International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

Dual-branch CrossViT for ovarian cancer diagnosis: Integrating and explainable AI for real-time clinical applications

Anamul Haque Sakib [1], Md Ismail Hossain Siddiqui [2], Hasib Fardin [3], Jesika Debnath [4] and Abdullah Al Sakib [5,*]

[1] *Department of Business Administration, International American University, 3440 Wilshire Blvd STE 1000, Los Angeles, CA 90010, USA.*
[2] *Department of Engineering/Industrial Management, Westcliff University, Irvine, CA 92614, USA.*
[3] *Department of Engineering Management, Westcliff University, Irvine, CA 92614, USA.*
[4] *Department of Computer Science, Westcliff University, Irvine, CA 92614, USA.*
[5] *Department of Information Technology, Westcliff University, Irvine, CA 92614, USA.*

International Journal of Science and Research Archive, 2025, 15(01), 1834-1847

Publication history: Received on 13 March 2025; revised on 22 April 2025; accepted on 24 April 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.15.1.1164

Abstract

Early and accurate detection of ovarian cancer significantly improves patient outcomes by allowing for timely treatment. This study introduces a deep learning (DL) framework using a dual-branch Cross Vision Transformer (CrossViT) for classifying ovarian cancer subtypes through high-resolution histopathological images. Unlike traditional convolutional neural networks (CNNs), which struggle with capturing global dependencies, CrossViT utilizes multi-scale self-attention to extract detailed textural patterns and broader contextual information. This design addresses class imbalance and enhances feature learning, leading to improved diagnostic accuracy. A dataset of 100,000 histopathological images representing five ovarian cancer subtypes was compiled from Kaggle. The images underwent preprocessing, including noise reduction, data augmentation to balance class sizes, and pixel normalization for uniformity. The model also uses Gradient-weighted Class Activation Mapping (Grad-CAM) to highlight important image regions for classification, ensuring transparency and clinical reliability. Results show that the CrossViT model outperforms existing CNN models, achieving a classification accuracy of 99.24% and superior scores in F1, specificity, Matthews Correlation Coefficient (MCC), and Precision-Recall AUC (PR AUC). Additionally, a real-time web application has been developed for clinicians to quickly classify subtypes from histological samples. Future work will focus on improving computational efficiency and using more diverse datasets to enhance generalizability and clinical use.

Keywords: Ovarian cancer; Deep learning; Histopathological imaging; Explainable AI; Medical imaging

1. Introduction

Ovarian cancer presents a significant challenge in modern oncology due to its asymptomatic nature and high mortality rate. It ranks as the eighth most common cause of cancer-related death among women globally, with approximately 313,000 new cases and 207,000 deaths reported each year [1], [2]. Despite advancements in screening and detection methods, over 70% of ovarian cancer cases are diagnosed at an advanced stage (Stage III or IV), where the five-year survival rate drops below 30% [3], [4]. In contrast, the survival rate exceeds 90% for early-stage detections. Traditional diagnostic methods, which rely on manual histopathological evaluations and biochemical markers, can be labor-intensive and are often subject to interpretation variability between practitioners [5]. However, the integration of artificial intelligence in medical imaging has brought about a transformative shift in early disease detection. In particular, DL techniques have shown promise in enhancing diagnostic accuracy by automating the analysis of complex

* Corresponding author: Abdullah Al Sakib

imaging data. This approach reduces human error and supports more effective, timely, and personalized treatment strategies.

Recent studies have showcased the potential of CNNs and hybrid architectures for classifying ovarian cancer subtypes [6], [7]. For instance, previous research [8], [9], [10] integrating models such as InceptionV3, VGG19, and EfficientNet has achieved remarkably high accuracy rates. However, these approaches face several significant challenges. Many models suffer from severe class imbalance issues due to the uneven distribution of histopathological images among ovarian cancer subtypes [11]. Moreover, while high prediction accuracies are reported, the inherent black-box nature of DL models often hinders clinical trust and interpretability a critical factor when implementing AI systems in healthcare settings [3], [12]. Additionally, many existing models lack robust external validation, limiting their generalizability across diverse clinical scenarios [5], [9].

To address these limitations, our study introduces an innovative framework based on a Vision Transformer (ViT) model, specifically tailored to the multi-class classification of ovarian cancer subtypes. Unlike conventional CNN-based approaches, the proposed model leverages a dual-branch architecture, enabling it to capture both local textural details and global contextual information from histopathological images. By integrating a multi-scale self-attention mechanism, our framework effectively mitigates the effects of class imbalance and enhances diagnostic accuracy. Furthermore, to ensure that the model's decision-making process is transparent and interpretable, we incorporate Grad-CAM—a well-established explainable AI (XAI) technique. This integration not only builds clinical confidence but also provides insights into which regions of the tissue images are most critical for classification, thereby aligning model outputs with established diagnostic criteria.

Our study presents a comprehensive evaluation of the proposed model, validated on a large-scale, diverse dataset. The dataset includes high-resolution histopathological images representing various ovarian cancer subtypes, preprocessed through rigorous normalization, augmentation, and balancing techniques to overcome dataset skewness. We benchmark our results against state-of-the-art methods, demonstrating significant improvements across multiple performance metrics, including F1 score, specificity, MCC, and PR-AUC. The experimental results underscore the robustness and superior generalization of the proposed model in comparison to traditional approaches. Our contributions can be summarized as follows:

- Proposed a dual-branch ViT (CrossViT) model that captures both fine-grained textures and global contextual patterns, essential for accurate ovarian cancer subtype classification.
- Implemented advanced data augmentation and normalization strategies, we address significant class imbalance and ensure enhanced model robustness.
- Integrated Grad-CAM within our framework provides transparent and interpretable insights into the model's decision process, fostering clinical trust.
- Experimented on a large and diverse dataset demonstrates that our model achieves superior performance over traditional CNN-based methods, with improvements across key evaluation metrics.
- Developed a user-centric web application that delivers real-time diagnostic predictions, bridging the gap between research and practical clinical deployment.

The remainder of this paper is organized as follows: Section 2 reviews related work in ovarian cancer classification and the limitations of existing approaches. Section 3 details the dataset, preprocessing techniques, and the proposed CrossViT model architecture. Section 4 presents comprehensive experimental results and analyses. Finally, Section 5 and 6 concludes the paper with a discussion and conclusion with the implications of our findings and future research directions.

2. Related Works

DL is transforming smart healthcare by improving early detection and personalized treatment, particularly in medical imaging [13], [14], [15] and cancer informatics [16], [17], [18]. Researchers are utilizing DL and medical imaging to boost diagnostic accuracy and survival rates in ovarian cancer.

