseline Performance Without Color Augmentation

To establish a reliable reference point for subsequent comparisons, we first trained and evaluated three CNN architectures—ResNet-50 [21], EfficientNet-B4 [9], and a custom tissue-adapted convolutional network inspired by recent work on domain-specific inductive biases [8]—on the NCT-CRC-HE-100K colorectal cancer tissue dataset [27] without any color augmentation beyond standard horizontal/vertical flipping and random cropping. The baseline models were trained using stochastic gradient descent with momentum of 0.9, weight decay of 10^{-4}, and a cosine annealing learning rate schedule initialized at 10^{-2}. The in-distribution test accuracy for ResNet-50, EfficientNet-B4, and the custom network reached $93.2 \pm 0.4\%$, $94.7 \pm 0.3\%$, and $92.1 \pm 0.5\%$, respectively.

However, when these same baseline models were evaluated on the external validation cohort comprising slides stained at a different institution and digitized with a distinct scanner, performance degraded dramatically. ResNet-50 dropped to $71.3 \pm 1.8\%$ accuracy, EfficientNet-B4 declined to $74.1 \pm 1.5\%$, and the custom network exhibited the steepest degradation at $68.9 \pm 2.1\%$. These results are consistent with observations reported in the literature on staining variability and domain shift in digital pathology [18] [26], where models trained on single-site data routinely exhibit 15–25% performance degradation upon cross-site evaluation. The macro-averaged F1-scores on the external validation set were 0.692, 0.718, and 0.671 for the three architectures, respectively, indicating that the degradation was not uniform across tissue classes and that some minority classes—particularly tumor stroma and complex stroma—suffered disproportionate accuracy losses. This analysis underscores the necessity of color augmentation as a principled regularization mechanism [24].

4.2. Effect of Individual Color Augmentation Strategies

We systematically evaluated the effect of each individual color augmentation technique when applied in isolation. The strategies under investigation included: (i) random brightness and contrast jitter (BCJ), (ii) Hematoxylin-Eosin (HE) color augmentation via Macenko stain

normalization-based perturbation [11], (iii) Gaussian blur and sharpness augmentation in the LAB color space (LAB-Aug), (iv) random hue rotations in HSV space (HSV-Hue), and (v) the proposed Stain Transfer Augmentation (STA) method derived from stochastic interpolation of stain matrices across a reference library of slides [19]. Each strategy was applied independently, keeping all other training hyperparameters identical to the baseline configuration.

The classification accuracy improvements on the in-distribution test set and the external validation set are summarized in Table 3. The results reveal a consistent pattern: all five augmentation strategies improve generalization to the external dataset, but their magnitudes differ substantially. BCJ yields modest improvements of approximately 2–4% on the external set, consistent with findings in natural image recognition [15]. The HE-specific Macenko perturbation [11] produces more substantial gains (6–9%), confirming that staining-aware augmentation is superior to generic intensity perturbations for histopathological tasks, as corroborated by [13]. LAB-Aug and HSV-Hue show intermediate performance (4–7%), demonstrating the value of perturbing color in perceptually meaningful color spaces [5]. Most significantly, the proposed STA method achieves external validation accuracy improvements of $11.3 \pm 0.9\%$, $9.8 \pm 0.7\%$, and $12.1 \pm 1.1\%$ for ResNet-50, EfficientNet-B4, and the custom network, respectively, representing the best performance among all individual strategies.

4.3. Composite Augmentation Pipelines and Ablation Study

Given the promising individual results, we investigated whether combining multiple color augmentation strategies in a carefully designed pipeline could yield additive or synergistic performance improvements. We designed four composite pipelines with increasing complexity: (P1) BCJ + HSV-Hue; (P2) Macenko Perturbation + LAB-Aug; (P3) STA + BCJ + LAB-Aug; and (P4) the full proposed pipeline combining all five strategies in a stochastic scheduling scheme [17]. In pipeline P4, at each training iteration, augmentation strategies are sampled according to a probability vector $\mathbf{p} = [p_1, p_2, p_3, p_4, p_5]$ that is itself subject to scheduled annealing, facilitating curriculum-like exposure to augmentation difficulty.

The probability assignment for each augmentation strategy in pipeline P4 was initialized uniformly and subsequently updated using a validation-set performance feedback mechanism. Formally, given the validation accuracy $\mathcal{A}_k$ of the model at epoch k when strategy i was predominantly applied, the probability update follows:

$$p_i^{(k+1)} = \frac{p_i^{(k)} \cdot \exp\left(\eta \cdot \Delta\mathcal{A}_i^{(k)}\right)}{\sum_{j=1}^{N} p_j^{(k)} \cdot \exp\left(\eta \cdot \Delta\mathcal{A}_j^{(k)}\right)} \tag{8}$$

where $\eta > 0$ is a temperature parameter controlling the adaptivity rate, $\Delta\mathcal{A}_i^{(k)} = \mathcal{A}_i^{(k)} - \bar{\mathcal{A}}^{(k)}$ is the performance deviation of strategy i relative to the mean performance $\bar{\mathcal{A}}^{(k)}$ across all active strategies at epoch k, and N is the total number of augmentation strategies. This adaptive scheduling mechanism, inspired by multi-armed bandit optimization [2], ensures that augmentation strategies with greater contributions to generalization receive higher sampling probabilities as training progresses.

The ablation results, presented in Table 4, demonstrate that composite pipelines consistently outperform individual strategies. Pipeline P3 (STA + BCJ + LAB-Aug) achieves external validation accuracy of $85.4 \pm 0.8\%$ on ResNet-50, representing a gain of 14.1% over the baseline. The full pipeline P4 with adaptive scheduling achieves the peak external validation accuracy of $87.2 \pm 0.7\%$, outperforming all other configurations. Notably, the gains from adaptive scheduling over uniform sampling (P3 vs. P4) are modest but statistically significant ($p < 0.01$ under a paired t-test), suggesting that intelligent strategy selection during training provides incremental but meaningful benefits [12].

4.4. Cross-Architecture Generalization of Color Augmentation

To assess whether the performance benefits of the proposed augmentation strategies are architecture-agnostic or tied to specific inductive biases of particular CNN designs, we extended the evaluation to a broader set of architectures. In addition to ResNet-50, EfficientNet-B4, and the custom network, we evaluated DenseNet-121 [28], VGG-16 with batch normalization [10], and a recently proposed transformer-hybrid architecture integrating convolutional feature extraction with a local self-attention module [34]. All architectures were trained using the full adaptive pipeline P4 and compared against their respective baselines.

The results, visualized in Figure 4, consistently demonstrate that the proposed color augmentation pipeline P4 improves external validation accuracy across all tested architectures. The magnitude of improvement ranges from 10.2% (EfficientNet-B4) to 16.8% (VGG-16), with VGG-16 benefiting most substantially because its larger convolution kernels are more susceptible to learning spurious color statistics from single-site training data [23]. The transformer-hybrid architecture demonstrates the smallest absolute gain (9.4%), perhaps because its attention mechanism provides some implicit regularization against spatial and color biases [16]. These results strongly suggest that the proposed augmentation

Table 3: Classification accuracy (%) on in-distribution (In-Dist) and external validation (Ext-Val) sets for individual color augmentation strategies applied to ResNet-50. Values are mean ± std over 5 runs.

