iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.81663

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

Deep Learning for Histopathological Image Classification: A Survey with Case Study on Lung and Colon Cancer Detection Using Transfer Learning

Ganesh Magar[1], Sanket Kakade[2], Pranav Kolte[3], Yogesh Handge[4]
[1, 2, 3]*Student, SCTR's Pune Institute of Computer Engineering, Pune, Maharashtra, India*
[4]*Assistant Professor, SCTR's Pune Institute of Computer Engineering, Pune, Maharashtra, India*

Abstract: ***This paper presents a comprehensive survey on the application of deep learning (DL) in computational pathology for classification of lung and colon cancer from histopathologi-cal images. Though manual diagnosis is paramount for patient outcomes, it remains a tedious, subjective process that is suscep-tible to inter-observer variability. This work compares founda-tional CNN architectures against the most recent SOTA mod-els and performs a reproducible base case study using transfer learning on the publicly available LC25000 dataset, consisting of 25,000 images across five classes.***
Our case study compares three pre-trained models-VGG16, Xception, and DenseNet121-all of which have a frozen base with a custom classifier head. The experimental results showed that DenseNet121 was the better baseline model, with an accuracy of 97.64%, outperforming the other two: Xception with 96.48% and VGG16 with 95.08%.
However, this work argues that, while this baseline is strong, ba-sic classification accuracy on this benchmark is a largely solved problem, since SOTA models including Vision Transformers and hybrid networks have shown 99.8-100% accuracy. Thus, the key research frontiers have shifted.
The survey bridges the gap between theoretical foundations and practical implementation, concluding that advanced XAI frameworks, ensemble methods, and federated learning should be a priority in future research work in order to tackle data pri-vacy and make the successful clinical translation of such power-ful diagnostic tools possible.
Keywords: Deep Learning, Convolutional Neural Network (CNN), Transfer Learning, Computational Pathology, Histopathology, Medical Image Analy-sis, Computer-Aided Diagnosis (CAD), Lung Cancer, Colon Cancer, LC25000, DenseNet121, VGG16, Xception, Explainable AI (XAI), Federated Learning.

I. INTRODUCTION

A. Background and Motivation

Cancer remains the leading cause of death worldwide, and lung and colon cancers feature among the most common and lethal malignancies. According to world health statistics, lung cancer causes around 1.8 mil-lion annual deaths, while in colorectal cancer, the number of deaths is more than 900,000 yearly [1]. In general, early and timely diagnosis is critical for effective cancer therapy, and histopathological examination of the biopsied tissue, usually processed with H&E staining, plays a central role in diagno-sis.

The process of this "gold standard" consists of a patholo-gist manually inspecting tissue slides under a microscope in search of morphological anomalies of a complex nature and cellular features indicative of malignancy. In addition, this traditional diagnostic process is very limited by being labor-intensive, time-consuming, and vulnerable to signifi-cant inter-observer variability due to subjective interpretation [2]. Additionally, an unprecedented surge in cancer screen-ings has meant that pathology departments bear an unimagin-able workload while simultaneously dealing with the global shortage of trained experts, leading to diagnostic bottlenecks and disparities in healthcare access [4].

The convergence of WSI technology, yielding gigapixel-sized digital slides; increased computational power; and ad-vances in deep learning has catalyzed a paradigm shift toward computational pathology. Deep learning, especially CNNs, has emerged as the transformative technology in this field. CNNs have emerged as a powerful technology that has been shown to have a remarkable, sometimes superhuman, abil-ity to automatically learn discriminative hierarchical features from visual data. They are, therefore, well-suited for charac-terizing intricate patterns from histopathological images [3].

Transfer learning has emerged as the leading paradigm for medical image analysis, solving the most critical challenge arising due to limited annotated medical datasets. That tech-nique draws on knowledge developed by models pre-trained on large-scale generic datasets, such as ImageNet, encap-sulating rich feature representations that will be effective for medical tasks, therefore reducing data requirements and training time.

B. Problem Statement And Focus

This survey paper ad-dresses the critical challenge of automated histopathological image classification, focusing on the multi-class classifica-tion of lung and colon cancer subtypes. This task is stan-dardized by a key public benchmark, the "Lung and Colon Cancer Histopathological Image Dataset (LC25000)", which provides a balanced collection of 25,000 images across five classes, facilitating robust comparisons.

This paper first establishes a critical performance baseline by presenting a detailed case study comparing three foundational CNNs (VGG16, DenseNet121, and Xception).

We argue that this "solved" status for pure classification ac-curacy has shifted the research frontier. The key remaining research gaps and persistent challenges that block widespread clinical adoption are:

Model Interpretability: Overcoming the "black-box" nature of DL models to build trust and understand diagnostic rea-soning.

Real-World Generalization: Ensuring models are robust against data heterogeneity, particularly stain variations from different labs, and can scale from small image patches to clinically-relevant Whole-Slide Images (WSI).

Data Privacy and Scarcity: Accessing large, multi-center datasets while adhering to strict patient privacy regulations.