Radhakrishnan et al. [19] introduced a DL framework integrating InceptionV3 with XAI methods such as Grad-CAM and DeepLift for classifying five ovarian cancer subtypes from histopathological images. The model achieved 97.96% accuracy but lacked external validation. Fahim et al. [20] presented OVANet, a dual attention-enhanced CNN combining VGG19 and InceptionV3, and reached 99.01% accuracy using augmented images. However, the study was limited by its single-modality input and reliance on a small original dataset. In another effort, Behera et al. [21] combined EfficientNet-

B0 and fine-KNN to extract features and classify ovarian cancer subtypes, achieving 100% accuracy on 725 images. Despite the impressive results, the model's sensitivity for certain subtypes was compromised due to class imbalance.

Wen et al. [22] proposed MsaMIL-Net, an end-to-end multi-scale MIL network using UNet++ segmentation and patch-level classification for whole slide images. It achieved a 93.2% accuracy and 0.955 AUC but was computationally intensive and lacked real-time adaptability. Alshdaifat et al. [23] introduced a hybrid Xception-ViT model for CT-based classification of ovarian neoplasms, achieving up to 98.73% accuracy across multiple datasets, although the data originated from a single institution, limiting its generalizability. Similarly, Shetty et al. [24] focused on enhancing prediction reliability through data preprocessing on a large-scale numeric ovarian tumor dataset. After applying normalization, feature selection, and class balancing via stratified sampling, classifiers like SVM and Logistic Regression achieved up to 92% accuracy. However, their work emphasized data preparation over architectural innovation.

While these studies have laid critical groundwork, limitations persist. Most existing models either underperform in imbalanced clinical scenarios or lack explainability, which is vital for trust in healthcare AI. Moreover, state-of-the-art comparisons show that ensemble and CNN-based models often overlook transformer-based architectures, which excel at capturing long-range dependencies and contextual features.

To overcome these gaps, our study introduces a ViT-based model for multi-class ovarian cancer classification. The model is trained on a balanced dataset and integrated with Grad-CAM for interpretability. We address class imbalance using advanced sampling techniques and validate our results against state-of-the-art benchmarks. This unified framework improves not only prediction accuracy but also transparency providing clinicians with interpretable insights into the diagnostic process. Our approach contributes a more robust, explainable, and generalizable diagnostic tool for smart healthcare environments.

3. Materials and Methods

The training process involves experimenting with multiple architectures as shown in Figure 1. The model is evaluated using metrics like Specificity**,** MCC**,** PR AUC**,** and F1-Score, and performance is visualized through Grad-CAM**,** confusion matrix**,** and learning curves. The best performing model is also deployed in a web application for real-time predictions.

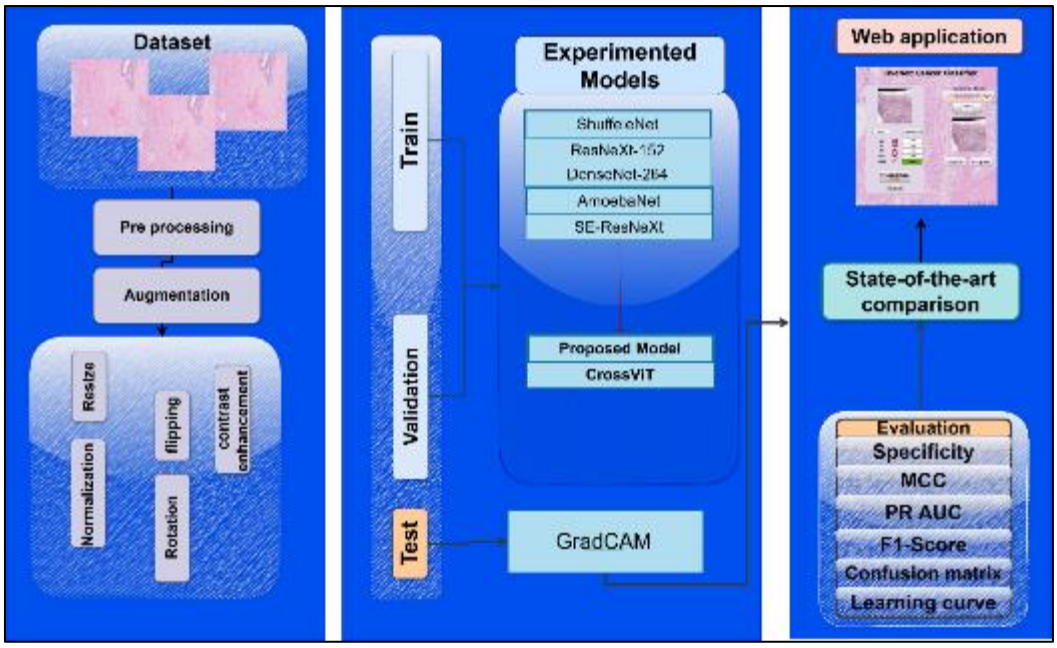

Figure 1 Overview of the proposed workflow for ovarian cancer classification

3.1. Data Description

The dataset used in this study [25] was obtained from Kaggle and contains high-resolution histopathological images representing five subtypes of ovarian cancer. These subtypes are classified based on cellular morphology, genetic profiles, and clinical behavior, aiding in personalized treatment planning. The subtypes include High-Grade Serous

Carcinoma (HGSC), Clear-Cell Carcinoma (CC), Endometrioid Carcinoma (EC), Low-Grade Serous Carcinoma (LGSC), and Mucinous Carcinoma (MC). The HGSC subtype has the largest number of images, totaling 13,188 samples, including 12,000 training and 1,188 testing images. The EC subtype includes 8,154 images, with 7,421 in the training set and 733 in the test set. The CC category comprises 6,130 images in total, with 5,579 for training and 551 for testing. LGSC, which is less represented, contains 3,195 images—2,908 training and 287 testing samples. Lastly, the MC subtype includes 3,599 images, distributed as 3,276 in training and 323 in testing. This uneven distribution highlights a moderate class imbalance across subtypes. Each image in the dataset depicts a tissue patch stained and scanned at high magnification, enabling precise visual assessment for automated subtype classification. These images serve as input for DL models to differentiate between subtypes and enhance diagnostic accuracy. A representative image sample from this dataset is shown in Figure 2.

