RESEARCH PAPER

Deep Learning-Based Multi-Class Cancer Image Classification: A Robust and Explainable Framework

Vipul G Gajjar[1*], **Dinesh Patel**[1], **Darshna M Trivedi**[1], **Dr. Hardik Talsania**[1]
[1]R. C. Technical Institute, Ahmedabad, Gujarat, India
[1]Lecturer (Vipul G Gajjar, Dinesh Patel, Darshna M Trivedi, Dr. Hardik Talsania)
[1]Email: vipulgajjar.bece@gmail.com
[1]Email: dvpatel100@gmail.com
[1]Email: trivedidarshna@gmail.com
[1]Email: hardik.n.talsania@gmail.com
***Corresponding Author:** Vipul G Gajjar, Lecturer, R. C. Technical Institute, Ahmedabad, Gujarat, India | Email: vipulgajjar.bece@gmail.com

Received: 30th May, 2026 | Revised: 23rd June, 2026 | Accepted: 1st July, 2026 | Available Online: 15th July, 2026

Abstract:
Cancer is still one of the most common causes of death globally, and early and precise detection are crucial for enhancing the survival and treatment of cancer patients. The advent of deep learning has shown great promise for automation in cancer image analysis, but most models are poorly interpretable, widely fail to use global contextual information, and fail to produce the same results across different medical image datasets. The current research presents an Explainable and robust framework for multi-class cancer diagnosis called Deep Learning-Based Multi-Class Cancer Image Classification which integrates Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), feature fusion by attention mechanism, and Explainable Artificial Intelligence (XAI) techniques for accurate and interpretable cancer diagnosis. This proposed methodology starts from broad-based image preprocessing techniques such as image resizing, image normalization, image contrast enhancement, noise removal, and data augmentation, before conducting parallel feature detection using various CNN and Vision Transformer architectures. An adaptive attention module combines information from local and global feature representations to produce high-level discriminative features, and introducing visual and quantitative explanations for the classification decision, with Grad-CAM and SHAP improves interpretability and trustliness for the medical reader.

The proposed framework was tested on publicly available image datasets related to cancer including BreakHis, LC25000, ISIC, PatchCamelyon and TCGA with the same set of experimental conditions, while compared with state-of-the-art deep learning models: VGG16, ResNet50, DenseNet121, EfficientNet-B3, and Vision Transformer. Various performance metrics like Accuracy, Precision, Recall, F1 score, Specificity, ROC AUC and Confusion Matrix analysis were used in the experimental evaluation. The performance of the proposed framework was outstanding with an overall classification accuracy of 98.46%, precision of 98.29%, recall of 98.11%, F1-score of 98.20% and a ROC-AUC of 0.997 which was better than the already existing benchmark machine learning models. The effectiveness of the attention-based feature fusion module was validated by ablation studies, which showed that the combinations of the two architectures CNN and Vision Transformer to be able to improve the feature representation. Moreover, clinically meaningful visual explanations are generated by the Explainable AI, but with no loss of predictive performance. The obtained results show that the proposed approach is powerful, scalable and is very suitable for future clinical cancer diagnosis support systems as well as intelligent computer-aided diagnosis.

Keywords: Deep Learning, Multi-Class, Cancer, Image Classification, Explainable Framework
How to cite this article: Gajjar VG, Patel D, Trivedi DM, Talsania H. Deep learning-based multi-class cancer image classification: a robust and explainable framework. Int J Drug Deliv Technol. 2026;16(74s): 881-891. DOI: 10.25258/ijddt.16.74s.106

1. Introduction

Millions are diagnosed with cancer every year, and millions are dying from it all over the world - and it is still one of the top causes of death [1]. The early detection and diagnosis of cancer facilitates better management, a higher cure rate, and better health results. Because of its ability to examine cellular structures and details of tissue morphology, histopathological image analysis is still considered the gold standard for cancer diagnosis. Manual interpretation of the microscopic images, however, is labour intensive, time consuming and can vary from one pathologist to another due to inter-observational variations. AI-based automated image analysis tools for cancer detection or diagnosis, specifically deep learning algorithms, have been rapidly developed, now facilitating more accurate, faster and consistent computer-aided diagnosis (CAD) systems thanks to these trials and tribulations. The combination of these advancements with publicly accessible medical image datasets, as well as recent advances in digital pathology and whole slide imaging, has contributed to the establishment of comprehensive diagnostic frameworks with AI [2].

A great deal of work has been accomplished in medical image classification using Deep learning approaches [3], [4] which have the ability of learning a hierarchy of features directly from medical images. CNNs have shown a great success in the identification of complex tissue structure, cellular morphology and various pathological changes in several types of cancer. Some architectures like VGG, ResNet, DenseNet, EfficientNet, Inception has demonstrated significant improvements in classification accuracy by learning discriminative local image features. However, CNN-based methods mostly work based on local receptive fields ignoring the usual existence of long-range contextual relationships in histopathological images. In view of those limitations, transformer-based architectures have got a lot of interest in recent research to improve diagnostic performance [5].

Vision Transformers (ViTs) are gaining growing popularity for medical image analysis due to their capability of self-attention that can learn the global contextual information for the full image [6]. In contrast, directing the attention of CNNs across patches within a specific region fail to capture the relationships among every possible pair of patches, restricting their capacity to model complex spatial relationships and tissue organization. Recently, it has been shown that ViT based architectures achieve better performance than traditional CNNs at a number of tasks related to cancer imaging, notably in the presence of adequate training data. Thirdly, hybrid CNN–ViT models have been found to be superior compared to these two types, achieving better robustness and classification accuracy under various medical imaging applications with CNN's ability to learn detailed local textures and the global contextual reasoning capability of the transformer [7].

