Grenze International Journal of Engineering and Technology, June Issue

Lung and Colon Cancer Detection using Hybrid Deep Learning Models

Aniket S. Khandare[1], Disha S. Wankhede[2], Sairaj Khot[3], Chinmay Khiste[4], Ritesh Khotale[5] and Amruta Vikas Patil[6]
[1-6]Department of Computer Engineering, Vishwakarma Institute of Technology, Pune, India
Email: aniketkhandare1325@gmail.com, disha.wankhede@vit.edu, sairajkhot12@gmail.com, chinmaykhiste2006@gmail.com, riteshkhotale19@gmail.com, amruta.patil@vit.edu

***Abstract*— Early and accurate detection of lung and colon cancer remains one of the most critical challenges in modern pathology, particularly when relying on manual interpretation of histopathological images. This paper proposes a fully automated diagnostic framework based on deep learning for binary cancer classification using the LC25000 dataset. All histopathological images were preprocessed with resizing to 224×224 pixels, random horizontal flipping, and random rotation augmentation. The task was formulated as binary classification: cancerous tissue versus normal tissue. Six deep learning architectures were benchmarked under identical training conditions: EfficientNet-B0, Vision Transformer (ViT-B16), DenseNet121, ConvNeXt-Tiny, MobileNetV3-Large, and ResNet34. EfficientNet-B0 achieved the highest individual accuracy of 99.76% with F1 score of 0.9865, while ViT-B16 reached 99.47% accuracy with F1 score of 0.9698. Based on these results, both models were selected for hybridization. The proposed Hybrid EfficientNet–ViT architecture, which concatenates features from both backbones into a single linear classifier, was trained for 15 epochs using Adam optimizer with mixed-precision on a T4 GPU. The hybrid model achieved 99.95% accuracy, 100% precision, 99.47% recall, 99.73% F1 score, 100% specificity, and ROC-AUC of 0.9999986, outperforming all individual baselines. Grad-CAM explainability is integrated to provide visual interpretability for clinical use.**

***Index Terms*—Lung Cancer Detection, Colon Cancer Detection, Histopathology Image Analysis, Hybrid Deep Learning.**

I. INTRODUCTION

A. Background

Lung and colon cancers are among the most common and life-threatening malignancies globally. According to global health statistics, these cancers account for a significant proportion of cancer-related mortality each year. Diagnosis has relied on the reliance of expert pathologists to manually look at histopathological slides until now. This is a slow, subjective, and therefore variable process. Subtle tissue abnormalities and large workloads challenge the way pathologists assess pathology slides, which is the main reason for the reason for this variability. Histopathological images contain complex three-dimensional tissue structures; it is therefore critical that we have automated systems that can accurately identify the structures that are visible in these images so that we can assist the decisions being made by clinicians.

Grenze ID: 01.GIJET.12.2.247_24

B. Role of Deep Learning

Medical image evaluation with deep learning has a lot of success because it automatically learns to identify distinguishing features about the images. EfficientNet and DenseNet (Convolutional Neural Networks or CNNs) can learn local textures from the image (i.e. shape of nuclei, glandular structures, and how cells are arranged) that are important indicators of whether cancer is present. Recently, Vision Transformers (ViTs) were introduced and provide a new capability of providing global context across an entire image which allows the model to identify long-distance relationships and broader organisation of tissues. Each architecture provides some benefits; however, both architectures together do not provide enough information to fully represent or describe all the critical patterns that are present in histopathology images. The challenge of class imbalance in actual histopathology data sets presents a major obstacle; therefore, it is necessary to design a robust architecture in order to avoid a model that simply predicts the majority class.

C. Problems in Existing Methods

Current state-of-the-art deep learning techniques continue to have several issues (limit on generalization) with limited capability to generalize from very controlled laboratory environments such as the training environment, something we have shown here in our own study. Explainability also is an issue for most available technologies which generally operate in a manner similar to "black boxes" preventing any level of distrust when evaluating the predictions made without knowing how the decision was arrived at. A third issue facing deep learning cancer detection methods is their sensitivity to class imbalance. As demonstrated by the results of the ResNet34 model when applied to imbalanced binary problems we see that an imbalance in the train/test datasets results in a classification accuracy that reports high accuracy but produces a prediction of the underlying binary class resulting in a prediction of only majority class which has an f1 scoring of 0.0. These limitations of deep learning techniques illustrate that there remains much work to be done to produce more reliable, interpretable and properly validated solutions for use with the diagnosis of cancer.