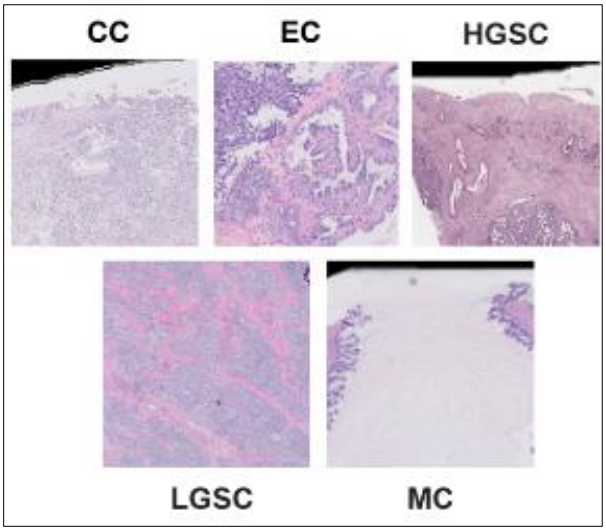

Figure 2 Sample images from each class of the experimental dataset

3.2. Data Preprocessing

To ensure uniformity and enhance the quality of the histopathological images, several preprocessing techniques were applied prior to training the DL model. First, all images were resized to 224 × 224 pixels to maintain consistency in input dimensions across the model pipeline. This resizing step is crucial for compatibility with pre-trained convolutional and transformer-based architectures. Next, noise reduction was performed using the Bilateral Filter, which smooths the image while preserving edges shown in Equation 1. Where, (f_r) is the range kernel for intensity difference, (f_s) is the spatial kernel for distance, and (W_p) is the normalization factor. This technique effectively eliminates high-frequency noise without blurring important boundaries in tissue structures.

$$I^{\text{filtered}}(x) = \frac{1}{W_p} \sum_{i \in \Omega} I(i) \cdot f_r(|I(i) - I(x)|) \cdot f_s(|i - x|) \tag{1}$$

To improve the visibility of important features, Contrast Limited Adaptive Histogram Equalization (CLAHE) was applied for contrast enhancement. CLAHE divides the image into small tiles and applies histogram equalization with a contrast limit, avoiding noise amplification in Equation 2. Here, $C(i)$ is the cumulative histogram value for intensity level (i), ($M \times N$) is the tile size, and (L) is the number of gray levels.

$$P(i) = \frac{C(i) - C_{min}}{(M \times N) - C_{min}} \times (L - 1) \tag{2}$$

Finally, Min-Max Normalization was employed to scale pixel intensity values into a range between 0 and 1, facilitating faster convergence during training. The normalization formula is given in Equation 3. This transformation ensures that each pixel's intensity contributes equally to the learning process, reducing the dominance of higher magnitude values.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{3}$$

3.3. Data Augmentation and Balancing

To address class imbalance and increase dataset diversity, we applied a range of augmentation techniques as summarized in Table 1. These included geometric transformations (e.g., flips, rotations, zooming), photometric adjustments (e.g., brightness/contrast changes, Gaussian noise), and advanced transformations like elastic distortions and cutout.

Table 1 Data augmentation strategies for enhanced model performance in ovarian cancer detection

Augmentation Type	Parameters/Range	Apply to
Horizontal Flip	50% chance	All classes
Vertical Flip	50% chance	All classes
Rotation	±20° to ±45°	All classes
Zoom In/Out	0.9x – 1.1x	All classes
Shift (Width/Height)	±10%	All classes
Shearing	10° to 20°	Minority classes (LGSC, MC)
Brightness/Contrast	0.8 – 1.2 range	All classes
Gaussian Noise	σ = 0.01–0.05	LGSC, MC
Blur (Gaussian/Median)	Kernel size = 3×3 or 5×5	Moderate use for all
Cutout / Random Erasing	Small patches randomly removed	Optional for robustness
Elastic Transform	Small alpha and sigma values	use on minority classes

By combining these techniques across the original dataset, we synthetically expanded minority classes (e.g., LGSC and MC) up to 20,000 images each as shown in table 2. This approach ensured balanced class distributions, enriched feature variability, and enhanced the model's robustness to real-world image variations during training. This imbalance can bias model learning. Post-augmentation, each subtype was standardized to 20,000 samples, divided into 80% training, 15% testing, and 5% validation sets. This balanced distribution (as shown in the table) ensures the model learns equally from all classes, reduces prediction bias, and improves generalization—especially for underrepresented subtypes like LGSC and MC.

Table 2 Distribution of original and augmented samples across ovarian cancer subtypes

Subtype	Non-Aug Samples	Train (80%)	Test (15%)	Val (5%)
HGSC	13,188	10,550	1,978	660
EC	8,154	6,523	1,223	408
CC	6,130	4,904	920	306
LGSC	3,195	2,556	479	160
MC	3,599	2,879	540	180
Subtype	Augmented Samples	Train (80%)	Test (15%)	Val (5%)
HGSC	20,000	16,000	3,000	1,000
EC	20,000	16,000	3,000	1,000
CC	20,000	16,000	3,000	1,000
LGSC	20,000	16,000	3,000	1,000
MC	20,000	16,000	3,000	1,000

3.4. Baseline Models

To ensure comprehensive subtype classification, we evaluated five state-of-the-art DL models alongside our proposed CrossViT. The baseline models include:

ResNeXt-152: This model utilizes a split-transform-merge strategy with grouped convolutions, allowing it to extract rich and diverse features while enhancing computational efficiency. Its high cardinality increases representational power without significantly raising the number of parameters [26]. We selected ResNeXt-152 due to its consistent performance in fine-grained image recognition tasks and its balance between depth, accuracy, and computational cost, making it particularly suitable for differentiating complex subtypes in histopathological images.

DenseNet-264: It introduces dense connectivity by connecting each layer to every other layer in a feed-forward manner. This design promotes feature reuse, enhances gradient flow, and reduces the risk of overfitting. DenseNet-264, in particular, offers greater network depth with fewer parameters. It was chosen for its proven effectiveness in handling small and imbalanced datasets [27], which are common in medical imaging, as well as for its ability to generalize well in deeply layered architectures.

ShuffleNet: The model is optimized for lightweight and high-speed computation, utilizing pointwise group convolutions and channel shuffling to minimize computational overhead. Despite its compact size, it maintains strong accuracy. We included ShuffleNet to assess the performance of efficient models in resource-constrained settings and to explore the feasibility of deploying cancer classification systems on portable, real-time clinical devices [28].