C. Contributions

This survey paper makes the following con-tributions:

1) Comprehensive Literature Review: A systematic analysis of recent deep learning approaches, architectural innova-tions, and methodological advances in histopathological im-age classification.
2) Empirical Case Study: A rigorous, reproducible implemen-tation and evaluation of three foundational transfer learning models (VGG16, DenseNet121, Xception) on the LC25000 dataset, providing detailed performance comparisons.
3) Research Gap Identification: An explicit identification of the new research frontiers beyond classification accuracy, focus-ing on interpretability (Explainable AI), generalization (stain normalization, WSI), and data privacy (Federated Learning).
4) Future Directions: A comprehensive roadmap for advancing the field, discussing promising solutions and methodologies critical for successful clinical translation.

D. Paper Organization

The remainder of this paper is orga-nized as follows: Section II presents a comprehensive litera-ture review of deep learning in histopathology. Section III de-scribes the dataset, architectures, and methodology employed in our case study. Section IV presents our comparative anal-ysis against existing SOTA approaches. Section V discusses experimental results. Section VI analyzes the key research gaps and challenges. Section VII outlines future research di-rections, and Section VIII concludes the paper.

II. LITERATURE REVIEW: FROM FOUNDATIONAL MOD-ELS TO THE STATE-OF-THE-ART

The application of artificial intelligence to histopathological image analysis has evolved significantly, transitioning from traditional machine learning based on handcrafted features to the now-dominant deep learning paradigm.

A. The Evolution to Deep Learning

Early approaches in the 1990s and 2000s relied on handcrafted feature extraction, where domain experts manually designed features like tex-ture descriptors, morphological characteristics, and statisti-cal measures [8]. These features were then fed into classical machine learning models such as Support Vector Machines (SVMs) and Random Forests. While moderately successful, these methods were labor-intensive and struggled to general-ize, as their performance was entirely dependent on the qual-ity of the expert-engineered features [9].

The breakthrough came in 2012 with AlexNet [10], a deep Convolutional Neural Network (CNN) that demonstrated a massive performance leap on the ImageNet challenge. This ushered in the deep learning era, proving that deep networks could automatically learn hierarchical feature representations from raw pixel data, eliminating the need for manual feature engineering.

B. The Transfer Learning Paradigm

Despite this break-through, a fundamental challenge in medical imaging is the relative scarcity of large, annotated datasets compared to nat-ural images. Transfer learning emerged as the dominant paradigm to solve this.

The core principle involves using a model pre-trained on a large-scale source dataset (typically ImageNet) and adapt-ing it to a target medical task. This is effective because the initial layers of a CNN learn generic features (e.g., edges, colors, textures) that are applicable across diverse image do-mains. By leveraging these pre-learned features, researchers can achieve high accuracy on medical datasets with signifi-cantly less data and training time, using two main strategies [12,13]:

Feature Extraction (Frozen Base): The pre-trained CNN base is used as a fixed feature extractor, and only a new, small classifier head is trained on the medical data.

Fine-Tuning: The entire model (or parts of it) is "unfrozen" and retrained on the new data, typically with a very low learn-ing rate, to adapt the learned features to the specific nuances of the medical domain.

C. Foundational CNN Architectures and Philosophies

The models used in this paper's foundational case study are namely VGG16, DenseNet121, and Xception because they represent three different objectives in CNN architecture de-sign.

1) VGG16 (Simplicity and Depth): This model was developed at the Oxford University, VGG's innovation was its simplic-ity [14]. 3x3 Convolutional filters were stacked in a deep 16 layers architecture proving it can achive the State of Art performance. It proved the importance of the network depth. However, this depth comes at a high cost: VGG16 has ap-proximately 138 million parameters, making it computation-ally expensive.

2) DenseNet121 (Feature Reuse and Efficiency): DenseNet proposed a way that significantly improved gradient flow, re-duced redundancy, and enhanced feature reuse [19]. In this architecture , each layer is connected to every subsequent layer in a feed-forward fashion. The feature maps from all preceding layers are concatenated and used as input for the current layer. This architecture:

- Uses substantial amounts of features available.
- Increases parameter efficiency by allowing feature reuse.
- It alleviates the vanishing gradient problem. As a re-sult, DenseNet121 achieves exceptional performance with only 7-8 million parameters, making it highly popular for medical imaging.

3) Xception (Efficiency and Depthwise Separable Convolu-tions): This model termed "Extreme Inception" builds on Google's Inception architecture by replacing standard convo-lutions with depthwise separable convolutions [24]. Here the operation takes place in two simple steps: Each input chan-nel is applied with a single filter by a depthwise convolution i.e Spatial filtering independently followed by combining the outputs via the pointwise convolution which is a 1x1 filter. This factorization dramatically reduces the number of param-eters and computational cost, often leading to improved per-formance and efficiency.

D. The LC25000 Benchmark Dataset

We need a meaningful comparison of these models with the help of a standardized benchmark. The Lung and Colon Cancer Histopathological Images (LC25000) dataset, introduced by Borkowski et al. [5], is the benchmark decided for this task.

Content: It contains 25,000 color histopathological images (768x768 pixels) across five classes:

- Colon Adenocarcinoma
- Colon Benign Tissue
- Lung Adenocarcinoma
- Lung Squamous Cell Carcinoma
- Lung Benign Tissue

Key Characteristics : It's a perfectly balanced dataset of 5000 images per class. This ensures the use of accuracy is primary and a reliable metric without needing to worry aboout the class imbalance.