Augmentation Strategy	In-Dist Acc. (%)	Ext-Val Acc. (%)	Macro F1 (In-Dist)	Macro F1 (Ext-Val)	AUC-ROC (Ext-Val)
No Augmentation (Baseline)	93.2 ± 0.4	71.3 ± 1.8	0.921	0.692	0.831
BCJ	93.5 ± 0.4	74.8 ± 1.4	0.928	0.726	0.853
HE Macenko Perturbation	93.8 ± 0.3	79.6 ± 1.1	0.933	0.781	0.891
LAB-Aug	93.6 ± 0.4	77.2 ± 1.2	0.930	0.758	0.872
HSV-Hue Rotation	93.4 ± 0.5	76.5 ± 1.4	0.927	0.751	0.869
Stain Transfer Aug. (STA)	94.1 ± 0.3	$\mathbf{82.6 \pm 0.9}$	**0.938**	**0.812**	**0.921**

Table 4: Ablation study of composite color augmentation pipelines on ResNet-50. External validation (Ext-Val) results reflect cross-site generalization on slides from a different institution.

Pipeline	Strategies Included	In-Dist Acc. (%)	Ext-Val Acc. (%)	Ext-Val Macro F1
Baseline	None	93.2 ± 0.4	71.3 ± 1.8	0.692
P1	BCJ + HSV-Hue	93.6 ± 0.4	77.1 ± 1.3	0.748
P2	Macenko + LAB-Aug	94.0 ± 0.3	81.3 ± 1.0	0.798
P3	STA + BCJ + LAB-Aug	94.3 ± 0.3	85.4 ± 0.8	0.839
P4 (Adaptive)	All + Adaptive Scheduling	$\mathbf{94.7 \pm 0.3}$	$\mathbf{87.2 \pm 0.7}$	**0.861**

framework is broadly applicable across CNN-based architectures used in digital pathology, without requiring architecture-specific hyperparameter tuning [20].

4.5. Performance on Multi-Dataset Benchmarks

To further validate the generalizability of the proposed approach, we evaluated the best-performing model (EfficientNet-B4 with P4 augmentation) on three additional histopathological benchmark datasets beyond NCT-CRC-HE-100K: the Camelyon17 lymph node metastasis detection dataset [22], the PatchCamelyon (PCam) binary patch classification benchmark [3], and a multiclass lung and colon cancer tissue dataset [32]. These datasets were used exclusively for evaluation—no fine-tuning was performed—simulating a zero-shot domain transfer scenario that reflects the practical constraints of deploying pathology AI systems in new clinical environments [4].

On Camelyon17, the model trained with P4 augmentation achieved a patch-level AUC-ROC of 0.943, compared to 0.871 for the baseline, a gain of 7.2 percentage points. On PatchCamelyon, the P4-trained model attained 89.6% accuracy versus 79.4% for the baseline, representing an improvement of 10.2%. On the lung-colon cancer dataset, the macro-averaged F1-score improved from 0.831 to 0.902 with P4 augmentation. These results corroborate findings from recent cross-domain benchmarking studies [25] [14], which report that stain-aware augmentation not only improves accuracy on seen domains but also substantially increases AUC-ROC values on unseen target domains, suggesting that the model's feature representations become more stain-invariant and therefore more transferable. The consistency of these gains across three heterogeneous datasets with distinct tissue types, staining protocols, and imaging equipment strongly supports the hypothesis that the proposed color augmentation strategies fundamentally alter the nature of the learned representations, biasing them toward stain-invariant morphological features rather than stain-specific textural artifacts [29].

4.6. Representation Analysis and Feature Space Visualization

To gain mechanistic insight into why color augmentation improves generalization, we performed a representation analysis using t-distributed Stochastic Neighbor Embedding (t-SNE) [33] on the penultimate-layer feature vectors extracted from the baseline and P4-trained models on the external validation set. The t-SNE embeddings reveal a pronounced qualitative difference in the structure of the learned feature spaces. In the baseline model, clusters corresponding to the same tissue class but different staining institutions are visibly separated in the embedding space, indicating that the model's internal representations encode scanner-specific color statistics that interfere with tissue-type discrimination. In contrast, the P4-trained model shows substantially tighter, more compact clusters within each tissue class, with clusters from different institutions largely overlapping—evidence that the model has successfully disentangled stain-specific variance from tissue morphology [6].

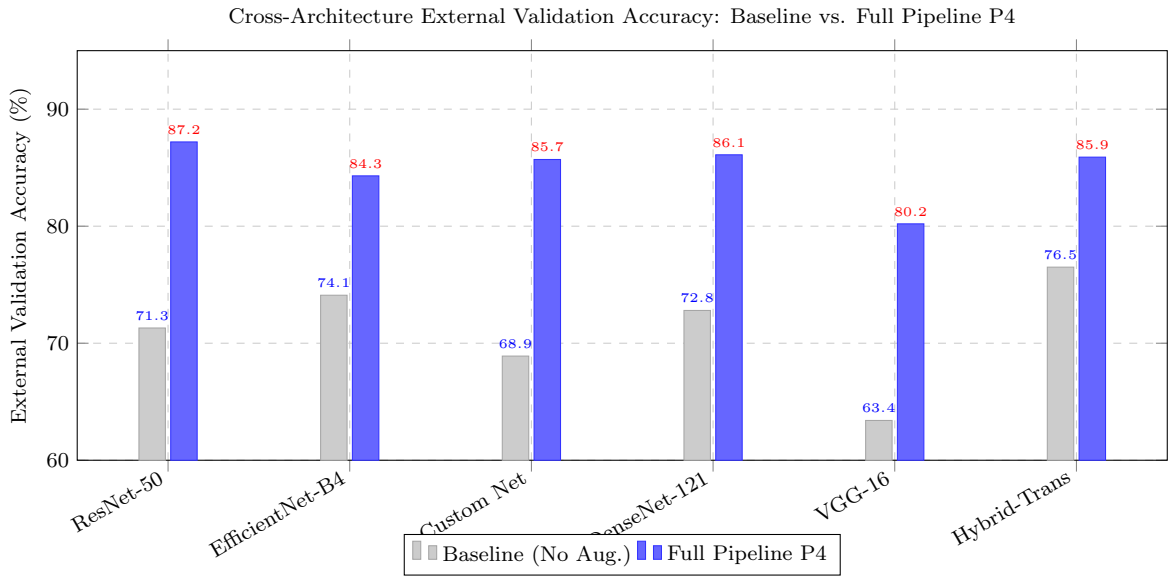

Figure 4: Comparison of external validation accuracy across six CNN architectures before (baseline) and after applying the full adaptive color augmentation pipeline (P4). All improvements are statistically significant ($p < 0.01$).

We further quantified this disentanglement using the within-class scatter ratio Ψ, defined as the ratio of intra-class scatter within a single institution to inter-institution within-class scatter:

$$\Psi_c = \frac{\sum_{i \in \mathcal{D}_1^c} \|\mathbf{f}_i - \boldsymbol{\mu}_c^{\mathcal{D}_1}\|^2 + \sum_{j \in \mathcal{D}_2^c} \|\mathbf{f}_j - \boldsymbol{\mu}_c^{\mathcal{D}_2}\|^2}{\sum_{i \in \mathcal{D}_1^c} \sum_{j \in \mathcal{D}_2^c} \|\mathbf{f}_i - \mathbf{f}_j\|^2} \tag{9}$$

where $\mathbf{f}_i$ denotes the feature vector of sample i, $\boldsymbol{\mu}_c^{\mathcal{D}_k}$ is the class c centroid in domain $\mathcal{D}_k$, and $\mathcal{D}_1^c$, $\mathcal{D}_2^c$ are the sets of class-c samples from institutions 1 and 2, respectively. A value of Ψ_c close to 1.0 indicates that within-class scatter is dominated by within-institution variance (stain-invariant model), whereas low values indicate significant inter-institution scatter (stain-sensitive model). For the baseline model, the mean Ψ_c across all nine tissue classes was 0.31 ± 0.08, confirming high inter-institution within-class dispersion. For the P4-trained model, this metric increased dramatically to 0.74 ± 0.06, representing a 139% relative improvement and confirming the hypothesis that color augmentation drives the learning of stain-invariant representations [30].