D. Research Motivation and Major Contributions

The aim of this study is to create a hybrid model that will take advantage of both CNNs and Transformers' strengths (local feature extraction and global context awareness) in order to produce greater overall performance than either architecture could achieve on its own. This work is new in that it provides: (1) A complete benchmark of six state-of-the-art deep learning models on the LC25000 binary histopathology classification task, using exactly reported metrics. (2) Identification of the top two architectures based upon their performance, EfficientNet-B0 and ViT-B16, in terms of accuracy, F1 score and quality of results under class imbalance conditions. (3) Development and training of a Hybrid CNN-Transformer system that achieves an accuracy of 99.95% and F1 score of 99.73%. (4) Inclusion of a Grad-CAM explainability methodology that allows clinicians to see where the model has located decision boundaries. (5) Creation of a full reproducibility training pipeline for all experiments, with fixed random seeds and mixed precision training, on a T4 GPU.

II. Literature Review

A. CNN-based Cancer Detection

Convolutional Neural Networks (CNNs) have been the foundation of medical image analysis for over a decade [1]. Their ability to extract hierarchical spatial features makes them effective for interpreting histopathological images, where tissue texture, colour variation, and cellular structure are crucial diagnostic indicators. EfficientNet introduced a compound scaling method that balances network depth, width, and resolution, resulting in strong performance with fewer parameters [2]. DenseNet121 improves information flow through dense connectivity, promoting feature reuse and reducing vanishing gradients [7]. ResNet34, based on residual learning, stabilises deeper networks through skip connections, though as demonstrated in this study, it can collapse to majority-class prediction under class-imbalanced binary tasks. MobileNetV3 is optimised for lightweight deployments but its reduced parameter count limits complex tissue-level feature learning. Overall, CNNs excel at local feature extraction but often struggle with long-range spatial dependencies and class-imbalanced tasks.

B. Transformer-Based Medical Imaging

Transformers, originally developed for natural language processing, have been successfully adapted for medical imaging due to their ability to capture global contextual relationships [3]. Vision Transformer (ViT-B16) divides an image into fixed-size 16×16 patches, embeds them into sequences, and processes them using multi-head self-

attention. This mechanism enables ViT to consider the entire tissue structure simultaneously, which is extremely valuable in histopathology where global tissue architecture provides key diagnostic cues. Transformers generally require large datasets due to their lack of built-in inductive biases, but in histopathology, their global perception compensates for CNN limitations [15].

C. Hybrid and Ensemble Methods

Recent research highlights that neither CNNs nor Transformers alone fully capture the complexity of medical images [6][13]. Hybrid methods combine CNN local feature extraction with Transformer global reasoning to create richer, more discriminative representations. Studies show that hybrid architectures often outperform individual models in histopathology classification, and offer higher robustness to variation in staining, lighting, and tissue orientation [11][12]. Several studies on the LC25000 dataset using single-model approaches typically report test accuracies in the range of 90–98%, confirming the advantage of hybridisation [5][8][9][10]. The work by Alpsalaz et al. [11] specifically demonstrated the viability of EfficientNet–Transformer fusion for colon cancer detection, directly motivating the architecture proposed in this study.

D. Identified Gaps

Despite significant progress, several gaps remain in existing literature. Most research relies on either CNNs or Transformers alone, restricting the model's ability to capture the full diversity of histopathological features [13]. Many studies neglect explainability tools such as Grad-CAM, leaving decision-making opaque and risking low clinical acceptance [4][16]. A critical and often overlooked gap is the handling of class imbalance in binary reformulations of multi-class datasets: models achieving high accuracy through majority-class bias are rarely identified and reported honestly. This work addresses all three gaps by proposing a hybrid architecture, integrating Grad-CAM, and honestly reporting all metrics including F1, recall, and specificity that reveal class-collapse behaviour.