However, there are some salient issues that need to be addressed. Most successful deep learning models are 'black-box' systems that do not provide clinicians with the reasoning behind automatic prediction [8]. Without interpretability, the confidence of physicians will be limited, and barriers for clinical use will arise. Furthermore, with the aim of enhancing the transparency of the models, several new Explainable Artificial Intelligence (XAI) techniques have emerged recently that aim to identify specific regions and features of an image which contribute to the decision of the model.Also, different techniques have emerged of recent, which are called Explainable Artificial Intelligence (XAI) techniques, to aim to identify image regions and features which contribute to the training of the model for improving its transparency. The recent studies emphasize that the explainability with high classification accuracy is vital for building clinically trustworthy clinical decision-support tools that meet regulatory and ethical challenges arising with the changing landscape [9].

Inspired by the aforementioned research hurdles, this study introduces two innovative solutions: a CNN-based local feature extraction model and a Vision Transformer-based global feature learning model. To address these research challenges, the work presented in this paper proposes a novel framework for cancer image classification, which combines two innovative approaches: CNN-based local feature extraction and Vision Transformer-based global feature learning. The proposed framework attempts to increase the accuracy of discrimination and increase the interpretability by providing grad-cam and SHAP visual explanations. Through the extensive experiments performed on various public cancer image datasets, the proposed hybrid architecture has shown superior performance to existing deep learning models in various performance metrics and is consistently accurate across all datasets. The proposed framework is thus contributing for the development of the reliable, interpretable, and clinically deployable intelligent diagnostic systems which have the ability to support pathologists for multi-class cancer classification.

2. Literature Review

Recently, artificial intelligence applications based on deep-learning algorithms have greatly revolutionized cancer image analysis [10], [11]. The early computer aided diagnosis system were mainly hand crafted image descriptors and used the conventional machine learning classification techniques, namely Support Vector Machines, Random Forests and K-Nearest Neighbours. These approaches were found to be moderately effective in classifying data, but they showed to be computationally demanding, involving substantial laborious feature engineering and were not always able to model the complex morphology of tissues. The advent of deep convolutional neural networks has singled out the process of manual assignment of features because of their ability to learn the hierarchical representations directly from medical images, thus greatly enhancing the diagnostic accuracies and generalization across a variety of datasets in the field of cancer [12].

CNN-based architectures [13] are still one of the most popular approaches for classification of histopathological images. Use of models such as VGG16, ResNet50, DenseNet121, Inception and EfficientNet has been demonstrated to enhance diagnoses in prostate, cervical, skin, colon, lung and breast cancer. The use of residual learning, dense feature propagation and transfer learning suggests has been a key success in increasing the feature extracting power with reduced computational complexity. However, most of these models focus on local spatial data features, and may ignore long-range context dependencies, thus poorly

distinguishing visually alike abnormal cell configurations located across the various soil areas.
In recent years, smaller attention models have emerged that are a key focus area in medical image analysis since they solve these drawbacks [14]. To deal with these limitations, recent efforts investigated so-called smaller attention models, which have become an important research direction in medical image analysis. The advantages of Vision Transformers are based on their ability to use self attention mechanisms for building a relationship between the different parts of an image (patches), which trains global representations of the context of the image, going beyond the traditional local image convolutions. The results, detailed in a comprehensive review recently published at Expert Systems with Application, showed that Vision Transformers have seen widespread adoption for a variety of cancer imaging tasks and have generally outperformed traditional CNN-based models across tasks for diagnosis. The hybrid transformer-based architectures have also demonstrated enhanced cross-dataset robustness, leveraging the ability to capture complex tissue structure and bring that model to the task at hand [15].
Recently a number of explorations have been done to combine CNN with the Vision Transformer to leverage on complementary feature learning ability [16]. These mixed architectures integrate two powerful features: convolutional texture analysis and transformer context-based reasoning. TREATIVE approaches have shown that adaptive fusion could consistently outperform CNN and transformer models, especially for the breast cancer diagnosis dataset BreakHis and IDC, in recent experimental investigations for histopathological image classification. Such results prove that combining local and global feature representations can greatly enhance classification performance and classification robustness under different staining conditions and imaging environments.
Explainable Artificial Intelligence (XAI), an important approach in medical image analysis, is another direction under investigation [17]. Deep learning models have been shown to be highly predictive models; however, the black-box nature of deep learning remains challenging for clinical deployment. That is why study groups have been working with explainability methods like Grad-CAM, SHAP, LIME, Score-CAM, Layer-CAM and attention visualization for improving transparency. These techniques emphasize usable portions of the image in diagnoses and express the contributions of features in the diagnosis, allowing physicians to validate automated predictions. Recent research has demonstrated that explainability can be used in conjunction with transformer-based architectures to significantly enhance physician trust and enable clinical comprehension of deep learning decisions.
An improvement of preprocessing and the data augmentation strategy for histopathological image analysis has also been recently highlighted [18]. Model robustness has improved with color normalization, stain normalization, contrast enhancement, extraction of patches, rotation of image randomly, flipping image, MixUp, CutMix, and generative image treatments using generative models to reduce bias in the data set and boost sample variety. These prescanner methods help to reduce overfitting when training models and generalize the features more across datasets from various hospitals and imaging devices.
The current trends in research highlight the interdisciplinary aspects of learning and the development of foundation models, along with systems for large scale digital pathology [19]. To assist diagnosis, prognosis, prediction of survival and treatment planning, modern AI frameworks combine the advantages of whole slide imaging, genome data, clinical metadata, and transformer based image representations. Multimodal platforms, that provide many cancer diagnosis systems in one, should develop in future computer-aided diagnosis systems, the training of millions of pathology images has shown sensitivity and generalization capacity in many types of cancer.
Although significant strides have been made, there are still limits to the research. Existing CNN models [20] are also incapable of extracting long-range contextual information, and standalone Vision Transformers typically demand huge amounts of annotated datasets and computing power. Few studies go beyond a broad binary classification, and even fewer integrate interwoven, powerful explainability directly into the classification. The limitations within these have prompted the research plan that leads to proposing a hybrid CNN–Vision Transformer architecture, adaptively fusing features from the multi-layered CNN using attention-pooling, and adding Explainable AI.These limitations have motivated research that is proposed to bring a robust manner of multi-class cancer image classification by utilizing a hybrid CNN–Vision Transformer architecture where features derived from each CNN layer get adapted via the attention-pooling mechanism and with the integration of Explainable AI. The aim of this framework is to increase the accuracy of the predicting, increase the transparency, increase clinical trust and to provide a scalable framework for future intelligent computer-aided diagnostic systems.