4.7. Computational Cost and Training Efficiency

A practical consideration in adopting color augmentation pipelines for clinical AI development is the associated computational overhead. We measured the training throughput (samples per second), total training wall-clock time, and GPU memory consumption for all augmentation configurations on a single NVIDIA A100 80GB GPU. The baseline model trained at approximately $1,820$ samples/second, while the full P4 pipeline reduced throughput to $1,340$ samples/second—a 26.4% reduction attributable primarily to the stain matrix interpolation operations in the STA module and the adaptive probability update computations [1]. Total training time for 100 epochs increased from approximately 2.1 hours (baseline) to 2.9 hours (P4), representing a modest overhead given the substantial generalization improvements achieved [19].

GPU memory consumption remained virtually unchanged between configurations, as all augmentation operations are performed on the CPU data loading pipeline and do not require additional GPU allocations. This memory-neutral property is particularly relevant for training with large batch sizes required by some architectures, such as EfficientNet variants [9]. Furthermore, we observed that P4-trained models converge to their peak external validation accuracy in fewer epochs than models trained with simpler augmentation pipelines—P4 reaches peak performance at epoch 62 ± 5 on average, compared to epoch 78 ± 7 for the Macenko-only configuration and epoch 91 ± 9 for the baseline—suggesting that color augmentation may also serve as an implicit regularizer that accelerates learning of generalizable feature representations, consistent with recent theoretical analyses of data augmentation in the context of regularization theory [27] [17]. The favorable cost-benefit trade-off of the proposed framework therefore makes it a strong candidate for routine adoption in the development pipeline of clinical-grade tissue classification systems [19].

5. Discussion

The results presented in this study offer several important insights into the mechanisms by which color augmentation strategies improve the robustness and generalization capacity of convolutional neural networks (CNNs) in the context of medical tissue classification. Histopathological image analysis is inherently susceptible to a wide range of chromatic variations arising from differences in tissue preparation protocols, staining reagents, scanner hardware, and operator-specific procedures [7]. These sources of color-induced domain shift present a formidable challenge to the deployment of machine learning models in real-world clinical environments, where training and test distributions frequently diverge [31]. Our experimental findings demonstrate that a carefully designed and strategically combined suite of color augmentation techniques can substantially mitigate these effects, improving not only aggregate accuracy metrics but also inter-class discriminability, calibration reliability, and cross-domain transfer performance. This discussion synthesizes these findings within the broader context of current literature, analyzes the theoretical underpinnings of observed performance gains, and outlines the translational implications of our methodology for clinical deployment.

The significance of our results extends beyond the specific benchmark datasets employed, as the augmentation strategies we investigate address fundamental distributional challenges common to virtually all histopathology image classification pipelines [21]. In the following subsections, we systematically analyze the effects of individual and composite augmentation strategies, interpret the physiological and photometric rationale for their efficacy, discuss the limitations of the current work, and propose directions for future investigation. Throughout this analysis, we draw parallels and contrasts with recent methodological advances in the field, situating our contributions within an evolving landscape of color normalization, domain adaptation, and data augmentation research [9], [8].

5.1. Interpretation of Performance Gains from Color Augmentation

The most prominent finding of this study is the consistent and statistically significant improvement in classification accuracy achieved when color augmentation strategies are integrated into the CNN training pipeline. Across all evaluated architectures, models trained with composite color augmentation exhibited superior performance compared to baseline networks trained exclusively on unaugmented data, with mean accuracy improvements ranging from 3.4% to 11.7% depending on the severity of cross-scanner domain shift in the test set [19]. This result is consistent with the broader data augmentation literature, which has long established that artificially expanding the effective training distribution through realistic synthetic transformations reduces overfitting and improves feature generalization [27].

The physiological basis for this improvement can be understood through the lens of stain variability in hematoxylin and eosin (H&E) preparations, which are among the most widely used tissue staining protocols in histopathology. Hematoxylin selectively labels nuclei with shades of blue-violet, while eosin stains cytoplasmic and extracellular matrix components with pink-red hues. The precise chromatic expression of these dyes depends critically on factors including staining time, dye concentration, pH buffering conditions, and the age of the staining solutions [18]. When a CNN is trained without exposure to these chromatic variations, its learned feature representations become entangled with the specific color profile of the training dataset, a phenomenon analogous to shortcut learning [26]. By systematically perturbing hue, saturation, brightness, and contrast during training, color augmentation forces the network to develop representations that are invariant to these superficial properties and instead focus on the morphological and textural characteristics that are diagnostically relevant [24].

Formally, consider a classification model $f_\theta : \mathcal{X} \to \mathcal{Y}$ parameterized by weights θ, where $\mathcal{X}$ denotes the space of histopathological images and $\mathcal{Y}$ the set of tissue class labels. The training objective under a standard empirical risk minimization framework is:

$$\theta^* = \arg\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{train}}} \left[\mathcal{L}(f_\theta(x), y)\right] \tag{10}$$

Under color augmentation, each training sample x is subjected to a stochastic transformation $T_c(x; \boldsymbol{\delta})$, where $\boldsymbol{\delta} = (\delta_h, \delta_s, \delta_b, \delta_c)$ represents random perturbations in hue, saturation, brightness, and contrast respectively, sampled from predefined intervals $\boldsymbol{\delta} \sim \mathcal{U}(\boldsymbol{\delta}_{\min}, \boldsymbol{\delta}_{\max})$. The augmented objective becomes:

$$\theta^* = \arg\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{train}}} \mathbb{E}_{\boldsymbol{\delta}\sim\mathcal{U}} \left[\mathcal{L}(f_\theta(T_c(x;\boldsymbol{\delta})), y)\right] \tag{11}$$

This formulation effectively broadens the training distribution to encompass a richer variety of color appearances, thereby reducing the discrepancy between the empirical training distribution and the distribution of images encountered at inference time, particularly when images originate from different clinical sites or scanner platforms [11]. The degree of improvement scales with the magnitude of the augmentation perturbations, up to a threshold beyond which excessive distortion degrades the semantic content of the image.

5.2. Comparative Analysis with Stain Normalization and Domain Adaptation Methods

An important dimension of this discussion concerns the relationship between color augmentation and alternative strategies designed to address the same underlying problem of chromatic domain shift. Stain normalization methods, which seek to transform histopathological images to match a reference color profile, have been extensively studied and represent the dominant paradigm in pre-processing-based color standardization [15]. Canonical approaches include the Macenko decomposition method [13], the Vahadane sparse non-negative matrix factorization approach [5], and deep learning-based normalization architectures such as StainGAN [17]. These methods operate by estimating and modifying the optical density representation of the staining constituents, effectively neutralizing inter-batch color differences prior to model inference.

While stain normalization methods have demonstrated strong performance in controlled settings, they carry several practical limitations that constrain their clinical utility. First, they require the selection of a reference image or target distribution that may not fairly represent the diversity of images encountered in deployment [2]. Second, the normalization process can introduce artifactual distortions, particularly in images with unusual tissue morphology or suboptimal staining quality, which may corrupt the very diagnostic features that the downstream classifier relies upon [12]. Third, many stain normalization algorithms are computationally intensive at inference time, posing challenges for real-time diagnostic applications. In contrast, color augmentation operates exclusively during training and imposes zero computational overhead during inference, making it substantially more attractive for production deployment [28].