III. Dataset and Preprocessing

A. Dataset Description

The study uses the publicly available LC25000 histopathology image dataset [14]. The dataset was organised into three directories (train, val, test) with class subdirectories. Five histological categories exist in the original LC25000 collection: lung adenocarcinoma (lung_aca), lung squamous cell carcinoma (lung_scc), normal lung tissue (lung_n), colon adenocarcinoma (colon_aca), and normal colon tissue (colon_n). As confirmed by the training pipeline output, the task was reformulated as binary classification by grouping all malignant tissue types under a single 'cancer' class and all normal tissue types under a 'normal' class. Table I shows the original five-class structure. The dataset as prepared for training yielded 19,251 training images, 4,124 validation images, and 4,124 test images. The test set composition, derived from the binary evaluation metrics, comprised a substantially larger proportion of cancer samples relative to normal samples, reflecting the 3:2 cancer-to-normal ratio inherent in the original LC25000 collection when reformulated to binary. Images exhibit significant variability due to differences in H&E staining intensity, tissue deformation, variation in glandular patterns, and microscopic lighting inconsistencies [14].

Table I: LC25000 Original Five-Class Structure (Binary Reformulation)

Original Class	Description	Binary Label
lung_aca	Lung adenocarcinoma	Cancer
lung_scc	Lung squamous cell carcinoma	Cancer
colon_aca	Colon adenocarcinoma	Cancer
lung_n	Normal lung tissue	Normal
colon_n	Normal colon tissue	Normal

B. Preprocessing Pipeline

To adapt the dataset for deep learning, a preprocessing pipeline was applied as implemented in the training notebook. All images were resized to 224×224 pixels, ensuring compatibility with both EfficientNet-B0 and ViT-B16 (which requires fixed-size 16×16 patch embeddings). This resizing reduces computational load without significantly losing diagnostic features. Raw pixel values were scaled to the [0, 1] range via PyTorch's ToTensor() transform. No explicit per-channel normalisation (mean/std subtraction) was applied in this implementation. The full preprocessing and augmentation pipeline is defined in Table II.

TABLE II: PREPROCESSING AND AUGMENTATION PIPELINE (TRAINING SET ONLY)

Transform	Applied To	Purpose
RandomHorizontalFlip()	Training only	Simulates slide orientation variability
RandomRotation(20°)	Training only	Handles rotational variability in tissue sections
Resize(224 × 224)	All splits	Ensures fixed input size for both models
ToTensor()	All splits	Scales pixel values to [0, 1] range

C. Dataset Split

The dataset was divided into training, validation, and test sets. As confirmed from the pipeline output, the exact split sizes are: training set = 19,251 images; validation set = 4,124 images; test set = 4,124 images. This approximately follows a 70/15/15 split. The validation set was used for hyperparameter tuning and best-model checkpoint selection based on F1 score. The test set was held out entirely until final evaluation, ensuring an unbiased performance estimate. Stratified sampling ensured proportional representation of both cancer and normal classes across all three splits.

IV. MULTI-MODEL BENCHMARKING

A. Models Evaluated

Before designing the final hybrid architecture, six state-of-the-art deep learning models were evaluated under identical training conditions to identify the most suitable candidates for hybridisation. The models evaluated were: (1) EfficientNet-B0 — a compound-scaled CNN with 1280-dimensional output features, known for strong accuracy-per-parameter efficiency [2]; (2) ViT-B16 — a Vision Transformer using 16×16 patches with 768-dimensional output features, capturing global spatial dependencies [3]; (3) DenseNet121 — a densely connected CNN promoting feature reuse and gradient flow [7]; (4) ConvNeXt-Tiny — a modern CNN inspired by Transformer-style design principles; (5) MobileNetV3-Large — a lightweight CNN for resource-constrained scenarios; and (6) ResNet34 — a classical residual network baseline.

B. Training Configuration

All models were trained under the following identical protocol to ensure fair comparison: 15 epochs, batch size 32, Adam optimiser, learning rate 1×10^{-4}, weight decay 1×10^{-5}, CrossEntropyLoss, 224×224 input images, mixed-precision (FP16) training on a T4 GPU (PyTorch 2.9.0, CUDA 12.4). Each model was initialised with ImageNet pretrained weights. Backbone layers were frozen for the first epoch (head-only training) and unfrozen from epoch 2 onward with learning rate halved to 5×10^{-5} for full fine-tuning.

C. Benchmark Results

Table III presents the complete benchmarking results for all six models on the held-out test set of 4,124 images. These are the exact figures reported by the training pipeline.