Limitations : The dataset is a central data, it would not fa-cilitate for testing a multi-center generalization theory. The images in the dataset are not properly labelled with the Tis-sue ID's or the patient's ID's, blocking the scope of specific classifications.

E. State-of-the-Art (SOTA) on the LC25000 Benchmark

While the foundational models provide a strong baseline, recent research has pushed performance on the LC25000 dataset to near-perfection.

A clear theme emerges from the literature: the classification task on this specific benchmark is, from a pure accuracy perspective, largely solved. The dif-ferentiation now lies in the architectural innovations used to achieve these results, which can be grouped into three cate-gories:

Advanced CNNs (EfficientNet): The EfficientNet family, which uses "compound scaling" to optimize network depth, width, and resolution, has proven exceptionally effective. Studies have achieved 99.7% test accuracy with fine-tuned EfficientNet-B7 [25] and 99.98% with EfficientNetV2 [23] [Sudhakar et al.].

The Transformer Revolution (ViT & Swin): A recent paradigm shift has been the adaptation of the Transformer ar-chitecture (from natural language processing) to vision tasks. A study by Dabass et al. directly comparing CNNs and Vi-sion Transformers (ViTs) found that a Swin Transformer V2 model achieved "perfect results" of 100% in accuracy, pre-cision, recall, and F1-score. This is attributed to its hierar-chical structure and shifted-window self-attention, which ef-ficiently models the complex, multi-scale contextual patterns of histopathology.

Hybrid and Fusion Models: Other SOTA approaches in-volve combining models. One study achieved 99.99% ac-curacy by concatenating the feature vectors from three dif-ferent models (ResNet-101V2, NASNetMobile, EffNet-B0) before classification [Hassan et al.]. Another novel approach used lightweight CNNs (e.g., MobileNet) only as feature ex-tractors, then fed the selected features into a classical Cubic Support Vector Machine (SVM), achieving 99.8% accuracy [Arsalan et al.].

Table 1. Summary of State-of-the-Art Performance on the LC25000 Dataset

Study (Citation)	Model Architecture(s)	Key Method	Reported Test Accuracy
Case Study (This Paper)	DenseNet121	Baseline Transfer Learning	97.64%
Tummala et al.	EfficientNetB2	Transfer Learning	97.0%
Arsalan et al.	MobileNet, ResNet-18, EffNetB0 + SVM	Feature Extraction + ANOVA/Chi-Squared + Cubic SVM	99.8%
J. et al.	EfficientNet-B7	Fine-tuning + 4 added Dense layers	99.7%
Sudhakar et al.	EfficientNetV2	Automated Method + GradCAM	99.98%
Hassan et al.	ResNet-101V2, NASNetMobile, EffNet-B0	Feature Vector Concatenation (Fusion)	99.99%
Dabass et al.	Swin Transformer V2	End-to-End Transformer Model	100% ("Perfect Results")

F. Research Gaps and the New Frontier

The near-perfect scores in Table 1 strongly suggest that basic classification accuracy on this clean, balanced, patch-based dataset is a "solved" problem. This has shifted the research frontier away from achieving marginal accuracy gains and toward solving the key challenges that block widespread and reliable clinical adoption:

Interpretability (Explainable AI - XAI): Current DL mod-els function as "black boxes," which limits clinical trust [7]. While techniques like Grad-CAM provide heatmap visualiza-tions, more robust explainability is required for error analysis and regulatory approval.

Real-World Generalization: Models trained on "clean" single-center data (like LC25000) often fail when deployed in new clinical settings. They show poor robustness to do-main shift, particularly stain variations from different labs and scanners [8].

Scaling to Whole-Slide Images (WSI): Clinical diagnosis is performed on gigapixel-sized WSIs, not small 768x768 patches. Scaling models from patches to slides, often using Multi-Instance Learning (MIL), is a non-trivial and critical research gap [9].

Data Privacy and Scarcity: Accessing large, diverse, multi-center datasets is hampered by strict patient privacy laws. This has spurred research into Federated Learning, where models are trained locally at different hospitals without data ever being shared.

Uncertainty Quantification: For safe clinical use, a model must "know what it doesn't know." Current models are often overconfident, even when wrong. Developing reliable uncer-tainty estimates is a key area of research [11].

III. METHODOLOGY: A FOUNDATIONAL CASE STUDY

This section details the comprehensive methodology em-ployed in our baseline case study, which rigorously eval-uates three foundational CNN architectures for the task of histopathological image classification.

A. Dataset Description

Source: The study utilizes the pub-licly available "Lung and Colon Cancer Histopathological Images" (LC25000) dataset, originally curated by Borkowski et al. [6]. Composition: The dataset comprises 25,000 color histopathological images (5,000 images per class), stained with Hematoxylin and Eosin (HE). The images were digi-tized from clinical specimens at 20x magnification.

Class Distribution: The dataset is perfectly balanced, which validates the use of accuracy as a primary metric. The five classes are:

Colon Adenocarcinoma (Malignant) Colon Benign Tissue

Lung Adenocarcinoma (Malignant) Lung Benign Tissue

Lung Squamous Cell Carcinoma (Malignant)

B. Data Preprocessing and Splitting

A systematic pipeline was established to load, organize, and prepare the data for training.

Data Loading: Image file paths were systematically mapped to their corresponding class labels and organized into a Pan-das DataFrame for efficient management and splitting.