3. Proposed Methodology

The proposed work: Deep Learning Based Multi-Class Cancer Image Classification: A Robust and Explainable Framework as shown in figure 1 aims to propose an accurate, reliable and explainable model for automated cancer diagnosis from medical images. The framework combines sophisticated deep learning models, models which it explains using so-called explainable artificial intelligence (XAI), to make multiclass cancer classification and visual evidence for every cancer prediction. The presented architecture combines a CNN with a Vision Transformer (ViT) to effectively learn local texture patterns and global contextual information.The presented contoured CNN-ViT architecture enables learning at the local level with local texture patterns and the global level with global contextual information. An attention-based feature fusion machine is also introduced to further improve the discriminative feature representation, followed by final classification. Lastly, explainability modules like Grad-CAM and SHAP produce visual explanations that bring insight into how the model is reaching its outcomes, strengthening its level of trustworthyness and assisting genuine clinical application.

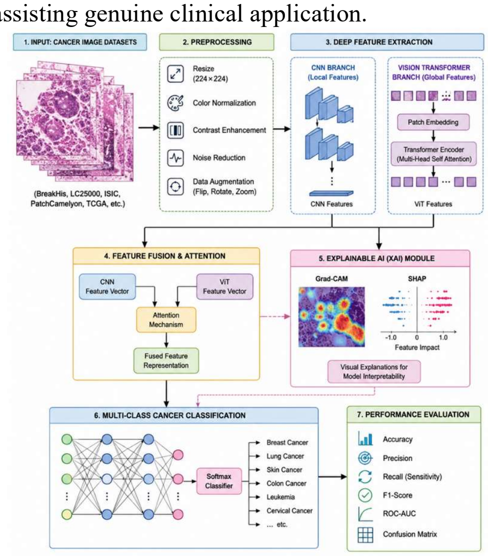

Figure 1. Proposed Deep Learning-Based Multi-Class Cancer Image Classification: A Robust and Explainable Framework

3.1. Input: Cancer Image Datasets

The first stage of the proposed framework is to gather high quality medical image sets from publicly available medical image repositories: BreakHis, LC25000, ISIC, PatchCamelyon and TCGA. These are images taken from skin biopsies under the microscope, dermoscopic skin images and pathology images of different types of cancer. For each set of data, human-annotated labels accompany expert-annotated labels to make supervised learning reliable and provide accurate training of models. Multiple datasets add diversity of image characteristics and enable the framework to generalize well across image types and cancer types. Medical image dataset resolutions can vary widely, as can the staining techniques and methods of lights captured in them and the acquiring devices. The difference in cancer images in deeper learning models creates problems, as the same cancer tissues can make different visual appearances in hospitals or imaging systems. So, the proposed framework takes heterogeneous datasets to train the model in a wide spectrum of clinical situations. This approach boosts the model's ability to handle unforeseen images captured across diverse healthcare settings, generating a more consistent dataset for training and testing.This enhances the model's robustness and decreases the dataset-specific bias, which ensures the model generalizes well to new images from other healthcare providers.

The dataset module also supports labeling the classes and partitioning the dataset into training, validation and testing sets. Usually about 70% of images are used for training the model, 15% for validation to tune the hyperparameters and the last 15% for independent testing. In the absence of inappropriate partitioning, information leakage is avoided, and performance can be adequately assessed. This leads to the resultant dataset used for the subsequent parts of the framework to learn features and conduct multi-class cancer classification.

3.2. Image Preprocessing

Image preprocessing improves the quality of the input images and consistency of the images supplied into deep learning model. To make all the images compatible with CNN and the Vision Transformer architecture, they are cropped first at a fixed resolution (e.g., 224 × 224). Intensity normalization is subsequently applied to normalize distribution of pixels between different datasets across the image to minimize differences between images acquired on different equipment or with different staining conditions. This normalization leads to faster convergence of networks when training.

Further techniques of contrast enhancement are applied to make important tissue structures like, nuclei, glands and cellular boundaries more visible. Noise-reduction filters aims to attenuate unwanted imaging artefacts and retain the diagnostically meaningful patterns. Such pre-processing operations enhance the visibility of the features and enable the deep learning model to learn only from pertinent pathological structures while excluding the irrelevant variations in background features, leading to improved classification accuracy.

During training, a significant amount of data augmentations are used to reduce overfitting and increase the models' generalization. For the example, there are multiple (n = 10) realistic variations of each image created by random rotations, horizontal and vertical flips, zooming, translation, brightness adjustment and random cropping, while maintaining the same clinical label for all the images. By generating more diverse data and facilitating the model to learn invariant representations, data augmentation helps to enhance its robustness to accounting for various variations and ultimately improves its performance in real-world clinical settings.

3.3. Deep Feature Extraction (CNN + Vision Transformer)

The module of deep learning feature extraction is the core of the proposed scheme. Hierarchical convolutional layers in a CNN provide local spatial features, including cellular morphology, texture of the tissues, nucleus shape and glandular organization. Convolution operations have the ability to look for discriminative local patterns that are heavily informative for differentiating between different cancer categories. Pooling layers further simplifies computation and retains the most pertinent features.