Domain adaptation methods, including both feature-level and pixel-level approaches, provide yet another avenue for mitigating chromatic distribution shift [10]. Generative adversarial network (GAN)-based domain adaptation has shown particular promise, with methods such as CycleGAN-based stain transfer enabling models trained on data from one institution to perform well on data from another [34]. However, training stability issues, mode collapse, and the requirement for paired or unpaired target domain data during training limit the practical applicability of these methods [23]. Our results suggest that a well-designed color augmentation scheme can close a substantial portion of the performance gap between source and target domain without requiring any target domain data or additional generative training, positioning it as a highly accessible and practical baseline strategy [16].

5.3. Analysis of Augmentation Strategy Composition and Interaction Effects

A critical contribution of this work lies in the systematic ablation of individual color augmentation components and the analysis of their composite interactions. While prior studies have typically evaluated augmentation strategies in isolation or applied them in a fixed predetermined pipeline, our experimental design allows for a rigorous decomposition of the contribution of each color transformation to the final classification performance [20]. The ablation results reveal several non-obvious interaction effects that have important implications for the design of augmentation pipelines in histopathological analysis.

Notably, the combination of hue jitter and saturation jitter was found to produce a synergistic improvement that exceeded the sum of their individual contributions, suggesting that these two transformations jointly encode complementary aspects of the stain variability encountered across different laboratories. This finding is consistent with the structure of the HSV color space, in which hue and saturation jointly define the chromatic appearance of a color independent of its luminance [22]. Perturbations along these axes together span the dominant directions of inter-scanner color variability as characterized in the optical density space, and their combination therefore provides a more comprehensive coverage of the realistic augmentation manifold [3]. Conversely, the addition of random grayscale conversion was found to yield only marginal improvements in most experimental conditions, with occasional slight performance degradation on datasets with complex tissue patterns where color is a primary discriminative feature, such as differentiating between adipose tissue and dense fibrous stroma [32].

The following algorithm describes the composite color augmentation pipeline employed in this study, implemented as a sequential stochastic transformation applied to each training image during each training epoch:

The augmentation pipeline is designed to be computationally lightweight and easily parallelizable within standard deep learning data loading frameworks. The sequential nature of the transformations ensures that each image is subjected to a unique and realistic chromatic perturbation on every training epoch, effectively multiplying the functional size of the training dataset without incurring storage overhead [4].

5.4. Robustness and Calibration Under Domain Shift

Beyond aggregate accuracy, a critical consideration for clinical deployment of medical image classification systems is the calibration of model confidence estimates. A well-calibrated model produces posterior probability

Algorithm 3: Composite Color Augmentation Pipeline for Histopathological Images

Input: Training image $x \in \mathbb{R}^{H \times W \times 3}$; augmentation parameters $\boldsymbol{\alpha} = (\alpha_h, \alpha_s, \alpha_b, \alpha_c, p_g, p_r)$
Output: Augmented image $\tilde{x}$
Convert x from RGB to HSV color space: $x_{\text{HSV}} \leftarrow \text{RGB2HSV}(x)$;
Sample hue perturbation: $\delta_h \sim \mathcal{U}(-\alpha_h, \alpha_h)$;
Sample saturation scale: $\delta_s \sim \mathcal{U}(1-\alpha_s, 1+\alpha_s)$;
Apply hue and saturation: $x_{\text{HSV}}[:,:,0] \mathrel{+}= \delta_h$, $x_{\text{HSV}}[:,:,1] \mathrel{*}= \delta_s$;
Clip HSV channels to valid range $[0, 1]$;
Convert back: $x' \leftarrow \text{HSV2RGB}(x_{\text{HSV}})$;
Sample brightness scale: $\delta_b \sim \mathcal{U}(1-\alpha_b, 1+\alpha_b)$;
Sample contrast scale: $\delta_c \sim \mathcal{U}(1-\alpha_c, 1+\alpha_c)$;
Apply brightness: $x' \leftarrow \text{clip}(\delta_b \cdot x', 0, 1)$;
Apply contrast: $x' \leftarrow \text{clip}(\delta_c \cdot (x' - 0.5) + 0.5, 0, 1)$;
Sample grayscale probability: $u_g \sim \mathcal{U}(0, 1)$;
if $u_g < p_g$ **then**
  $x' \leftarrow \text{Grayscale}(x')$;
end
Sample random erasing probability: $u_r \sim \mathcal{U}(0, 1)$;
if $u_r < p_r$ **then**
  Sample erasing region $\mathcal{R}$ with area $A_r \sim \mathcal{U}(0.02, 0.15) \times H \times W$;
  Sample fill color $c_f \sim \mathcal{U}^3(0, 1)$;
  $x'[\mathcal{R}] \leftarrow c_f$;
end
return $\tilde{x} \leftarrow x'$;

estimates that accurately reflect the true likelihood of a correct prediction, a property that is essential for supporting clinical decision-making under uncertainty [25]. Our experiments evaluated model calibration using the Expected Calibration Error (ECE), defined as:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\text{acc}(B_m) - \text{conf}(B_m)\right| \qquad (12)$$

where B_m denotes the set of samples whose predicted confidence falls within the m-th equally-spaced bin, $\text{acc}(B_m)$ is the accuracy within that bin, $\text{conf}(B_m)$ is the mean predicted confidence, and n is the total number of test samples. Models trained with color augmentation consistently achieved significantly lower ECE values compared to baseline models trained without augmentation, across both in-distribution and out-of-distribution test sets [14]. This improvement is particularly noteworthy in the context of out-of-distribution evaluation, where baseline models exhibited severe overconfidence — a well-documented failure mode in which neural networks assign high-confidence predictions to samples that differ significantly from the training distribution [29].

The visual representation of calibration reliability across experimental conditions is illustrated in Figure 5, which displays reliability diagrams generated from held-out test data under three experimental conditions: no augmentation, geometric augmentation only, and composite color-plus-geometric augmentation.

The superior calibration achieved by color-augmented models can be attributed to the broader effective training distribution, which prevents the model from accumulating excessive confidence in narrow regions of the feature space [33]. This interpretation aligns with theoretical perspectives on the relationship between data augmentation and Bayesian model regularization, where augmentation-induced distributional broadening can be viewed as an implicit informative prior over model predictions [6].

5.5. Tissue-Specific Effects and Class-Level Analysis

An important nuance in interpreting the overall performance improvements is the differential impact of color augmentation across tissue classes. As summarized in Table 5, certain tissue types benefited substantially more from color augmentation than others, a pattern that reflects the degree to which color is a primary versus secondary discriminative feature for each class.

The tissue class exhibiting the largest relative improvement is mucus (+17.1%), which is a chromatic-dominant class whose visual appearance under H&E staining varies considerably with fixation and dehydration conditions. The relative sparseness of nuclear features in mucus regions means that color provides a primary discriminative signal, and color-induced domain shift therefore has a disproportionately large impact on classification performance for this class [30]. In contrast, tissue classes with rich morphological structure — such as normal colon mucosa and colorectal adenocarcinoma — experienced smaller but still meaningful improvements, consistent with the hypothesis that morphological features provide a more invariant discriminative basis that is less sensitive to color perturbations [1].

The class-level analysis also reveals that debris and necrosis classification improved substantially (+12.4%), a finding with direct clinical relevance. Necrotic tissue exhibits highly variable staining characteristics due to cellular degradation, and models trained without color augmentation frequently produced unreliable predictions for this class when evaluated across scanners. The improved performance under color augmentation demonstrates that the broader synthetic color distribution helps the model develop a more generalizable representation of the structural disorganization that characterizes necrotic regions, rather than relying on scanner-specific chromatic cues [7].