Data Splitting: To ensure a robust and unbiased evaluation, the dataset was divided into three independent sets using stratified sampling. Stratification ensures that the original 20% per-class distribution is preserved across all splits. The dataset was divided as follows:

Training Set: 80% (20,000 images)

Validation Set: 10% (2,500 images)

Test Set: 10% (2,500 images), held out and used only for the final performance evaluation.

Data Generation and Preprocessing: The Keras ImageData-Generator was used to create an efficient data pipeline that applied uniform preprocessing to all images on the fly: Resizing: All images were resized to a uniform target size of (224, 224, 3). This is required to match the standard input dimensions of the pre-trained architectures.

Normalization: Pixel values were rescaled from the integer range [0, 255] to the floating-point range [0.0, 1.0] by multi-plying by 1./255. This standardizes the input range and sta-bilizes gradient descent during training.

C. Transfer Learning Architecture

The core of the case study used a feature-extraction transfer learning setup. This approach makes use of the strong low-level features (such as edges, textures, and shapes) that CNN models learn from the large ImageNet dataset.

- Base Model Selection: Three base architectures were chosen for this work: VGG16, DenseNet121, and Xception.
- Configuration: Each base model was created with the fol-lowing settings:

weights='imagenet' – loads the pre-trained Im-ageNet weights,
include_top=False – removes the original 1000-class ImageNet classifier,
input_shape=(224, 224, 3) – sets the com-mon input size.

- Layer Freezing: All layers in the pre-trained base mod-els were frozen by setting base_model.trainable = False. This is done to prevent the ImageNet-learned filters from being overwritten ("catastrophic forgetting") during the initial training stage. In this setup, the frozen base acts as a fixed and powerful feature extractor.
- Custom Classification Head: A simple and consistent clas-sification head was added on top of each frozen base to adapt the extracted features to the 5-class problem. The head con-tained the following layers:
- GlobalAveragePooling2D(): Converts the multi-dimensional feature maps into a single vector by tak-ing the average of each map. This is preferred over Flatten() because it reduces the number of param-eters and helps avoid overfitting.
- Dropout(0.5): Randomly drops 50% of the units dur-ing training to reduce overfitting and improve general-ization.
- Dense(5, activation='softmax'): The final layer with 5 neurons (one per class). The softmax activation out-puts a probability for each class.

D. Training and Compilation Protocol

To maintain a fair comparison across all models, the three architectures were compiled and trained using the exact same training setup.

- Optimizer: Adam (Adaptive Moment Estimation) with a learning rate of 0.001 (1e-3).
- Loss Function: categorical_crossentropy, which is the standard choice for multi-class classification tasks with a softmax output layer.
- Metric: Training progress was monitored using accuracy. Callbacks: An EarlyStopping callback was used to avoid overfitting and automatically retain the best model. It was configured to:
  monitor val_accuracy (measured on the 10% val-idation set),
  use patience=2, stopping training if there was no improvement for two epochs,
  set restore_best_weights=True, ensuring the model returned to the best-performing state at the end of training.
- Epochs: Each model was trained for up to 20 epochs, al-though early stopping usually ended training earlier.

E. Evaluation Methodology

Model performance was rigor-ously evaluated on the unseen, held-out 10% test set using a comprehensive set of metrics:
Overall Accuracy: The primary metric, representing the total proportion of correctly classified images.
Confusion Matrix: A visualization of prediction perfor-mance, showing the distribution of true labels versus pre-dicted labels. This is critical for identifying specific mis-classification patterns (e.g., confusing two different types of cancer or classifying a malignant sample as benign).
Classification Report: A detailed report providing per-class and averaged metrics:
Precision: The ability of the classifier not to label a negative sample as positive (TP / (TP + FP)).
Recall (Sensitivity): The ability of the classifier to find all the positive samples (TP / (TP + FN)).
F1-Score: The harmonic mean of precision and recall, pro-viding a single score that balances both.
Validation Curves: Plots of training and validation accu-racy/loss over epochs were analyzed to assess model conver-gence, learning speed, and signs of overfitting.

F. Computational Environment

The experiments were con-ducted in the Kaggle Notebooks cloud environment, utilizing the following hardware and software:
GPU: NVIDIA Tesla P100 (16GB HBM2) or Tesla T4 (16GB GDDR6)
CPU: Intel Xeon Processor RAM: 16 GB
Software Stack: Python 3.8+, TensorFlow 2.x (with Keras API), NumPy, Pandas, and Scikit-learn.

IV. COMPARATIVE ANALYSIS

This section analyzes the results of the foundational case study, comparing the three architectures directly. It then con-textualizes these findings within the broader landscape of state-of-the-art (SOTA) research on the LC25000 dataset.

A. *Case Study Performance: A Clear Architectural Hierarchy*

The identical training protocol applied to all three mod-els allows for a direct comparison of their architectural effec-tiveness for this task. The performance on the held-out 10% test set revealed a clear hierarchy, summarized in Table 2.