AmoebaNet: It is based on neural architecture search (NAS) and evolves network architectures through reinforcement learning and evolutionary strategies to achieve optimal accuracy. Its automated design exploration uncovers unconventional but highly effective structures [29]. AmoebaNet was selected for its outstanding performance in image classification challenges and its ability to adaptively learn complex visual patterns, making it ideal for nuanced subtype distinctions in cancer pathology.

SE-ResNeXt: This model integrates Squeeze-and-Excitation (SE) blocks into the ResNeXt framework, introducing a channel-wise attention mechanism that dynamically recalibrates feature responses. This enables the network to emphasize the most informative features, improving focus on lesion-specific regions in histopathological slides [30]. We chose SE-ResNeXt for its superior capacity to model channel interdependencies, which is crucial for identifying subtle textural variations among cancer subtypes.

3.5. Proposed CrossViT

We propose CrossViT to learn multi-scale visual features as shown in Figure 3, by processing image patches at different resolutions in parallel. Unlike single-scale ViTs, CrossViT effectively captures both local textures and global context, critical for histopathological pattern recognition [31]. Each branch applies self-attention to its respective patch size, and cross-attention is used to integrate complementary features [32], [33]. Shown in Equation 4-5, Where (MSA) is the multi-head self-attention, and (Z^l)denotes patch embeddings at layer(l). This bidirectional exchange fuses fine- and coarse-grained features, improving subtype differentiation.

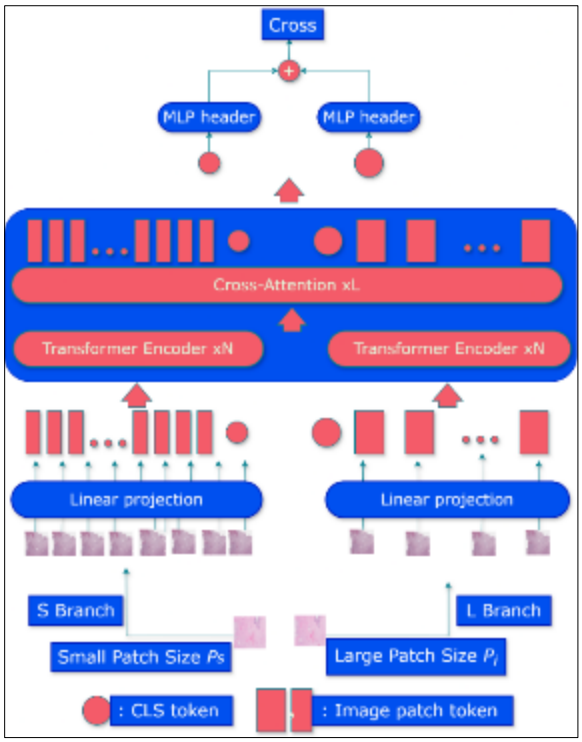

Figure 3 Architecture of our proposed CrossViT model

CrossViT was chosen for its superior ability to handle high-resolution pathological textures and model long-range dependencies across scales, achieving improved generalization and interpretability in our classification task.

$$Z^{l}_{\text{small}\to\text{large}} = \text{MSA}\left(Z^{l}_{\text{large}}, Q^{l}_{\text{small}}, K^{l}_{\text{large}}, V^{l}_{\text{large}}\right) \tag{4}$$

$$Z^{l}_{\text{large}\to\text{small}} = \text{MSA}\left(Z^{l}_{\text{small}}, Q^{l}_{\text{large}}, K^{l}_{\text{small}}, V^{l}_{\text{small}}\right) \tag{5}$$

Evaluation and Hyperparameters Settings

Table 3 The hyperparameter settings used for model training

Hyperparameter	**Range / Candidates**	**Selected Value**
Learning Rate	{1e-5, 5e-5, 1e-4, 5e-4}	1e-4
Batch Size	{32, 64, 128}	64
Dropout Rate	{0.2, 0.3, 0.5}	0.5
Optimizer	{SGD, Adam, AdamW}	AdamW
Weight Decay	{1e-5, 1e-4, 5e-4}	1e-4
Learning Rate Schedule	{Constant, Step, Cosine Annealing}	Cosine Annealing
Warm-up Steps	{0, 200, 500}	500
Max Epochs	{30, 50, 100}	100
Early Stopping	{3, 5, 10}	5

The evaluation was conducted using several key metrics: F1 Score, Specificity, PR AUC, and MCC. The F1 Score balanced precision and recall addressing class imbalance among ovarian cancer subtypes. Specificity measured the model's accuracy in identifying negative instances, reducing false positives in clinical decisions. PR AUC highlighted the balance between precision and recall at different thresholds, particularly in imbalanced scenarios. MCC combined true and false positives and negatives into a single score, making it suitable for multi-class classification. A confusion matrix visualized

prediction outcome, revealing misclassification patterns, especially among similar classes like LGSC and MC. Learning curves for accuracy and loss were plotted over 30 epochs to track training dynamics and detect overfitting or underfitting. We also used 10-fold stratified cross-validation to ensure unbiased performance estimates, maintaining proper class distribution in each fold. These strategies confirmed the model's stability, reliability, and readiness for real-world use.

To optimize model generalization further, we employed the hyperparameter optimization settings that are summarized in Table 3. We explored a range of values for key parameters such as learning rate, batch size, optimizer, and dropout rate through empirical tuning, selecting configurations that achieved optimal performance on the validation set.

4. Results and Discussion

Table 4 presents the performance comparison of six DL models on ovarian cancer subtype classification before and after data augmentation. The evaluation metrics include F1 Score, Specificity, Matthews Correlation Coefficient (MCC), and PR AUC. Before augmentation, CrossViT achieved the highest performance across all metrics, recording an F1 score of 95.82% ± 0.6, MCC of 91.33% ± 0.6, and PR AUC of 94.21% ± 0.8. Other models, such as SE-ResNeXt and ResNeXt-152, showed competitive results but underperformed slightly due to the impact of class imbalance.