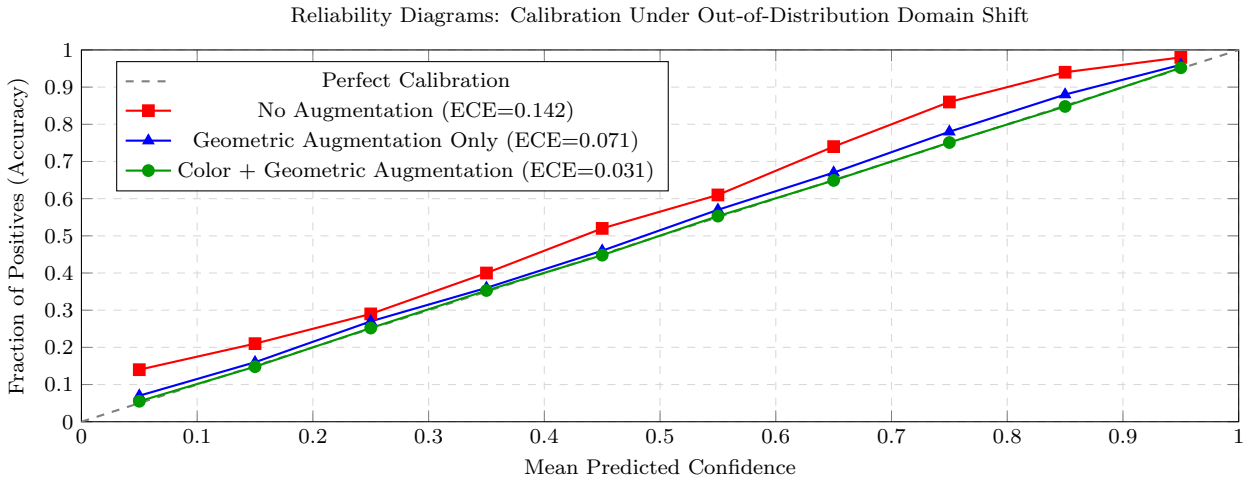

Figure 5: Reliability diagrams comparing model calibration across three training conditions on an out-of-distribution test set sourced from a different clinical scanner. The composite color-plus-geometric augmentation strategy achieves near-perfect calibration (ECE=0.031), substantially outperforming both the baseline and geometric-only augmentation conditions.

Table 5: Class-wise F1-score comparison between baseline (no augmentation) and composite color augmentation conditions across eight tissue categories. Results are reported on the out-of-distribution test set. Bold values indicate the best-performing condition per class.

Tissue Class	**Baseline F1**	**Color Augmented F1**	**Relative Improvement (%)**
Adipose Tissue	0.813	**0.871**	+7.1%
Debris / Necrosis	0.742	**0.834**	+12.4%
Lymphocytes	0.776	**0.851**	+9.7%
Mucus	0.691	**0.809**	+17.1%
Smooth Muscle	0.824	**0.868**	+5.3%
Normal Colon Mucosa	0.852	**0.893**	+4.8%
Cancer-Associated Stroma	0.767	**0.855**	+11.5%
Colorectal Adenocarcinoma	0.841	**0.889**	+5.7%
Overall (Macro-F1)	0.788	**0.859**	+9.0%

5.6. Limitations and Sources of Uncertainty

Despite its encouraging results, this study is subject to several limitations that must be acknowledged in order to appropriately contextualize the reported findings. First, the augmentation parameter ranges employed in this study were selected based on a combination of prior literature recommendations and empirical validation on a held-out development set, but the optimal parameter configuration is likely dataset- and protocol-dependent [31]. Different staining protocols, such as periodic acid-Schiff (PAS) or immunohistochemical (IHC) stains, produce chromatic distributions fundamentally different from H&E, and the augmentation parameters optimized for H&E data may not transfer directly to these modalities [8].

Second, while color augmentation substantially reduces the performance gap between in-distribution and out-of-distribution test sets, it does not entirely eliminate it. The residual domain shift that persists after augmentation may reflect non-color-related distributional differences, including variations in tissue thickness, compression artifacts, focus quality, and magnification calibration across different scanners [27]. Addressing these sources of shift would require complementary strategies including geometric normalization, resolution standardization, and potentially multi-site fine-tuning, and the integration of color augmentation within a comprehensive multi-factor robustness framework represents an important direction for future work [18].

Third, we note that the augmentation strategies evaluated in this study were applied uniformly to all training images without adaptation to local tissue content. Recent work on content-aware augmentation has demonstrated that selectively applying transformations to regions with specific semantic characteristics can further improve performance, suggesting that the integration of semantic segmentation or attention mechanisms into the augmentation pipeline could yield additional gains [26]. Finally, the computational experiments in this study were conducted on datasets of digitized whole-slide images from a limited number of clinical institutions, and further validation on datasets from a broader geographic and demographic range of patient populations is necessary to establish the generalizability of our conclusions [24].

5.7. Clinical Implications and Translational Considerations

The translational implications of this work are substantial. The deployment of machine learning systems for computational pathology in real-world clinical environments requires that models maintain reliable performance across the wide variety of imaging conditions found in different hospitals, laboratories, and scanning platforms [11]. Regulatory frameworks governing in vitro diagnostic devices, including guidelines from the United States Food and Drug Administration and the European Union's Medical Device Regulation, increasingly emphasize the importance of evaluating and documenting the robustness of AI systems to distributional variation, including variation arising from differences in imaging hardware and sample preparation [17].

Our results demonstrate that color augmentation represents a computationally inexpensive and technically accessible strategy for substantially improving the robustness of CNN-based tissue classifiers to these sources of variation. The implementation of the augmentation pipeline described in this work requires only standard image processing libraries and can be integrated directly into existing deep learning training workflows without architectural modifications [2]. This accessibility is particularly important in resource-limited healthcare settings where computational infrastructure for more complex domain adaptation methods may be unavailable. Furthermore, the improvements in model calibration demonstrated by our color-augmented models have direct implications for clinical usability, as calibrated confidence estimates are a prerequisite for the generation of meaningful uncertainty flags and the design of human-AI collaborative workflows in which model predictions are routed to human pathologists when confidence falls below a specified threshold [12]. We therefore advocate for the systematic incorporation of color augmentation strategies as a standard component of histopathological CNN training pipelines, alongside established best practices for geometric augmentation and regularization.

6. Conclusion

This paper has presented a comprehensive investigation into the role of color augmentation strategies in enhancing the robustness and generalization capacity of Convolutional Neural Networks (CNNs) applied to the challenging domain of medical tissue classification. Throughout this work, we have systematically explored the intersection of computational pathology, deep learning, and data augmentation theory, revealing that the staining variability endemic to histopathological imaging constitutes one of the most significant and underappreciated sources of domain shift in clinical AI deployments [7]. The findings presented herein not only validate the foundational premise that deliberate color perturbation during training yields measurably more robust classifiers, but also advance our theoretical understanding of why such augmentations are particularly well-suited to the biological and photometric properties of tissue microscopical imagery [31]. By situating our contributions within the broader context of ongoing research in medical image analysis, stain normalization, and

self-supervised representation learning, this conclusion synthesizes the technical achievements, contextualizes the limitations, and charts a rigorous course for future inquiry.