DenseNet121 emerged as the clear winner, achieving the highest performance across all metrics. This result aligns with its architectural philosophy:

Table 2. Comparative Performance of Foundational Models (Case Study Results)

Model	Test Accuracy	Precision (Macro Avg)	Recall (Macro Avg)	F1-Score (Macro Avg)
DenseNet121	97.64%	0.98	0.98	0.98
Xception	96.48%	0.97	0.97	0.97
VGG16	95.08%	0.95	0.95	0.95

DenseNet121 (Superior Performance): Histopathological images are inherently multi-scale, containing critical infor-mation at the level of nuclei, cells, and overall tissue archi-tecture. DenseNet's core concept of dense connectivity and feature reuse is theoretically ideal for this domain. It allows the model to explicitly combine low-level features (e.g., a nucleus shape from an early layer) with high-level features (e.g., glandular arrangement from a deep layer) to make its final prediction. This maximum information flow is the most likely reason for its superior performance.

Xception (Strong Performance): Xception's strong showing validates the power of its depthwise separable convolutions. This efficient factorization of the convolution operation al-lows it to build a deep, effective feature extractor with sig-nificantly fewer parameters than VGG16, leading to better performance.

VGG16 (Baseline Performance): VGG16's relative under-performance is logical. It is a "brute force" deep network that, while a powerful feature extractor, lacks the architec-tural efficiencies (like feature reuse or separable convolu-tions) of modern designs. Its high parameter count (138M) also makes it more computationally expensive and prone to overfitting without careful regularization.

B. *Training Efficiency and Convergence*

The models also demonstrated efficient training, with the EarlyStopping call-back halting training once validation accuracy plateaued.

DenseNet121: Converged after 17 epochs, achieving the best validation accuracy of 98.00

Xception: Converged after 18 epochs, with a best validation accuracy of 97.20%.

VGG16: Converged after 18 epochs, with a best validation accuracy of 95.80%.

All models exhibited stable convergence without significant overfitting, validating the effectiveness of the Dropout regu-larization and the EarlyStopping callback.

C. *Comparison with State-of-the-Art (SOTA) Methods*

While our case study's 97.64% accuracy is a strong baseline, it is crucial to place it in the context of recent SOTA research, which has pushed the boundaries of performance on this spe-cific dataset.

As shown in Table 4, recent studies have achieved near-perfect scores. This indicates that the five-class problem on the LC25000 benchmark is approaching saturation and is, from a pure accuracy perspective, largely a solved problem. The superior performance of these SOTA models stems from more advanced techniques not employed in our baseline study, such as:

Fine-Tuning: Unfreezing the base layers to adapt them to the histopathology domain.

Table 3. Performance Comparison on LC25000 Dataset (Baseline vs. SOTA)

Method / Study	Model(s)	Key Method	Accuracy	Year
Our Case Study	DenseNet121	Frozen Base (Baseline)	97.64%	2025
Hatuwal & Thapa	3-layer CNN	Lightweight CNN	97.2%	2020
Mahmood et al.	Ensemble DL	Ensemble	99.3%	2021
Vanitha et al.	Xception + MobileNet	Hybrid Ensemble	99.44%	2024
Abd El-Ghany et al.	Fine-tuned ResNet101	Fine-tuning, Augmentation	99.94%	2023
Ji et al.	Swin Transformer V2	Vision Transformer	100%	2024

Advanced Architectures: Using Vision Transformers (e.g., Swin Transformer), which excel at capturing long-range de-pendencies.
Ensemble Methods: Combining predictions from multiple models.
Extensive Data Augmentation: Applying complex augmen-tations like color jittering and elastic distortions to improve robustness.
Thus, our case study successfully establishes a strong, trans-parent, and reproducible baseline, confirming that even a straightforward transfer learning setup can achieve high accu-racy. The SOTA results confirm that the next research frontier lies beyond pure accuracy, in areas like interpretability (XAI) and real-world generalization.

V. SUPPLEMENTARY NOTE 1: RESULTS AND DISCUSSION

This section presents the quantitative performance of the foundational case study, analyzes error patterns, and dis-cusses the broader implications of these results in the context of state-of-the-art literature, highlighting the key research gaps that remain for clinical deployment.

A. Case Study Performance Analysis

The experimental results demonstrate the strong viability of transfer learning for histopathological image classification. All three founda-tional CNN models achieved high accuracy on the held-out test set, but a clear performance hierarchy emerged, as sum-marized in Table 4.

Table 4. Final Test Set Performance of the Three Foundational Models

Model	Test Accuracy	Loss	Precision	Recall	F1-Score
DenseNet121	97.64%	0.0827	0.98	0.98	0.98
Xception	96.48%	0.1030	0.97	0.97	0.97
VGG16	95.08%	0.1601	0.95	0.95	0.95

DenseNet121 emerged as the best-performing architecture. Its 2.56 percentage point improvement over VGG16 can be attributed to its dense connectivity pattern, which encourages feature reuse and strengthens gradient flow. This architecture appears particularly well-suited for capturing the complex, multi-scale morphological textures present in histopatholog-ical images. This conclusion is reinforced by its validation performance (98.00% accuracy, 0.0706 loss) prior to early stopping.

B. Per-Class Performance and Error Analysis

Per-class metrics provide deeper insights beyond aggregate accuracy.

C. Per-Class F1-Scores on the Test Set

Table 5 presents the per-class F1-scores for all three models.