Table 4 Performance comparison of each model

Before Augmentation				
Model	F1	Specificity	MCC	PR AUC
ResNeXt-152	93.21 ± 0.4	92.45 ± 0.9	88.69 ± 1.7	90.12 ± 0.9
DenseNet-264	91.88 ± 0.5	90.77 ± 1.9	86.55 ± 1.8	88.90 ± 1.1
ShuffleNet	89.75 ± 1.6	88.12 ± 0.8	83.72 ± 1.9	86.43 ± 1.8
AmoebaNet	92.40 ± 1.4	91.29 ± 1.7	87.44 ± 1.4	89.15 ± 1.0
SE-ResNeXt	94.10 ± 0.9	93.26 ± 0.4	89.71 ± 0.9	91.64 ± 0.8
CrossViT	95.82 ± 0.6	94.85 ± 0.7	91.33 ± 0.6	94.21 ± 0.8
After Augmentation				
Model	F1	Specificity	MCC	PR AUC
ResNeXt-152	97.01 ± 1.3	96.25 ± 1.1	94.13 ± 1.1	96.03 ± 1.0
DenseNet-264	96.28 ± 0.9	95.40 ± 0.5	93.26 ± 0.7	95.12 ± 0.7
ShuffleNet	94.32 ± 1.4	93.18 ± 1.5	90.08 ± 1.5	92.85 ± 1.1
AmoebaNet	95.75 ± 0.4	94.86 ± 0.4	92.58 ± 0.8	94.02 ± 0.8
SE-ResNeXt	96.84 ± 0.3	96.03 ± 0.8	94.51 ± 0.9	95.74 ± 0.6
CrossViT	98.81 ± 0.4	98.77 ± 0.3	97.66 ± 0.6	99.24 ± 0.2

After augmentation, performance improvements were observed across all models, with CrossViT again leading significantly. It attained the best overall metrics, including F1 score of 98.81% ± 0.4, Specificity of 98.77% ± 0.3, MCC of 97.66% ± 0.6, and PR AUC of 99.24% ± 0.2. These results demonstrate that data augmentation helped mitigate class imbalance and improved generalization. CrossViT's multi-scale self-attention mechanism proved particularly effective in capturing complex histopathological patterns, outperforming CNN-based architectures even in the post-augmentation scenario.

Figure 4 presents the confusion matrices for ovarian cancer subtype classification using CrossViT, comparing performance before and after data augmentation. Prior to augmentation, the model exhibited significant misclassifications, particularly for minority classes such as LGSC and MC, due to skewed class distribution. The confusion matrix showed off-diagonal entries indicating overlapping feature representations and inadequate generalization. Post-augmentation, the matrices show a more pronounced diagonal structure, reflecting enhanced class separability and reduced inter-class confusion. Correct predictions notably increased for all subtypes (e.g., HGSC: 1,960

→ 2,961), confirming that augmentation improved representation learning and mitigated bias. This refinement is consistent with observed improvements in F1 Score and MCC, affirming the model's robustness in a balanced and diverse learning environment.

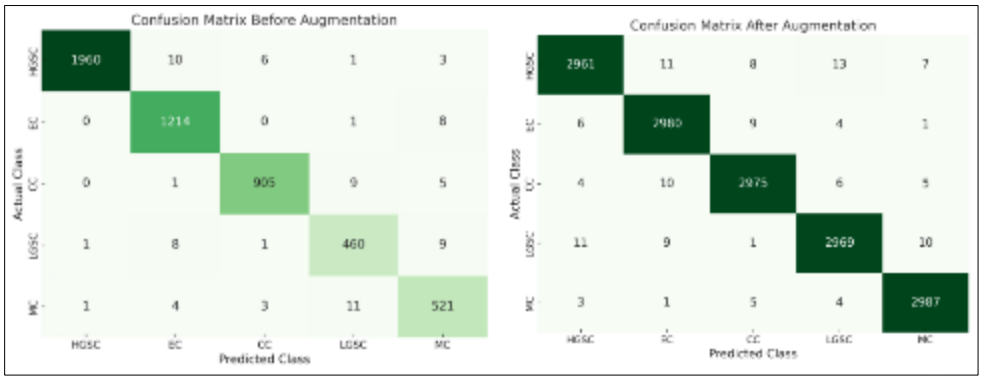

Figure 4 Confusion matrix of proposed model before and after augmentation

Figure 5 shows the learning curves of the CrossViT model for ovarian cancer classification, comparing training and validation performance before and after data augmentation. Before augmentation, the model exhibits signs of overfitting; the training accuracy steadily increases, but the validation accuracy fluctuates, and the validation loss remains higher, indicating poor generalization. In contrast, after augmentation, the learning curves show much improved generalization. The validation accuracy closely tracks the training accuracy, achieving near 99% by the end of training, while the validation loss decreases significantly. This demonstrates that data augmentation helps the model overcome class imbalance, stabilizing the learning process and improving overall performance on unseen data.

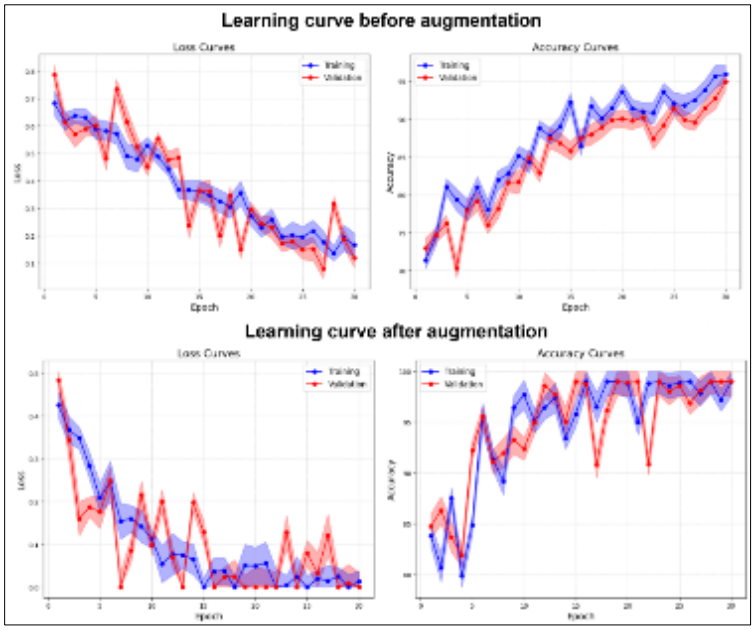

Figure 5 Learning curves before and after data augmentation for the CrossViT model

The web application provided in Figure 6, OveNet: Cancer Classifier, allows users to upload histopathological images and classify them into one of five ovarian cancer subtypes: HGSC, MC, CC, EC, or LGSC. Upon uploading an image, the system utilizes the trained CrossViT model to generate predictions. The classification result is displayed with a prediction bar showing the probability for each class. In this example, the HGSC subtype is predicted with 99.00% certainty, while the other subtypes have very low prediction probabilities (below 1%). The interface provides an easy-to-use experience with options to reload the application or re-upload a new image. This user-centric design, combined with the high accuracy of the CrossViT model, offers a powerful, real-time diagnostic tool for clinical settings. The prediction bar and result display are visually intuitive, ensuring that users can easily interpret model outputs. Moreover, the web application supports the deployment of DL models with real-time prediction capabilities, integrating Grad-CAM

visualizations for explainability. This helps clinicians understand which regions of the tissue image contributed to the model's decision.