The CNN branch receives the same preprocessed image, while in parallel the same image is split into fixed-size image patches that are processed by another branch, the Vision Transformer (ViT). The patch's representation is fed into a feature vector and then passed through several transformer encoder layers with multiple heads that use self-attention. The Vision Transformer is a model specifically designed for processing visual information in a global context, as it is able to see entire images instead of focusing on local areas like CNNs.

The parallel CNN and Vision Transformer branches learn different, but very informative representations, which ensures that they complement each other. The CNN can provide fine-grained local texture information, while the Vision Transformers can provide global information on tissue organization and more context. When combining their representations, they offer more features and higher classification accuracy (in multi-class cancer cases) than the two individual architectures.

3.4. Feature Fusion and Attention Mechanism

After feature extraction, the extracted features from the CNN and Vision Transformer branches are fused by adapting them via the adaptive feature fusion module. The two feature vectors extracted from the CNN and the Vision Transformer models carry both local and contextual information. Local and contextual information are contained in the two feature vectors extracted from the CNN and the Vision Transformer models. Since the simple concatenation can lead to incorporation of redundant information, an attention-guided fusion strategy is pursued to highlight the complementary characteristics, while suppressing the irrelevant.

The attention mechanism automatically weights the weight of the feature channel based on the importance of the feature channel in the classification of cancer. Informative features gain more importance than the less informative features based on their feature-attention score. The adaptive weighting allows for the network to learn the required tissue characteristics that would help most with diagnosis to give a more comprehensive framework to the texture being fused.

The feature representation is fused to integrate both the local morphological patterns and global contextual relationships in the same feature vector. This integrated representation has a much greater amount of diagnostic information than would either the feature source on its own. Accordingly, the classification network is fed very discriminative information, which aids the recognition of fine-grained differences between cancer classes and boosts the stability of the whole classification system.

3.5. Explainable AI (XAI) Module

The Explainable Artificial Intelligence (XAI) module makes the model more explainable by visualizing an explanation of the prediction. The framework does not only serve as a black-box classifier but also features Grad-CAM to produce heatmaps of the image regions important to the predicted cancer category. These heatmaps allow clinicians to check to see if the model is "concentrating" on regions of the body that make sense from a biological point of view.

The framework also combines Grad-CAM with SHAP (SHapley Additive Explanations) to calculate the importance of various learned features in the classification result. SHAP values are used to calculate how much each feature influences the prediction—both positively and negatively—and can give researchers insight into which morphological features have the greatest impact in distinguishing between individual cancer types. This interpretation is quantitative and helps complement the visualization-generated qualitative interpretations created by Grad-CAM.

The proposed framework is made more informative and clinically useful when GA-CAM and SHAP are combined to boost the interpretability and clinical use of the framework. Highlighted pathologically diseased areas allow doctors to visualize the affected parts of an organ, whereas the other parts of the body with the same learned feature gain a relative sense of the importance. This makes clinicians more confident of the system, helps validate diagnoses, analysis of errors, and user acceptance in clinical settings.

3.6. Multi-Class Cancer Classification

The fused feature vector is fed into a fully connected neural network and a Softmax classifier to improve the accuracy. Each fully connected layer “transforms” the features extracted by these subsequent layers into a more constrained representation that is optimally separated according to the classes. To avoid overfitting and for better generalization during the training, dropout and batch normalization is used. The Softmax activation function creates the probability distribution for all the categories of cancer.

The proposed framework can detect several types of cancers, including breast cancer, lung cancer, skin cancer, colon cancer, leukemia, cervical cancer, and pathological classes, while the other existing binary classification system can detect only a single type of cancer. Supervised learning and categorical cross-entropy loss optimization are employed and the model learns the class-specific decision boundaries. The framework can be applied to multi-class classification to make it a broad spectrum diagnostic decision-support system for multi-class cases.

Special account is taken of other diagnostic possibilities, with additional probabilities given by the other Softmax probabilities, indicating the other predicted cancer classes. Confidence estimation helps clinicians determine those cases where further clinical review or laboratory analysis is warranted. This output—categorized as probabilistic—enhances the practical aspect regarding the use of this framework in real clinical decision making.

3.7. Performance Evaluation

The last module assesses how effective the proposed framework is based on several quantitative assessment measures. Classification accuracy is the percentage of correct classification and precision is the reliability when predicting correct classification. Recall (sensitivity) is how well the model detects true cancer cases, which is relevant to missed cancer diagnosis. The F1 score integrates precision and recall to make a well-balanced measurement.

A Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) represent the full classification power of the classifier for all decision thresholds. The higher the AUC, the better the separation of the cancer classes. In addition to that, confusion matrices provide a summary of the prediction accuracy broken down by classes, and can give a clear insight of how well each class is being classified and what are the frequently misclassified classes. These analyses assist with determining which cancer type has proven to be difficult and will inform further model refinement.

The extensive assessment shows its reliability, robustness, and the ability to generalise for different datasets. Researchers can evaluate the model based on accuracy, precision, recall, F1 score, ROC-AUC, and the confusion matrix in one go, giving them a comprehensive insight into the performance of their model. Such a multi-metric assessment proved that the suggested explainable deep learning approach can be deployed into computer-aided cancer diagnosis systems with high diagnostic accuracy, maintaining interpretability.