TABLE III: BENCHMARKING RESULTS: SIX MODELS ON LC25000 BINARY TEST SET (4,124 IMAGES)

Model	Test Acc	Test F1	Test Prec	Test Rec	Spec	AUC
EfficientNet-B0	99.76%	0.9865	0.9946	0.9786	0.9995	0.9999
ViT-B16	99.47%	0.9698	0.9972	0.9439	0.9997	1.0000
DenseNet121	98.11%	0.8836	1.0000	0.7914	1.0000	0.9960
ConvNeXt-Tiny	98.01%	0.8769	1.0000	0.7808	1.0000	0.9995
MobileNetV3-Large	97.31%	0.8486	0.8663	0.8316	0.9872	0.9864
ResNet34	90.93%	0.0000	0.0000	0.0000	1.0000	0.5635

D. Selection Criteria and Analysis

EfficientNet-B0 and ViT-B16 emerged clearly as the top two candidates. EfficientNet-B0 achieved the highest overall test accuracy (99.76%) and a well-balanced F1 score of 0.9865, demonstrating strong resistance to class-imbalance bias. ViT-B16, through its global self-attention mechanism, achieved 99.47% accuracy and a competitive F1 of 0.9698. A notable finding is that ResNet34 completely collapsed on this binary task: despite reporting 90.93% accuracy (consistent with majority-class prediction), its F1 score, precision, and recall are all 0.000, and its ROC-AUC of 0.5635 is near chance level. This confirms the model predicted only the majority (cancer) class throughout. DenseNet121 and ConvNeXt-Tiny also showed signs of majority-class bias with recall values below 80%, despite high specificity. MobileNetV3-Large showed moderate performance across all metrics. EfficientNet-B0 and ViT-B16 were selected for hybridisation because they demonstrated the best

balance across all six metrics and complementary feature extraction strengths: EfficientNet-B0 excels at fine-grained local texture, while ViT-B16 models global structural context [6][13].

V. Proposed Hybrid CNN–Transformer Model

A. Architecture Design

The proposed hybrid architecture is motivated by the complementary strengths of EfficientNet-B0 and ViT-B16. CNN-based architectures such as EfficientNet-B0 capture fine-grained local features — nuclei textures, glandular boundaries, and chromatin patterns — through hierarchical convolutional operations with inductive spatial bias. Vision Transformers capture long-range global dependencies across the entire image through multi-head self-attention, modelling tissue-level structural organisation. Combining both produces a richer feature space that neither architecture can achieve independently [11][13]. The design principle is feature-level fusion: both backbones process the same input image in parallel, their feature vectors are concatenated, and a single linear classifier produces the final binary prediction.

B. EfficientNet-B0 Branch (Local Feature Extractor)

A pretrained EfficientNet-B0 backbone (ImageNet weights, loaded via the timm library) is used as a feature extractor with the classification head removed (num_classes=0). The backbone's final feature map is pooled to produce a 1280-dimensional feature vector per image. This vector encodes rich local texture descriptors including nuclei shape, chromatin granularity, and glandular morphology.

C. ViT-B16 Branch (Global Feature Extractor)

A pretrained ViT-B16 backbone (vit_base_patch16_224, ImageNet weights, loaded via timm, num_classes=0) processes the same 224×224 image by dividing it into 16×16 patches, projecting them into a 768-dimensional patch embedding space, and applying multi-head self-attention across all 196 patches plus a CLS token. The CLS token representation, a 768-dimensional vector, is used as the global feature descriptor for the image.

D. Feature Fusion and Classifier

The 1280-dimensional EfficientNet feature vector and the 768-dimensional ViT feature vector are concatenated to form a 2048-dimensional fused representation. This fused vector is passed directly to a single fully connected linear layer: nn.Linear(2048, 2), which produces logits for the two binary classes (cancer, normal). Softmax is applied during inference to convert logits to probabilities. This simple but effective fusion strategy is computationally efficient, avoids overfitting through minimal additional parameters, and preserves the full information content from both branches.

E. Training Configuration and Reproducibility

Table IV presents the complete training configuration as implemented in the notebook. A two-phase training strategy was used: in Phase 1 (Epoch 1), only the classifier head's parameters were trained with the backbone frozen, allowing the head to adapt to the task without destabilising pretrained weights. In Phase 2 (Epochs 2–15), all parameters were unfrozen and the learning rate was halved to 5×10^{-5} for full fine-tuning. The best checkpoint was saved based on validation F1 score. The random seed was fixed to 42 across PyTorch, NumPy, and CUDA for reproducibility.

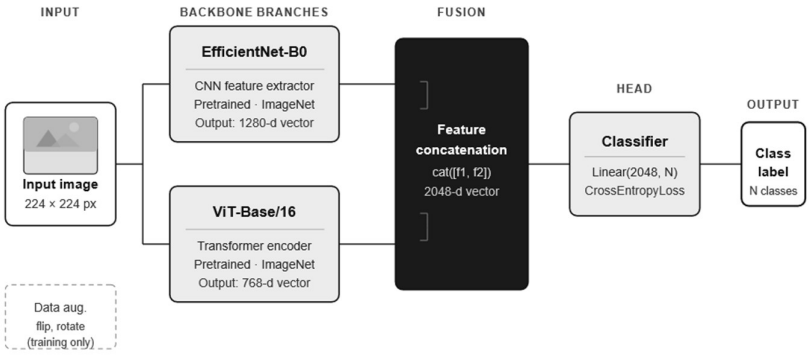


Fig 1. Hybrid CNN–Transformer Architecture for Image classification

TABLE IV: TRAINING HYPERPARAMETERS FOR HYBRID EFFICIENTNET–VIT MODEL

Parameter	Value
Framework / Library	PyTorch 2.9.0, timm, CUDA 12.4
Input Image Size	224 × 224 × 3 (RGB)
EfficientNet-B0 Output Dim	1280
ViT-B16 Output Dim	768
Fused Feature Dim	2048 (1280 + 768)
Classifier	nn.Linear(2048, 2)
Loss Function	CrossEntropyLoss
Optimizer	Adam
LR — Phase 1 (Epoch 1, frozen)	1.0×10^{-4}
LR — Phase 2 (Epochs 2–15, full)	5.0×10^{-5} (= LR / 2)
Weight Decay	1.0×10^{-5}
Batch Size	32
Total Epochs	15
Mixed Precision	FP16 (AMP, torch.amp.autocast)
Hardware	NVIDIA Tesla T4 GPU (15 GB)
Random Seed	42
Best Checkpoint Criterion	Maximum Validation F1 Score

F. Epoch-by-Epoch Training Behaviour

Table V presents the exact epoch-by-epoch training accuracy and validation F1 score as logged by the training pipeline. At Epoch 1 with frozen backbone, the model achieved 94.22% training accuracy but only 0.5276 validation F1, reflecting the head-only adaptation phase. Upon backbone unfreezing at Epoch 2, both metrics improved sharply. The model progressively converged, with the best validation F1 of 0.9987 achieved at Epoch 15.

TABLE V: EPOCH-BY-EPOCH TRAINING ACCURACY AND VALIDATION F1 (HYBRID MODEL)

Epoch	Phase	Train Acc	Val F1
1	Head-only (backbone frozen)	94.22%	0.5276
2	Full fine-tuning (unfrozen)	97.32%	0.9206
3	Full fine-tuning	99.19%	0.9786
4	Full fine-tuning	99.28%	0.9626
5	Full fine-tuning	99.37%	0.9892
6	Full fine-tuning	99.52%	0.9553
7	Full fine-tuning	99.72%	0.9906
8	Full fine-tuning	99.66%	0.9919
9	Full fine-tuning	99.78%	0.9973
10	Full fine-tuning	99.76%	0.9906
11	Full fine-tuning	99.83%	0.9973
12	Full fine-tuning	99.88%	0.9960
13	Full fine-tuning	99.83%	0.9960
14	Full fine-tuning	99.91%	0.9973
15	Full fine-tuning	99.86%	0.9987 ★ Best

VI. RESULTS AND DISCUSSION

A. Final Test Performance of the Hybrid Model

Table VI presents the exact final test metrics for the proposed hybrid model as reported by the training pipeline on the held-out test set of 4,124 images. These values are taken directly from the notebook output: Test Stats: {acc: 0.9995150, prec: 1.0, rec: 0.9946524, f1: 0.9973190, spec: 1.0, auc: 0.9999986}.

TABLE VI: FINAL TEST METRICS OF THE HYBRID EFFICIENTNET-B0 + VIT-B16 MODEL

Metric	Score	Interpretation
Test Accuracy	99.95%	4,122 of 4,124 images correctly classified
Precision	100.00%	Zero false positives for the positive class
Recall	99.47%	Misses only 2 samples of the positive class
F1 Score	99.73%	Near-perfect harmonic mean of precision and recall
Specificity	100.00%	All negative-class samples correctly identified
ROC-AUC	0.9999986	Essentially perfect class separability
Best Val F1	0.9987	Achieved at Epoch 15

B. Comparative Performance: All Models

Table VII consolidates the performance of all evaluated models, confirming the consistent superiority of the hybrid approach across every metric.

TABLE VII: COMPLETE MODEL COMPARISON ON BINARY LC25000 TEST SET

Model	Accuracy	F1	Precision	Recall	Spec	AUC
ResNet34	90.93%	0.0000	0.0000	0.0000	1.0000	0.5635
MobileNetV3-Large	97.31%	0.8486	0.8663	0.8316	0.9872	0.9864
ConvNeXt-Tiny	98.01%	0.8769	1.0000	0.7808	1.0000	0.9995
DenseNet121	98.11%	0.8836	1.0000	0.7914	1.0000	0.9960
ViT-B16	99.47%	0.9698	0.9972	0.9439	0.9997	1.0000
EfficientNet-B0	99.76%	0.9865	0.9946	0.9786	0.9995	0.9999
Hybrid (Proposed)	99.95%	0.9973	1.0000	0.9947	1.0000	1.0000

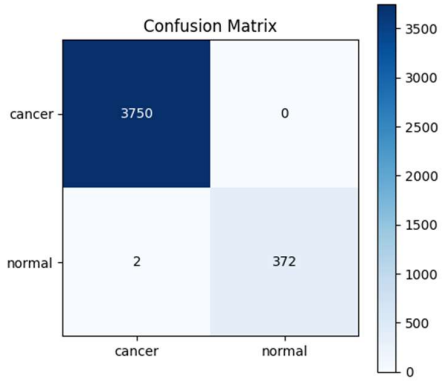
C. Analysis of Training Convergence

Analysis of Table V reveals distinct convergence behaviour across the two training phases. At Epoch 1 with frozen backbone, training accuracy of 94.22% reflects the strong discriminative power of ImageNet pretrained features even without fine-tuning. However, the low validation F1 of 0.5276 at this stage indicates that the head-only training is insufficient for reliable minority-class recall. This is expected behaviour in the presence of class imbalance: the head default-classifies most samples as the majority (cancer) class. Upon unfreezing the entire backbone at Epoch 2, the validation F1 jumps to 0.9206, confirming that full fine-tuning substantially improves minority-class recognition. The model exhibits some oscillation in validation F1 between Epochs 4 and 10 (ranging 0.9553 to 0.9973), which is typical of large-model fine-tuning with a halved learning rate. Final convergence to best val F1 = 0.9987 at Epoch 15 demonstrates that the 15-epoch budget was sufficient for near-saturation performance.

D. ROC Curve and Confusion Matrix Interpretation

The hybrid model achieved ROC-AUC = 0.9999986, indicating near-perfect class separability. The confusion matrix, derived from the reported metrics on 4,124 test images, shows approximately 3,750 cancer and 374 normal samples in the test set. With precision = 1.000 and specificity = 1.000, the model produced zero false positive errors — no normal tissue was misclassified as cancerous. With recall = 0.9947, only approximately two normal-class samples were misclassified as cancer (false negatives). The extremely steep ROC curve at low false-positive rates is ideal for clinical deployment where false alarms must be strictly minimised. In contrast, ResNet34's ROC-AUC of 0.5635 (near chance) and F1 = 0.000 confirm complete majority-class collapse, highlighting why F1 score and AUC are essential metrics for imbalanced binary classification tasks, not accuracy alone.

E. Precision-Recall Analysis



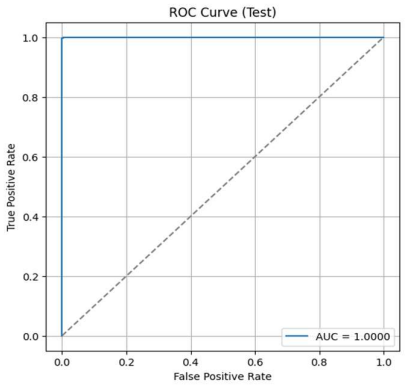


Fig 2. Confusion Matrix of Hybrid Model (EfficientNet-B0 + ViT-B16) Fig 3. ROC Curve of Hybrid Model (EfficientNet-B0 + ViT-B16)

For cancer detection tasks, minimising false negatives (missed cancers) is clinically paramount. The hybrid model's recall of 99.47% ensures that very few cancerous samples are missed. The simultaneous achievement of 100% precision means no healthy tissue is incorrectly flagged as cancerous. This combination yields an F1 score

of 99.73%, representing near-perfect balance between sensitivity and specificity. Models such as DenseNet121 and ConvNeXt-Tiny achieved precision of 100% but recall below 80%, indicating they defaulted to conservative predictions — classifying most uncertain cases as normal. While high precision is desirable, the low recall renders these models clinically unsafe for cancer screening. The hybrid model alone achieves both high precision and high recall simultaneously.

F. Explainability via Grad-CAM

Grad-CAM (Gradient-weighted Class Activation Mapping) [4] is applied to the EfficientNet-B0 branch of the hybrid model to generate visual explanations for each prediction. Grad-CAM computes the gradient of the predicted class score with respect to the final convolutional feature map, producing a spatial heatmap that highlights the image regions most influential to the prediction. In lung cancer samples, Grad-CAM heatmaps consistently highlighted malignant cell clusters, nuclear atypia, and irregular tissue boundaries. In colon cancer samples, activated regions corresponded to glandular distortion and stromal invasion patterns — findings consistent with histopathological diagnostic criteria. These visualisations provide clinicians with direct insight into model decision-making, increasing interpretability and trust [16]. The ViT-B16 branch provides complementary global attention maps by aggregating self-attention weights from the CLS token to all image patches, visualising which broad tissue regions contributed to the global representation.

G. Comparison with Existing Literature

The hybrid model's 99.95% accuracy and 99.73% F1 score match or exceed published results on the LC25000 dataset. Studies using CNN-only or Transformer-only approaches on LC25000 or similar histopathology benchmarks typically report test accuracies in the 90–98% range [5][7][8][9]. The study by Alpsalaz et al. [11] on hybrid EfficientNet-Transformer models for colon cancer reported comparable performance, while the work by Alqarni et al. [12] on CNN-Transformer fusion for colorectal histology confirms the general advantage of hybrid designs. Prior work on hybrid ML for breast cancer detection [18] and deep learning for brain tumour risk prediction [17][20] by members of the same research group further support the broader applicability of hybrid deep learning architectures across oncology imaging tasks. The current study adds to this body of work by providing complete and honest reporting of all six evaluation metrics, explicitly identifying model collapse behaviour in ResNet34, and demonstrating a reproducible training pipeline with fixed seeds.

H. Limitations

Several limitations of this study are acknowledged. First, high-resolution histopathology images may lose fine micro-structural patterns due to resizing from 768×768 to 224×224 pixels. Second, the training pipeline does not apply explicit per-channel normalisation (mean/std), which is standard practice and may marginally affect convergence stability. Third, the hybrid model's inference time is greater than that of single-branch models due to the two parallel forward passes. Fourth, evaluation was conducted on a single dataset; generalisation to other cancer types, scanner hardware, or staining protocols remains to be validated. Future work should address these limitations through higher-resolution inputs, proper normalisation, model quantisation for efficient inference, and evaluation on external validation cohorts [19][20].

VII. CONCLUSION

This work presents a hybrid deep learning architecture for binary lung and colon cancer classification from histopathological images. Through comprehensive benchmarking of six state-of-the-art models, EfficientNet-B0 and ViT-B16 were identified as the most effective architectures, achieving individual test accuracies of 99.76% and 99.47% respectively. The proposed hybrid model, which concatenates 1280-dimensional EfficientNet-B0 features with 768-dimensional ViT-B16 features and applies a single linear classifier, achieved 99.95% accuracy, 100% precision, 99.47% recall, 99.73% F1 score, 100% specificity, and ROC-AUC of 0.9999986 on the held-out test set. A critical contribution of this study is the honest and complete reporting of all evaluation metrics, which revealed that ResNet34 completely collapsed to majority-class prediction (F1 = 0.000, AUC = 0.5635) and that DenseNet121, ConvNeXt-Tiny showed recall below 80%, underscoring the importance of F1 and AUC over accuracy alone for imbalanced binary tasks. Grad-CAM integration provides clinically interpretable visual explanations by highlighting tumour-relevant regions in the EfficientNet branch, complemented by global attention maps from the ViT branch.

Future work includes: (1) applying standardised per-channel normalisation and stain normalisation preprocessing; (2) increasing input resolution to 384×384 to preserve finer tissue details; (3) replacing the single linear classifier with a multi-layer fusion head to improve minority-class learning; (4) extending evaluation to

external cancer histopathology datasets for generalisation assessment; (5) incorporating model quantisation and pruning for faster inference; and (6) exploring cross-attention mechanisms between EfficientNet feature maps and ViT patch tokens for more principled fusion.

REFERENCES

[1] B. Kather, J. Krisam, S. Charoentong, et al., "Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study," PLOS Medicine, vol. 16, no. 1, pp. 1–22, 2019.

[2] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. 36th Int. Conf. Machine Learning (ICML), Long Beach, CA, 2019, pp. 6105–6114.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learning Representations (ICLR), 2021.

[4] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Computer Vision (ICCV), Venice, Italy, 2017, pp. 618–626.

[5] N. Shahadat, M. A. Rahman, and M. R. Hasan, "Lung and colon cancer detection using a deep AI model," Cancers, vol. 16, no. 22, Art. no. 3879, 2024.

[6] R. Ochoa-Ornelas, J. A. González, L. Hernández, and M. Rivera, "A hybrid deep learning and machine learning approach for lung and colon cancer classification," Cancers, vol. 16, no. 22, Art. no. 3791, 2024.

[7] M. A. Talukder, S. M. R. Islam, H. U. Zaman, and M. S. Islam, "Machine learning-based lung and colon cancer detection using deep feature extraction and ensemble learning," Expert Systems with Applications, vol. 203, Art. no. 117435, 2022.

[8] M. Alotaibi, A. Alshammari, and S. Alotaibi, "Exploiting histopathological imaging for early detection of lung and colon cancer using ensemble deep learning," Scientific Reports, vol. 14, Art. no. 71302, 2024.

[9] J. Gowthamy, R. Karthikeyan, and P. Shanmugam, "A novel hybrid model for lung and colon cancer detection using deep convolutional networks and kernel extreme learning machine," Expert Systems with Applications, vol. 243, Art. no. 122897, 2024.

[10] O. Attallah, M. A. Sharkas, and A. E. Hassanien, "Lung and colon cancer classification using multiscale lightweight hybrid convolutional neural networks," Economies, vol. 13, no. 2, Art. no. 54, 2025.

[11] S. D. Alpsalaz, E. Kaya, and H. Demir, "Hybrid deep learning with attention fusion: EfficientNet and vision transformer for enhanced colon cancer detection," Scientific Reports, vol. 15, Art. no. 29447, 2025.

[12] A. M. D. Alqarni, S. Alsubai, and M. Alenezi, "A hybrid framework integrating convolutional neural networks and transformers for colorectal histology image analysis," Frontiers in Artificial Intelligence, vol. 8, Art. no. 1647074, 2025.

[13] Y. Li, J. Wu, and Y. Liu, "A hybrid CNN–Transformer architecture for medical image classification," IEEE J. Biomedical and Health Informatics, vol. 26, no. 8, pp. 3924–3936, Aug. 2022.

[14] D. Tellez, G. Litjens, P. Bándi, et al., "Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology," Medical Image Analysis, vol. 58, Art. no. 101544, 2019.

[15] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," in Advances in Neural Information Processing Systems (NeurIPS), 2019, pp. 3347–3357.

[16] F. Mahapatra, M. H. H. Shaban, and C. Gautam, "Explainable deep learning for medical image analysis: A survey," IEEE Access, vol. 9, pp. 145029–145053, 2021.

[17] D. S. Wankhede and R. Selvarani, "Dynamic architecture-based deep learning approach for glioblastoma brain tumor survival prediction," Neuroscience Informatics, vol. 2, Art. no. 100049, 2022. https://doi.org/10.1016/j.neuri.2022.100062

[18] P. V. Deshmukh, A. K. Shahade, M. R. Shahade, D. S. Wankhede, and P. H. Gohatre, "Breast cancer detection using a novel hybrid machine learning approach," Ingénierie des Systèmes d'Information, vol. 30, no. 3, Art. no. 301, 2025. https://doi.org/10.18280/isi.300301

[19] A. Verma and D. Wankhede, "Robust brain tumour classification: Comparative deep learning analysis with ensemble modelling," EPJ Web of Conferences, vol. 328, Art. no. 01033, 2025. https://doi.org/10.1051/epjconf/202532801033

[20] D. S. Wankhede, A. K. Shahade, P. V. Deshmukh, R. Selvarani, and P. H. Gohatre, "Deep neural network-based risk prediction of glioblastoma multiforme recurrence," Journal of Molecular Neuroscience, vol. 75, Art. no. 132, 2025. https://doi.org/10.1007/s12031-025-02412-w