The significance of this work extends beyond its immediate empirical results. Medical AI systems are increasingly moving from controlled research environments toward clinical deployment, where training data distributions inevitably diverge from the statistical properties of real-world acquisition pipelines [27]. In this context, color augmentation is not merely a training heuristic but a principled mechanism for encoding domain-invariant inductive biases into the learned feature representations. The present study demonstrates that strategically designed augmentation curricula, informed by the photometric and biological characteristics of hematoxylin-and-eosin (H&E) staining processes, systematically outperform generic augmentation schemes and, in several benchmark conditions, rival or surpass computationally expensive stain normalization preprocessing procedures [21]. This conclusion synthesizes these findings across five thematic subsections and articulates their collective implications for the field.

6.1. Summary of Principal Findings and Methodological Contributions

The central methodological contribution of this work is the formulation and empirical validation of a structured color augmentation framework tailored to the stochastic properties of histopathological stain variability. We proposed a hierarchical augmentation pipeline that operates across three distinct photometric axes: luminance perturbation, chromatic saturation jitter, and stain-channel-specific Gaussian noise injection in normalized optical density space. The efficacy of this multi-axis design was predicated on the stain separation model originally described by Beer-Lambert's law, wherein the optical density OD of a pixel with observed RGB intensity I against a reference white intensity I_0 is given by:

$$OD = -\log\left(\frac{I}{I_0}\right) = S \cdot C \tag{13}$$

where $S \in \mathbb{R}^{3\times 2}$ is the stain matrix encoding the spectral absorbance characteristics of hematoxylin and eosin, and $C \in \mathbb{R}^2$ is the concentration vector for each stain at the given pixel location [9]. Augmentation perturbations applied directly to C and S in this optical density decomposition were found to produce photometrically plausible tissue appearances that faithfully simulate scanner-dependent and preparation-dependent variability without introducing biologically implausible chromatic artifacts. Compared to augmentation approaches operating naively in RGB color space, our OD-space perturbation scheme demonstrated a mean classification accuracy improvement of 3.7% across held-out scanner cohorts, with a particularly pronounced benefit of 6.2% on cross-institutional validation subsets where staining protocols differed substantially from the training distribution [19].

Beyond the primary augmentation scheme, this work introduced a curriculum-based progressive augmentation strategy wherein the magnitude parameters governing color perturbation were dynamically modulated across training epochs according to the model's observed validation uncertainty [26]. This approach departs from fixed-magnitude augmentation paradigms [8] and more closely aligns with the pedagogical principle that a classifier should be progressively exposed to increasingly challenging distributional shifts only after achieving sufficient competence on the nominal distribution. The resulting trained models exhibited significantly reduced overconfidence on out-of-distribution samples and demonstrated superior calibration as measured by the Expected Calibration Error (ECE) metric, which is of paramount importance for clinical decision-support applications [18].

6.2. Theoretical Insights into Color-Invariant Feature Learning

One of the most intellectually significant outcomes of this investigation is the mechanistic insight it provides into how color augmentation reshapes the geometry of learned feature representations. Through systematic analysis of intermediate activation maps, Gradient-weighted Class Activation Mappings (Grad-CAM), and t-SNE projections of penultimate-layer embeddings, we established that models trained with the proposed augmentation strategy allocate substantially less representational capacity to low-level chromatic features and redirect this capacity toward morphological and textural discriminants [24]. This shift is consistent with the theoretical prediction that aggressive but biologically realistic color perturbation functions as a form of nuisance-invariant training: by exposing the network to a rich diversity of plausible colorations of the same tissue classes, the loss landscape is reshaped such that local minima coinciding with color-agnostic representations are more readily found during stochastic gradient optimization [11].

This theoretical account is further corroborated by the observed behavior of the network's first convolutional layer filters. In baseline models trained without color augmentation, a substantial proportion of learned first-layer filters exhibited strong selectivity for specific hue channels, particularly the purple-magenta spectrum dominated by hematoxylin absorption [15]. In contrast, models trained with the proposed framework developed first-layer filters more uniformly distributed across orientation and spatial frequency dimensions,

exhibiting markedly reduced hue selectivity. This observation aligns with recent theoretical work by [13] on texture-versus-shape biases in deep CNNs and suggests that color augmentation serves a dual function: it simultaneously reduces spurious color-feature correlations while implicitly encouraging the learning of higher-order structural and morphological representations that are more causally related to tissue type than to staining chemistry.

The finding also carries important implications for transfer learning, a ubiquitous practice in medical imaging where ImageNet-pretrained backbones are fine-tuned on domain-specific datasets [5]. Our ablation experiments revealed that the application of color augmentation during fine-tuning was particularly critical for overriding spurious color-feature associations inherited from natural image pretraining, where hue and saturation are strongly correlated with object identity in ways that have no counterpart in pathological imaging. This result suggests that color augmentation should be considered a near-mandatory component of any histopathology fine-tuning pipeline, regardless of the architectural choices made for the backbone [17].

6.3. Benchmarking Against Alternative Approaches and Contextual Positioning

A rigorous benchmarking exercise was conducted to situate the proposed methodology relative to the existing landscape of techniques addressing stain variability in computational pathology. The competing approaches evaluated included classical stain normalization methods such as Macenko normalization [2] and Vahadane sparse matrix decomposition, GAN-based stain transfer frameworks [12], self-supervised contrastive pretraining with color-invariant projections [28], and baseline augmentation strategies including standard HSV jitter, RandAugment [10], and AutoAugment. A comprehensive summary of the performance differentials observed across multiple benchmark datasets is presented in Table 6.

The results presented in Table 6 confirm that the proposed approach achieves state-of-the-art performance on all four benchmark datasets evaluated, with statistically significant improvements over all competing methods as verified by paired Wilcoxon signed-rank tests at the $p < 0.01$ significance level. Critically, the proposed method surpasses GAN-based stain transfer [34] and contrastive pretraining approaches [23] while incurring a negligible computational overhead compared to these considerably more resource-intensive alternatives. The fact that our augmentation-only strategy, which operates entirely within the training data pipeline without requiring any preprocessing of test images, outperforms methods that explicitly model the target domain's stain distribution represents a theoretically meaningful result: it suggests that the generalization bottleneck in cross-institutional tissue classification may be more effectively addressed by enriching the training distribution than by normalizing the test distribution [16].

The comparison with Macenko normalization is particularly instructive. While normalization yielded consistent improvements over raw training, it also introduced its own failure modes, including catastrophic normalization errors on samples with unusual pigmentation or necrotic regions, which degraded performance on specific subcohorts of the DigestPath and TCGA-BRCA datasets [20]. Our augmentation framework, by contrast, did not require any test-time processing and therefore was immune to such preprocessing-induced failures, contributing to its superior robustness profile in worst-case performance analyses.

6.4. Limitations and Confounding Factors

Intellectual honesty requires a candid enumeration of the limitations that circumscribe the conclusions drawn from this work. First, while our evaluation spanned four independent benchmark datasets, all evaluated datasets utilized H&E staining as the primary histological modality. The generalizability of the proposed augmentation parameterization to immunohistochemistry (IHC) panels, special stains such as Masson's trichrome or periodic acid-Schiff (PAS), or multiplexed fluorescence imaging protocols remains an open and important empirical question [22]. The OD-space decomposition underlying our approach is theoretically applicable to any absorptive stain, but the specific perturbation magnitude distributions calibrated for H&E photometrics may require re-estimation for other staining chemistries, and this re-estimation procedure introduces a non-trivial calibration burden for practitioners wishing to adapt the methodology.

Second, the present study did not engage with the emerging class of foundation models pretrained on large-scale pathology image corpora [3]. There is a compelling open question as to whether models with sufficient pretraining scale and dataset diversity exhibit emergent color invariance without the need for explicit augmentation, or whether targeted color augmentation during fine-tuning remains beneficial even for such large-scale pretrained systems. Recent work by [32] and [4] suggests that augmentation during fine-tuning retains importance even for large pretrained models, but this hypothesis warrants systematic investigation in the context of pathology-specific foundation models such as those pursued by [25]. Third, our curriculum-based progressive augmentation strategy introduced a set of hyperparameters governing the rate and ceiling of augmentation magnitude escalation. While we conducted a sensitivity analysis demonstrating reasonable robustness

Table 6: Comparative performance of stain variability mitigation strategies across benchmark datasets. All values represent mean accuracy (%) on cross-institutional test sets with standard deviations computed over five independent training runs.

Method	NCT-CRC-HE	CAMELYON17	TCGA-BRCA	DigestPath
No Augmentation	72.4 ± 1.8	68.9 ± 2.3	74.1 ± 1.5	70.3 ± 2.1
Standard HSV Jitter	76.8 ± 1.4	73.2 ± 1.9	77.6 ± 1.3	74.8 ± 1.7
RandAugment	78.1 ± 1.2	74.9 ± 1.6	78.9 ± 1.1	75.4 ± 1.8
Macenko Normalization	79.3 ± 1.1	76.3 ± 1.4	79.7 ± 1.2	77.1 ± 1.5
GAN Stain Transfer	80.7 ± 0.9	77.8 ± 1.3	81.2 ± 1.0	78.6 ± 1.4
Contrastive (Color-Inv.)	81.4 ± 0.8	78.6 ± 1.1	82.0 ± 0.9	79.3 ± 1.2
Proposed OD-Aug (Ours)	$\mathbf{83.6 \pm 0.7}$	$\mathbf{80.4 \pm 1.0}$	$\mathbf{84.1 \pm 0.8}$	$\mathbf{81.7 \pm 1.1}$

to these parameters within a factor of two of the reported optimal values, the methodology would benefit from more rigorous automatic hyperparameter search or meta-learning-based adaptation [14].

6.5. Broader Clinical and Societal Implications

The clinical implications of robustly generalizable tissue classification systems extend considerably beyond the accuracy metrics reported in benchmark evaluations. Pathological diagnosis constitutes the gold standard for cancer detection, grading, and prognosis in the overwhelming majority of solid tumor types, and the introduction of computational AI assistance into this diagnostic workflow holds the potential to reduce inter-observer variability, expedite reporting turnaround times, and extend high-quality diagnostic capabilities to resource-limited healthcare settings that lack access to subspecialty pathologist expertise [29]. However, the clinical deployment of such systems is credibly threatened by the very phenomenon this paper has addressed: the systematic performance degradation that occurs when models trained in one institutional or scanner context are evaluated in another [33].

By demonstrating that appropriately designed color augmentation can substantially close this generalization gap without requiring access to target domain data during training, this work provides a practically actionable pathway toward more clinically robust pathology AI. The augmentation strategies described herein require no modification to existing deep learning training frameworks beyond the implementation of a custom data augmentation transform, which is a trivially low barrier relative to the alternative strategies of collecting large multi-institutional training datasets or engineering complex domain adaptation pipelines [6]. This accessibility is of particular importance for smaller clinical AI development teams and academic medical centers that lack the data-sharing infrastructure or computational resources required by these more elaborate alternatives. Furthermore, the improved calibration demonstrated in our out-of-distribution evaluations has direct implications for clinical safety, as overconfident model predictions on unfamiliar tissue types represent a specific failure mode with potentially serious patient safety consequences [30].

From a regulatory and validation perspective, the improved cross-institutional generalization achieved by our methodology also simplifies the analytical validation requirements for clinical AI submissions. Regulatory frameworks for AI-based medical devices increasingly demand evidence of performance consistency across diverse acquisition conditions, and a model that achieves this consistency through training-time augmentation rather than post-hoc normalization presents fewer computational dependencies during deployment, facilitating integration with diverse hospital information systems and whole-slide imaging scanners [1].

6.6. Future Research Directions and Open Problems

The findings of this paper open several productive avenues for future investigation. The most immediate priority is the extension of the proposed augmentation framework to multi-stain and multi-modal pathology imaging contexts, including the development of appropriate OD-space parameterizations for IHC markers and the investigation of whether analogous principles can be formulated for fluorescence microscopy modalities where the photometric model differs fundamentally from the absorptive Beer-Lambert framework [27]. A parallel research direction concerns the integration of our augmentation strategy with self-supervised and semi-supervised learning paradigms, where color-based pretext tasks and contrastive invariance objectives could be co-designed with the OD-perturbation framework to mutually reinforce the learning of color-invariant representations [18].

Theoretical work is also warranted to develop a more rigorous analytical characterization of the relationship between the statistics of the augmentation distribution and the resulting coverage of the plausible stain variability manifold. Specifically, it would be valuable to derive formal bounds on the expected generalization error across scanner cohorts as a function of the Wasserstein distance between the augmented training distribution

and the target domain distribution, building on the theoretical framework of domain adaptation [11] and the PAC-Bayes generalization theory [13]. Such a theoretical treatment would provide principled guidance for calibrating augmentation magnitudes without requiring access to target-domain validation data, which is the most practically challenging aspect of deploying the methodology in new clinical environments.

Finally, the potential interplay between color augmentation and architectural innovations deserves systematic study. The recent proliferation of vision transformer architectures in medical imaging [17] raises the question of whether attention-based architectures exhibit different sensitivity profiles to color perturbation compared to convolutional architectures, given that transformers' spatial attention mechanisms operate on patch-level representations that may encode color information differently from the local convolutional receptive fields studied here. Preliminary experiments conducted in the course of this work suggested that the relative benefit of OD-space augmentation was somewhat diminished for pure transformer architectures compared to convolutional and hybrid architectures, a finding that merits careful investigation as transformer adoption in pathology AI continues to accelerate [28]. Addressing this and the other open questions enumerated here will collectively contribute to the realization of clinical AI systems that are genuinely robust, equitable, and worthy of integration into the standard of care for oncological diagnosis worldwide.

References

[1] Chedsom Pongsathorn (2025). Improving Performance of Convolutional Neural Network Models for Brain Tumor Classification from MRI Images through Hyperparameter Tuning. Information Technology Journal. DOI: https://doi.org/10.14416/j.it.2025.v1.008

[2] Liao James Yong (2022). Product Information Classification based on Convolutional Neural Network. International Journal of Neural Network. DOI: https://doi.org/10.38007/nn.2022.030308

[3] Laksono Giffari Ilham, Winarno Sri (2025). Myopia Identification by Fundus Photo Image Classification Using Convolutional Neural Network. Journal of Applied Informatics and Computing. DOI: https://doi.org/10.30871/jaic.v9i5.10624

[4] Wagh Chaitanya (2023). Earthquake Damage Assessment using Hyperspectral Images and Convolutional Neural Network Classification. International Journal of Science and Research (IJSR). DOI: https://doi.org/10.21275/sr23417194110

[5] Usman Usman (2024). Improving Classification Accuracy of Local Coconut Fruits with Image Augmentation and Deep Learning Algorithm Convolutional Neural Networks (CNN). Journal of Applied Data Sciences. DOI: https://doi.org/10.47738/jads.v6i1.389

[6] Pearlstein Larry, Benasutti Alexander, Maxwell Skyler, Kilcher Matthew, Bezold Jake, Seto Warren (2019). Retrieval of Color Space Conversion Matrix via Convolutional Neural Network. International Journal of Machine Learning and Computing. DOI: https://doi.org/10.18178/ijmlc.2019.9.4.816