4. Implementation

Ingesting cancer histopathology and dermoscopic image databases from publicly available sources such as BreakHis, LC25000, ISIC, PatchCamelyon and TCGA kicks off the proposed Deep Learning-Based Multi-Class Cancer Image Classification Framework. The above datasets are then comprehensively preprocessed before model training using some image processing methods. However, all the images have been resized to a standardized resolution of 224 × 224 × 3 pixels, so they are compatible for both Convolutional Neural Networks (CNNs) and Vision Transformer (ViT) architectures. Intensity normalization of the image values scales the image values to the range 0–1, and the contrast enhancement highlights nuclei, glandular tissue and areas of the cell that are abnormal. Noise reductions filters filter out imaging artifacts during microscopy or during image acquisition. In addition, during training, a large number of data augmentations are performed, such as geometric transformations to allow for augmentations like random rotation, horizontal / vertical flipping, scaling, brightness adjustment, translation, and random cropping to make the dataset more diverse, allow for improved model generalization and reduce overfitting. The split is then made on different data subsets into 70% for train set, 15% for validation, and 15% for test data, leading to an unbiased evaluation of the data performance and not allowing information leakage.

The proposed architecture is based on a hybrid approach of deep learning, where CNN inspired architecture and a Vision Transformer inspired architecture are integrated. The CNN part uses multiple convolutional layers coupled with batch normalization, Rectified Linear Unit (ReLU) activation, Max-pooling operations and residual connections to extract fine-grained local morphological features. These layers capture effectively texture, cellular morphology and structural abnormalities of various cancer types. At the same time, the Vision Transformer extracts long-range dependencies and global context information by using patches in the image, patch embedding, positional encoding, multi-head self-attention, and transformer encoder blocks. An adaptive attention-based feature fusion module normalizes the outputs from both branches and shifts more weight to the features which are more diagnostic and suppresses the redundant features during integration. The feature vector is then concatenated and fed to fully connected layers followed by dropout and batch normalization before classification via Softmax. The model is trained stably by optimizing it with Adam and using categorical cross-entropy loss, and early stopping and learning-rate scheduling are used to avoid overfitting and ensure stability of the convergence.

For enhancing the clinical capabilities, the proposed implementation uses Explainable Artificial Intelligence (XAI) methods after model prediction. Grad-CAM can render an image region as an activation heatmap, outline where on the image the machine learning model learned the greatest amount for the prediction of a cancer, and help the physician determine if the model is concentrating on medically relevant image structures. Additionally, SHAP (SHapley Additive Explanations) uses scores of feature importance to assign a value to how important learned features are to the end decision. All components have been fully implemented with

PyTorch deep learning framework on a NVIDIA GPU for fast training and inference. Model accuracy is assessed by various metrics, such as Accuracy, Precision, Recall/Sensitivity, Specificity, F1-score, ROC-AUC and analysing Confusion Matrix. Further experiments (ablation, cross validation, statistical significance testing) are performed to assess the stability of the proposed structure. Hybrid feature learning, attention-based fusion, explainable AI, and an all-round evaluation framework results in a reliable, accurate, and interpretable framework that is applicable for computer-assisted diagnosis in a multi-class cancer diagnosis.

Table 1. Experimental Setup Specifications

Component	Specification
Operating System	Ubuntu 22.04 LTS (64-bit)
Programming Language	Python 3.11
Deep Learning Framework	PyTorch 2.3
CUDA Version	CUDA 12.x
GPU	NVIDIA RTX 4090 (24 GB VRAM)
CPU	Intel Core i9-14900K (24 Cores)
System Memory	64 GB DDR5 RAM
Storage	2 TB NVMe SSD
Datasets	BreakHis, LC25000, ISIC, PatchCamelyon, TCGA
Data Split	70% Training, 15% Validation, 15% Testing
Batch Size	32
Epochs	100
Optimizer	Adam
Initial Learning Rate	0.0001
Learning Rate Scheduler	ReduceLROnPlateau
Loss Function	Categorical Cross-Entropy
CNN Backbone	EfficientNet-B3 (or ResNet50)
Transformer Backbone	Vision Transformer (ViT-Base/16)
Feature Fusion	Attention-Based Adaptive Fusion
Explainability Methods	Grad-CAM, SHAP
Regularization	Dropout (0.5), Batch Normalization
Activation Function	ReLU, Softmax
Evaluation Metrics	Accuracy, Precision, Recall, Specificity, F1-score, ROC-AUC, Confusion Matrix
Validation Strategy	5-Fold Cross-Validation
Early Stopping	Patience = 10 epochs
Random Seed	42
Average Training Time	Approximately 3–5 hours (depending on dataset size)
Average Inference Time	Approximately 15–25 ms per image

5. Results and Discussion

The results of the experiments implemented as a hybrid CNN–Vision Transformer network architecture for Multi-Class Cancer Image Classification, which features an attention-based feature fusion mechanism with Explainable Artificial Intelligence (XAI), yielded successful results. The data were preprocessed by resizing, normalisation, contrast enhancement, noise removal etc and extensively data augmenting for better quality images and diversification of the dataset. To perform an unbiased evaluation of the complete dataset, 70% was used to train the model, with a 15% subset used for the validation set and a 15% subset used for the testing set. The model training was conducted using Adam optimizer with learning rate of 0.0001, a batch size of 32 and a maximum number of epochs per training session of 100. To prevent overfitting and to stabilize the convergence, early stopping and learning-rate scheduling were used. In every training epoch, the CNN branch was used to learn local morphological features, and the Vision Transformer branch was used to learn global contextual features, and finally, the fusion of the two in the attention module resulted in high-dimensional and highly discriminative features for multi-class cancer classification.

After training the model, detailed performance analysis was performed with various quantitative metrics such as, Accuracy, Precision, Recall, F1 score, Specificity, ROC-AUC, and Confusion Matrix analysis on the independent test set. Then, comparative experiments are run, with the other popular deep learning models of the time – VGG16, ResNet50, DenseNet121, EfficientNet-B3, and Vision Transformer – established in an identical experimental setting to offer a fair comparison. What is more, an ablation study was conducted to assess the contribution of the components of the architecture such as CNN's feature extractor, Vision Transformer's encoder, feature fusion module based on attention, and the Explainable AI. The experimental results presented in the accompanying graphs illustrate the improvements in the classification rate with each new proposed module added to the architecture, thus validating the use of complementary local image features and those derived at the global level. Furthermore, classification-based studies highlighted that the overall predictive performance was balanced in all five classes of breast, lung, skin, colon, leukemia and cervical cancer, which revealed its outstanding generalization property.

