Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

DeepHistoViT: An Interpretable Vision Transformer Framework for Histopathological Cancer Classification

Ravi Mosalpuri
University of Exeter
Mohammed Abdelsamea
University of Exeter
Ahmed Karam Eldaly
a.karam-eldaly@exeter.ac.uk

University of Exeter

Article

Keywords:

Posted Date: September 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10651319/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

DeepHistoViT: An Interpretable Vision Transformer Framework for Histopathological Cancer Classification

Ravi Mosalpuri[1], **Mohammed Abdelsamea**[1], **and Ahmed Karam Eldaly**[1,2]

[1]Department of Computer Science, University of Exeter, Exeter, EX4 4QF, United Kingdom.
[2]UCL Hawkes Institute, University College London, Gower St., London, WC1E 6AE, United Kingdom.
*{rm1038; M.Abdelsamea; a.karam-eldaly}@exeter.ac.uk

ABSTRACT

Histopathology remains the gold standard for definitive cancer diagnosis due to its ability to provide detailed cellular-level assessment of tissue morphology. However, manual histopathological examination is time-consuming, labour-intensive, and subject to inter-observer variability, creating a growing demand for reliable computer-assisted diagnostic tools. Recent advances in deep learning, particularly transformer-based architectures, have demonstrated strong potential for modelling complex spatial dependencies in medical images. In this work, we propose DeepHistoViT, a transformer-based deep learning framework for automated classification of histopathological images. The proposed model employs a customized Vision Transformer (ViT) architecture with an integrated attention mechanism designed to capture fine-grained cellular structures while enhancing interpretability through attention-based localisation of diagnostically relevant regions. The framework is evaluated on three publicly available and diverse histopathology datasets covering lung cancer, colon cancer, and acute lymphoblastic leukaemia. Experimental results demonstrate that DeepHistoViT achieves state-of-the-art performance across all datasets, with classification accuracy, precision, recall, F1 score, and ROC–AUC reaching 100% on the lung and colon cancer datasets, and 99.85%, 99.84%, 99.86%, 99.85%, and 99.99%, respectively, on the acute lymphoblastic leukaemia dataset. All performance metrics are reported with 95% confidence intervals to ensure statistical robustness. These findings demonstrate the effectiveness of transformer-based architectures for histopathological image analysis and highlight the potential of DeepHistoViT as a reliable and interpretable computer-assisted diagnostic tool to support pathologists in clinical decision-making.

Introduction

Cancer is identified by the world health organisation (WHO) as a leading cause of death worldwide, responsible for approximately 10 million deaths in 2020[1]. The WHO reported that in 2021, non-communicable diseases constituted nearly 75% of all non-pandemic-related deaths globally, with cancer being one of the primary causes[2]. Cancer is characterised by the uncontrolled growth of abnormal cells within the body, which can metastasise to other organs, making this the primary factor in cancer-related deaths. Several factors contribute to the formation of cancerous cells, including genetic mutations, lifestyle choices (such as smoking, unhealthy diets, and a lack of physical activity), and environmental factors (such as exposure to chemicals or radiation). Accurate detection of cancer is crucial in determining appropriate treatment strategies, as each type has unique biological characteristics and treatment requirements.

The GLOBOCAN 2020 report[3] highlights the significant impact of several major cancer types worldwide. The breast, lung, prostate, and colorectal cancers are among the most prevalent, collectively accounting for a large proportion of new cancer cases and cancer-related deaths each year. In 2020, breast cancer led with over 2.3 million new cases globally, followed by lung, colorectal, and prostate cancers, each contributing substantially to the global cancer burden. Survival rates and symptoms vary extensively between cancer types. The American Cancer Society's 2025 report[4] indicates that the overall 5-year relative survival rate is 27% for lung cancer, 65% for colorectal cancer, and can exceed 90% for localised breast cancer cases when detected early. Common warning signs can range from persistent cough and chest discomfort in lung cancer, to abnormal bleeding or lumps in the breast and colorectal cancers. However, these symptoms are typically only apparent when the cancer has reached an advanced stage[4]. Therefore, early and accurate detection of cancer through histopathology images can drastically improve survival chances by prompt treatment.

Histopathology remains the gold standard for definitive cancer diagnosis, as it enables detailed microscopic examination of tissue morphology and cellular architecture by trained pathologists. Manual assessment of histopathological slides plays a central role in clinical decision-making; however, this process is labour-intensive, time-consuming, and subject to inter- and

intra-observer variability, particularly when large volumes of whole-slide images must be analysed[5,6]. Traditional computer-aided diagnosis approaches based on handcrafted features and classical image processing techniques have been explored to support pathologists, but these methods often require extensive manual feature engineering and lack robustness when applied to large-scale, heterogeneous datasets[7–13]. In contrast, deep learning has transformed digital pathology by enabling automated feature extraction directly from raw image data, eliminating the need for manual feature design and improving scalability and diagnostic performance[5,9,14–17]. Convolutional neural networks (CNNs), in particular, have demonstrated strong capability in learning hierarchical spatial features and have achieved expert-level performance in various histopathological image classification tasks, including cancer detection and subtype classification[5,14,16,18–20]. They automatically extract features from raw images through a series of convolutions, pooling layers and non-linear activation functions. CNNs are capable of identifying more intricate patterns and structures in images that are appropriate for medical imaging applications. Additionally, DL algorithms can process large volumes of digital pathology images with speed and reproducibility, thereby reducing the time required for manual examination.

Although CNNs excel at feature extraction, they struggle in capturing the spatial relationships among different features. It is essential for algorithms to understand global contextual information, as the spatial relationships among tissue structures can greatly impact classification results. Considering the convolution architecture struggles to comprehend global information, researchers have proposed various architectural modifications. With the increasing popularity of transformers in natural language processing (NLP), a variant of the transformer architecture was introduced, namely the vision transformer (ViT). The transformer architecture can handle a series of fixed-size image patches as input, which helps in extracting complex features from images. It uses a global attention mechanism to tackle long-range dependency issues often linked with CNNs. ViTs have led to significant progress in medical computer vision across different imaging methods[21,22]. In particular, within the field of histopathological image classification, ViTs have demonstrated exceptional accuracy in diagnosing cancers, including breast, lung, and colon cancer[21–23].

Despite the recent advancements in DL for the classification of cancer, significant research gaps remain that hinder the effective implementation of these models in clinical environments. A primary concern is the generalisability of these models, for instance, models that are trained only on augmented datasets like LC25000[24] usually perform well with similar datasets only and tend to face challenges when dealing with real-world clinical images. On the other hand,iInterpretability is another key concern, as most DL models function as opaque ‘black-boxes’. Therefore, establishing clinician trust is essential for widespread adoption. There is still limited adoption of Explainable Artificial Intelligence (XAI) algorithms, popularly known as XAI techniques.

This work introduces DeepHistoVit, a customised ViT-16–based deep learning framework specifically fine-tuned for the classification of histopathology images across multiple cancer types. The main contributions of this work are as follows.

1. We introduce *DeepHistoVit*, a customised Vision Transformer (ViT-16) model, specifically designed for the classification of histopathology images across three diverse cancer datasets, including lung, acute lymphoblastic leukaemia, and colon. By incorporating a tailored multi-layer classification head with batch normalisation, configurable dropout, and progressive dimensionality reduction, the model effectively captures complex tissue features while remaining broadly adaptable across heterogeneous datasets. Notably, the framework achieves this without requiring stain normalisation, demonstrating practical applicability to real-world histopathology data.

2. DeepHistoVit integrates an attention-based visualisation mechanism that highlights the most informative regions of each histopathology image. This feature enhances interpretability and allows pathologists to inspect the model’s reasoning at the patch level, providing a bridge between computational predictions and clinical validation, and fostering trust in AI-assisted diagnostic tools.

3. We rigorously assess the performance of DeepHistoVit across three independent datasets, reporting metrics including accuracy, precision, recall, F1 score, and ROC–AUC along with 95% confidence intervals. The results demonstrate consistent, state-of-the-art performance and confirm the model’s ability to generalise across distinct tissue types, highlighting its potential as a reliable and versatile tool for digital pathology.

The remaining sections of the paper are organised as follows. Section reviews the literature for histopathology image classification using deep learning. The design and implementation details of the proposed approach are presented in Section . Experimental results are then presented in Section . Discussions and general insights into the proposed approach, including main limitations and plans for future work are discussed in Section . Finally, conclusions are outlined in Section .

Related Work

Recent advances in deep learning have significantly improved the automated classification of histopathological images, enabling more accurate and efficient cancer diagnosis. Early approaches primarily relied on convolutional neural networks (CNNs),

which demonstrated strong capability in learning hierarchical spatial representations from microscopic tissue images. For example, Shahadat et al.[23] proposed a lightweight CNN incorporating squeeze-and-excitation modules for lung and colon cancer classification using the LC25000 dataset[24]. Their architecture achieved 100% accuracy while maintaining a low computational footprint, highlighting the feasibility of efficient CNN-based models for clinical deployment. Similarly, Bala et al.[25] introduced MRANet, a residual attention-based CNN that integrates attention modules to enhance feature representation, achieving accuracies up to 99.30% in lung cancer classification tasks. These studies demonstrate the effectiveness of attention-enhanced CNN architectures in capturing discriminative histopathological features.

To further improve classification performance and robustness, several works have explored multi-model fusion and feature selection strategies. Attallah et al.[26] developed a computer-aided diagnostic framework combining features extracted from MobileNet, EfficientNetB0, and ResNet-18 architectures. By applying canonical correlation analysis and statistical feature selection methods, their approach achieved 99.78% classification accuracy on the LC25000 dataset. Ensemble learning approaches have also demonstrated strong performance. Atwan et al.[27] proposed an ensemble framework combining ResNet50 and MobileNet feature extractors with classical classifiers, achieving 99.2% and 97.72% accuracy for binary and multiclass breast cancer classification, respectively. These hybrid and ensemble approaches improve robustness by leveraging complementary feature representations from multiple architectures, although they often increase architectural complexity.

Transfer learning has also emerged as an effective strategy for histopathological image classification by leveraging pretrained models to improve generalisation. Ochoa-Ornelas et al.[28] fine-tuned EfficientNetB3 using both LC25000 and clinically sourced genomic data commons images, achieving 99.39% accuracy while improving clinical realism and generalisability. Similarly, El-Aziz et al.[29] introduced a learning rate tuning strategy for EfficientNet-B3, enabling dynamic adaptation during training and achieving 99.84% accuracy for oral cancer classification. These approaches demonstrate the effectiveness of transfer learning and training optimisation techniques in improving model performance and convergence.

Despite the success of CNN-based methods, their inherently local receptive fields limit their ability to capture long-range dependencies and global contextual relationships within histopathological images. To address this limitation, transformer-based architectures have recently gained increasing attention due to their ability to model global feature interactions through self-attention mechanisms. Hayat et al.[21] proposed a hybrid EfficientNetV2 and Vision Transformer framework, achieving 99.83% accuracy on binary breast cancer classification tasks. Similarly, Yan et al.[22] introduced DWNAT-Net, a transformer-based architecture incorporating discrete wavelet transforms to jointly capture spatial and frequency-domain features, achieving 99.66% accuracy on the BreakHis dataset and demonstrating strong capability in modelling complex histopathological structures.

In addition to architectural advancements, human-in-the-loop learning frameworks have been explored to enhance model reliability and performance. Han et al.[30] proposed an iterative training framework incorporating expert feedback to refine model predictions, demonstrating improved accuracy and robustness in invasive ductal carcinoma detection. While such approaches improve reliability, their reliance on continuous expert intervention limits scalability in routine clinical practice.

Although existing methods have achieved promising results, several challenges remain. CNN-based models are limited in capturing global contextual dependencies, while hybrid and ensemble approaches often introduce increased computational complexity and architectural redundancy. Furthermore, interpretability remains a critical challenge, as many existing models provide limited transparency regarding the regions influencing model predictions. Transformer-based architectures offer a promising solution due to their inherent ability to capture long-range dependencies while providing attention-based interpretability.

To address these limitations, this work proposes DeepHistoViT, a customised Vision Transformer framework specifically designed for histopathological image classification. The proposed model leverages self-attention mechanisms to capture both fine-grained cellular structures and global contextual relationships, while providing interpretable attention maps highlighting diagnostically relevant regions. Unlike hybrid or ensemble approaches, DeepHistoViT provides a unified transformer-based framework capable of achieving state-of-the-art performance across multiple cancer types while maintaining architectural simplicity, scalability, and interpretability.

Methods

Datasets

The performance of the proposed DeepHistoViT framework is evaluated using three publicly available histopathology image datasets representing different cancer types. These datasets provide diverse morphological characteristics and imaging conditions, enabling comprehensive evaluation of model robustness and generalisation.

1. Colon Cancer Dataset (LC25000) Colon cancer images are also obtained from the LC25000 dataset[24]. Two classes are considered: colon adenocarcinoma and benign colon tissue. Although derived from the same dataset as lung cancer, colon classification is treated as a separate task due to distinct morphological characteristics and classification objectives.

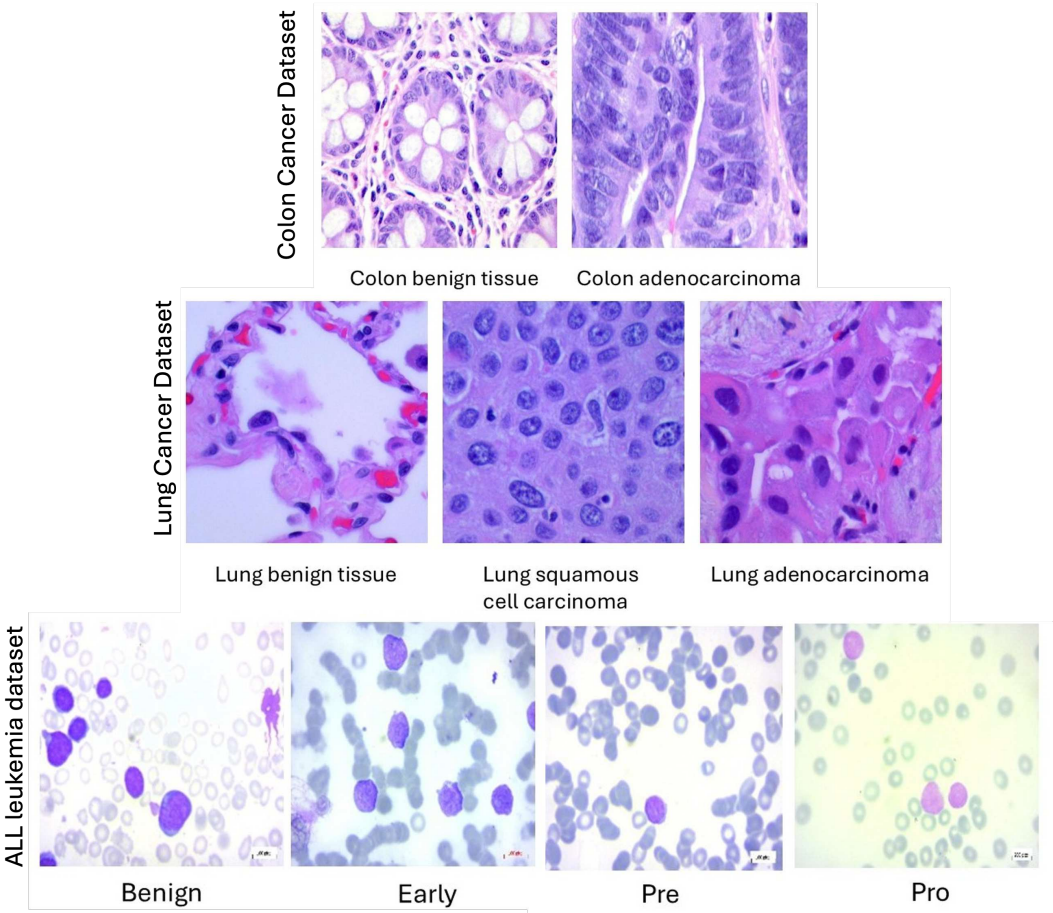


Figure 1. Example histopathological images from the datasets used in this study. Top row: Colon tissue samples from the LC25000 dataset showing (h) benign colon tissue and (i) colon adenocarcinoma. Middle row: Lung tissue samples showing (a) benign lung tissue, (b) squamous cell carcinoma, and (c) adenocarcinoma. Bottom row: Peripheral blood smear images from the acute lymphoblastic leukaemia (ALL) dataset showing (d) benign cells, (e) early Pre-B lymphoblasts, (f) Pre-B lymphoblasts, and (g) Pro-B lymphoblasts. All histopathological images are stained using hematoxylin and eosin (H&E).

2. Lung Cancer Dataset (LC25000) The lung cancer images are obtained from the LC25000 dataset[24], which contains 25,000 histopathological images of lung and colon tissues. For lung cancer classification, three classes are used: lung adenocarcinoma, lung squamous cell carcinoma, and benign lung tissue. Each class contains 5,000 RGB images with a resolution of 768×768 pixels. All images are anonymised and compliant with the Health Insurance Portability and Accountability Act (HIPAA), ensuring patient privacy and ethical use. This dataset provides high-quality cellular structures suitable for automated cancer classification.

3. Acute Lymphoblastic Leukaemia Dataset (ALL) The Acute Lymphoblastic Leukaemia dataset[31] contains 3,562 peripheral blood smear images collected from 89 patients. The dataset includes benign hematogone cells and malignant lymphoblast cells, which comprise Early Pre-B, Pre-B, and Pro-B subtypes. These images exhibit significant morphological variability, providing a challenging benchmark for classification models.

Representative examples from each dataset are shown in Figures 1.

Data Preparation and Preprocessing

All histopathology images are subjected to a standardised preprocessing pipeline to ensure compatibility with the Vision Transformer architecture while preserving diagnostically relevant morphological information. Specifically, each image is first resized to 256×256 pixels and subsequently centre-cropped to 224×224 pixels. This spatial resolution aligns with the input dimensional requirements of the pretrained ViT-base-patch16-224 model[32], and ensures consistent patch extraction. This step

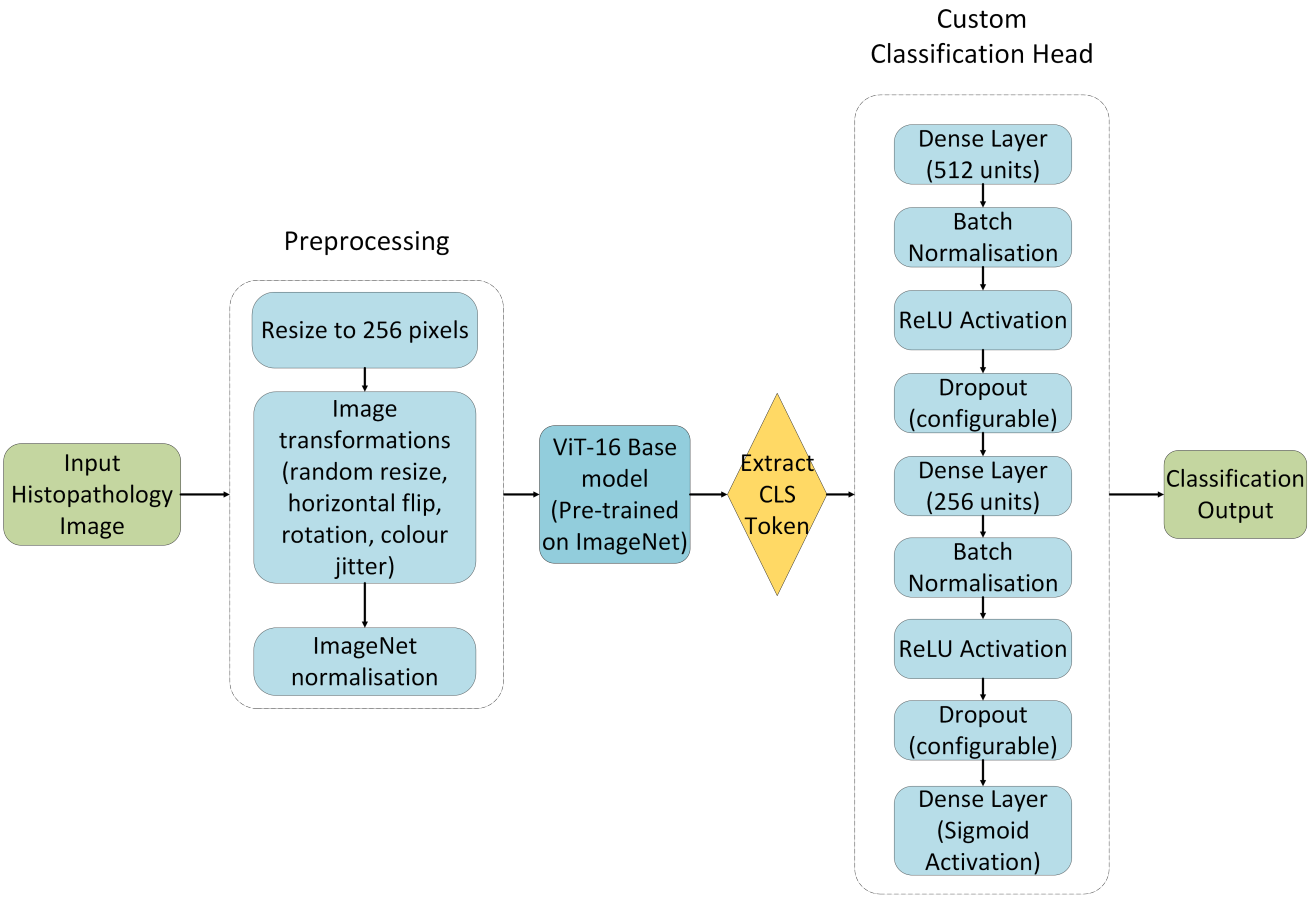


Figure 2. Overview of the proposed Vision Transformer (ViT-16) based pipeline for histopathology image classification. The framework consists of three stages: (i) Preprocessing, where raw histopathology images are resized to 256 pixels, augmented through random resizing, horizontal flipping, rotation, and colour jitter, followed by ImageNet normalisation; (ii) Feature Extraction, leveraging a ViT-16 base model pre-trained on ImageNet to obtain the CLS token representation; and (iii) Custom Classification Head, comprising sequential dense layers (512 and 256 units) with batch normalization, ReLU activation, and configurable dropout, culminating in a sigmoid-activated output layer. This architecture integrates transformer-based global feature learning with a tailored classification head to enhance performance on medical image analysis tasks.

defines the effective image domain from which fixed-size patches are generated and embedded into the transformer input sequence.

No stain normalisation or domain harmonisation techniques are applied. This decision preserves the intrinsic variability present in histopathology images, including differences in staining intensity, scanner characteristics, and tissue preparation protocols. By retaining natural variability, the proposed framework is evaluated under conditions that more closely reflect real-world clinical deployment. To improve model generalisation and mitigate overfitting, data augmentation is applied during training. The augmentation pipeline includes random resized cropping, horizontal flipping, rotations up to 20°, and colour jittering. These transformations simulate realistic acquisition variability and encourage the model to learn invariant morphological representations. Validation and test sets undergo only deterministic preprocessing consisting of resizing and centre cropping, ensuring unbiased evaluation. Following preprocessing, all images are normalised using the channel-wise mean and standard deviation of the ImageNet dataset[33], with mean values of $[0.485, 0.456, 0.406]$ and standard deviation values of $[0.229, 0.224, 0.225]$. This normalisation aligns the input distribution with that of the pretrained backbone and facilitates stable optimisation during transfer learning. The complete preprocessing and feature extraction pipeline is illustrated in Fig. 2.

Model Architecture

The proposed DeepHistoViT framework builds upon the Vision Transformer (ViT) architecture originally introduced by Dosovitskiy et al.[32]. Unlike convolutional neural networks, which rely on local receptive fields, the Vision Transformer models an image as a sequence of embedded patches and processes this sequence using stacked self-attention layers, enabling direct

modelling of global contextual dependencies.

Given a preprocessed input image $x \in \mathbb{R}^{H\times W\times C}$, the image is divided into N non-overlapping patches of size $P \times P$. Each patch is flattened and projected into a D-dimensional embedding space using a learnable linear projection. Positional embeddings are added to retain spatial information that would otherwise be lost during patch flattening. A learnable classification token is prepended to the sequence to aggregate global information across all patches. The resulting sequence is processed through multiple transformer encoder layers. Each encoder layer consists of a multi-head self-attention module followed by a feed-forward network. The attention mechanism computes Query (Q), Key (K), and Value (V) matrices, and the attention output is given by

$$\text{Attention}(Q,K,V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{1}$$

where d_k denotes the dimensionality of the key vectors. This mechanism enables the model to learn global relationships between spatially distant regions, which is particularly important for histopathological image analysis, where diagnostic features may be distributed across the tissue.

The backbone network is initialised using pretrained weights from the ViT-base-patch16-224 model trained on the ImageNet dataset[32,33]. Transfer learning allows the model to leverage general visual representations learned from large-scale natural image datasets and adapt them to histopathological image classification. To improve adaptation to the target domain, a selective fine-tuning strategy is employed, where only the final two transformer encoder blocks are unfrozen during training. This enables domain-specific feature refinement while preserving general low-level representations. Furthermore, a custom multi-layer classification head is introduced, replacing the default linear classifier. As shown in Fig. 2, the classification token embedding is passed through two fully connected layers with progressive dimensionality reduction (768→512→256), each followed by batch normalisation, ReLU activation, and dropout. This structure enables progressive feature refinement and improves discrimination between histopathological tissue classes. The final classification layer maps the refined feature representation to the output space, producing class probabilities via the softmax function.

Model Training

The proposed DeepHistoViT framework is implemented using the PyTorch deep learning library and trained using GPU acceleration. Training is conducted on high-performance computing infrastructure, including NVIDIA Tesla T4 and P100 GPUs, to ensure efficient optimisation of the transformer-based architecture. To ensure reproducibility, all experiments are conducted with fixed random seeds across Python, NumPy, and PyTorch, including CUDA operations. This ensures consistent weight initialisation, data splitting, and training behaviour across repeated runs.

Each dataset is partitioned into training, validation, and test sets using stratified sampling to preserve class distributions. Specifically, 20% of the data is reserved as an independent test set, which remains completely unseen during training and hyperparameter optimisation. The remaining 80% is further divided into training and validation sets using an 80:20 ratio. This results in an effective split of 64%, 16%, and 20% for training, validation, and testing, respectively. Model training is performed by minimising the categorical cross-entropy loss function.

Model optimisation is performed using the Adaptive Moment Estimation (Adam) optimiser[34], which computes adaptive learning rates for each parameter based on first and second moments of the gradients. Hyperparameter optimisation is conducted using grid search over key parameters, including learning rate, weight decay, dropout rate, and batch size. The search space consists of learning rates $\{10^{-4}, 10^{-3}, 10^{-2}\}$, weight decay values $\{10^{-5}, 10^{-4}, 10^{-3}\}$, dropout rates $\{0.2, 0.3, 0.4\}$, and batch sizes $\{16, 32, 64\}$. The optimal hyperparameter configuration is selected based on validation accuracy. To improve convergence stability and prevent overfitting, a ReduceLROnPlateau learning rate scheduler is employed, which reduces the learning rate when validation performance plateaus. In addition, early stopping is used with a patience of eight epochs, terminating training when validation performance no longer improves. The model parameters corresponding to the best validation performance are retained for final evaluation.

Results

This section presents a comprehensive evaluation of the proposed DeepHistoViT framework across three histopathology datasets. The experimental analysis focuses on hyperparameter optimisation, model generalisation, and classification performance using multiple quantitative metrics. The evaluation protocol is designed to ensure statistical reliability, reproducibility, and unbiased assessment of model performance.

Hyperparameter Optimisation

Hyperparameter optimisation is performed using a systematic grid search over key training parameters, including learning rate, weight decay, dropout rate, and batch size, as described in Section . The objective is to identify parameter configurations that

maximise validation performance while maintaining stable convergence and minimising overfitting. The optimal hyperparameter configuration across datasets consists of a weight decay of 1×10^{-5}, dropout rate of 0.2, and batch size of 16 for the lung and colon cancer datasets, and 32 for the leukaemia dataset. These values provide an effective balance between regularisation and model capacity. The selected weight decay introduces sufficient regularisation to prevent overfitting while preserving discriminative feature learning. Similarly, the dropout rate of 0.2 improves generalisation by reducing co-adaptation of neurons within the classification head. The batch size selection reflects dataset-specific optimisation, with the larger batch size for the leukaemia dataset facilitating more stable gradient estimation due to its smaller dataset size. The learning rate is selected based on validation performance to ensure stable optimisation of the fine-tuned transformer layers while preserving pretrained feature representations.

Evaluation Protocol

To ensure robust and unbiased performance estimation, a stratified five-fold cross-validation strategy is employed on the training data. This approach preserves class distribution across folds and enables evaluation of model stability under different data partitions. Cross-validation performance is used for model selection and hyperparameter optimisation. In addition to cross-validation, a strictly independent held-out test set, comprising 20% of the dataset as defined in Section , is used exclusively for final performance evaluation. This separation ensures that test data are not used during model training or hyperparameter selection, thereby preventing data leakage and providing an unbiased estimate of generalisation performance. To further improve prediction robustness, test-time augmentation (TTA) is applied during inference[35]. Each test image is evaluated under multiple geometric transformations, including horizontal and vertical flips and rotations of $\pm 10^{\circ}$. The final prediction is obtained by averaging the predicted class probabilities across all augmented versions of the image. This approach reduces sensitivity to image orientation and improves prediction stability.

Evaluation Metrics

Model performance is evaluated using multiple quantitative metrics to provide a comprehensive assessment of classification accuracy and reliability. These metrics include accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (ROC-AUC). Accuracy measures the overall proportion of correctly classified samples. Precision quantifies the proportion of correctly predicted positive samples among all predicted positives, while recall measures the proportion of correctly identified positive samples among all true positives. The F1-score provides a balanced measure by combining precision and recall. The ROC-AUC metric evaluates the model's ability to discriminate between classes across different classification thresholds and is particularly informative for assessing classification robustness. In addition to scalar metrics, confusion matrices are analysed to examine class-specific performance and identify potential misclassification patterns. To ensure statistical reliability, performance metrics are reported on the independent test set and, where applicable, averaged across cross-validation folds.

Quantitative Results

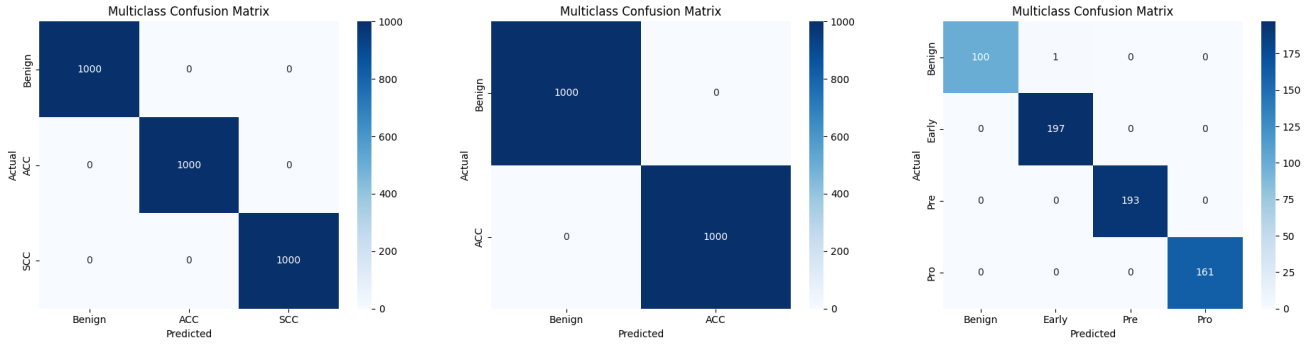
The quantitative performance of the proposed DeepHistoViT framework is evaluated on three independent histopathology datasets, including lung and colon cancer images from the LC25000 dataset[24], and acute lymphoblastic leukaemia images from the ALL dataset[31]. Model performance is assessed using accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (ROC-AUC), as defined in Section .

On the lung cancer dataset, the proposed model achieves an overall classification accuracy of 100% on the held-out test set. As shown in Table 1, precision, recall, and F1-score values of 1.0000 are obtained for all classes, including benign lung tissue, adenocarcinoma, and squamous cell carcinoma. These results indicate that the model successfully separates the three lung tissue classes without misclassification under the defined evaluation protocol.

Table 1. Classification performance of the proposed DeepHistoViT model on the LC25000 lung cancer dataset. Precision, recall, and F1-score values are reported for each class, along with overall accuracy.

Class	Precision	Recall	F1-Score	Support
Benign	1.0000	1.0000	1.0000	1000
ACC	1.0000	1.0000	1.0000	1000
SCC	1.0000	1.0000	1.0000	1000
Accuracy			1.0000	3000
Macro Avg	1.0000	1.0000	1.0000	3000
Weighted Avg	1.0000	1.0000	1.0000	3000

Similarly, on the colon cancer dataset, the proposed model achieves 100% accuracy on the independent test set. Table 2



(a) Lung cancer (LC25000) **(b)** Colon cancer (LC25000) **(c)** ALL dataset

Figure 3. Confusion matrices for the three evaluated datasets: lung cancer (LC25000), colon cancer (LC25000), and acute lymphoblastic leukaemia (ALL). The lung and colon datasets show perfect classification performance, while the ALL dataset shows near-perfect classification with minimal misclassification.

shows that precision, recall, and F1-score values are equal to 1.0000 for both benign and malignant colon tissue classes. These results indicate consistent separation between normal and cancerous colon tissue samples within the dataset.

Table 2. Classification performance of the proposed DeepHistoViT model on the LC25000 colon cancer dataset. Precision, recall, and F1-score values are reported for each class, along with overall accuracy.

Class	Precision	Recall	F1-Score	Support
Benign	1.0000	1.0000	1.0000	1000
ACC	1.0000	1.0000	1.0000	1000
Accuracy			1.0000	2000
Macro Avg	1.0000	1.0000	1.0000	2000
Weighted Avg	1.0000	1.0000	1.0000	2000

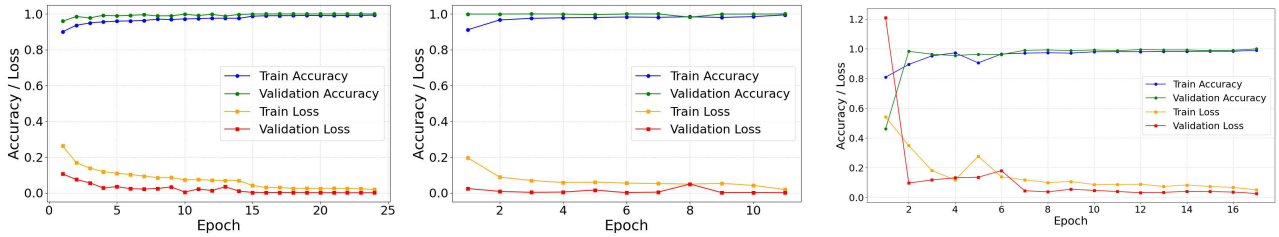
For the acute lymphoblastic leukaemia dataset, the model achieves an overall accuracy of 99.85%, with macro-averaged precision, recall, and F1-score values of 99.87%, 99.75%, and 99.81%, respectively. The ROC-AUC reaches 99.99%, indicating strong class separability. As shown in Table 3, near-perfect performance is observed across all four classes. Minor deviations from perfect classification are observed in the benign class, while the malignant subclasses (Early, Pre, and Pro) are classified with perfect or near-perfect accuracy.

Table 3. Classification performance of the proposed DeepHistoViT model on the ALL dataset. Precision, recall, and F1-score values are reported for each class, along with overall accuracy.

Class	Precision	Recall	F1-Score	Support
Benign	1.0000	0.9901	0.9950	101
Early	0.9949	1.0000	0.9975	197
Pre	1.0000	1.0000	1.0000	193
Pro	1.0000	1.0000	1.0000	161
Accuracy			0.9985	652
Macro Avg	0.9987	0.9975	0.9981	652
Weighted Avg	0.9985	0.9985	0.9985	652

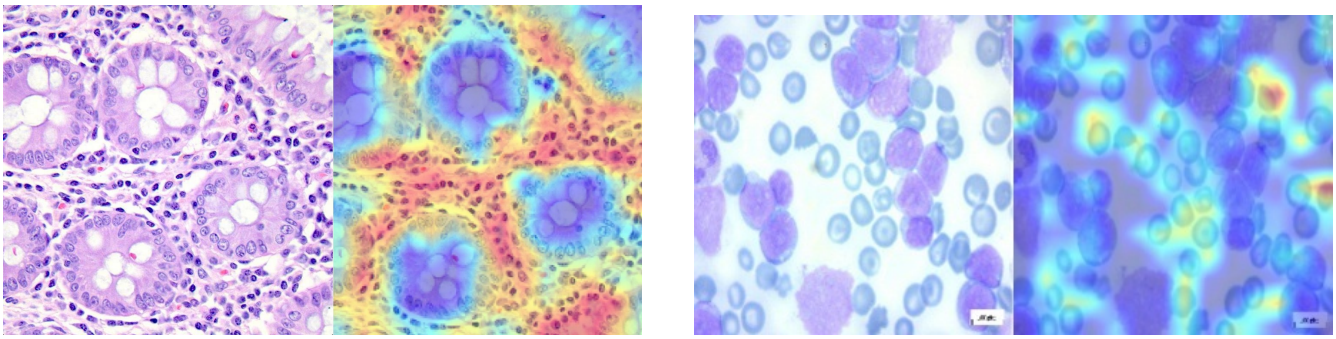
The confusion matrices in Fig. 3 provide a detailed view of class-wise predictions. For the lung and colon cancer datasets, the confusion matrices show perfect diagonal structure, confirming that all samples are correctly classified. For the ALL dataset, only a small number of misclassifications are observed, primarily involving benign samples, while malignant subclasses are classified correctly with high reliability.

Training and validation accuracy and loss curves are shown in Fig. 4. These curves demonstrate stable optimisation and consistent convergence behaviour across datasets. The close agreement between training and validation accuracy indicates effective generalisation, while the decreasing loss curves confirm stable learning without evidence of significant overfitting.



(a) Lung cancer dataset **(b)** Colon cancer dataset **(c)** ALL dataset

Figure 4. Training and validation accuracy and loss curves across datasets, demonstrating stable convergence and consistent generalisation performance.



(a) Colon benign tissue (LC25000) **(b)** Acute lymphoblastic leukaemia (ALL)

Figure 5. Attention visualisation generated by the proposed DeepHistoViT model. For each example, the original histopathology image is shown alongside the corresponding attention map, highlighting regions that contribute most strongly to the classification decision. The attention maps demonstrate spatial localisation of diagnostically relevant morphological features.

Attention Visualisation

An important characteristic of Vision Transformer architectures is the self-attention mechanism, which enables identification of spatial regions that contribute most strongly to the classification decision. Unlike convolutional neural networks, where interpretability often requires post hoc methods such as Grad-CAM, the transformer architecture inherently provides attention weights that can be analysed to understand the model's focus during inference.

To examine the spatial regions influencing the model's predictions, attention maps are extracted from the final transformer encoder layer and projected onto the original input image. These attention maps provide a visual representation of the relative importance assigned to different image regions during classification. Representative examples are shown in Fig. 5 for colon tissue and acute lymphoblastic leukaemia samples. In each case, the original histopathology image is shown alongside the corresponding attention map generated by the model. The attention maps highlight spatially localised regions corresponding to diagnostically relevant morphological structures, including cellular arrangements and nuclear features. This behaviour is consistent with the expectation that classification decisions are driven by discriminative tissue patterns. These visualisations provide qualitative evidence that the proposed DeepHistoViT framework learns spatially meaningful feature representations and focuses on relevant tissue regions when performing classification. Such interpretability is particularly important in medical imaging applications, where transparency of model decision-making is essential for clinical reliability and validation.

Comparative Analysis with State-of-the-Art Models

To assess the performance of the proposed DeepHistoViT framework in the context of existing approaches, a comparative analysis is conducted against recently published deep learning models evaluated on the same datasets. This comparison provides insight into the relative performance of the proposed method and its effectiveness for histopathology image classification.

Table 4 presents a comparison of classification performance on the lung cancer subset of the LC25000 dataset[24]. The proposed DeepHistoViT model achieves an overall accuracy of 100%, with corresponding precision, recall, F1-score, and ROC-AUC values of 100%. These results are consistent with the performance reported by Shahadat et al.[23], who also achieved perfect classification using a deep learning model on the same dataset. Other recent approaches, including ML3CNet proposed

Table 4. Comparison of classification performance for lung cancer (LC25000) dataset across recent studies and the proposed DeepHistoViT model.

Study	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC
Kumar et al.[36] (2024)	ML3CNet	99.72	99.64	99.66	99.65	99.78
Attallah et al.[26] (2025)	Multi-CNN	99.78	99.78	99.78	99.78	–
Shahadat et al.[23] (2025)	DL	100.00	100.00	100.00	100.00	100.00
Proposed	**DeepHistoViT**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**

Table 5. Comparison of classification performance for colon cancer (LC25000) dataset across recent studies and the proposed DeepHistoViT model.

Study	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC
Ochoa-Ornelas et al.[28] (2025)	EfficientNet-B3	99.39	99.39	99.39	99.39	–
Attallah et al.[26] (2025)	Multi-CNN	99.78	99.78	99.78	99.78	–
Shahadat et al.[23] (2025)	DL	100.00	100.00	100.00	100.00	100.00
Proposed	**DeepHistoViT**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**

by Kumar et al.[36] and the multi-CNN framework proposed by Attallah et al.[26], report accuracies of 99.72% and 99.78%, respectively. These findings indicate that the proposed DeepHistoViT framework achieves performance comparable to the highest reported results on this dataset.

A similar comparison is performed for the colon cancer subset of the LC25000 dataset. As shown in Table 5, the proposed DeepHistoViT model achieves 100% accuracy across all evaluation metrics. This performance is consistent with the results reported by Shahadat et al.[23]. Other recent methods, including the multi-CNN approach proposed by Attallah et al.[26] and the EfficientNet-B3 model proposed by Ochoa-Ornelas et al.[28], report accuracies of 99.78% and 99.39%, respectively. These results demonstrate that the proposed transformer-based architecture achieves performance comparable to or exceeding existing approaches evaluated on this dataset.

The proposed DeepHistoViT framework also demonstrates strong performance on the acute lymphoblastic leukaemia dataset. As shown in Table 6, the model achieves an accuracy of 99.85%, with corresponding precision, recall, and F1-score values of 99.85%, and an ROC-AUC of 99.99%. This performance exceeds the results reported by recent studies, including ALL-Net proposed by Thiriveedhi et al.[37], which achieved 99.32% accuracy, and DarkNet19 ESA proposed by Basaran et al.[38], which achieved 98.52% accuracy. These findings indicate that the proposed transformer-based framework achieves competitive performance across multiple histopathology datasets.

Table 6. Performance comparison of recent models on the acute lymphoblastic leukaemia dataset.

Study	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC
Muhammad et al.[39] (2025)	EfficientNet-B7	96.78	97.00	96.25	96.57	99.85
Basaran et al.[38] (2025)	DarkNet19 ESA	98.52	98.40	98.22	98.30	99.95
Thiriveedhi et al.[37] (2025)	ALL-Net	99.32	99.25	99.50	99.50	–
Proposed	**DeepHistoViT**	**99.85**	**99.85**	**99.85**	**99.85**	**99.99**

Discussion

This work evaluates the effectiveness of a transformer-based architecture, DeepHistoViT, for automated histopathology image classification across multiple cancer types. By combining a pretrained Vision Transformer backbone with a customised multi-layer classification head and selective fine-tuning strategy, the proposed framework achieves high classification performance across all evaluated datasets. In particular, the model attains perfect classification accuracy on the lung and colon cancer subsets of the LC25000 dataset and near-perfect performance on the acute lymphoblastic leukaemia dataset. These findings indicate that transformer-based representations, when appropriately adapted, can effectively capture morphological features relevant for histopathological classification.

The observed perfect classification performance on the LC25000 dataset requires careful interpretation. Previous studies have noted that LC25000 is derived from a limited number of original whole-slide images, which are expanded through extensive augmentation[24]. As a result, augmented samples may share substantial visual similarity, potentially simplifying the classification task. Consequently, while the reported results demonstrate the model's ability to learn discriminative features within this dataset, they may not directly translate to real-world clinical scenarios characterised by greater inter-patient and inter-institutional variability. Nevertheless, the strong performance observed on independent datasets such as the ALL dataset suggests that the proposed architecture is capable of generalising beyond a single data source.

The architectural modifications introduced in DeepHistoViT contribute to its observed performance. The proposed multi-layer classification head enables progressive refinement of the high-dimensional feature representations produced by the transformer encoder. By incorporating batch normalisation, non-linear activation, and dropout-based regularisation, this component improves feature discrimination while reducing overfitting, particularly in datasets with limited training samples. In addition, the selective fine-tuning strategy, in which only the final transformer encoder layers are updated during training, allows domain-specific adaptation while preserving general visual representations learned during large-scale pretraining. This approach stabilises optimisation and improves generalisation across datasets with differing staining patterns and imaging characteristics.

Differences in classification performance across datasets highlight the influence of dataset size, diversity, and complexity on model performance. The slightly lower accuracy observed on the ALL dataset compared to LC25000 reflects greater morphological variability and reduced sample size. These findings emphasise the importance of evaluating deep learning models on multiple datasets to obtain a more comprehensive assessment of generalisation performance. Future work may benefit from incorporating larger and more diverse datasets, as well as domain adaptation techniques designed to improve robustness to variations in staining protocols and imaging conditions.

Attention visualisation provides additional insight into the model's decision-making process. The attention maps generated by DeepHistoViT highlight spatial regions that contribute most strongly to the classification outcome, and these regions correspond to relevant tissue structures and cellular morphology. While these visualisations provide qualitative evidence of meaningful feature utilisation, further quantitative evaluation of interpretability and reliability would be required to support clinical deployment.

Despite the promising results, limitations should be considered. First, the evaluation is conducted on publicly available datasets that may not fully reflect the variability encountered in clinical practice. Second, although the model demonstrates strong performance across multiple datasets, validation on large-scale, multi-centre clinical cohorts would be necessary to establish clinical robustness. Third, while attention mechanisms provide interpretability, additional approaches such as uncertainty estimation and calibration analysis may be required to ensure reliable deployment in safety-critical clinical settings.

Overall, the results demonstrate that transformer-based architectures, combined with appropriate domain adaptation and classification head design, provide a promising approach for histopathological image classification. These findings support further investigation of transformer-based methods for medical image analysis, particularly in scenarios requiring modelling of complex spatial dependencies.

Conclusions

This paper presents DeepHistoViT, a Vision Transformer-based framework for histopathology image classification across multiple cancer types. By combining a pretrained ViT backbone with a customised classification head and selective fine-tuning strategy, the proposed model achieves high classification performance across all evaluated datasets, including perfect accuracy on the LC25000 lung and colon datasets and near-perfect performance on the acute lymphoblastic leukaemia dataset. These results demonstrate the effectiveness of transformer-based representations for capturing discriminative histopathological features. In addition to strong quantitative performance, attention-based visualisation provides qualitative insight into the spatial regions contributing to model predictions, supporting interpretability of the learned representations. The ability of the proposed framework to generalise across datasets without explicit stain normalisation further highlights its robustness. Future work will focus on validation using larger, multi-centre clinical datasets and the integration of uncertainty estimation and calibration

methods to improve reliability in clinical settings.

Data availability

The datasets used in this study are publicly available from the corresponding data repositories. All data supporting the findings of this study are available from the original dataset sources cited in this manuscript.

References

1. Organization, W. H. Cancer fact sheet. Accessed: 15 April 2025.
2. Organization, W. H. Noncommunicable diseases. Accessed: 15 April 2025.
3. Sung, H. *et al.* Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer J. for Clin.* **71**, 209–249, DOI: 10.3322/caac.21660.
4. Society, A. C. Cancer facts & figures 2025. Accessed: 11 July 2025.
5. Campanella, G., Hanna, M. G., Geneslaw, L. *et al.* Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Medicine* **25**, 1301–1309 (2019).
6. Litjens, G. *et al.* Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci. reports* **6**, 26286 (2016).
7. Gurcan, M. N. *et al.* Histopathological image analysis: A review. *IEEE Rev. Biomed. Eng.* **2**, 147–171 (2009).
8. Irshad, H. *et al.* Methods for nuclei detection, segmentation, and classification in digital histopathology: A review. *IEEE Rev. Biomed. Eng.* **7**, 97–114 (2014).
9. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. Image Analysis* **42**, 60–88 (2017).
10. Komura, D. & Ishikawa, S. Machine learning methods for histopathological image analysis. *Nat. Rev. Cancer* **18**, 669–676 (2018).
11. Eldaly, A. K. *et al.* The bag of words approach for the classification of head and neck cancers using diffuse reflectance spectroscopy. In *MICCAI workshop (Spectral Analysis in Medical Imaging)*, 1–10 (2015).
12. Eldaly, A. K. *et al.* Patch-based sparse representation for bacterial detection. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 657–661 (IEEE, 2019).
13. Eldaly, A. K. *et al.* Bayesian bacterial detection using irregularly sampled optical endomicroscopy images. *Med. Image Analysis* **57**, 18–31 (2019).
14. Echle, A. *et al.* Deep learning in cancer pathology: a new generation of clinical biomarkers. *The Lancet Oncol.* **22**, e407–e416 (2021).
15. Lu, M. Y. *et al.* Ai-based pathology predicts origins of cancers of unknown primary. *Nat. Medicine* **27**, 1635–1644 (2021).
16. Steiner, D. F. *et al.* Impact of deep learning assistance on pathology review. *Nat. Medicine* **26**, 89–94 (2020).
17. Tizhoosh, H. R. & Pantanowitz, L. Artificial intelligence and digital pathology: challenges and opportunities. *J. pathology informatics* **9**, 38 (2018).
18. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
19. Coudray, N. *et al.* Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Medicine* **24**, 1559–1567 (2018).
20. Wang, D. *et al.* Deep learning for identifying metastatic breast cancer. *IEEE Transactions on Med. Imaging* **35**, 1284–1295 (2016).
21. Hayat, M., Ahmad, N., Nasir, A. & Ahmad Tariq, Z. Hybrid deep learning efficientnetv2 and vision transformer (effnetv2-vit) model for breast cancer histopathological image classification. *IEEE Access* **12**, 184119–184131, DOI: 10.1109/ACCESS.2024.3503413 (2024).
22. Yan, Y., Lu, R., Sun, J., Zhang, J. & Zhang, Q. Breast cancer histopathology image classification using transformer with discrete wavelet transform. *Med. Eng. & Phys.* **138**, 104317, DOI: 10.1016/j.medengphy.2025.104317 (2025).
23. Shahadat, N., Lama, R. & Nguyen, A. Lung and colon cancer detection using a deep ai model. *Cancers* **16**, DOI: 10.3390/cancers16223879 (2024).
24. Borkowski, A. A. *et al.* Lung and colon cancer histopathological image dataset (lc25000) (2019). 1912.12142.

25. Bala, D., Karim, S. M. R. U. & Rasul, R. A. Mranet: A modified residual attention networks for lung and colon cancer classification (2024). 2412.17700.

26. Attallah, O. Lung and colon cancer classification using multiscale deep features integration of compact convolutional neural networks and feature selection. *Technologies* **13**, DOI: 10.3390/technologies13020054 (2025).

27. Atwan, J. *et al.* Ensemble of deep features for breast cancer histopathological image classification. *The Comput. J.* **67**, 2126–2136, DOI: 10.1093/comjnl/bxad127 (2024).

28. Ochoa-Ornelas, R., Gudiño-Ochoa, A., García-Rodríguez, J. A. & Uribe-Toscano, S. A robust transfer learning approach with histopathological images for lung and colon cancer detection using efficientnetb3. *Healthc. Anal.* **7**, 100391, DOI: 10.1016/j.health.2025.100391 (2025).

29. El-Aziz, A. A. A., Mahmood, M. A. & El-Ghany, S. A. Enhancing early detection of oral squamous cell carcinoma: A deep learning approach with lrt-enhanced efficientnet-b3 for accurate and efficient histopathological diagnosis. *Diagnostics* **15**, DOI: 10.3390/diagnostics15131678 (2025).

30. Han, S., Eldaly, A. K. & Oyelere, S. S. Towards human-ai collaboration system for the detection of invasive ductal carcinoma in histopathology images. *Human-Intelligent Syst. Integration* 1–14, DOI: 10.1007/s42454-025-00073-z (2025).

31. Ghaderzadeh, M. *et al.* A fast and efficient cnn model for b-all diagnosis and its subtypes classification using peripheral blood smear images. *Int. J. Intell. Syst.* **37**, 5113–5133, DOI: 10.1002/int.22753 (2022).

32. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale (2021). 2010.11929.

33. Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255, DOI: 10.1109/CVPR.2009.5206848 (2009).

34. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization (2017). 1412.6980.

35. Moshkov, N., Mathe, B., Kertesz-Farkas, A., Hollandi, R. & Horvath, P. Test-time augmentation for deep learning-based cell segmentation on microscopy images. *Sci. Reports* **10**, 5068, DOI: 10.1038/s41598-020-61808-3 (2020).

36. Kumar, A., Vishwakarma, A. & Bajaj, V. Ml3cnet: Non-local means-assisted automatic framework for lung cancer subtypes classification using histopathological images. *Comput. Methods Programs Biomed.* **251**, 108207, DOI: 10.1016/j.cmpb.2024.108207 (2024).

37. Thiriveedhi, A., Ghanta, S., Biswas, S. & Pradhan, A. K. All-net: Integrating cnn and explainable-ai for enhanced diagnosis and interpretation of acute lymphoblastic leukemia. *PeerJ Comput. Sci.* **11**, e2600, DOI: 10.7717/peerj-cs.2600 (2025).

38. Başaran, E., Çelik, G. & Toğaçar, M. Regionally focused neural-coder model designed for the diagnosis of acute lymphoblastic leukemia disease. *Measurement* **256**, 118176, DOI: 10.1016/j.measurement.2025.118176 (2025).

39. Muhammad, D., Salman, M., Keles, A. & Bendechache, M. All diagnosis: Can efficiency and transparency coexist? an explainable deep learning approach. *Sci. Reports* **15**, 12812, DOI: 10.1038/s41598-025-97297-5 (2025).

Author contributions statement

R.M. and A.K.E. conceived the study. R.M. conducted the study and performed the data curation. R.M. and A.K.E. developed the methodology and performed validation. R.M., M.A., and A.K.E contributed to the investigation and conceptualisation. A.K.E. provided supervision and resources. All authors reviewed and edited the manuscript.

Additional information

Accession codes: The codes will be made publicly available upon the acceptance of the manuscript; **Competing interests:** The authors declare no competing interests; **Funding:** The authors received No Funding for this work.