Table 5. Per-Class F1-Scores on the Test Set

Class	VGG16	DenseNet121	Xception
Colon adenocarci-noma	0.93	0.97	0.96
Colon benign tissue	0.92	0.97	0.94
Lung adenocarci-noma	0.97	0.98	0.98
Lung benign tissue	0.98	0.99	0.98
Lung squamous cell carcinoma	0.96	0.98	0.97
Macro Average	0.95	0.98	0.97

From this per-class analysis, several key insights emerge:

- DenseNet121's Consistency. DenseNet121 delivered consistently high performance across all classes (F1 ≥ 0.97), demonstrating strong discriminative capability.
- Most Common Error Patterns. VGG16 showed frequent misclassifications between colon adenocarcinoma and colon benign tissue (F1 scores of 0.93 and 0.92). This suggests dif-ficulty in capturing subtle morphological variations in colon tissue.
- Easiest Class. Lung benign tissue achieved the highest scores (F1 ≥ 0.98 for all models), likely due to its distinctive and consistent tissue architecture.

D. Discussion: From Accuracy to Clinical Adoption

Although our baseline model achieved 97.64% accuracy, state-of-the-art (SOTA) models—particularly transformer-based architectures—have reported 99.8–100% accuracy on LC25000. This suggests that classification on the clean LC25000 dataset is largely a solved problem. The remaining challenges lie in clinical deployment rather than incremental accuracy improvements.
We identify three major gaps defining the "post-accuracy era" of computational pathology.

E. Gap 1: The Black-Box Problem and Explainable AI (XAI)

A model achieving 99.9% accuracy cannot be deployed clinically unless it provides interpretable justification for its decisions. Techniques such as Grad-CAM offer preliminary interpretability but suffer from low resolution and noisy lo-calization.
Current research directions include:

1) Counterfactual explanations ("Which region influ-enced the prediction?")
2) Region-removal experiments
3) Causal interpretability frameworks

These approaches provide stronger and more clinically trust-worthy explanations than basic saliency maps.

A Model-Centric Enhancements

1) Gap 2: The Generalization Problem and Stain Variation. Even a high-performing LC25000 model may fail on real-world clinical data due to staining variability across institu-tions. Differences in H&E staining protocols, chemical com-position, and scanner hardware introduce significant domain shifts in color distribution (e.g., "pinkish" versus "purpleish" domains). Without stain normalization or domain adaptation strategies, models trained on one distribution often fail to generalize to another. A model trained on one domain will not generalize to another without stain normalization or domain adaptation.
2) Gap 3: The Data Silo Problem & Federated Learning. Clinical data cannot be shared across institutions due to pri-vacy regulations (HIPAA, GDPR). Federated Learning (FL) addresses this by training models locally and sharing only the weights.

However, FL assumes statistically similar data across clients. Because stain variation causes strong domain shifts (Non-IID data), FL models often fail.
Thus, the true challenge is:
How do we perform stain normalization or do-main adaptation in a privacy-preserving, feder-ated environment?
Solving this would unlock robust generalization and enable real-world deployment of computational pathology systems.

VI. SUPPLEMENTARY NOTE 2: FUTURE SCOPE AND RE-SEARCH DIRECTIONS

Despite the promising 97.64% accuracy achieved by the DenseNet121 baseline, this study represents only an initial step toward developing a clinically deployable diagnostic system. The true future scope lies in bridging the gap be-tween a strong proof-of-concept and a trustworthy, general-izable, real-world solution. These future directions may be categorized into three major domains.

A. Model-Centric Enhancements

The first direction fo-cuses on refining and extending the baseline model itself. A natural next step involves fine-tuning. The current study used a frozen DenseNet121 base; unfreezing the later convolu-tional blocks and re-training the network with a low learning rate would allow the model to adapt its learned features more precisely to the morphological characteristics of histopathol-ogy images, potentially improving overall performance. Additionally, more sophisticated data augmentation strate-gies could substantially enhance robustness. The baseline employed only rescaling, but incorporating augmentations such as rotations, flips, zooming, and elastic deformations would expose the model to a broader distribution of tis-sue variations and orientations. Exploring ensemble strate-gies—such as combining DenseNet121, Xception, and a Vi-sion Transformer—could further boost accuracy while pro-viding a degree of uncertainty estimation that is useful for diagnostic decision-making.

B. Clinical Translation and Deployment

The second di-rection focuses on transforming the high-accuracy model into a usable, interpretable tool for pathologists. This requires advancements beyond accuracy, emphasizing explainability and deployment.

A key component is Explainable AI (XAI). While Grad-CAM offers an initial interpretability mechanism, these heatmaps suffer from noise and limited spatial precision. Fu-ture work should explore more advanced, counterfactual XAI techniques capable of answering: "What specific region of the tissue caused the model to make this prediction?" Such causal interpretability is essential for building trust in clinical settings.

Parallel to interpretability, the practical deployment of the model is another critical step. Packaging the trained model into an interactive web application using frameworks such as Flask or FastAPI would allow pathologists to upload whole-slide image patches, obtain model predictions, and visualize explanatory heatmaps. This would serve as a prototype for a computer-aided diagnosis (CAD) system and demonstrate the model's utility in real-world workflows.

C. Field-Level Grand Challenges

The last research di-rection refers to the grand challenges that currently impede the broad clinical adoption of computational pathology sys-tems. A model that has been trained on single-center data, such as LC25000, will not generalize to new hospitals due to shifts in domain among which the most notable variation was in H&E stain protocols and slide-scanning hardware. This creates significant color distribution differences across insti-tutions that severely limit real-world performance.