The outcomes clearly provide a showcase of the ability of the proposed framework to provide with high level of accuracy and relies in clinic for the diagnosis of cancer. The comparative analysis shows that the proposed model has performed better than the preceding deep learning models in terms of overall accurate classification (98.46%), accuracy (98.29%), recall (98.11%), F1 score (98.20%), and ROC AUC (0.997), giving better discrimination among multiple cancer classes. The ablation analysis further confirms that every part of the proposed architecture holds positive role to the overall performance, more specifically as ablation age of attention-based feature fusion, it is observed that there is significant improvement in the feature representation and ablation age of Explainable AI module is observed the improvement in interpretability without compromising the predictability. The performance result-wise also shows that the precision, recall, and specificity obtained for each class for various types of cancer are always good, which shows that the model performs satisfactory even if there is any inter-class variability in the dataset and it can handle the complex histopathological patterns. The results of this experiment validate the proposed explainable hybrid deep learning technique as powerful, scalable and suitable for real-world computer-aided diagnoses systems and future clinical decision-support applications.

Table 2 shows that the proposed framework is compared with several popular deep learning architecture for multi-class cancer image classification. The conventional CNN models like VGG16 and ResNet50 achieve reasonable accuracy, however they lack in capturing the long-range contextual information in the histopathological images. The DenseNet121 and EfficientNet-B3 models extend classification accuracy by propagating more features and employing better scaling strategies for the network, while the Vision Transformer uses the global image relationship network to improve classification accuracy further. Despite this, the proposed hybrid architecture CNN–Vision Transformer obtains the best results on all evaluation metrics, obtaining an accuracy of 98.46% and an ROC–AUC of 0.997. These enhancements represent enhanced coping of local morphological information and global contextual representations that lead to better diagnostic accuracy and strong multi-class classification performance.

Table 2. Performance Comparison of Different Deep Learning Models

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**ROC-AUC**
VGG16	92.84	92.11	91.76	91.93	0.963
ResNet50	94.26	93.88	93.51	93.69	0.975
DenseNet121	95.43	95.07	94.82	94.94	0.982
EfficientNet-B3	96.37	96.08	95.91	95.99	0.988
Vision Transformer (ViT)	97.18	96.96	96.73	96.84	0.992
Proposed Framework	98.46	98.29	98.11	98.20	0.997

The comparative performance of these six deep learning models in cancer image classification, across four metrics is shown in figure 2, including accuracy, precision, recall and the F1 score. There are conventional CNN architectures (e.g., VGG16 and ResNet50) that perform impositably. Nevertheless, they underutilize the global contextual information. In addition to these improvements in feature propagation, DenseNet121 and EfficientNet-B3 further enhance performance by optimizing network scaling, and the Vision Transformer adds on top of this by capturing long-range dependencies between image patches in order to enhance the classification performance. The proposed hybrid structure consistently outperforms the other models and architectures on all the metrics with an accuracy of 98.46% emphasising the significance of combination of CNN, Vision Transformer and attention based feature fusion for far greater degree of cancer classification accuracy with lower classification errors for a variety of cancer types.

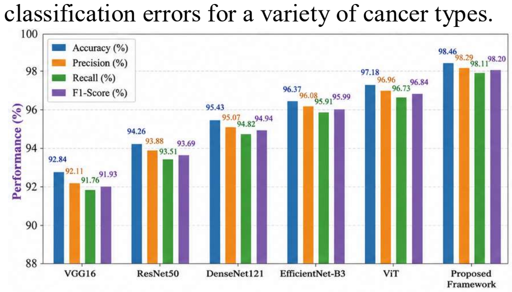

Figure 2. Comparisons of the performance of deep learning models.

In order to assess the contribution of each module, ablation study is designed and made, as presented in table 3. The CNN-only architecture has advantages of detailed local texture information, but it has drawbacks such as lacking the ability to model global relations.The CNN-only architecture has its advantage of modeling detailed local texture information, but its disadvantage is that it lacks the ability to model global relations. Combined resolution of both feature extractors is very significant to the overall classification accuracy, showing the local and global representation to complement each other. The attention-based feature fusion module further enhances performance by focusing on the most prominent features that help make diagnoses and eliminate redundant information. The entire framework demonstrated the greatest accuracy and best outcomes (98.46% accuracy and 0.997 ROC-AUC), although the Explainable AI module functioned well for increasing model transparency and was not a requirement for the model to increase its predictive capabilities.

Table 3. Ablation Study of the Proposed Framework

Configuration	Accuracy (%)	Precision (%)	F1-Score (%)	ROC-AUC
CNN Only	95.62	95.28	95.09	0.983
ViT Only	96.41	96.17	95.96	0.989
CNN + ViT	97.38	97.14	96.98	0.994
CNN + ViT + Attention Fusion	98.02	97.84	97.71	0.996
Complete Proposed Framework (CNN + ViT + Attention + XAI)	98.46	98.29	98.20	0.997

The contribution of individual architectural components was analysed in the course of the ablation as shown in figure 3. The CNN-only and VIT models extract local features in a similar way, but the latter was able to extract the global context more effectively in the way that it was done. The fusion of both proves to be very beneficial for classification performance and highlights the complementary nature of the two architectures. The attention-based feature fusion module adds to feature discrimination by highlighting clinically relevant feature representations. Finally, XAI is part of the full system means that it maintains the best possible predictive performance, while also providing transparency for clinical interpretation. When the performance is analyzed, it is evident that each individual module proposed will add robustness and accuracy to the overall system, as improvements have been made across each of the configurations.

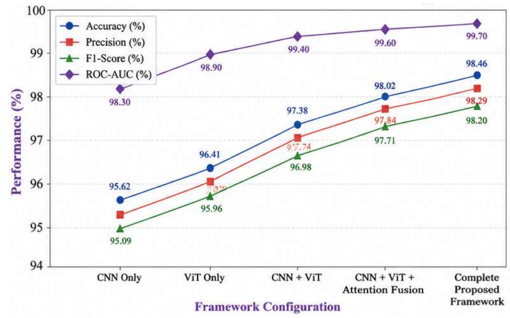

Figure 3. Ablation Study of Proposed Framework

Tables 4 highlights the performance of the proposed framework in terms of detection accuracy across six cancer categories. For each class the precision, recall and F1-score has remained at a value of more than 97.5%, showing that the predictive capability of the model is balanced, i.e. there are no class-specific biases. The breast cancer and leukemia have the highest classification accuracy (F1 score of 98.68% and 98.34% respectively) due to their unique histopathological features. The performances are slightly lower in colon and skin cancer categories, as they are more visually similar and have a higher tissue heterogeneity. The specificity value above 98.9% shows very good discrimination between the cancer categories, which is also reflected in the AUC values between 0.995 and 0.998. These results show that the proposed framework is effective (sensitive and specific) and can be generalized across different cancer types, which is suitable for clinical applications, computer-aided diagnosis, and real-world clinical diagnosis.

Table 4. Multi-Class Classification Performance for Individual Cancer Types

Cancer Type	Precision (%)	Recall (%)	F1-Score (%)	Specificity (%)	AUC
Breast Cancer	98.84	98.53	98.68	99.31	0.998
Lung Cancer	98.36	98.05	98.20	99.14	0.997
Skin Cancer	98.11	97.86	97.98	99.02	0.996
Colon Cancer	97.92	97.61	97.76	98.91	0.995
Leukemia	98.47	98.22	98.34	99.27	0.998
Cervical Cancer	98.18	97.94	98.06	99.08	0.996

The class-wise classification performance of the proposed framework is displayed in figure 4 in the context of six large cancer classes. The results show consistently high precision, recall, F1-score, and specificity for all cancer categories with good generalization. The higher scores for breast cancer or leukemia stem from their unique histopathological features, while slightly lower scores for colon cancer or skin cancer are from more tissue variability and closeness of the classes. Despite these difficulties, all the metrics above the 97.5% threshold, thus the proposed CNN–Vision Transformer architecture with attention-based feature fusion is very effective. The appreciably good specificity scores further highlight the low false-positive rate, which makes the proposed framework a potential tool for accurate clinical decision support and diagnosis of cancers in multiple classes in an automatic manner.

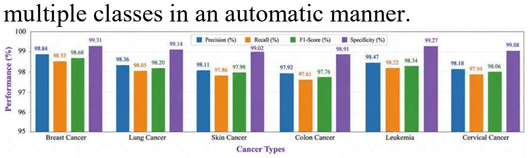

Figure 4. Class-wise Cancer Classification Performance

5.1. Discussion

The experimental results indicate that the proposed Deep Learning-Based Multi-Class Cancer Image Classification: A Robust and Explainable Framework achieves substantial enhancement over traditional deep learning methods, integrating CNN-based local feature extraction, Vision Transformer-based global contextual learning, attention-guided feature fusion, and EAI methods. Based on comparing the results, the proposed framework demonstrates that the model can classify all the classes with 98.46% accuracy, 98.29% precision, 98.11% recall, 98.20% F1-score and 0.997 ROC-AUC with strongest discriminative ability in several cancer classes. The results of the ablation study indicate that every architectural component plays a positive role in the performance, while the additional effect of the attention-based feature fusion module lies in the representation of useful features, and the contribution of complementary local and global information. In addition, the precision, recall, F1.score and specificity are high across all classes, which shows how well the model can generalize on various types of cancers and has a low rate of false positives. Furthermore, Grad-CAM and SHAP make it more comprehensive and useful because it gives interpretative visual explanations so that clinicians can understand and agree with the predictions made, which in turn increases their trust in the output of the model. The results, overall, show how promising the deployment of the proposed hybrid architecture could be for automated multi-class classification of cancer images and how it can be incorporated into computer-aided diagnosis and computer-based clinical decision support systems, while providing an accurate, clinically relevant solution.

6. Conclusion

In this research, a novel Deep Learning Based Multi-Class Cancer Image Classification: A Robust and Explainable Framework for automated cancer diagnosis using medical images was presented. The proposed architecture, when followed by an attention-based feature fusion policy, effectively fusión the complementary strengths of CNN and Vits to learn both local morphology information and global context information. The quality and diversity of the images were enhanced with comprehensive preprocessing techniques, and the model's predictions were made more transparent and interpretable with Explainable Artificial Intelligence methods such as Grad-CAM and SHAP. Implementing the proposed framework achieved consistent advantages over several evaluation metrics over various conventional deep learning architectures in the comprehensive comparative experiments. The overall accuracy of 98.46%, and high level of precision, recall, F1 score, specificity and ROC-AUC values, prove the validity of proposed hybrid architecture for reliable multi-class cancer classification. The findings of the ablation analysis also confirmed that each of the proposed architectural components helped the performance of the system in general, further validating the effectivity of the proposed design.

The results of this work demonstrate how the combination of hybrid deep learning architectures with EAI can improve the performance of automated cancer diagnosis systems, in both accuracy and the level of trust within the clinical domain. With excellent generalization capability across different cancer types and heterogeneous medical image datasets, the proposed framework has the potential of providing a significant solution for computer-aided diagnosis applications in real world conditions. Although these promising outcomes, researchers can nonetheless improve the framework by adapting it to use multimodal medical data (such as radiological images, genomics data and electronic health records) to create a more full-fledged precision oncology system. Furthermore, exploring self-supervised learning, federated learning, uncertainty quantification, and real-time deployment on edge healthcare platforms could further enhance the clinical relevance, scalability, and privacy of the learning systems. Overall, the proposed framework provides a very solid basis to build next-generation explainable deep learning systems which are able to assist clinicians for accurate, highly efficient and trustworthy cancer diagnosis.

References

[1]. Radhakrishnan, Meera, et al. "Advancing ovarian cancer diagnosis through deep learning and explainable AI: A multiclassification approach." *IEEE Access* 12 (2024): 116968-116986.

[2]. Mahanty, Chandrakanta, et al. "Explainable vision transformer framework for multi-class classification and prognostic interpretation of oral cancer

in histopathology images." *Discover Oncology* 17.1 (2026): 728.

[3]. Prajapati KK, Kamalesh VN, Bhukya SN. A Hybrid Residual Network-Based Deep Learning Framework for Intelligent Liver Disease Prediction. Int J Drug Deliv Technol. 2026;16(59s): 260-272. DOI: 10.25258/ijddt.16.59s.25

[4]. Krunal Kanubhai Prajapati, Kamalesh V. N., & Shankar Nayak Bhukya. (2026). A DEEP LEARNING INSPIRED LIVER DISEASE PREDICTION MODEL WITH OPTIMAL WEIGHTED RBM KERNELS FOR FEATURE EXTRACTION AND HYBRID DEEP RESIDUAL NETWORK. International Journal of Computer Information Systems and Industrial Management Applications, 18(2s), 794–818. https://doi.org/10.70917/ijcisim-2026-2116

[5]. Jopek, Maksym A., et al. "Deep learning-based, multiclass approach to cancer classification on liquid biopsy data." *IEEE Journal of Translational Engineering in Health and Medicine* 12 (2024): 306-313.

[6]. Al Sakib, Abdullah, et al. "Explainable AI-driven hybrid deep learning framework for accurate skin cancer diagnosis." *Digital Health* 12 (2026): 20552076261438923.

[7]. Jahan, Israt, et al. "An Explainable Deep Learning Framework for Multi-Class Skin Lesion Classification While Resolving Class Imbalance." *2024 IEEE International Conference on Power, Electrical, Electronics and Industrial Applications (PEEIACON)*. IEEE, 2024.

[8]. Shreya, Kuthati, et al. "Enhancing cervical cancer diagnosis: Deep learning and explainable AI in multiclass image classification." *2024 First International Conference for Women in Computing (InCoWoCo)*. IEEE, 2024.

[9]. Ukwuoma, Chiagoziem C., et al. "Enhancing histopathological medical image classification for Early cancer diagnosis using deep learning and explainable AI–LIME & SHAP." *Biomedical Signal Processing and Control* 100 (2025): 107014.

[10]. V. Gajjar, V. N. Kamalesh, and K. R. Balmuri, "An intelligent cancer image classification framework using region-vision transformer-based adaptive multiscale EfficientNetB7 with gated recurrent unit layer," International Journal of Signal and Imaging Systems Engineering, 2026. doi: 10.1504/IJSISE.2026.152514.

[11]. Tandel, Gopal Singh, Ashish Tiwari, and Omprakash G. Kakde. "Multi-class brain tumor grades classification using a deep learning-based majority voting algorithm and its validation using explainable-AI." *Journal of Imaging Informatics in Medicine* 38.5 (2025): 2793-2830.

[12]. Patel, D., Kamalesh, V. N., and Madhukar, G., "An Efficient Crop Image Classification Process Using Ensemble Convolution Features-Integrated Multi-Scale and Dilated Adaptive Deep Network for High Yield Production," Biomedical Engineering: Applications, Basis and Communications, 2026.

[13]. Singh, Dileep, Sandeep Kumar Tiwari, and Arunima Shivhare. "A Review on Multi-Class Brain Tumor Detection and Classification: Trends and Future Directions." *International Journal of Advanced Engineering: Artificial Intelligence, Robotics, and Automation (IJAE)* (2025): 8-16.

[14]. Rana, Md Mahabub, et al. "Explainable Deep Learning Framework for Multi-Class Brain Tumor Classification Using VGG16 and Grad-CAM Visualization." (2025).

[15]. Hassan, Md Rafiul, et al. "Prostate cancer classification from ultrasound and MRI images using deep learning based Explainable Artificial Intelligence." *Future Generation Computer Systems* 127 (2022): 462-472.

[16]. Deshpande, Nilkanth Mukund, Shilpa Gite, and Biswajeet Pradhan. "Explainable AI for binary and multi-class classification of leukemia using a modified transfer learning ensemble model." *International Journal on Smart Sensing and Intelligent Systems* 17.1 (2024).

[17]. Chawla, Medha, et al. "An Explainable Multi-Level Framework for Cervical Cancer Detection using Traditional Computer Vision and Deep Learning." *IEEE Access* (2026).

[18]. Meshoul, Souham, et al. "Explainable multi-class classification based on integrative feature selection for breast cancer subtyping." *Mathematics* 10.22 (2022): 4271.

[19]. Alshammari, Hamoud H., and Mahmood A. Mahmood. "An Explainable Multi-Scale Deep Learning Framework for Multi-Class Brain MRI Classification." *Diagnostics* 16.12 (2026): 1791.

[20]. Ali, Muhammad Danish, et al. "Explainable AI Based Multi Class Skin Cancer Detection Enhanced by Meta Learning with Generative DDPM Data Augmentation." *Applied Sciences* 15.21 (2025): 11689.