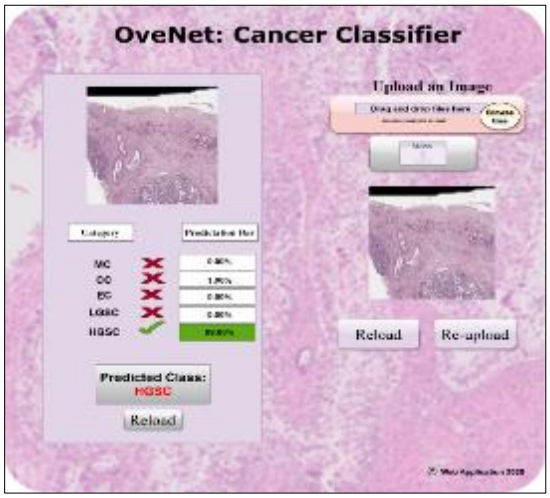

Figure 6 Demo of web application

The state-of-the-art (SOTA) comparison in Table 5 summarizes the performance of various models on ovarian cancer subtype classification across different datasets. Among the studies reviewed, Fahim et al. [2] achieved the highest accuracy of 99.01% using the OVANet model on an augmented dataset of 13,024 images. Alshdaifat et al. [8] achieved 98.73% using the Xception_ViT model with 4,070 CT images, while our proposed CrossViT model outperformed all others, achieving an impressive 99.24% on a large and diverse dataset of 100,000 samples.

Despite the large size of our dataset, CrossViT's performance can be attributed to its advanced multi-scale self-attention mechanism, which effectively captures both fine-grained textures and global patterns crucial for histopathological classification. Additionally, data augmentation significantly reduced class imbalance, allowing the model to generalize better, while maintaining high interpretability through Grad-CAM, enabling clinicians to trust its predictions. The combination of a strong architecture and balanced dataset made CrossViT the best-performing model, showcasing its ability to handle large-scale, complex data while maintaining high accuracy.

Table 5 Comparison of our results with previous studies

Reference	Sample Size	Model	Highest Result
Radhakrishnan et al. [19]	1,470 images	InceptionV3	97.96%
Fahim et al. [20]	13,024 images (augmented)	OVANet	99.01%
Wen et al. [22]	UBC-OCEAN, BCNB, DigestPath2019	MsaMIL-Net	93.2%
Macis et al. [34]	238 CE-CT image volumes	VGG19 + SVM	0.97
Alshdaifat et al. [23]	4,070 CT images	Xception_ViT	98.73%
Shetty et al. [24]	39,900 patient records	SVM, LR	91.7%
Ours	100,000 samples	CrossViT	99.24%

Figure 7 shows Grad-CAM visualizations for five ovarian cancer subtypes: Clear-Cell Carcinoma (CC), Endometrioid Carcinoma (EC), High-Grade Serous Carcinoma (HGSC), Low-Grade Serous Carcinoma (LGSC), and Mucinous Carcinoma (MC). The top row displays the original histopathological images, while the bottom row features Grad-CAM heatmaps from the CrossViT model, highlighting important regions for classification. Warmer colors (red and yellow) indicate high importance, while cooler colors (blue and purple) show areas of less influence. For CC and EC, the model focuses on dense epithelial and glandular tissues. In HGSC, it highlights broader areas with papillary structures and nuclear atypia. For LGSC and MC, the model captures subtler cues, demonstrating its ability to recognize fine details even in less common subtypes. These visualizations confirm the effectiveness of the CrossViT model's dual-branch, multi-scale self-attention mechanism and enhance transparency in its decision-making process. By identifying the tissue regions that

influenced the classification, Grad-CAM builds clinician trust and aligns the model with diagnostic criteria necessary for real-world medical applications.

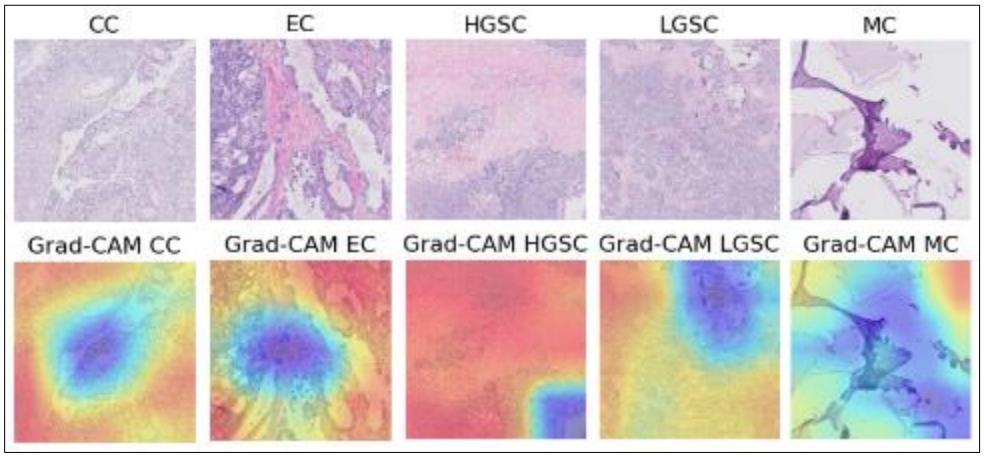

Figure 7 Grad-CAM heatmaps of the CrossViT model's classification across ovarian cancer subtypes

Our experimental results indicate that CrossViT outperformed alternative architecture due primarily to its dual-branch design and multi-scale self-attention mechanism. These features enabled effective capture of both fine-grained textures and global contextual patterns from histopathological images. In addition, our comprehensive preprocessing pipeline—comprising rigorous data augmentation, precise noise reduction, and hybrid feature extraction—played a pivotal role in enhancing model performance. This pipeline ensured balanced data distribution, minimized image artifacts, and prioritized the most discriminative features, thereby optimizing the learning process.

The broad implications of our work extend beyond technical performance. Clinically, our model's high accuracy and integrated explainability pave the way for earlier ovarian cancer detection, which can significantly improve patient outcomes. Economically, deploying a reliable automated diagnostic tool may reduce healthcare expenditures by decreasing reliance on extensive manual evaluations and enabling timely therapeutic interventions. Socially, enhanced diagnostic capabilities foster greater trust among clinicians and patients, ultimately contributing to improved public health and more efficient resource allocation.

However, several technical limitations must be acknowledged. The computational demands of the transformer-based CrossViT model are substantial, posing challenges for deployment in resource-constrained settings. The increased complexity also results in longer training times and necessitates specialized hardware for real-time applications. Furthermore, despite our robust preprocessing, some domain-specific variations and residual noise may still affect the model's generalizability across diverse clinical environments.

Future work should address these limitations by exploring model compression and lightweight transformer alternatives to reduce computational overhead. Additional studies are needed to evaluate domain adaptation strategies, advanced hyperparameter optimization, and extended validation on larger, multi-institutional datasets to further enhance the model's applicability in clinical practice. Moreover, refining feature selection techniques and integrating additional explainability methods will be crucial for ensuring that the model's predictions are both accurate and transparent. These efforts will bridge the gap between research and application.

5. Conclusion

Our study demonstrates that integrating a dual-branch ViT architecture with advanced preprocessing techniques results in a highly robust model for ovarian cancer subtype classification. Through a meticulous experimental process, we have shown that our approach achieves superior accuracy and generalizability on a large dataset of histopathological images, marking a significant advancement in automated diagnostic systems.

A pivotal component of our work is the development of a real-time web application that leverages our model. This tool not only offers immediate diagnostic insights but also incorporates visual explanations, thereby enhancing transparency and supporting clinical decision-making. The webapp has the potential to streamline diagnostic workflows by reducing

turnaround times and minimizing human error, ultimately contributing to improved patient management and healthcare efficiency.

Despite these promising outcomes, certain limitations remain. The complex architecture of CrossViT, along with the associated heavy computational requirements, may limit its immediate deployment in low-resource settings. Additionally, while our dataset is extensive, further validation across more diverse clinical environments is needed to cement the model's robustness and adaptability. Looking ahead, future efforts will focus on optimizing computational efficiency through model compression and the exploration of alternative more lightweight architectures. Broader multi-institutional studies and enhanced domain adaptation strategies will also be vital to ensure the model's scalability in real-world clinical applications. Efforts to further improve explainability will help in bridging the gap between cutting-edge AI methodologies and everyday clinical practice.

In summary, our research is clear by combining innovative model architectures with state-of-the-art preprocessing and a user-centric web application, we pave the way for next-generation diagnostic tools that can revolutionize how ovarian cancer is detected and managed. This work lays the foundation for impactful, real-time clinical implementations, driving forward the potential for AI to transform healthcare delivery and patient outcomes.

Compliance with ethical standards

Disclosure of conflict of interest

There is not conflict of interests.

References

[1] S. J. K. J. Kumar, G. P. Kanna, D. P. Raja, and Y. Kumar, "A Comprehensive Study on Deep Learning Models for the Detection of Ovarian Cancer and Glomerular Kidney Disease using Histopathological Images," Archives of Computational Methods in Engineering, vol. 32, no. 1, pp. 35–61, Jan. 2024, doi: 10.1007/S11831-024-10130-6/METRICS.

[2] P. Rustamadji et al., "A Decade of Ovarian Cancer in Indonesia: Epidemiology and Survival Analysis from 2010 to 2020," Journal of Clinical Medicine 2025, Vol. 14, Page 1692, vol. 14, no. 5, p. 1692, Mar. 2025, doi: 10.3390/JCM14051692.

[3] S. Lavanya J M and S. P, "Innovative approach towards early prediction of ovarian cancer: Machine learning-enabled XAI techniques," Heliyon, vol. 10, no. 9, May 2024, doi: 10.1016/J.HELIYON.2024.E29197/ASSET/6B20E58F-DA95-465F-9BCE-16631E8E559B/MAIN.ASSETS/GR12.JPG.

[4] J. Y. Lee et al., "Changes in ovarian cancer survival during the 20 years before the era of targeted therapy," BMC Cancer, vol. 18, no. 1, pp. 1–8, May 2018, doi: 10.1186/S12885-018-4498-Z/TABLES/4.

[5] D. Ban et al., "A personalized probabilistic approach to ovarian cancer diagnostics," Gynecol Oncol, vol. 182, pp. 168–175, Mar. 2024, doi: 10.1016/J.YGYNO.2023.12.030.

[6] Y. Zhang et al., "Suppression of CYLD by HER3 confers ovarian cancer platinum resistance via inhibiting apoptosis and by inducing drug efflux," Exp Hematol Oncol, vol. 14, no. 1, Dec. 2025, doi: 10.1186/s40164-025-00620-z.

[7] J. Li et al., "An Aptamer-Based Nanoflow Cytometry Method for the Molecular Detection and Classification of Ovarian Cancers through Profiling of Tumor Markers on Small Extracellular Vesicles," Angewandte Chemie International Edition, vol. 63, no. 4, p. e202314262, Jan. 2024, doi: 10.1002/ANIE.202314262.

[8] Y. Long, H. Shi, J. Ye, and X. Qi, "Exploring Strategies to Prevent and Treat Ovarian Cancer in Terms of Oxidative Stress and Antioxidants," Jan. 01, 2025, Multidisciplinary Digital Publishing Institute (MDPI). doi: 10.3390/antiox14010114.

[9] T. A. Fahim, F. B. Alam, and K. T. Ahmmed, "OVANet: Dual Attention Mechanism based New Deep Learning Framework for Diagnosis and Classification of Ovarian Cancer Subtypes from Histopathological Images," IEEE Access, 2024, doi: 10.1109/ACCESS.2024.3460077.

[10] L. Wang et al., "Drug resistance in ovarian cancer: from mechanism to clinical trial," Molecular Cancer 2024 23:1, vol. 23, no. 1, pp. 1–26, Mar. 2024, doi: 10.1186/S12943-024-01967-3.

[11] H. S. Zelisse et al., "Improving histotyping precision: The impact of immunohistochemical algorithms on epithelial ovarian cancer classification," Hum Pathol, vol. 151, p. 105631, Sep. 2024, doi: 10.1016/J.HUMPATH.2024.105631.

[12] S. R. Kongara et al., "Performance evaluation of optimized convolutional neural network mechanism in the detection and classification of ovarian cancer," Multimed Tools Appl, vol. 83, no. 28, pp. 71311–71334, Aug. 2024, doi: 10.1007/S11042-024-18115-0/METRICS.

[13] R. Haque et al., "A Scalable Solution for Pneumonia Diagnosis: Transfer Learning for Chest X-ray Analysis," 2024 7th International Conference on Contemporary Computing and Informatics (IC3I), pp. 255–262, Sep. 2024, doi: 10.1109/IC3I61595.2024.10829132.

[14] A. Al Noman et al., "Monkeypox Lesion Classification: A Transfer Learning Approach for Early Diagnosis and Intervention," 2024 7th International Conference on Contemporary Computing and Informatics (IC3I), pp. 247–254, Sep. 2024, doi: 10.1109/IC3I61595.2024.10828678.

[15] M. D. Hosen et al., "Parasitology Unveiled: Revolutionizing Microorganism Classification Through Deep Learning," 2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT), pp. 1163–1168, May 2024, doi: 10.1109/ICEEICT62016.2024.10534322.

[16] R. Haque et al., "A transfer learning-based computer-aided lung cancer detection system in smart healthcare," IET Conference Proceedings, vol. 2024, no. 37, pp. 594–601, Mar. 2025, doi: 10.1049/ICP.2025.0858.

[17] M. Sohaib, M. J. Hasan, and Z. Zheng, "A multichannel analysis of imbalanced computed tomography data for lung cancer classification," Meas Sci Technol, vol. 35, no. 8, p. 085401, May 2024, doi: 10.1088/1361-6501/AD437F.

[18] S. Ahmmed et al., "Enhancing Brain Tumor Classification with Transfer Learning across Multiple Classes: An In-Depth Analysis," BioMedInformatics 2023, Vol. 3, Pages 1124-1144, vol. 3, no. 4, pp. 1124–1144, Dec. 2023, doi: 10.3390/BIOMEDINFORMATICS3040068.

[19] M. Radhakrishnan, N. Sampathila, H. Muralikrishna, and K. S. Swathi, "Advancing Ovarian Cancer Diagnosis Through Deep Learning and eXplainable AI: A Multiclassification Approach," IEEE Access, vol. 12, pp. 116968–116986, 2024, doi: 10.1109/ACCESS.2024.3448219.

[20] T. A. Fahim, F. B. Alam, and K. T. Ahmmed, "OVANet: Dual Attention Mechanism based New Deep Learning Framework for Diagnosis and Classification of Ovarian Cancer Subtypes from Histopathological Images," IEEE Access, 2024, doi: 10.1109/ACCESS.2024.3460077.

[21] S. K. Behera, A. Das, and P. K. Sethy, "Deep fine-KNN classification of ovarian cancer subtypes using efficientNet-B0 extracted features: a comprehensive analysis," J Cancer Res Clin Oncol, vol. 150, no. 7, Jul. 2024, doi: 10.1007/s00432-024-05879-z.

[22] J. Wen, J. Wen, and M. Fang, "MsaMIL-Net: An End-to-End Multi-Scale Aware Multiple Instance Learning Network for Efficient Whole Slide Image Classification," Mar. 2025, [Online]. Available: http://arxiv.org/abs/2503.08581

[23] E. H. Alshdaifat et al., "Hybrid vision transformer and Xception model for reliable CT-based ovarian neoplasms diagnosis," Intell Based Med, vol. 11, Jan. 2025, doi: 10.1016/j.ibmed.2025.100227.

[24] R. Shetty, M. Geetha, U. Dinesh Acharya, and G. Shyamala, "Enhancing Ovarian Tumor Dataset Analysis through Data Mining Preprocessing Techniques," IEEE Access, 2024, doi: 10.1109/ACCESS.2024.3450520.

[25] Thite Sunil, "Ovarian Cancer Classification Dataset," Kaggle, 2024, Accessed: Apr. 09, 2025. [Online]. Available: https://www.kaggle.com/datasets/sunilthite/ovarian-cancer-classification-dataset

[26] G. Lakshmi G and P. Nagaraj, "Lung cancer detection and classification using optimized CNN features and Squeeze-Inception-ResNeXt model," Comput Biol Chem, vol. 117, p. 108437, Aug. 2025, doi: 10.1016/J.COMPBIOLCHEM.2025.108437.

[27] M. Z. Hasan, M. A. H. Rony, S. S. Chowa, M. R. I. Bhuiyan, and A. A. Moustafa, "GBCHV an advanced deep learning anatomy aware model for accurate classification of gallbladder cancer utilizing ultrasound images," Scientific Reports 2025 15:1, vol. 15, no. 1, pp. 1–20, Feb. 2025, doi: 10.1038/s41598-025-89232-5.

[28] R. Haque et al., "Advancing Early Leukemia Diagnostics: A Comprehensive Study Incorporating Image Processing and Transfer Learning," BioMedInformatics 2024, Vol. 4, Pages 966-991, vol. 4, no. 2, pp. 966–991, Apr. 2024, doi: 10.3390/BIOMEDINFORMATICS4020054.

[29] S. Feng, Z. Li, B. Zhang, T. Chen, and B. Wang, "DSF2-NAS: Dual-Stage Feature Fusion via Network Architecture Search for Classification of Multimodal Remote Sensing Images," IEEE J Sel Top Appl Earth Obs Remote Sens, 2025, doi: 10.1109/JSTARS.2025.3545831.

[30] A. Priya and P. Shyamala Bharathi, "SE-ResNeXt-50-CNN: A deep learning model for lung cancer classification," Appl Soft Comput, vol. 171, p. 112696, Mar. 2025, doi: 10.1016/J.ASOC.2025.112696.

[31] J. Kang, M. Mpabulungi, and H. Hong, "Multi-modal CrossViT using 3D spatial information for visual localization," Multimed Tools Appl, vol. 84, no. 5, pp. 2059–2083, Feb. 2025, doi: 10.1007/S11042-024-20382-W/METRICS.

[32] F. Siddiqui, J. Yang, S. Xiao, and M. Fahad, "Enhanced deepfake detection with DenseNet and Cross-ViT," Expert Syst Appl, vol. 267, p. 126150, Apr. 2025, doi: 10.1016/J.ESWA.2024.126150.

[33] W. Panyarak et al., "CrossViT with ECAP: Enhanced deep learning for jaw lesion classification," Int J Med Inform, vol. 193, p. 105666, Jan. 2025, doi: 10.1016/J.IJMEDINF.2024.105666.

[34] C. Macis et al., "A Convolutional Neural Network Tool for Early Diagnosis and Precision Surgery in Endometriosis-Associated Ovarian Cancer," Applied Sciences (Switzerland), vol. 15, no. 6, Mar. 2025, doi: 10.3390/app15063070.