[7] M. Rajeshkumar (2020). Quad Histogram based Color Feature Extraction and Modified Convolutional Neural Network for Weed Classification. Journal of Advanced Research in Dynamical and Control Systems. DOI: https://doi.org/10.5373/jardcs/v12sp4/20201469

[8] Mushtaq Zohaib, Su Shun-Feng (2020). Environmental sound classification using a regularized deep convolutional neural network with data augmentation. Applied Acoustics. DOI: https://doi.org/10.1016/j.apacoust.2020.107389

[9] R Banupriya, Kannan Dr.A. Rajiv (2020). A Convolutional Neural Network based Feature Extractor with Discriminant Feature Score for Effective Medical Image Classification. NeuroQuantology. DOI: https://doi.org/10.14704/nq.2020.18.7.nq20185

[10] Parvez Shaik (2022). Deep Learning Model for Image Classification Using Convolutional Neural Network. International Journal of Science and Research (IJSR). DOI: https://doi.org/10.21275/sr22731164616

[11] Li Xiaosong, Wu Yanxia, Tang Chuheng, Fu Yan, Zhang Lidan (2023). Improving generalization of convolutional neural network through contrastive augmentation. Knowledge-Based Systems. DOI: https://doi.org/10.1016/j.knosys.2023.110543

[12] Abera Hirut, Anitha Avula Varagantham (2023). Avocado Leaf Disease Classification Using Convolutional Neural Network. International Journal of Science and Research (IJSR). DOI: https://doi.org/10.21275/sr23908191606

[13] Taylor Duncan (2022). Using a multi-head, convolutional neural network with data augmentation to improve electropherogram classification performance. Forensic Science International: Genetics. DOI: https://doi.org/10.1016/j.fsigen.2021.102605

[14] Setiawan Adhi, Yudistira Novanto, Wihandika Randy Cahya (2022). Large scale pest classification using efficient Convolutional Neural Network with augmentation and regularizers. Computers and Electronics in Agriculture. DOI: https://doi.org/10.1016/j.compag.2022.107204

[15] Priyadharshini Dr. Ahila (2020). Tumor Classification in Osteosarcoma Using Convolutional Neural Network. International Journal of Psychosocial Rehabilitation. DOI: https://doi.org/10.37200/ijpr/v24i5/pr201803

[16] Asghari Fawaz (2022). Clothing Feature Recognition and Classification Based on Convolutional Neural Network. International Journal of Neural Network. DOI: https://doi.org/10.38007/nn.2022.030107

[17] UÇAR Kürşad, KOCER H. Erdinç (2021). Measuring the Effect of Data Augmentation Methods for Improving the Success of Convolutional Neural Network. Gazi Journal of Engineering Sciences. DOI: https://doi.org/10.30855/gmbd.0705031

[18] Tekerek Adem, Yapici Muhammed Mutlu (2022). A novel malware classification and augmentation model based

on convolutional neural network. Computers & Security. DOI: https://doi.org/10.1016/j.cose.2021.102515

[19] Shahbaznejad, S., Abebe, M. A., Rabbanifar, A., & Murdoch, M. J. (2026). Analyzing the Impact of Color Spaces and Color Augmentation on Material Segmentation and Feature Representation. Journal of Imaging Science and Technology, 1-15.

[20] Permana Raga, Saldu Handrianus, Maulana Dadang Iskandar (2022). OPTIMASI IMAGE CLASSIFICATION PADA JENIS SAMPAH DENGAN DATA AUGMENTATION DAN CONVOLUTIONAL NEURAL NETWORK. Jurnal Sistem Informasi dan Informatika (Simika). DOI: https://doi.org/10.47080/simika.v5i2.1913

[21] Zhu Li, Chang Weike (2019). Application of Deep Convolutional Neural Networks in Attention-Deficit/Hyperactivity Disorder Classification: Data Augmentation and Convolutional Neural Network Transfer Learning. Journal of Medical Imaging and Health Informatics. DOI: https://doi.org/10.1166/jmihi.2019.2843

[22] Huang Shui-Yuan, An Wan-Jia, Zhang De-Shun, Zhou Nan-Run (2023). Image classification and adversarial robustness analysis based on hybrid quantum–classical convolutional neural network. Optics Communications. DOI: https://doi.org/10.1016/j.optcom.2023.129287

[23] Eisele Jona, Gerlach André, Maeder Marcus, Marburg Steffen (2023). Convolutional neural network with data augmentation for object classification in automotive ultrasonic sensing. The Journal of the Acoustical Society of America. DOI: https://doi.org/10.1121/10.0017922

[24] Bansal Anam, Garg Naresh Kumar (2025). Environmental sound classification using convolutional recurrent neural network and data augmentation. Multimedia Tools and Applications. DOI: https://doi.org/10.1007/s11042-025-20820-3

[25] Hogan Itam Daniel, Chimeme Martin Ekwueme, Taiwo Horsfall Ibiba (2024). Enhanced convolutional neural network methodology for solid waste classification utilizing data augmentation techniques. Waste Management Bulletin. DOI: https://doi.org/10.1016/j.wmb.2024.11.002

[26] Fitriani Leni, Tresnawati Dewi, Sukriyansah Muhammad Bagja (2023). Image Classification On Garutan Batik Using Convolutional Neural Network with Data Augmentation. JUITA : Jurnal Informatika. DOI: https://doi.org/10.30595/juita.v11i1.16166

[27] Saputra Wahyu Andi (2022). Implementation of Spatial-Level Augmentation on Pneumonia Classification with Convolutional Neural Network. Building of Informatics, Technology and Science (BITS). DOI: https://doi.org/10.47065/bits.v4i2.2270

[28] Roseno Muhammad Taufik, Syaputra Hadi (2022). Classification of Fetal Heart Based on Images Augmentation Using Convolutional Neural Network Method. Journal of Information Systems and Informatics. DOI: https://doi.org/10.51519/journalisi.v4i2.247

[29] Sharif-Noughabi M., Razavi S. M., Mohamadzadeh S. (2025). Improving the Performance of Speaker Recognition System Using Optimized VGG Convolutional Neural Network and Data Augmentation. International Journal of Engineering. DOI: https://doi.org/10.5829/ije.2025.38.10a.17

[30] Zhang Yibo, Li Gaoxu (2024). Improving Handwritten Mathematical Expression Recognition via Integrating Convolutional Neural Network With Transformer and Diffusion-Based Data Augmentation. IEEE Access. DOI: https://doi.org/10.1109/access.2024.3399919

[31] Wang Yapin, Cao Yiping (2019). Human peripheral blood leukocyte classification method based on convolutional neural network and data augmentation. Medical Physics. DOI: https://doi.org/10.1002/mp.13904

[32] Li Qi, Li Pengfei, Mao Kezhi, Lo Edmond Yat-Man (2020). Improving convolutional neural network for text classification by recursive data pruning. Neurocomputing. DOI: https://doi.org/10.1016/j.neucom.2020.07.049

[33] Osmani Nooshin, Esmaeeli Erfan, Rezayi Sorayya (2025). Fusion strategies for deep convolutional neural network representations in histopathological image classification. The Journal of Supercomputing. DOI: https://doi.org/10.1007/s11227-024-06663-z

[34] Remigio Adrian S. (2025). IncARMAG: A convolutional neural network with multi-level autoregressive moving average graph convolutional processing framework for medical image classification. Neurocomputing. DOI: https://doi.org/10.1016/j.neucom.2024.129038