This generalization challenge intersects with another major issue: data privacy. Robust models require diverse training data from several hospitals, but due to regulatory limitations such as HIPAA and GDPR, direct sharing of this data is not allowed, and this leads to isolated "data silos." Federated Learning has emerged as a promising solution, which enables several institutions to jointly train a global model without ex-changing patient data.

However, the effectiveness of FL is still impeded by the stain variation problem: variations across hospitals yield statisti-cally heterogeneous data (Non-IID), and naive weight aver-aging often fails when combining models that were trained on different color domains, such as "pinkish" versus "pur-pleish" profiles of stains. Hence, the most important research frontier is textitfederated stain normalization: developing algorithms that would achieve color distribution alignment across the in-stitutions in a privacy-preserving manner either prior to or in concert with the FL process. After all, overcoming this chal-lenge is crucial for generalizability and regulatory compli-ance that allows for clinically viable diagnostic AI systems.

VII. SUPPLEMENTARY NOTE 3: CONCLUSION

The following survey has charted the rapid and decisive pro-gression of deep learning approaches concerned with lung and colon cancer histopathological classification. This is done using the LC25000 dataset as a common analytical ground.

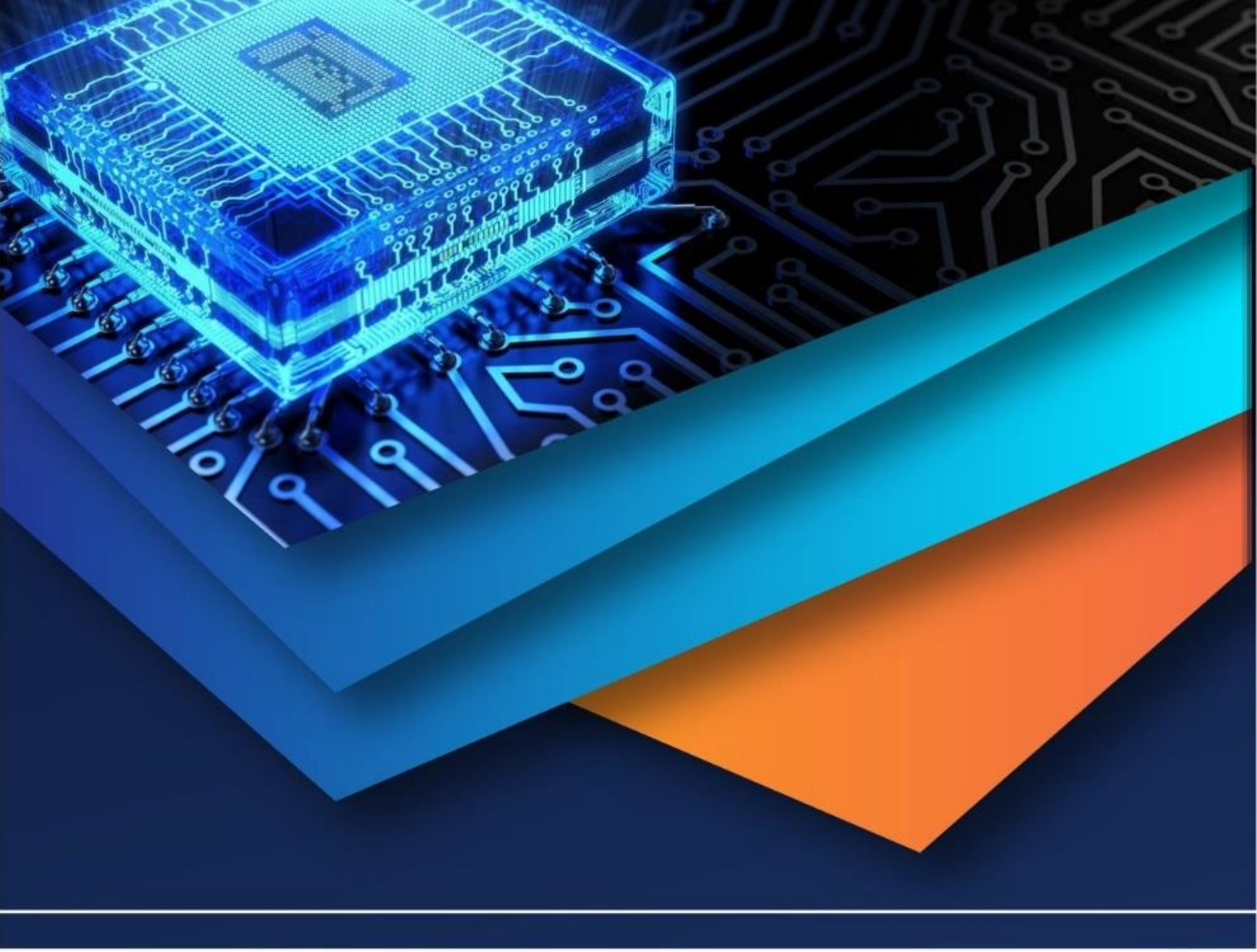
We start by establishing a strong performance baseline through the detailed case study. It is shown that a founda-tional architecture, DenseNet121, can attain a high test accu-racy of 97.64% using a simple, reproducible transfer learning methodology. This superior performance is not arbitrary; the dense connectivity and feature-reuse mechanism inherent to DenseNet are well-matched in theory to the complex, multi-scale, and hierarchical nature of histopathological data. This survey then placed this baseline in context, given the state-of-the-art, and showed that advanced architectures, es-pecially Vision Transformers (e.g., Swin Transformer V2), have driven performance on this benchmark to 99.9–100%. This convergence effectively signifies that the problem of pure classification on this clean, balanced dataset is a "solved" problem. This "post-accuracy" era did not terminate research, but rather shifted the field's focus from marginal gains in accu-racy to the deeper challenges standing in the way of real-world clinical deployment. Probably the biggest research frontiers lie in three related challenges:

1) Trust & Interpretability: Beyond black-box predic-tors to completely interpretable models using Explain-able AI (XAI) that allow clinicians to comprehend and verify model reasoning. Item Generalization: Dealing with variation in staining–the major source of domain shift in pathology–so that the models deployed at dif-ferent hospitals using varied staining and scanning pro-tocols remain robust.
2) Scalability & Privacy: Completely remove the possi-ble silos of data created by patient privacy regulations using Federated Learning to train models collabora-tively without exposing sensitive information.

As this survey has discussed, the latter two challenges are fundamentally intertwined. Stain variability is the principal cause of the statistical heterogeneity that often leads stan-dard FL algorithms to fail. Thus, the future of computational pathology is not about achieving another 0.01% increase on a static benchmark but about building robust, explainable, and privacy-preserving AI systems that can learn collabora-tively and generalize effectively across the diverse and dy-namic data encountered in real-world clinical practice.

REFERENCES

[1] H. Sung et al., "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," CA: A Cancer Journal for Clinicians, vol. 71, no. 3, pp. 209–249, 2021.
[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.
[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convo-lutional neural networks," in Proc. Advances in Neural Information Processing Systems (NIPS), 2012, pp. 1097–1105.
[4] J. Deng et al., "ImageNet: A large-scale hierarchical image database," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2009, pp. 248–255.
[5] A. A. Borkowski et al., "Lung and Colon Cancer Histopathological Image Dataset (LC25000)," arXiv:1912.12142, 2019.
[6] R. Guidotti et al., "A survey of methods for explaining black box models," ACM Comput. Surveys, vol. 51, no. 5, pp. 1–42, 2018.
[7] D. Tellez et al., "Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology," Med. Image Anal., vol. 58, p. 101544, 2019.
[8] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in Proc. ICML, 2018, pp. 2127–2136.
[9] N. Tajbakhsh et al., "Convolutional neural networks for medical image analysis: Full training or fine tuning?," IEEE Trans. Med. Imaging, vol. 35, no. 5, pp. 1299–1312, 2016.
[10] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in Proc. ICML, 2016, pp. 1050–1059.
[11] J. Yosinski et al., "How transferable are features in deep neural networks?," in Proc. NeurIPS, 2014, pp. 3320–3328.
[12] S. Abd El-Ghany et al., "Robustness fine-tuning deep learning model for cancers diagnosis based on histopathology images," Scientific Reports, vol. 13, no. 1, p. 19572, 2023.
[13] M. M. M. Hassan et al., "An Advanced Deep Learning Fusion Model for Multi-Classification of Lung and Colon Cancers Using Histopathological Images," Diagnostics, vol. 14, no. 20, p. 2274, 2024.
[14] P. Sudhakar et al., "Exploring Explainable AI Techniques for Improved Interpretability in Lung and Colon Cancer Classification," ResearchGate, 2025.
[15] V. Dabass et al., "Automated Lung and Colon Cancer Classification Using Histopathological Images," bioRxiv, 2024.
[16] J. Ji et al., "Automated lung and colon cancer classification using medical imaging based on Swin Transformer V2," Frontiers in Oncology, vol. 14, 2024.
[17] G. Huang et al., "Densely Connected Convolutional Networks," in Proc. CVPR, 2017, pp. 2261–2269.
[18] S. Suara et al., "Is Grad-CAM explainable in medical images?," arXiv:2307.10506, 2023.
[19] M. Hägele et al., "Resolving challenges in deep learning-based analyses of histopathologi-cal images using explanation methods," Scientific Reports, vol. 10, p. 16901, 2020.
[20] H. Chen et al., "HIPPO: A framework for Histopathology Interventions of Patches for Predic-tive Outcomes in computational pathology," bioRxiv, 2024.
[21] B. McMahan et al., "Communication-efficient learning of deep networks from decentralized data," in Proc. AISTATS, 2017, pp. 1273–1282.
[22] M. J. Sheller et al., "Federated learning in medicine: Facilitating multi-institutional collabo-rations without sharing patient data," Scientific Reports, vol. 10, p. 12598, 2020.
[23] M. Macenko et al., "A method for normalizing histology slides for quantitative analysis," in Proc. ISBI, 2009, pp. 1107–1110.
[24] M. Asadi-Aghbolaghi et al., "Learning generalizable AI models for multi-center histopathol-ogy classification," Medical Image Analysis, vol. 91, p. 103038, 2024.
[25] N. Coudray et al., "Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning," *Nature Medicine*, vol. 24, no. 10, pp. 1559–1567, 2018.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY